TRANSFORMATIONS
A DARIAH Journal

Introducing *Transformations: A DARIAH Journal*

Toma Tasovac
Belgrade Center for Digital Humanities / Digital Research Infrastructure for the Arts and Humanities

Humanists are no strangers to transformations: we study how ideas, beliefs and convictions evolve over time; how social and cultural practices shift; and how artistic forms adapt under historical pressures. We trace how languages emerge, grow, flourish and, sometimes, die. We examine how texts, images and cultural artefacts are conceived, structured and then given new meanings in various historical periods and across different media. We explore the interplay of tradition and innovation, continuity and rupture, memory and creativity. Because, like Heraclitus, we know that the only true constant is change.

But what's in a name? When we chose to call our new publication *Transformations: A DARIAH Journal*, we were, generally speaking, evoking these deep, timeless currents of humanistic thought, while, at the same time, and in a very specific sense, referring to the ways in which the more recent embrace of digital methods has reshaped our disciplines. The shift from analogue to digital, from paper to screens, and from close to distant reading techniques has once again brought into focus for us the important role that technology plays in the production of knowledge.

These days, in particular, with artificial intelligence looming everywhere we turn, we are caught—in yet another variation on the theme of relentless change—between the promises of spectacular scientific discoveries and the threat of human obsolescence. In such a precarious landscape, we cannot afford to treat our own methodological, conceptual and aesthetic transformations as mere rhetorical flourishes in some grand narrative of technological progress. Instead, we should be mindful of them and embrace them as the necessary precondition for maintaining the relevance of humanistic inquiry.

EPI sciences

To cite this article: Tasovac T. (2025). Introducing *Transformations: A DARIAH Journal*. *Transformations. A DARIAH Journal*, vol. 1, p. 1-3. https://doi.org/10.46298/transformations.15787

Choice of name aside, why would a research infrastructure like DARIAH decide to start a scholarly journal in the first place? We have argued elsewhere that research infrastructures can only thrive in the long term if the researchers who contribute to their development and growth receive academic credit for the kind of work they do in and around these facilities (Tasovac et al. 2023). The mission of *Transformations*, which we launched as DARIAH celebrated its 10-year anniversary as a European Research Infrastructure Consortium (ERIC), is, therefore, unabashedly programmatic. We seek to foster open science in the arts and humanities, enable fairer research assessment practices and support the recognition of a broader range of research outputs—beyond articles and monographs—that have long remained undervalued by conventional academic metrics. We do this because we are a founding member of the Coalition for Advancing Research Assessment (CoARA), and because we recognise that traditional research outputs often conceal the infrastructural labour that underpins them. *Transformations* will strive to highlight these overlooked contributions and give them the credit they deserve.

We are committed to publishing scholarly accounts of various "ephemeral" activities—detailed descriptions of datasets and workflows, annotation practices, software tools and training materials, to name just a few—alongside the more traditional narrative papers. Because we know that knowledge production is a multi-step process, and that excellence is not only found in polished monographs or dense journal articles at the end of a research journey. It is also to be sought in the careful design of textual corpora, the meticulous crafting of code, the complex modelling of 3D digital twins, and the iterative interplay of annotation and analysis, all of which deserve to be recognised and shared.

The medium of our journal itself aligns with its message. *Transformations* is hosted on Episciences, a diamond open-access overlay journal platform provided by our partner, the Centre for Direct Scientific Communication. Episciences allows us to set up a transparent publishing system on top of a community-governed, public infrastructure. Unlike traditional paywalled journals, which are tied to commercial publishers and driven by exorbitant profit margins, an overlay journal such as *Transformations* does not hide content behind proprietary systems. Instead, it operates by peer reviewing work that has been deposited in trusted open repositories such as HAL or arXiv, where it can be freely accessed, indexed and preserved—at no additional cost to the author or the reader. This commitment to public infrastructures is part of the DARIAH ethos. By embracing the overlay journal model, we show how research infrastructures can help move away from commercial gatekeepers and towards more equitable research ecosystems.

It is also somehow fitting that our first thematic issue should be dedicated to the topic of workflows. Workflows are the connective tissue of scholarship—the

series of steps and decisions that guide us from collecting primary sources to shaping hermeneutic perspectives about them. Too often, these steps are glossed over in scholarly publications. By foregrounding workflows, we want to stress that *how* we do research in the digital age is just as important as *what* we discover. Documenting and sharing workflows makes digitally enabled research more reproducible and transparent, allowing others to reuse, learn from and build upon these processes. A researcher may rely on a specific database, software library or digital model: by drawing attention to them, we implicitly recognise those who maintain datasets, build tools and design platforms as first-class contributors to knowledge creation.

All of these commitments—to open science, fairer assessment practices, public infrastructures and reproducible workflows—should help us respond to the challenges and opportunities of conducting research in the age of artificial intelligence. AI tools are already assisting arts and humanities scholars in numerous tasks, from transcribing archival texts to analysing cultural datasets at scale. But they also prompt urgent questions about the nature of authorship, credibility and ethics. When algorithms can generate entirely plausible texts or identify patterns faster than any human, the humanistic approach to research that *Transformations* stands for—with its emphasis on context, critical interpretation and ethical self-reflection—remains more important than ever. Without transparent tooling and a shared consensus on the systemic importance of public infrastructures, we would find it increasingly difficult to distinguish rigorous scholarship from automatically generated, hallucinogenic noise.

What remains in the end, I hope, will be what was always already there in the beginning: an enduring impulse to seek, to question, to wonder, and to try to understand what it means to be human.

References

Tasovac, Toma, Laurent Romary, Erzsébet Tóth-Czifra, et al. 2023. "The Role of Research Infrastructures in the Research Assessment Reform: A DARIAH Position Paper." HAL. Version 1, 21 June. https://hal.science/hal-04136772v1.

TRANSFORMATIONS
A DARIAH Journal

Workflows: Introduction (Part 1)

Anne Baillot[1, 2], Françoise Gouzi[2], Toma Tasovac[2, 3]

1 Le Mans Université
2 DARIAH-EU
3 Belgrade Center for Digital Humanities

In simple terms, a workflow is a structured sequence of tasks or processes designed to carry out a research activity. Workflows are key to our daily research practice: they shape how we collect, process, interpret and manage information relevant to our work (Goble et al. 2020). Yet despite their centrality, there is no one standard way of describing a workflow in the humanities, as opposed to other areas of research (Atkinson et al. 2017). Lacking a common language (Crusoe et al. 2022), workflows are currently difficult to share, replicate and reuse.

This question, however, is not wholly overlooked. As the inaugural volume of *Transformations: A DARIAH Journal* proves, there is an active community of researchers in the humanities and social sciences that strives to foster a culture of workflow reusability.[1] The published articles showcase how projects from a variety of scholarly horizons and areas of research share a common interest in carefully drafting contextualised workflow descriptions to facilitate their reuse (see also Marongiu, McGillivray, and Khan 2024). And they do not simply shed light on the methodologies developed; they reflect on how to present these methodologies in a replicable manner and on what their significance is for scholarship at large. In the context of open science, digital tools are critical to the implementation of reusability schemes. As such, they play a central role in the articles we present here.

Depending on their scope, workflows require technological tooling of varying complexity. Still, rather than organising this volume by the size of workflows

1. We are publishing this first batch of articles ahead of the DARIAH Annual Event 2025 to provide early access to these contributions and foster timely discussion and exchange. The second batch of articles, which will complete this first volume, is planned for publication in the autumn of 2025.

EPI sciences

To cite this article: Baillot, A., Gouzi, F. & Tasovac, T. (2025). Workflows: Introduction (Part 1). *Transformations. A DARIAH Journal*, vol. 1, p. 1-5, https://doi.org/10.46298/transformations.15780, CC BY 4.0.

or the tools employed, we have chosen to group the articles according to their primary objects of study: metadata-based workflows, textual data-based workflows, and image-based workflows. In all three areas, the articles grapple with the challenge of formalising research methodologies in ways that are not only transparent and rigorous, but also meaningfully reusable. The opening article by **Isto Huvila, "Paradata Conveys Understanding of Workflows and Facilitates the Reuse of Research Data in the Arts and Humanities,"** offers a methodological apparatus and a generic approach to workflow reusability based on the documentation of research data management. The paper reflects on the gaps between the data management information typically provided by research projects and the information users need to actually re-implement the corresponding methodology. It also establishes a core theme running through all the articles: the empowering role of the FAIR Principles[2] in elevating workflows as valuable research outputs.

The first section is dedicated to workflows related to metadata, an essential aspect in ensuring the findability and accessibility of research data in line with the FAIR Principles. Catalogues play a key role in this section. The article by **Gustavo Candela, Cezary Rosiński and Arkadiusz Margraf, "A Reproducible Framework to Publish and Reuse Collections as Data: The Case of the European Literary Bibliography,"** sheds light on bibliographic databases as an example of digital collections curated by galleries, libraries, archives and museums (GLAM institutions) and relevant to humanities scholarship. The workflow this article presents focuses on a specific literary bibliographic database, the European Literary Bibliography, and proposes concrete steps to make the corresponding metadata reusable in computational projects. The interaction between catalogues and (literary) research is also central in the article by **Iiro Tiihonen and Kira Hinderks, "Genre Classification Workflow for the English Short Title Catalogue (ESTC)."** This paper offers a solution to improve genre classification in the extraction of metadata from a catalogue. Based on a large English-language corpus, it presents both literary challenges and extraction techniques, while also addressing important questions of representation bias. The article by **Till Grallert, "Adding Every Arabic Periodical Published Before 1930 to Wikidata,"** expands the discussion to non-English corpora, highlighting issues of multilingualism and information transfer across languages. Here too, it is crucial that the workflow bridge the different types of metadata—along with their formats, curation and management practices—whether created by scholars or generated by cataloguers in the GLAM sector. This contribution underscores the advantages of relying on Wikidata to implement the FAIR Principles in this context. In **"A Workflow to Publish Collections as Data: Looking Back at Europeana.eu and Forward to the Common European Data Space for Cultural Heritage,"** by **Gustavo Candela, Sally Chambers, Alba Irollo,**

2. See https://www.go-fair.org/fair-principles/.

Nuno Freire, Vicky Dritsou, Antoine Isaac, Agiatis Benardou, Vicky Garnett and Toma Tasovac, the authors consider the role of research infrastructures in leveraging sustainability. They present existing resources and EU-wide initiatives to facilitate data transfer from GLAM institutions to research projects, demonstrating the continuity of heritage data as both cultural and scholarly assets. In this section, both Candela, Rosiński and Arkadiusz, and Candela et al. showcase Collections as Data as a powerful tool to facilitate workflow reusability at the interface between heritage and research.

The second section addresses approaches to text-based data. Digital scholarship relies on textual information, since text is the basis for the provision and curation of computer-readable information. It extends from metadata and annotation to actual textual content. **"The Sixth Generation of the Perseus Digital Library and a Workflow for Open Philology,"** by **Gregory Crane, James Tauber, Alison Babeu, Lisa Cerrato, Charles Pletcher, Clifford Wulfman, Sergiusz Kazmierski and Farnoosh Shamsian,** offer insights into one of the most important humanities projects of the past decade. The Perseus Digital Library has not only pioneered the constitution and dissemination of rich digital corpora for humanities research; as this article demonstrates, the project is also able to refine its workflows iteratively, adapt to change and ensure reusability for a wide array of users. It combines technical detail, especially around exporting annotations, with reflections on how a massive project can evolve to improve quality and accessibility over time. **Emiliano Degl'Innocenti, Francesco Pinna, Alessia Spadi and Federica Spinelli** also address the management of textual collections, albeit at a smaller scale, in **"Creating a Scientific Workflow to Manage Old Italian Texts."** Their article showcases the complementarity of publishing a workflow description on the SSH Open Marketplace and contextualising it in a journal publication, providing both technical detail and general reflections on reusability. Following approaches to corpora in Ancient Greek (Crane et al.) and Old Italian (Emiliano Degl'Innocenti et al.), the article by **Adeline Clarke, Krista Lagus and Maria Valaste, "finnsurveytext: Analysis of Open-Ended Survey Responses in R,"** moves the focus to the social sciences and to lesser-resourced contemporary languages. The workflow at the heart of this paper concerns itself with the analysis of open-ended survey answers—e.g. free text—in languages other than English. In addition to detailed guidance on using a dedicated R package, it offers insight into the improvement of text quality through semantic enrichment tools. In this sense, it converges with Crane et al. in the effort to contextualise highly technical semantic tools and make them accessible to humanities and social science practitioners less familiar with computational tools—which is precisely what reusable workflows aim to achieve.

The final section is dedicated to image-based workflows. In this context, annotation and semantic enrichment remain essential but are layered with

additional challenges related to the generation, curation and analysis of images. In **"Improving Workflows in Digital Art History: Sharing Annotations for Cultural Heritage Image Segmentation and Object Detection," Léa Maronet and Alice Truc** explore the challenges of using computational methods in art history. Their analysis provides valuable insights into how image-related work can be enhanced through standardisation—a crucial step to achieve interoperability and reusability, guided by the FAIR Principles. Accessibility and reproducibility are also key in **"A Reproducible Workflow for the Creation of Digital Twins in the Cultural Heritage Domain"** by **Sebastian Barzaghi, Alice Bordignon, Federica Collina, Francesca Fabbri, Bruno Fanini, Daniele Ferdani, Bianca Gualandi, Ivan Heibi, Nicola Mariniello, Arcangelo Massari, Marcello Massidda, Arianna Moretti, Silvio Peroni, Sofia Pescarin, Maria Felicia Rega, Giulia Renda and Mattia Sullini.** This article tackles 3D reconstruction at the intersection of research and dissemination. It offers a thorough overview of the many standards used in heritage-based 3D reconstructions and proposes a workflow to facilitate the processing of the corresponding metadata. It then describes the creation of a digital exhibition—a digital twin of a real-life exhibition—and reflects on both the technical and museological challenges involved. Beyond its technical contributions to digitisation processes in a heritage context, the paper argues in favour of openness, accessibility and reusability.

The contributions in this issue provide contextualised workflow descriptions in a series of areas highly relevant to research in the humanities and social sciences. They pave the way to what has the potential to become a major research output in its own right: scholarly articles focused on research methodologies—or what might be called *workflow papers*. Positioned at the crossroads of open science, state-of-the-art technologies and reflexive approaches to major fields of inquiry, workflow papers offer new perspectives for humanities research. They increase data power and digital literacy while helping to shape scholarly ethics in line with the FAIR Principles. By dedicating this issue of *Transformations: A DARIAH Journal* to workflows, we intend to root workflows as a recognised publication format in the scholarly landscape of the arts and humanities.

References

Atkinson, Malcolm, Sandra Gesing, Johan Montagnat, and Ian Taylor. 2017. "Scientific Workflows: Past, Present and Future." *Future Generation Computer Systems* 75 (October): 216–27. https://doi.org/10.1016/j.future.2017.05.041.

Crusoe, Michael R., Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, et al. 2022. "Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language." *Communications of the ACM* 65 (6): 54–63. https://doi.org/10.1145/3486897.

Goble, Carole, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters, and Daniel Schober. 2020. "FAIR Computational Workflows." *Data Intelligence* 2 (1-2): 108–21. https://doi.org/10.1162/dint_a_00033.

Marongiu, Paola, Barbara McGillivray, and Anas Fahad Khan. 2024. "Multilingual Workflows for Semantic Change Research." *Journal of Open Humanities Data* 10(1): 15. https://doi.org/10.5334/johd.179.

TRANSFORMATIONS
A DARIAH Journal

Paradata Conveys Understanding of Workflows and Facilitates the Reuse of Research Data in the Arts and Humanities: The CAPTURE Project

Project note

Isto Huvila

Department of ALM, Uppsala University

Transformations, A DARIAH Journal
Volume 1, 2025
https://transformations.episciences.org

Dates
Received:14/11/2024
Accepted: 03/04/2025
Published: 10/06/2025

DOI: 10.46298/transformations.14747

Abstract

Project CAPTURE, funded by the European Research Council, has investigated the relatively unexplored question of what information about the workflows involved in creating, managing and using research data is required for data to be reusable in the future. This cross-disciplinary, multi-methods research underlines the risk of producing documentation that is of little use. It also emphasises the importance of making existing information findable and generating new documentation only on such aspects of workflows that otherwise remain undocumented. The project's findings also underscore differences in the perspectives of data creators, managers and users, indicating that the documentation generated by data creators and kept in repositories is not always well understood by data users or usable. Finally, the work carried out by the CAPTURE project points to the need for conceptual clarity. We discuss paradata both as a form of information on workflows, processes and practices, and as a framework concept to discuss how people give and receive information about workflows, processes and practices in general, in research and beyond.

Keywords: paradata, workflows, processes, documentation, research data, data reuse

EPIsciences

To cite this article: Isto Huvila (2025). Paradata Conveys Understanding of Workflows and Facilitates the Reuse of Research Data in the Arts and Humanities: The CAPTURE Project. Project note. *Transformations. A DARIAH Journal*, vol. 1, p. 1-12. https://doi.org/10.46298/transformations.14747

Paradata förmedlar förståelse för arbetsflöden och underlättar återanvändning av forskningsdata inom humaniora: CAPTURE-projektet

Sammanfattning

Det av Europeiska forskningsrådet finansierade forskningsprojektet CAPTURE har undersökt den relativt outforskade frågan om vilken information om arbetsflöden för skapande, hantering och användning av forskningsdata som är nödvändig för att data ska kunna återanvändas i framtiden. Denna tvärvetenskapliga forskning understryker risken med att producera dokumentation som är av liten nytta. Den betonar också vikten av att göra den redan existerande informationen sökbar och att endast generera ny dokumentation om sådana aspekter av arbetsflöden som annars förblir odokumenterade. Vidare understryker projektets resultat skillnader i perspektiven hos dataskapare, förvaltare och användare, vilket indikerar att den dokumentation som genereras av dataskapare och förvaras i dataarkiv inte alltid är väl förstådd och användbar för dataanvändare. Slutligen pekar arbetet i CAPTURE-projektet på behovet av begreppsmässig klarhet. Vi diskuterar paradata både som en form av information om arbetsflöden, processer och praktiker, och som ett ramverksbegrepp för att diskutera hur människor informerar och blir informerade om arbetsflöden, processer och praktiker.

Nyckelord: paradata, arbetsflöden, processer, dokumentation, forskningsdata, återanvändning av data

This work received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement number 818210, as a part of the project CApturing Paradata for documenTing data creation and Use for the REsearch of the future (CAPTURE).

Introduction

Data reuse in the arts and humanities requires an understanding not only of what the data is about, but also how it was conceived, for what purposes it was collected or created, and how it has been processed and used since its creation (e.g. Voss 2012; Faniel and Yakel 2017; Huvila 2022a). Project CAPTURE, funded by the European Research Council, inquired into the previously relatively unexplored question of exactly what information about the workflows involved in creating, managing and using research data is necessary for data to be reusable in the future (Huvila and Ekman 2024). In parallel, CAPTURE investigated how this information can be captured in both efficient and sufficiently comprehensive ways to support data reuse. Data on data creation and manipulation processes is articulated in the project as paradata (Börjesson, Sköld, and Huvila 2020). This term has notably been used in survey research (Schenk and Reuß 2024) and cultural heritage visualisation. In the latter, the London Charter and Seville Principles (Carrillo Gea et al. 2013) have emerged as two influential documents that stipulate the foundations for documenting visual representations in archaeological and cultural heritage–related contexts, underlining the importance of documenting not only visualisations themselves, but also the intellectual decisions and actions taken to produce them. More recently, the concept has been introduced in the context of research data (Huvila 2022a), and archives and records management (Davet, Hamidzadeh, and Franks 2023).

The purpose of CAPTURE was to create new knowledge and develop our understanding of how paradata is created and used. It also aimed to develop and test methods to capture paradata in research workflows. The results provide a basis for creating standards, as well as tools, for managing paradata and advances in data-intensive research, especially in fields that use heterogeneous research data from different cultural, organisational, temporary and cultural origins.

Paradata as a framework concept and form of data

One of the first tasks in the CAPTURE project was to engage in conceptual and theoretical work (see for example Sköld, Börjesson, and Huvila 2022; Huvila 2022a) relating to the concept of paradata and the way it is used and understood in different domains. The aim was to contribute to the development of paradata as a theoretical and practical concept. While references to the notion of paradata across disciplines show many similarities, there are also differences stemming from the variety of concerns being addressed (Sköld, Börjesson, and Huvila 2022; Cameron, Franks, and Hamidzadeh 2023).

In a concrete sense, paradata can be used to refer to a specific type of data that differs from both data and metadata. The term can also be understood in a theoretical sense as a means to conceptualise diverse forms of data or information

that can contribute knowledge of different types of processes, practices and workflows (Andersson, Huvila, and Sköld 2024; Huvila, Andersson, Sköld, and Liu 2024). Both approaches are fruitful, directing attention to the general need to convey understanding of workflows and highlighting the diverse explanatory functions of specific pieces of information (or data) that help provide such an understanding. The two approaches do, however, serve distinct purposes, and as such, deserve to be kept apart from each other. The problem with the first approach is that the empirical explorations of how paradata works in practice demonstrate how one person's paradata is likely to be another person's data and vice versa. For example, a description of the process of analysing a tissue sample is paradata for a scientist, but can function as data for a scholar of scientific practices. Therefore, rather than defining paradata by dichotomising it with other types of data and data-about-data, in the latter sense as a category of information, it is more fruitful to understand paradata as a category of conceptual and physical objects that can be appropriated, insofar as they are informative of processes and practices (Huvila 2025), rather than as a distinct kind of information or data.

The importance of documenting just enough

One major challenge for capturing and preserving paradata is that, when it comes to understanding specific aspects of workflows, individual data users' needs vary across situations. Without sufficient documentation on how scholars and other data creators and users have created, understood and interpreted data, there is a risk of ending up in what has been described as a digital dark age (Bollacker 2010), a situation where hard-to-reach and hard-to-find "dark data" becomes dominant (Geser and Niccolucci 2016; Anichini and Gattiglia 2015). Lack of contextual information about the creation of research data—both digital and non-digital—can, in the worst-case scenario, lead to the proliferation of substandard and hard-to-interpret datasets that may not support future research or foster the creation of new scholarly knowledge.

The practical goal of CAPTURE has been to support the implementation of national, European and global policies on data management and open data by enhancing our understanding of paradata and research-process documentation (see for example Beck and Neylon 2012; DCC 2017; Kansa and Kansa 2011; Gilby et al. 2022). In parallel, case studies of paradata in a broad range of disciplinary contexts (for example in Huvila, Andersson, and Sköld 2024b; Bruseker, Doerr, and Theodoridou 2017) have contributed to more efficient sharing and reuse of data in discipline-specific, thematic and interdisciplinary knowledge ecosystems and repositories.

The empirical focus of the project was archaeology. As a highly cross-disciplinary field operating with a broad variety of data from diverse disciplinary origins, ranging from material remains to textual scholarship, biology and

geophysics, it provides a fruitful context to inquire into the complexities of sharing knowledge on workflows. Many of the key challenges relating to data reuse and open data are prominently visible in archaeology (D'Andrea 2023). The benefits of working towards greater intellectual accessibility to datasets are equally apparent. Collecting data through excavations is a destructive activity that cannot be redone. The data-generation process in archaeology is different to many other humanities disciplines, where data can be generated or collected anew (Anichini and Gattiglia 2015), and newer, more precise or up-to-date data often substitutes older data for other purposes than historical research. This also means that archaeological datasets seldom become completely outdated and, as such, their reuse potential is extensive.

In addition to exploring archaeological data generation and documentation in depth, CAPTURE has engaged scholars from a range of disciplines in the inquiry of paradata and workflow documentation (e.g. Huvila, Andersson, and Sköld 2024b; Huvila and Sinnamon 2022; Huvila, Enwald, et al. 2022; Huvila, Greenberg, et al. 2021). As part of this work, the project has developed critical understanding of the social contexts and use of infrastructures emphasised in newer research agendas (e.g. Aloia et al. 2017; Lambourne et al. 2014) and empirical research (e.g. Mayernik et al. 2017). In parallel, it has created new knowledge of what data creators and users find important to document about data-related workflows, what explicit and implicit needs there are for documentation, and how the diverse needs and wishes of data creators and users can be satisfied in practice (Huvila, Andersson, Friberg, et al., forthcoming).

A key premise and insight from data documentation research is that it is impossible to document everything. There is also no limit to how much documentation can be generated. As a consequence, there is a risk that a lot of effort is invested in producing documentation with little relevance. The diversity of needs and the difficulty of anticipating them complicates the documentation of data-related workflows. Like all data (Mayernik and Acker 2018), paradata is bound to be incomplete. One significant challenge is how to document and preserve just enough, avoiding both over- and under-documentation. It is important to focus on identifying how the documentation effort can be minimised. This is especially critical for documentation that only serves assumed future purposes rather than immediate needs relating to data creation and its primary use. Much paradata can be captured automatically, but some aspects of workflows, in particular decisions and intellectual work, need to be documented manually. Further, it is important to examine what information might already be embedded in the data itself (Huggett 2012; Gant and Reilly 2017) and what aspects of workflows future users themselves can deduce using various types of non-literal "archaeological" or "forensic" post hoc methods (Kirschenbaum, Ovenden, and Redwine 2010) of inquiring into existing data. Before the CAPTURE project was initiated, research investigating

these approaches was carried out in isolation. There was little research covering the entire paradata phenomenon and how it links to communicating a more comprehensive understanding of research workflows.

Investigations of data creation and workflow reuse by archaeologists

The CAPTURE project and its approach are multidisciplinary. The participating researchers had experience and backgrounds in archival science, information science, archaeology and science studies. They helped approach paradata and the documentation of data-related workflows from multiple perspectives. Crucial aspects of understanding paradata and workflow documentation include how they link to the everyday work of their creators, managers and users; how different types of paradata are used to inform and how they inform their users; how they work when approached from within specific fields in the arts and humanities, and beyond.

The project used different methods to investigate the workflows, processes and practices that underpin the creation and use of research data within and outside of archaeology. It also proposed and developed methods to capture them. Studies have investigated paradata elements present in archaeological field reports and research data (e.g. Huvila, Sköld, and Börjesson 2021; Huvila, Börjesson, and Sköld 2022; Börjesson et al. 2022) as well as citations to methods literature and their usefulness in understanding research workflows (Huvila, Andersson, and Sköld 2022). Methods-wise, the project conducted ethnographic research; reviewed and tested previously proposed and newly developed methods for documenting paradata; and conducted a survey (Huvila, Andersson, and Sköld 2024a; Huvila, Andersson, Sköld, and Liu 2024) and interviews with key stakeholders (e.g. Börjesson, Huvila, and Sköld 2022).

Besides committing to conducting rigorous research *about* paradata, a critical question at the outset was how the CAPTURE project should think about and deal with paradata relating to the project itself. There were two obvious alternatives: try to document as much as possible, as meticulously as possible; or proceed as a normal research project would and at the end engage in an introspective analysis of what happened, how much and what kind of paradata was generated and preserved. The latter approach was chosen, as the alternative would have come with the apparent risk of us trying to do too much, resulting in something that could be done in this specific project, but would likely be unfeasible for any other project to follow. The work analysing paradata and documentation from CAPTURE hence started at the conclusion of the project. So far, it seems that a considerable amount of paradata was produced, and many of the processes and the progress of the work can be tracked from the available evidence. At the same

time, it is apparent that many crucial decisions were made, for example, during meetings and face-to-face discussions. The resulting paradata is inconsistent and may be absent, as we found in our analysis of other researchers' work.

It requires both effort and competence to exploit them (see Huvila, Olsson, et al., forthcoming).

There is a lot of paradata already out there—make sure not to lose it

The results from interview and survey studies, and our investigations into research publications and data, show that a lot of information qualifying as paradata on workflows is already available in different parts of typical research documentation. In archaeology, field reports are an important source of paradata as they are expected to document both the results and the investigation process. In addition to regular work descriptions, they convey knowledge – both directly and indirectly – of work processes, for example through descriptions of those processes and by providing information on participating actors. Another central source of paradata in archaeology is personal diaries and notebooks containing notes, observations and reflections on archaeological work. They correspond to the notebooks held by researchers elsewhere, especially in fieldwork and archival disciplines.

Photographs are another important source of paradata, in particular those showing ongoing work, the working environment and conditions (Huvila, Sköld, and Börjesson 2021). As with photographical analysis, close reading of datasets can provide information about their underlying workflows. Choice of words, descriptions and timestamps are just a few examples of elements in databases that incorporate knowledge of workflows (Börjesson et al. 2022). Since many of the traces of workflows, both in data and other research documentation, are often peripheral to major research outputs (Huvila, Sköld, and Andersson 2023)—even though standardisation offers multiple benefits for the findability and technical interoperability of data (e.g. Lins and Becker 2013)—there is a considerable risk that they are cleaned out in the process of normalising and standardising data. Early versions of research material that might seem unnecessary will be discarded in such a process (Börjesson et al. 2022). Although standards can permit the inclusion of multiple viewpoints through the documentation of parallel perspectives (Durost, Reich, and Girard 2022), such planned post hoc inclusion of concurrent and contradicting viewpoints does not solve the problem of losing something like an organic diversity of traces (Huvila, Sköld, and Andersson 2023) in the research material.

The fact that a lot of paradata already exists in available documentation means that the primary challenge in documenting processes is not necessarily how to

expand its quantity or scope. A critical problem is that paradata is scattered. It can be difficult to get a comprehensive understanding of the paradata that is available. In this respect, rather than generating more, it is more important to find existing paradata and make it findable for others. This requires an understanding of what information on workflows already exists, what is missing, and how it can be supplemented with whatever additional information is needed to convey an adequate understanding of data-creation workflows.

Dissimilar frames of reference complicate documentation

In addition to the problem that paradata might not always be available, it may also not meet the needs of its potential users. Different individuals often have slightly different views of what information is necessary and what a user can be expected to know. Infrastructural forms of backend work are seldom recognised and properly documented. For example, information on information management, standards and data structures is rarely documented in detail.

In addition, the results from the CAPTURE project show that data creators and users often have differing views on the type of paradata that is needed (Börjesson, Huvila, and Sköld 2022; Huvila, Andersson, and Sköld 2024b; Huvila, Juneström, et al. 2024). This aligns with earlier observations that understanding of the context of research data differs between data users and data creators (Faniel, Frank, and Yakel 2019). When a data creator is documenting paradata, the focus is often on details like what tools were used, who did what and who funded the work – details that seem obvious to the data creator to document (Huvila, Andersson, and Sköld 2024b). The resulting paradata follows their understanding of what is central to know about data-creation workflows and what is easy to document. In contrast, data users expect and need paradata that would help them understand workflows from the outset (Huvila, Andersson, Sköld, and Liu 2024).

Closing the gap between data creators, managers and users

The gap between what data creators and users consider important, meaningful and possible to document is a major challenge for documenting workflows, as well as for producing and preserving relevant paradata. It seems obvious that data users need better insights into specific workflows through relevant documentation, but also a general understanding of how research workflows function, as well as their underpinnings in the disciplines and research fields they engage with.

Although it has become a truism to underline that data is never raw, the dream of a universally usable, context-free data still very much underpins the way many researchers and non-researchers conceptualise it. Abandoning this idea not only in theory but also in practice is crucial if we are to create and preserve usable data and paradata. Just as data users need better insights into data generation and workflow processing, it is important that data creators understand how users approach their work, and what paradata is likely to be relevant to them. In many cases, even when data is thought to be comprehensible and reusable, it is, as Rösler (2020) points out in reference to Eco (2009), "under pressure" of the context of its creation. Rather than assuming that the documentation produced when creating data is adequate for its users, it is important that data creators focus on actual usage when creating and documenting data.

Data managers also need to be mindful. Whereas their obvious role is to bridge the gap between creators and users, there are indications that managers serve more to tame data, and make it manageable within contemporary technical and organisational infrastructures, than to foster data reuse. From this standpoint, it is critical we acknowledge that no one working with data on the journey from creation to use is neutral. Creating useful data documentation for others is difficult and becomes increasingly complicated when data creators, managers and users share little common ground (Huvila 2020). Even when it is created and managed with care, "taking" paradata into use (Huvila 2022b) is a task that requires effort and continuous critical reflection. It should not be taken lightly.

Read more about CAPTURE

Updates on the activities of the CAPTURE project, including workshops, online CAPTURE seminars with invited speakers and research results can be found on CAPTURE's website (http:/www.u.se/en/research/capture) and X (Twitter) @CAPTURE_ERC.

Data availability statement

Research data and publications produced by the CAPTURE project are available on the Digitala Vetenskapliga Arkivet portal at https://www.diva-portal.org/smash/project.jsf?dswid=1404&pid=project%3A1948

References

Aloia, Nicola, Franca Debole, Achille Felicetti, Ilenia Galluccio, and Maria Theodoridou. 2017. "Mapping the ARIADNE Catalog Data Model to CIDOC CRM: Bridging Resource Discovery and Item-Level Access." *Ricerca Scientifica e Tecnologie Dell'Informazione* 7 (1): 1–8. https://doi.org/10.2423/i22394303v7n1p1.

Andersson, Lisa, Isto Huvila, and Olle Sköld. 2024. "An Introduction to Paradata." In *Perspectives on Paradata*, edited by Isto Huvila, Lisa Andersson, and Olle Sköld, 1–14. Knowledge Management and Organizational Learning. Cham: Springer. https://doi.org/10.1007/978-3-031-53946-6_1.

Anichini, Francesca, and Gabriele Gattiglia. 2015. "Verso la rivoluzione. Dall'Open Access all'Open Data: La pubblicazione aperta in archeologia." *PCA European Journal of Post-Classical Archaeologies*, 5:299–326. http://www.postclassical.it/PCA_Vol.5.html.

Beck, Anthony, and Cameron Neylon. 2012. "A Vision for Open Archaeology." *World Archaeology* 44 (4): 479–97. https://doi.org/10.1080/00438243.2012.737581.

Bollacker, Kurt D. 2010. "Avoiding a Digital Dark Age: Data Longevity Depends on Both the Storage Medium and the Ability to Decipher the Information." *American Scientist* 98 (2): 106–10. https://doi.org/10.1511/2010.83.106.

Börjesson, Lisa, Isto Huvila, and Olle Sköld. 2022. "Information Needs on Research Data Creation." *Information Research* 27 (Special Issue). https://doi.org/10.47989/irisic2208.

Börjesson, Lisa, Olle Sköld, Zanna Friberg, Daniel Löwenborg, Gísli Pálsson, and Isto Huvila. 2022. "Re-purposing Excavation Database Content as Paradata: An Explorative Analysis of Paradata Identification Challenges and Opportunities." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6 (3): 1–18. https://doi.org/10.18357/kula.221.

Börjesson, Lisa, Olle Sköld, and Isto Huvila. 2020. "Paradata in Documentation Standards and Recommendations for Digital Archaeological Visualisations." *Digital Culture and Society* 6 (2): 191–220. https://doi.org/10.14361/dcs-2020-0210.

Bruseker, George, Martin Doerr, and Maria Theodoridou. 2017. *D5.1. Report on the Common Semantic Framework*. PARTHENOS.

Cameron, Scott, Patricia Franks, and Babak Hamidzadeh. 2023. "Positioning Paradata: A Conceptual Frame for AI Processual Documentation in Archives and Recordkeeping Contexts." *Journal on Computing and Cultural Heritage* 16 (4): 1–19. https://doi.org/10.1145/3594728.

Carrillo Gea, Juan M., Ambrosio Toval, José L. Fernández Alemán, Joaquín Nicolás, and Mariano Flores Gutiérrez. 2013. "The London Charter and the Seville Principles as Sources of Requirements for E-archaeology Systems Development Purposes." *Virtual Archaeology Review* 4 (9): 205–11. https://doi.org/10.4995/var.2013.4275.

D'Andrea, A. 2023. *I dati archeologici nella società dell'informazione*. Naples: UniorPress.

Davet, Jeremy, Babak Hamidzadeh, and Patricia Franks. 2023. "Archivist in the Machine: Paradata for AI-based Automation in the Archives." *Archival Science* 23 (2): 275–95. https://doi.org/10.1007/s10502-023-09408-8.

DCC (Digital Curation Centre). 2017. *An Analysis of Open Data and Open Science Policies in Europe, May 2017*. Apeldoorn: SPARC Europe & DCC.

Durost, Sébastien, Guillaume Reich, and Jean-Pierre Girard. 2022. "Terminologies, modèles de données archéologiques et thésaurus documentaires: réflexions à partir d'une typologie de céramiques." *Humanités numériques* 6, article 6. https://doi.org/10.4000/revuehn.3119.

Eco, Umberto. 2009. *Vertigine della lista*. Milan: Bompiani.

Faniel, Ixchel M., and Elizabeth Yakel. 2017. "Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation." In *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository*, edited by Lisa R. Johnston, 103–26. Chicago: ACRL.

Faniel, Ixchel M., Rebecca D. Frank, and Elizabeth Yakel. 2019. "Context from the Data Reuser's Point of View." *Journal of Documentation* 75 (6): 1274–97. https://doi.org/10.1108/JD-08-2018-0133.

Gant, Stefan, and Paul Reilly. 2017. "Different Expressions of the Same Mode: A Recent Dialogue between Archaeological and Contemporary Drawing Practices." *Journal of Visual Art Practice* 17 (1): 100–120. https://doi.org/10.1080/14702029.2017.1384974.

Geser, Guntram, and Franco Niccolucci. 2016. *D2.4: Final Innovation Agenda and Action Plan*. Report. ARIADNE.

Gilby, Emma, Matthias Ammon, Rachel Leow, and Sam Moore. 2022. *Open Research and the Arts and Humanities: Opportunities and Challenges*. Report. University of Cambridge Working Group on Open Research in the Arts and Humanities. https://doi.org/10.17863/CAM.86734.

Huggett, Jeremy. 2012. "Promise and Paradox: Accessing Open Data in Archaeology." In *Proceedings of the Digital Humanities Congress 2012 (Sheffield, UK, 6–8 September 2012)*, edited by Clare Mills, Michael Pidd, and Esther Ward. Sheffield: Humanities Research Institute.

Huvila, Isto. 2020. "Information-making–related Information Needs and the Credibility of Information." *Information Research* 25 (4). https://doi.org/10.47989/irisic2002.

Huvila, Isto. 2022a. "Making and Taking Information." *JASIST* 73 (4): 528–541. https://doi.org/10.1002/asi.24599.

Huvila, Isto. 2022b. "Improving the Usefulness of Research Data with Better Paradata." *Open Information Science* 6 (1): 28–48. https://doi.org/10.1515/opis-2022-0129.

Huvila, Isto. 2025. "A Paradata Reference Model." In *Paradata: Documenting Data Creation, Curation and Use*, edited by Isto Huvila, Lisa Andersson, Zanna Friberg, Ying-Hsang Liu, and Olle Sköld. Cambridge: Cambridge University Press.

Huvila, Isto, Lisa Andersson, Zanna Friberg, Ying-Hsang Liu, and Olle Sköld. Forthcoming. *Paradata: Documenting Data Creation, Curation and Use*. Cambridge University Press.

Huvila, Isto, Lisa Andersson, and Olle Sköld. 2022. "Citing Methods Literature: Citations to Field Manuals as Paradata on Archaeological Fieldwork." *Information Research* 27 (3). https://doi.org/10.47989/irpaper941.

Huvila, Isto, Lisa Andersson, and Olle Sköld. 2024a. "Patterns in Paradata Preferences Among the Makers and Reusers of Archaeological Data." *Data and Information Management* 8 (4): 100077. https://doi.org/10.1016/j.dim.2024.100077.

Huvila, Isto, Lisa Andersson, and Olle Sköld, eds. 2024b. *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge*. Knowledge Management and Organizational Learning. Cham: Springer. https://doi.org/10.1007/978-3-031-53946-6.

Huvila, Isto, Lisa Andersson, Olle Sköld, and Ying-Hsang Liu. 2024. "Data Makers and Users' Views on Useful Paradata: Priorities in Documenting Data Creation, Curation, Manipulation and Use in Archaeology." *International Journal of Digital Curation* 19 (1): 1–24. https://doi.org/10.2218/ijdc.v19i1.892.

Huvila, Isto, Lisa Börjesson, and Olle Sköld. 2022. "Archaeological Information-making Activities According to Field Reports." *Library & Information Science Research* 44 (3): 101171. https://doi.org/10.1016/j.lisr.2022.101171.

Huvila, Isto, and Stefan Ekman. 2024. "Documentation of Data Making, Processing and Use Facilitates Future Reuse of Research Data: The CAPTURE Project." In *Huminfra Conference (Gothenburg, Sweden, 10–11 January)*, edited by Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn, Lars Ahrenberg, and Anna Blåder, 26–30. Linköping: Linköping University. https://doi.org/10.3384/ecp205004.

Huvila, Isto, Heidi Enwald, Kristina Eriksson-Backa, Ying-Hsang Liu, and Noora Hirvonen. 2022. "Information Behavior and Practices Research Informing Information Systems Design." *JASIST* 73 (7): 1043–57. https://doi.org/10.1002/asi.24611.

Huvila, Isto, Jane Greenberg, Olle Sköld, Andrea Thomer, Ciaran Trace, and Xintong Zhao. 2021. "Documenting Information Processes and Practices: Paradata, Provenance Metadata, Life-Cycles and Pipelines." *Proceedings of the Association for Information Science and Technology* 58 (1): 604–9. https://doi.org/10.1002/pra2.509.

Huvila, Isto, Amalia Juneström, Jessica Kaiser, and Olle Sköld. 2024. "Dokumentation av arkeologiska arbetsprocesser underlättar framtida återanvändning av forsknings- och

undersökningsdata: CAPTURE-projektet." *Gjallarhorn* 43 (1): 28–32.

Huvila, Isto, Michael Olsson, Olle Sköld, Jessica Kaiser, and Lisa Andersson. Forthcoming. "Being Literate on Data or Practices: How Paradata Functions in the Context of Literacy." In *Proceedings of the CoLIS 2025 Conference (Glasgow, Scotland, 2–5 June 2025)*.

Huvila, Isto, and Luanne Sinnamon. 2022. "Sharing Research Design, Methods and Process Information in and out of Academia." *Proceedings of the Association for Information Science and Technology* 59 (1): 132–44. https://doi.org/10.1002/pra2.611.

Huvila, Isto, Olle Sköld, and Lisa Andersson. 2023. "Knowing-in-Practice, Its Traces and Ingredients." In *The Posthumanist Epistemology of Practice Theory: Re-Imagining Method in Organization Studies and Beyond*, edited by Michela Cozza and Silvia Gherardi, 37–69. Cham: Palgrave MacMillan.

Huvila, Isto, Olle Sköld, and Lisa Börjesson. 2021. "Documenting Information Making in Archaeological Field Reports." *Journal of Documentation* 77 (5): 1107–27. https://doi.org/10.1108/JD-11-2020-0188.

Kansa, Eric, and Sarah Whitcher Kansa. 2011. "Toward a Do-It-Yourself Cyberinfrastructure: Open Data, Incentives, and Reducing Costs and Complexities of Data Sharing." In *Archaeology 2.0: New Approaches to Communication and Collaboration*, edited by Eric C. Kansa, Sarah Whitcher Kansa, and Ethan Watrall, 57–91. Los Angeles: UCLA Cotsen Institute of Archaeology.

Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, D.C.: CLIR.

Lambourne, Gail, Lyndsey Stoakes, May Cassar, Koenraad Van Balen, Martin Rhisiart, Meirion Thomas, Riel Miller, and Liz Burnell. 2014. *Strategic Research Agenda*. Rome: JPI Cultural Heritage and Global Change. https://www.nim.gov.pl/files/articles/217/SRA-def.pdf.

Lins, Maike, and Hans-Georg Becker. 2013. "Open Data und Linked Data in einem Informationssystem für die Archäologie." In *(Open) Linked Data in Bibliotheken*, edited by Patrick Danowski and Adrian Pohl, 201–23. Berlin, Boston: De Gruyter Saur. https://doi.org/10.1515/9783110278736.201.

Mayernik, Matthew S., and Amelia Acker. 2018. "Tracing the Traces: The Critical Role of Metadata within Networked Communications." *JASIST* 69 (1): 177–80. https://doi.org/10.1002/asi.23927.

Mayernik, Matthew S., David L. Hart, Keith E. Maull, and Nicholas M. Weber. 2017. "Assessing and Tracing the Outcomes and Impact of Research Infrastructures." *JASIST* 68 (6): 1341–59. https://doi.org/10.1002/asi.23721

Rösler, Katja. 2020. "Kontext, Funktion, Transparenz und Zugang: Anmerkungen zur Publikation von Datensammlungen im WWW am Beispiel von Artefacts.mom.fr." *Germania* 98: 210–14.

Schenk, Patrick Oliver, and Simone Reuß. 2024. "Paradata in Surveys." In *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge*, edited by Isto Huvila, Lisa Andersson, and Olle Sköld. Knowledge Management and Organizational Learning. Cham: Springer.

Sköld, Olle, Lisa Börjesson, and Isto Huvila. 2022. "Interrogating Paradata." *Information Research* 27 (Special Issue). https://doi.org/10.47989/colis2206.

Voss, Barbara L. 2012. "Curation as Research. A Case Study in Orphaned and Underreported Archaeological Collections." *Archaeological Dialogues* 19 (2): 145–69. https://doi.org/10.1017/s1380203812000219.

TRANSFORMATIONS
A DARIAH Journal

Genre Classification Workflow for the English Short Title Catalogue (ESTC)

Iiro Tiihonen and Kira Hinderks

University of Helsinki

Transformations, A DARIAH Journal
Volume 1, 2025
https://transformations.episciences.org

Dates
Received: 15/11/2024
Accepted: 07/04/2025
Published: 11/06/2025

DOI: 10.46298/transformations.14754

Abstract

This article introduces an open-box workflow for labelling 94 percent of the English Short Title Catalogue (ESTC) with a unified genre classification scheme, as well as an approach to evaluating the classifications. As the ESTC covers most of the surviving books published in the early modern Anglosphere, our categorisation offers new opportunities for large-scale quantitative research on the early modern book trade. Our evaluation process directly engages with the ambiguity of any genre labelling or annotation schemes of early modern books and highlights problematic boundaries between the categories. We also provide summary statistics about the genre composition of the ESTC, demonstrate how the new data can be used to detect biases in other datasets of early modern books and discuss further possibilities for genre-related computational work with the ESTC.

Keywords: book history, classification, computational history, digital humanities, evaluation process, workflow

EPI sciences

To cite this article: Iiro Tiihonen and Kira Hinderks (2025). Genre Classification Workflow for the English Short Title Catalogue (ESTC). *Transformations. A DARIAH Journal*, vol. 1, p. 1-33. https://doi.org/10.46298/transformations.14754

ESTC-tietueiden genreluokittamisen työvuo

Abstrakti

Tämä artikkeli esittelee avoimen työvuon 94%. ESTC-tietueista kategorisointiin yhtenäisellä luokittelujärjestelmällä sekä lähestymistavan näiden luokittelujen arviointiin. Kategorisointi mahdollistaa uusia laaja-alaisia analyyseja, sillä ESTC kattaa suurimman osan Britti-imperiumissa tai englanniksi varhaismodernina aikana julkaistuista kirjoista. Arviointiprosessimme käsittelee ongelmia, joita varhaismodernien teosten luokitteluun väistämättä liittyy. Lisäksi esittelemme tilastoja ESTC:n genrejakaumasta, osoitamme uuden aineston höydyn muiden varhaismodernien kirja-aineistojen vinoumien tutkimisessa ja keskustelemme siitä, millaiset laskennalliset lähestymistavat ESTC-teosten genrejen analysointiin voivat olla tulevaisuudessa mahdollisia.

Avainsanat: kirjahistoria, luokittelu, laskennallinen historia, digitaalinen historia, arviointiprosessi, työvuo

Data availability statement

The ESTC data was obtained by the COMHIS group via agreement from the British Library and University of California Riverside. In accordance with this agreement, the enriched and harmonised ESTC (including the full genre information described here) will be released in a separate data publication in the near future. The code (workflow, its evaluation and analysis) related to this article and the datasets that we can share publicly are available at https://github.com/COMHIS/Genre_classification_workflow_transformations_public_2025. Data not included in the public repository will be provided for replication purposes upon request.

We thank the Finnish Cultural Foundation and the Research Council of Finland (grants no. 347709 and 333716) for funding the work on this article.

Authorship statement

Iiro Tiihonen: Conceptualisation (lead), Data curation (equal), Formal analysis, Investigation (equal), Methodology, Software, Visualisation, Writing—original draft (equal), Writing—reviewing & editing (equal). Kira Hinderks: Conceptualisation (supporting), Data curation (equal), Investigation (equal), Writing—original draft (equal), Writing—reviewing & editing (equal). This authorship statement uses the CRediT NISO standard.

This article is part of the ongoing and future-focused efforts of the Helsinki Computational History Group (COMHIS) to harmonise, standardise and enrich the ESTC. This work builds on previous initiatives within the group and lays the groundwork for continued advancements in the field.

1 Introduction

The English Short Title Catalogue (ESTC) is a union catalogue that covers most of the surviving books from the start of the hand press era to the end of the eighteenth century (1454–1800), printed either in the British Isles, Ireland, the colonies of the British Empire or in English anywhere.[1] While some item types (e.g. bookplates; Raven 2018), are missing or unsystematically presented, the ESTC remains the most comprehensive source for conventional books printed in the Anglosphere during the early modern period. The main bulk of the almost 500,000 records represent seventeenth- and especially eighteenth-century books printed in Britain, in English, with the latter century also including tens of thousands of vernacular records from Ireland and North America. Publications in Latin, French or printed on the European continent also number in the several thousands (for a more detailed description of the contents, see for example Lahti et al. 2019). By default, ESTC records can be queried with an interface provided by the British Library. At the time of writing, the interface was inaccessible due to a cyberattack, but a collaborative project (Print and Probability) between researchers from Carnegie Mellon University and University of California San Diego provides access to a fairly up-to-date version of the ESTC data.[2] Additionally, a beta version of the ESTC, managed jointly by the British Library and ESTC North America, is now available online through the Consortium of European Research Libraries (CERL).[3]

Since books as material objects have survived exceptionally well,[4] and the number of copies per edition varies far less than in the modern era,[5] the data in the ESTC enables a far more comprehensive and interpretable analysis of the objects it represents than most historical datasets. This article presents a workflow for classifying 94 percent of ESTC records with a unified categorisation scheme, describes how the resulting data was evaluated and summarises useful information about the genre distribution of the ESTC and its subsets. We devote significant attention to a systematic close reading of the labels suggested by the workflow and demonstrate how errors and ambiguities can offer insights that are historically and methodologically valuable. We add new

1. A union catalogue covers the records of multiple libraries.
2. https://estc.printprobability.org/
3. https://datb.cerl.org/estc/
4. The claimed survival rates (defined as at least one copy of an edition surviving and making it to the ESTC) vary from 55 percent for the period before the English Civil War (Hill 2018) to 70–75 percent for 1640–1700 (Raymond 2003; McKenzie 2002) and 90 percent for the eighteenth century (Suarez 2009). Despite the varying levels of evidence, assumptions and pure guesswork behind these numbers, it is at least safe to say that books as material objects have survived exceptionally well.
5. There was significant variation in the number of copies printed per edition in the early modern English-speaking world (Jenner 2011; Watt 1990; Simmons 2002), but less so than in the modern era. Roughly 1,000 copies per edition is often presented as a typical number (Raymond 2011; Gants 2002; Hill 2018).

and comprehensive evidence for the argument that the proportion of religious publications declined over time, find a permanent jump in the proportion of political publications during the English Civil War that warrants closer scrutiny, and systematise the analysis of biases in Eighteenth Century Collections Online (ECCO), a significant subset of the ESTC.

Much of the information in the ESTC (publication year, publisher information, physical description of the edition, etc.) has already been extensively processed to a form suitable for data analysis, most notably by the Helsinki Computational History Group (COMHIS) (Lahti et al. 2019). This work on the ESTC is part of the COMHIS effort to harmonise and enrich early modern bibliographic (meta) data collections to make them more useful for historical research. Bibliographic data typically needs preprocessing before rigorous computational analysis is possible (Lahti et al. 2019). Additionally, the work done by COMHIS facilitates easier use of bibliographic data in traditional qualitative history research, such as analyses of text reuse in the early modern period (Rosson et al. 2023).

The workflow described here is a recent addition to this long-term effort by COMHIS. The HPC-HD project—a collaboration between COMHIS and several computer science research groups—has significantly extended the large-scale analysis of language in subsets of the ESTC (most notably ECCO) with state-of-the-art machine learning methods, and created computational approaches for various linguistic phenomena, genre included (Zhang et al. 2022). In addition to this wider context, an article that considered books as commodities that vary by type (Tiihonen, Lahti, and Tolonen 2024) prompted the development of this workflow and employed an earlier version of the outputs.[6] This article introduces the (revised and improved) workflow for the first time, extends evaluation from a short discussion point in the appendix to a topic in its own right, and provides new information about the generated data. As there is a plan to release the genre data as part of the COMHIS-version of the ESTC in the near future, this article also serves as documentation for the genre variables.

Our conceptualisation of early modern genres is intentionally a very simple one: genre is limited to a single main label per edition (a debatable choice as a unit of text in itself), taken from a pool of very limited choices, and we do not consider many forms of information like visual cues (e.g. images) or complex embedding representations of texts. As we will discuss in the article, many other conceptualisations and corresponding computational paths could and should be explored as well. However, this simplicity also enables effective interpretation of both the process and the results. Moreover, our evaluation of the outputs enables us not only to verify that the labels are largely correct, but

6. For example, almanacks and multi-volume dictionaries would have differed significantly in terms of price, scope of potential buyers and utility.

also to identify potentially fruitful directions for further computational work on early modern genres. Instead of aiming to establish a single permanent and fixed conceptualisation of genre, the workflow presented here is a specific tool for specific applications, which hopefully aids with the development of other classification workflows.

Our workflow has four major advantages compared to preceding efforts to categorise editions in the ESTC. First, it covers virtually the entirety of the ESTC. Second, it is eclectic and leverages the many possible sources of information available, ranging from a large pool of manually annotated documents to established conventions of early modern book titles (e.g. "sermon preached at X"), instead of aiming for a single solution of typically inferior trustworthiness that is intended to cover the entire population to be classified. Third, it is open box. All operations related to classification can be communicated to an audience with little technical expertise. Our eclectic and open-box classification offers an alternative to increasingly employed black-box methods which, although capable of constructing complex decision processes (e.g. about genre), have special challenges in generalising reliably outside of the training data (Kapoor and Narayanan 2023).

Fourth, our evaluation process is tailored to suit the ambiguous nature of many eighteenth-century publications. Instead of having a "gold standard" test set, as is typical in machine learning or statistics, we manually evaluated a large sample of records to assess the relative quality of the label. In addition to clearly right (1) or wrong (0) classifications, we added the category of plausible (0.5), which, as it turns out, covers most of the "errors" in the classified data. Our results highlight that certain genre and topic distinctions are not bugs but features of the early modern publishing landscape: documents at the intersection of politics, religion and law, for example, are inherent parts of the print culture of the era. Instead of forcibly resolving and smoothing historical genre ambiguities, we aim to quantify their size and historically situate their patterns. As analyses of historical bibliographic data are not limited to the ESTC, we hope that our workflow also encourages other researchers working with similar material to consider whether their needs could be served by a simple and transparent workflow for classification and evaluation utilising humanities and resource-specific expertise rather than (or in addition to) the latest computational methods.

This paper is structured as follows: section 2 highlights the limitations of the ESTC's native genre classification scheme and discusses previous scholarly efforts to mitigate these limitations, for example by creating new classification schemes. This section also examines why the use of unsupervised machine learning models can be problematic for genre classification and stresses the need for detailed process descriptions in articles that introduce new classification

schemes. Section 3 briefly explains the terminology required for understanding the composition and structure of the ESTC and introduces our genre-labelling scheme. Section 4 describes our classification workflow, namely the steps for creating category labels. Section 5 discusses our evaluation of the accuracy of the generated labels. This section also highlights difficulties and pertinent category overlaps, showing that what at first glance appeared to be an error was often a genuine reflection of early modern print culture, where genres were far less distinct compared to the twenty-first century. The sixth section analyses the genre distribution in the entire ESTC, while the concluding section offers suggestions for refinement of our workflow and discusses potential directions for future research on genre classification.

2 The complexity of genre: Previous classification schemes of the ESTC and its subsets

Genre is a complex and multifaceted phenomenon and there is no one correct classification scheme that captures all its nuances in different contexts. The most appropriate scheme is highly dependent on the material and research questions. For historical data, in almost all cases, it is better to use a classification scheme that tries to reflect as closely as possible how historical actors understood genre. However, the genre and subject categories present in early modern union catalogues are often based on library cataloguing schemes that were developed in the nineteenth century, such as the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC). These schemes were created first and foremost to fit the needs of libraries, which differ from the objectives of a scholar wishing to analyse genre distribution patterns in the early modern period. The DDC and LCC reflect nineteenth-century understandings of genre and may impose anachronistic categories on earlier periods or fail to capture genres and subjects that were prevalent in one period but fell out of fashion later. Taking existing classification schemes as is—especially without a systematic evaluation of their biases—can therefore lead to the reproduction of fundamental issues.

Early modern scholars are usually well aware of the limitations of the genre classification schemes native to the ESTC and other union catalogues, but may continue to use them because genres are not their main object of study or because there exists no other option that would be historically more accurate. For instance, one of the earliest quantitative analyses of genre distribution in early modern Britain was based on genre categories taken from the DDC (Feather 1986). The reason for this was that Feather's dataset, the Eighteenth Century British Books (ECBB) short-title union catalogue, used the DDC. A rival catalogue to the ESTC, the ECBB was born out of the Project for Historical Biobibliography (PHIBB) at the University of Newcastle and was completed in

1981. Due to its smaller scope and more limited coverage, the ECBB has been dubbed "the poor library's ESTC" (Barr 1981, 229) and reviewers—some of whom were involved in the then-ongoing development of the ESTC—were only too keen to point out the ECBB's flaws, including the anachronistic elements present in the DDC (Snyder 1983; Amory 1983). In his quantitative analysis of the ECBB, Feather sought to curb the DDC's worst excesses by only including high-level subject categories. Despite Feather's best efforts, some issues remain unresolved: the category of "social sciences," the second most frequent in Feather's analysis, is especially problematic. Its individual components are never explained, and the concept, although coined in France in the 1770s (*sciences sociales*), did not become a distinct tradition until the nineteenth century (see for example Baker 1964).

This type of anachronism is not only present in catalogues containing nineteenth-century genre classification schemes; it also affects catalogues whose schemes were developed in the twentieth century. One such example is Eighteenth Century Collections Online (ECCO), the most extensive subset of the ESTC, for which records have been classified by topic. ECCO covers roughly half of the eighteenth-century records of the ESTC and crucially includes digitised and machine-readable versions of the full text of the editions that it covers. Gale, the company that provides access to ECCO, has labelled all records in ECCO with its own classification scheme. Unfortunately, the classification scheme does not suit eighteenth-century records well, since many of the categories do not correspond with historical understandings of genre or fail to capture the granularity of topic differences of the period.[7] The categories of "social sciences" and "religion and philosophy" are especially problematic in an eighteenth-century context, as they amalgamate very heterogeneous text types.

Scholars have sought to address this problem in multiple ways, with some mitigation strategies more successful than others. Under the assumption that a fully data-driven process can eliminate human biases, unsupervised machine learning methods for classification purposes have become increasingly popular: recent examples include the use of topic modelling on the HathiTrust Digital Library (Almelhem et al. 2023), EEBO (Grajzl and Murrell 2024) and Goldsmiths'-Kress Collection (Tiihonen et al. 2022). The term "topic model" itself is misleading, as it can lead users not familiar with the mathematics of topic models to assume that there is something inherently topic-like in the distributions generated by a topic model. This is not the case, as traditional topic models only generate distributions over units of text, and to our knowledge, other unsupervised methods cannot make sweeping ontological claims either. The general sensitivity of traditional topic models to particulars of

7. ECCO also gives an unbalanced representation of topic distribution in the ESTC (Tolonen, Mäkelä, and Lahti 2022), and we extend the analysis of this issue in the paper.

data preprocessing and the dangers of "eyeballing" interpretations of "topics" (Shadrova 2021; Gillings and Hardie 2023) are highly undesirable qualities when working with historical data. It is hard to see how generating categories with unsupervised methods in general could fully escape these problems.

To better represent early modern conceptions of genre in a way that is most suitable for their own research questions, many scholars have created annotated subsets of the ESTC that focus on specific periods or certain groups of authors (Corns 1986; Gants 2002; Fielding and Rogers 2017; Hill 2018). However, not all publications report the extent to which—or even whether—harmonisation and unification have been performed. Among studies that introduce new classification schemes, a recurring issue is that descriptions of the workflow are sometimes absent or underdeveloped. Rationales for the creation of certain labels and an evaluation of the genre data quality in relation to the existing scholarship are also frequently omitted. For instance, Suarez (2009) proposes an 11-category genre classification scheme for the eighteenth-century part of the ESTC, but does not explain whether this scheme is a modification of the original ESTC subject classification or whether it was developed from scratch. Suarez notes that this genre annotation was performed on a relatively small subset of the ESTC (ca. 24,000 records), but does not systematically assess the coverage of this subset or the precision of his proposed genre categories. Such omissions make it more difficult for readers to understand which research questions can be answered with the help of the newly developed genre classification scheme. At the opposite end of the spectrum are schemes that are highly descriptive but far too reductive to meet the scholarly requirements of historians. Almelhem et. al (2023) map the most common (0.01% of all records using it) subject keywords in the ESTC to just three categories (religion, political economy, science) and use these to study the aggregate genre composition of the ESTC records published in Britain in English. However, as the aim of this study was to test for specific biases in the HathiTrust Digital Library (HTD) data rather than to label the ESTC,[8] it also demonstrates the increasingly varied and multi-disciplinary interest towards ESTC and its genre composition.

There are two long-standing scholarly traditions of developing and refining genre classification schemes in sync with other bibliographic work on the most important subsets of the ESTC: Early English Books Online (EEBO) at Lancaster University (e.g. Murphy 2019) and ECCO at the University of Helsinki (e.g. Zhang et al. 2022). Both traditions are examples of successfully integrating humanities domain knowledge with sophisticated computational methods. In contrast to unsupervised or Large Language Model-driven approaches, here, the humanities scholar is the prime mover of the process: for example by

8. Appendix figure C.1 in Almelhem (2023) suggests that less than 15% of the ESTC corpus maps to these three categories with the approach they used.

manually annotating large data samples with genre labels, drawing on one's domain knowledge and expertise. By providing a way of classifying genre in the whole ESTC, the workflow introduced in this paper builds on and contributes to this rich tradition.

3 Definitions and taxonomies

3.1 Data vocabulary

This article makes heavy use of terminology designed to distinguish different types of data in the ESTC. For clarity, this section briefly introduces that terminology. The basic unit of the ESTC is a record representing one "printing operation," that is, a set of copies of a given text printed as part of the same process. In some more complicated situations, this definition and record do not perfectly correspond to each other.[9] In our vocabulary, records are called editions. The original ESTC records follow the MARC (Machine-Readable Cataloging) format for bibliographic data.[10] The ESTC records have been harmonised, standardised and enriched (Lahti et al. 2019) to a format more suitable for data analysis. Of the raw MARC data fields, the classifications described here rely primarily on the fields that contain the title of an edition (245a and 245b), information about bibliographies that cite a given edition (510a), references to microfilm collections that include a copy of a record (533f), and any fields that contain keywords describing the genre, topic and format of an edition (600, 610, 650, 651, 655). The keywords and full titles were converted to lowercase text in the classification workflow.

In addition to editions, we sometimes refer to works. A work is a grouping of all editions with more or less the same textual content. So, for example, there are several editions of Adam Smith's *The Wealth of Nations* in the ESTC with their distinct edition-level-ID, but they should map to the same work and corresponding work-ID. The ESTC does not hold comprehensive information about works as such, but this information has been algorithmically derived by the COMHIS group based on the title and author information in the ESTC (Ijaz, Roivainen, and Lahti 2019).

Among other information, ESTC editions have a title. Early modern titles were often very long, some even taking the whole first page of the publication. For this reason, they often also have a "short title" (the "ST" in ESTC) for easier

9. For example, multiple issues of a periodical are often stacked under the same record and multi-volume works were sometimes printed in pieces over a long period of time. They are sometimes still grouped together under a single record.
10. https://www.loc.gov/marc/bibliographic/

communication. In the classification workflow, we used the longer version of the title, as it provided more information about the topic of an edition.

The ESTC does not have a unified topic or genre classification, but many of the editions have information that is very useful for assigning such a classification. Perhaps most importantly, there are a great number of keywords relating to genre, topic and format that are part of the information characterising records. These keywords like "sermon," "political controversy" or "anatomy" originate in library cataloguing schemes. As such, they are problematic. There are thousands of unique keywords that are used in the ESTC, yet they are used unsystematically (some records have no keywords and others have many) and their relevance for historical research varies. For example, the keyword "early works to 1800" is very common in the ESTC (at least 12,000 records), but virtually useless for our purposes (it applies to all records). Filtering, mapping and manual curation is required in order to construct historically motivated higher-level categories (e.g. from thousands of varying relevancy to a handful of very relevant ones) suitable for our purposes, which enables the creation of analyses like those displayed in figures 4 and 5. When referring to topic, genre and format keywords in the ESTC, we mean this information.

Additionally, editions in the ESTC often refer to bibliographies or microfilm collections that cite or contain photocopies of the edition. For example, the Goldsmiths'-Kress microfilm collection (Whitten 1978) covers publications relevant for economic history, while *History of English Drama 1660–1900: Volume 6, A Short-title Alphabetical Catalogue of Plays* by Allardyce Nicoll (1959) lists plays. When we refer to microfilm and bibliography collection information in the ESTC, we mean this kind of data.

3.2 Labelling scheme

There are three hierarchical levels in our classification scheme: the supercategory, the main category and the subcategory. Of these, the middle one **(main category)** is the most important. It is a modest modification of the categorisation scheme used in previous articles by the Helsinki Computational History Group (COMHIS) (Zhang et al. 2022; Ryan and Tolonen 2024), and which was also used to classify early modern editions of the ESTC. This categorisation scheme is the result of a thorough investigation of all known books published by Andrew Millar, a prominent eighteenth-century publisher. Rooted in scholarship about early modern book history and designed for the era and location comprising most of the ESTC data (eighteenth-century British Empire), we adopted it and only made minor modifications. The most important difference is that, in our scheme, the physical size of the edition affects its classification (ephemera-entertainment and ephemera-practical only include short editions), and the categories of literature and art have been merged (literature and arts).

In this sense, it is more extensive than most genre-like classification schemes (in comparison to strictly topic-based classifications), as it also considers the material cues (physical size) given by the publication about its category.

The reasoning behind the size-related classifications is that, in some cases, size is an important part of the publication's type. For example, short publications that dealt with scientific matters like astronomy tended to lean more towards immediate practical uses (e.g. almanacs) than longer works with superficially similar contents. As for literature and arts, we concluded that it was too difficult and often not meaningful to distinguish between texts related to performing arts (poetry, plays) and prose; they are often convoluted in eighteenth-century publications (e.g. different kind of miscellanies that contain both), making the distinction very artificial. The subcategories aim to compensate for this merger by specifying the type of publication (e.g. play, poem or novel) with more precision whenever possible.

Table 1: List of the main categories of the classification scheme

Main category	Includes the following publication types
Religion	Bibles, sermons, prayers, theology, religious debates (can also be categorised as politics in some instances), psalms and hymns.
Literature and arts	Periodicals, novels, stories, visual arts, plays, poems, songs, pictures, operas, satires.
Ephemera-entertainment	Chapbooks, broadside poems. Short items (pamphlets) that would otherwise be included in literature and arts, philosophy or history.
Politics	Political pamphlets, societal discussion, news about wars, foreign politics, etc. (note: periodicals are still in literature), commerce, social issues (e.g. poverty).
Law	Legal and administrative documents, legal theory. As a rule of thumb, commentary and debate of legislation are assigned to politics; and description and statement are mapped to law.
Science	Mathematics, natural sciences, agriculture, technology, military organisation and building, logistics (e.g. canals, etc.), natural history, travel literature. Overlaps with literature and arts, notably because of travel literature. Description of facts leans more towards science and satirical, and/or entertaining purposes towards literature.
Education	Manners, upbringing, manuals that instruct on doing something, grammars, primers, dictionaries, conduct of life. Overlaps with topic-related categories (e.g. introduction to mathematics).

Main category	Includes the following publication types
Philosophy	Logic, metaphysics, ethics, political theory, aesthetics, psychology. Overlaps with politics, religion and science in particular. Conceptual/theoretical style of argumentation favours inclusion in this category.
Ephemera-practical	Almanacs, calendars, advertisements, notifications, pamphlet-sized education and science.
History	Historical works, biographies, description of events. Overlaps with literature, as many literary works also take the form of a memoir or "history." Overlaps with politics, as there is no clear-cut difference between news and recent history.

It is non-exhaustive, as there are always editions that do not fall neatly into any category.

Subcategories provide additional information about the publication in the form of keywords that specify some of its aspects in more detail. There can be multiple subcategories per record and they are separated by semicolons. For example, "poetry" and "commerce" are keywords appearing in the data. In total, 216,000 editions have at least one subcategory. Subcategories are not strictly related to any specific main category, but they often have a strong affiliation with certain main categories. For example, there are 22,000 editions with the subcategory "poem" in the main category "ephemera-entertainment" and 11,000 in the main category "literature and arts." This makes sense, because by default poems belong to these categories. However, there are also 810 editions with this subcategory in the main category "religion"; this also makes sense in an eighteenth-century context, when religious texts with varying amounts of poetic elements were abundant. Table 2 lists examples of subcategories, with the complete list provided in appendix A.

Table 2: Examples of subcategories and their typical relations to main categories

Subcategory	Typical for main category
Act, bill	Law
Theatre programme, almanac	Ephemera-practical
Petition	Politics
Sermon, prayer book	Religion
Metaphysics, logic	Philosophy
Biography, ancient history	History
Agriculture, medicine	Science
Play, poem	Literature and arts
Broadside poem, chapbook	Ephemera-entertainment

Supercategories aggregate main categories to form even more extensive groups. By using supercategories, one can avoid certain problems caused by systematic convolutions between main categories. For example, if there is no reason to distinguish political pamphleteering from legislative documents, it can be useful to treat them as one category of jurisprudence, which is not affected by the heavy tendency of politics to "bleed" into the category of law at the level of the main category. Supercategories are listed in table 3.

Table 3: Supercategories of the classification scheme

Supercategory	Merges the main categories of
Jurisprudence	Law, politics
Practical	Science, ephemera-practical, education
Leisure	Literature and arts, ephemera-entertainment
Philosophy	Philosophy
History	History
Religion	Religion

4 Classification process

The dataset was created by mapping the ESTC records to main categories and subcategories (described below) in six steps. An edition that acquired a label from one step was not passed to the next one. All steps were looped via the work-ID. This means that if one edition of a work acquired a classification during any of the given steps, this classification was projected to other editions of that work as well. As a result, all editions of the same work should have the same label.

First, all works except periodicals (i.e. works with at least one labelled edition) that were manually annotated in previous COMHIS work with a similar classification scheme (e.g. Ryan and Tolonen 2024; Tiihonen, Lahti, and Tolonen 2024) or as part of this article were given a main category and, if available, subcategories based on the manually annotated data. This resulted in more than 100,000 editions with a main category. Many (30,000) of the manually annotated documents also received a subcategory.

Second, all periodicals detected based on keyword, microfilm and bibliographic data (e.g. in which bibliographies the record is being cited) in the ESTC were labelled as periodicals. That is to say, they were given the main category "literature and arts" and the subcategory "periodical."

Third, a hand-selected set of ESTC keywords about topic, genre and format of editions were mapped to corresponding main categories and subcategories, and these mappings were used to label works. In exploring potentially useful topic and genre keywords, statistical approaches were used as heuristic tools (see appendix C for details), but all keywords were manually checked. Some keywords were also mapped to a more specific subcategory (e.g. the keyword "medicine" to the subcategory "medicine" in addition to the main category "science").

Fourth, a similar mapping-to-categories approach was used with n-grams (from monograms to pentagrams) of full ESTC titles. For example, n-grams of the form "sermon preached at" or "proclamations of X" are expressions that tell us the type of a publication (religion and law) very accurately. Here too, statistical tools (see appendix C) were only used as heuristic finding aids; the final approval of an n-gram to the list of n-gram-to-category mappings was a human-made choice. Some n-grams also mapped to more specific subcategories (e.g. "sermon preached at X" to the subcategory "sermon").

Fifth, ESTC n-grams that were deemed "not good enough but potentially useful" in the fourth step were used. Sixth, keywords that were deemed "not good enough but potentially useful" in the third step were used. For example, the keyword "England" was used at the sixth step, as it often maps to political pamphlets. However, "England" itself can easily relate to non-political topics as well, so the keyword was deemed to carry greater risks than keywords like "sermon" used in the third step, which describe the edition very accurately. That is to say, the sixth step consists of riskier keywords (and the fifth step consists of risker n-grams) used as a last resort if a more precise describer of the edition is not available.

After the evaluation, main categories were mapped to supercategories that aggregated them further.

In steps three, four, five and six, it was possible for multiple plausible labels to emerge, reflecting the complex reality of early modern genres, which were often far more enmeshed than their modern equivalents. For example, an edition could have both the keywords "sermon" (religion) and "political controversy" (politics). Rather than a category error, this is an accurate reflection of the deliberate topicality of early modern sermons, which were often used to comment on political events. In these instances, the category of an edition was decided based on the prominence (prominence score) of keywords or n-grams related to different main categories (the more prominent the better) and by the number of topics and n-grams of the edition mapping to the category (the more the better). We describe this step in detail in appendix B. Each classified edition of a work then "votes" for the main category of the entire work, with its

vote being equal to the prominence score of the chosen main category. Works classified based on keywords inherit all subcategories to which the keywords of the work map and works classified based on title inherit all subcategories to which the n-grams of the title map.

Figure 1: The workflow to categorise records in the ESTC

This part of the workflow is followed by evaluation of the classifications.

5 Evaluation

5.1 Overview

The quality of the main categories has been evaluated using a mixed methods approach. Two researchers (the authors) went through a random sample of 1,000 records with main categories and 500 without (i.e. the process did not result in a main category for the record). The sample contained information about the size of the document (book, pamphlet or in-between), the part of the process that produced the label, the suggested main category (and, if available, the suggested subcategory), as well as the full title. We decided to use the full title, as eighteenth-century works often contained lengthy subtitles that acted like a table of contents and therefore added important details about the nature of a work. In this period, titles commonly described the contents of a book with genre-related words and expressions that directly map to our classification scheme (e.g. “sermon preached at” or “almanac for the year X”).

Using a metadata sheet with the titles and size information (which is needed for some classifications), the authors were able to classify many of the more straightforward cases much more quickly compared to consulting individual full texts via the ECCO interface, for example. Although the origin of the classification was included in the sheet, this information was not used as the basis for classification choices, and documents that had previously been manually annotated were also scrutinised. Sometimes, the title itself was not sufficient to determine the label. For example, ambiguous and short titles did not allow for inferences about the genre of a text. In these cases, the authors consulted the edition from the ESTC, ECCO and other sources for close reading.

Both annotators scored the labels of the 1,000-record sample using the following scheme: 0 (wrong main category), 0.5 (plausible, but not the best choice) or 1 (best choice or one among equally good choices). A score of 0 was applied when there was no good argument for assigning the particular category to a record. For example, a prayer book classified as science would have received a 0. The classification 0.5 (plausible) was used if there was a good argument for the classification, but the researcher felt that the argument was weaker than for some other category or categories. For example, a pamphlet on religious toleration may have obtained the main category "religion" in the workflow. This would not be a wrong classification per se, but (many would argue) "politics" would be even more appropriate if the pamphlet was more focused on political rights relating to religion rather than religious topics as such. In this case, 0.5 would be an appropriate score. A score of 1 was used if the category of a record was the best one available. A record could receive this score even if there was an equally good (but not better) alternative. For example, a comical poem about commercial policy could be categorised as either "politics" or "literature and arts" with equal plausibility. For records that obtained a score of 0.5 or lower, the researchers also suggested the best possible label. For records with missing labels, the researchers provided their best suggestion. If the metadata, such as size and title, gave no indication of the genre category, both annotators consulted the original text and discussed pertinent issues with each other. These strategies helped to reduce the number of records for which it was not possible to evaluate the assigned label or give an updated label. In the end, a missing evaluation score was a rare phenomenon, occurring only for 0.4 percent of our 1,000-record sample. For the 500-record sample of missing labels, we were unable (i.e. both annotators classified the record as "unknown") to assign an updated label to 5.2 percent of records.

An important aspect of the evaluation was the two authors' different degree of involvement in developing the labelling workflow. The six-step workflow was designed and implemented exclusively by the first author, while the initial role of the second author was to provide an independent evaluation of the results. The different relationships (dependent and independent) of the authors to the

workflow is a plausible explanation for the fact that the first author was somewhat more prone to "fully agree" with the annotations provided by the system compared to the second author (see table 4). Having both an "internal" and "external" evaluation perspective in terms of the workflow development led to constructive discussions about the limitations and subjectivities of the workflow.

This manually evaluated data was used to estimate the precision of the main categories (how often they are right) as well as their coverage (how well a main category covers the relevant records). Coverage also takes into account the estimated effect of unclassified and mislabelled editions. In calculating coverage, a classification was considered to be right if the sum score of the two annotators was "good" (at least 1.5/2). A "good" score can be thought of as the label receiving mostly favourable evaluations (at least one score of 1 and one score of 0.5). "Plausible" refers to a sum score that was at least 1 and "best" to a perfect 2/2 sum score. The distribution of genres among editions that did not receive a label from the workflow was calculated by averaging the suggested labels of the two annotators.[11]

Table 4: Editions from the evaluated sample that obtained a score from both annotators, aggregated by the score they received. N=996

Score (IT)	Score (KH)	*N*
1	1	807
0.5	0.5	94
0	0	44
1	0.5	41
0.5	0	5
0.5	1	4
1	0	1

The estimates of precision and coverage are summarised in table 5. For all categories, 88–100 percent of the assigned editions are at least plausible. Instances where one annotator gave a score of 0 (wholly incorrect category) and the other a score of 1 (best category) were exceedingly rare. Whenever this kind of scoring did occur, both annotators subsequently consulted the original text and afterwards discussed the reason for their respective scoring. Such discussions often revealed that one of the annotators had consulted the original text before assigning the score, whereas the other had given their score based solely on the title. These measures helped to flag and resolve extreme inconsistencies and acted as additional verification steps for borderline cases. The final distribution of scores is presented in table 4. There is more variation in terms of

11. Averaging was also used to solve instances in which the two annotators suggested a different main category for wrongly classified records.

the best category in table 5, as it fluctuates between 58 and 100 percent, with most categories having a precision greater than 80 percent and all except one greater than 75 percent.[12] The coverage of all categories is estimated to be between 57 and 89 percent. All categories mostly include the correct editions and cover most of the relevant records.

Table 5: Evaluation results of the main categories

Main category	Precision (plausible)	Precision (good)	Precision (best)	Coverage
Education	0.96	0.87	0.83	0.64
Ephemera-entertainment	0.97	0.93	0.87	0.89
Ephemera-practical	0.96	0.91	0.86	0.71
History	0.88	0.67	0.58	0.65
Law	0.93	0.82	0.78	0.89
Literature and arts	0.96	0.8	0.78	0.86
Philosophy	1	1	1	0.81
Politics	0.94	0.8	0.77	0.57
Religion	0.95	0.89	0.82	0.89
Science	1	0.88	0.88	0.76

In some categories, most or all "errors" are in fact ambiguities, as the majority are erased by extending the definition of correct category from "the best one possible" to "plausible." This is also true at an aggregate level: 81 percent of the editions in the sample obtained the best possible score, whereas 95 percent were at least plausible (see table 6 for more details). This means that almost a fifth of the assigned categories were at least somewhat contested by at least one annotator, but more than two-thirds of these contested labels were still considered to be reasonable (plausible) even if not ideal. These 0.5-point divergences often reflect the two annotators' slightly different views on whether a category label could be considered the best-fitting (and thus receive a score of 1) or whether they regarded another category as even more suitable, therefore assigning a 0.5. Overall, both annotators were of the opinion that this in-between score was a valuable addition to the scoring system, enabling a more finely grained evaluation of the precision of a category and accounting for the complexity of genre in the early modern period.

12. Philosophy, the most precise category, and history, the least precise one, are also very small categories in the sample and in the population they approximate (figure 5), making it possible that they are affected by the "curse of the small N" (i.e. their extremity is related to the small number of observations). However, the error analyses of the next section demonstrate that lack of precision in the history category likely relates to real problems in defining the boundaries of that category.

The conceptualisation of a single main category as being "right" was intended as a simplification that consciously ignores many of the complexities of early modern books. Nevertheless, we implemented tentative analyses on how often the notion of a single best-fitting category "failed." This helped us to understand how much information was lost with the simplification and how much working within its confines affected our calculations. By counting instances where the annotators had commented that an edition had multiple good choices for labels as well as other disagreements about the best label (1-0.5, 1-0 score pairs from the annotators), we found 74 editions out of 1,000 that could be assigned with equal plausibility to several main categories. Some 29 of the 500 editions from the sample without labels received different main categories from the two annotators. This is roughly 6 percent of the sample and is possibly somewhat inflated by the fact that the annotators discussed divergences less frequently compared to the other sample. Both measures probably also miss some editions that are problematic for a single-category schema, but the proportion would need to rise considerably from the current 6–7 percent in order to threaten the assumption that a single main category works well enough to be a useful way to think about and compute genre.[13]

Lastly, we probed the consistency of annotator agreement. Cohen's Kappa scores (table 6) indicate that the annotators' assessments tend to agree to a clearly non-random degree. The score is well-known for being difficult to interpret. Yet even if strict guidelines are used (McHugh 2012), the score for quality assessment indicates strong agreement, with the confidence interval extending to moderate agreement, while the confidence interval for agreement on assigned labels is wholly within the range of nearly perfect agreement. An imbalance in quality labels (both annotators gave a score of 1 in more than 80 percent of instances) might explain why the evaluation sample received a lower Kappa score (even at random, most quality annotations would have matched). Based on the raw match rate of annotations, the annotators agree on 95 percent of the quality assessments of the labels (table 4). The agreement on assigned labels is 94 percent by this measure. We conclude that the annotators' assessments of what is the right label converge to a very high degree, suggesting that the conceptualisation of genre works well in most instances.

13. For example, averaging the number of editions belonging to different genres when the two annotators disagreed is a concession to the fact that the conceptualisation of a single main category per edition was not always applicable when doing calculations.

Table 6: Annotator agreement (Aa) and label quality estimated; 95 percent confidence intervals

Sample	Measure	2.5% bound	Estimate	97.5% bound	*N*
Sample with genre labels	Cohen's Kappa (Aa)	0.78	0.82	0.87	996
Sample without genre labels	Cohen's Kappa (Aa)	0.91	0.93	0.96	500
Sample with genre labels	Prob. of agreement (Aa)	0.94	0.95	0.96	996
Sample without genre labels	Prob. of agreement (Aa)	0.92	0.94	0.96	500
Sample with genre labels	Best labels	0.79	0.81	0.83	996
Sample with genre labels	Good labels	0.83	0.86	0.88	996
Sample with genre labels	Plausible labels	0.94	0.95	0.96	996

Cohen's Kappa was calculated with the psych R package (Revelle 2024).

5.2 Distant and close reading of correlated categories

Our aim is not to downplay the errors in the data. In fact, they are rather interesting as objects of analysis in themselves, since they demonstrate the problematic aspects of any attempted classification scheme of early modern books and pamphlets. Although the data we have about errors is not extensive (less than 200 observations "produced" information about errors in labels), a combination of distant and close reading enabled us to produce plausible explanations showing how some of the errors are related to structural overlaps that existed in early modern print culture. These overlaps also illustrate the differing literary conventions and traditions of the eighteenth century, when genres now considered distinct were combined much more frequently and loosely. As figure 2 demonstrates, the errors are very unequally distributed. Some pairs of real and assigned categories barely exist in the evaluated sample, but others are very significant. Notably, tens of percent of documents best classified as politics end up in law. In figure 2, history "bleeds" into politics, literature and arts, and religion to a considerable degree given the small size of the category[14], while figure 3 illustrates how history mistakenly includes editions better classified as literature, politics, education or science. In contrast, categories related to formalised and well-established discourses like law and religion are more resistant to losing a significant proportion of their records to other categories.

14. See for example figure 5.

Figure 2: Errors (label did not qualify as "good") in the evaluated sample compared to the estimated size (no. of "good" labels) of the real category

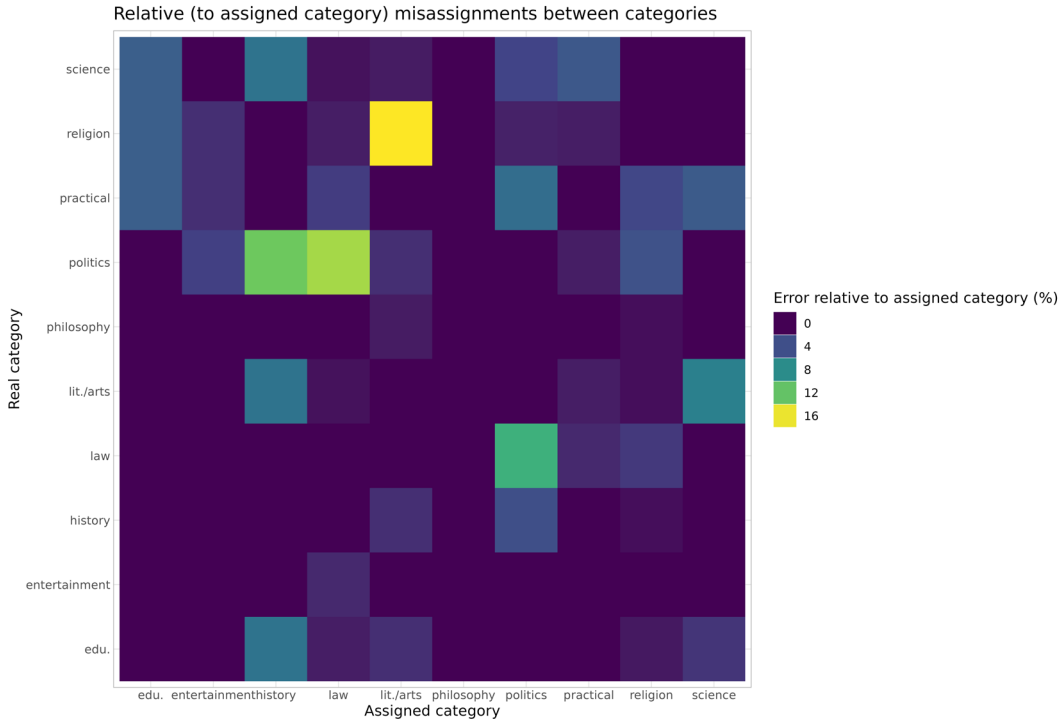

Figure 3: Errors (label did not qualify as "good") in the evaluated sample compared to the estimated size (no. of "good" labels) of the assigned category

The results of the quantitative error analysis prompted us to take additional steps, such as creating supercategories that combine multiple main categories. Our reasoning was that, in some instances, it is better to treat highly correlated

categories (especially politics and law) as one category. What these supercategories lose in granularity, they gain in robustness. The second step was to examine some of the problematic categories more closely. Our conclusion is that categorisation problems largely reflect problems in applying clear-cut modern categories to early modern books; however, they also provide tentative directions for further research on classifying the genre and topic of ESTC records.

In the close reading, we focused in particular on the connection between law and politics, the overlap between religion and politics, and the issue of history, the least precise category. The connection between politics and law is rather straightforward: politics as a category often deals with the law and is distinguished less by substance than by register or rhetoric. The following is the title of a record (T52822) from our sample that was classified as law:

> The worth of liberty consider'd: in a letter to a Member of the House of Commons upon the question, how far the late act against the immoderate use of spiritous liquors may affect the properties of all the people.

Both researchers gave this label a score of 0.5, as the document's content is an argumentative discourse about legislation rather than primary legislation as such. However, it certainly relates to legislation. This thematic overlap does not mean that the difference between the two categories would somehow be arbitrary: it is plausible that the argumentative language of early modern politics (or even its description, such as the one given in the title) could be distinguished from language strictly confined to legislative issues as such. For making such distinctions, state-of-the-art tools in machine learning or computational linguistics could be useful.

Another common (in relative terms) overlap occurs between religion and politics, which shows the complexity of applying modern genre categories to historical periods. In the early modern period, religion and politics were much more intertwined than is the case today. Sermons in particular were often very topical, especially during the English Civil War, and could touch on important political issues of the day. Debates about the rights of religious minorities—such as Dissenters (Protestants who were not members of the established Church of England) and Jews—contained a combination of religious and political arguments. These prevalent overlaps led the annotators to reflect on and discuss how to deal with cases where the outward form suggests one genre (here: religion) and the topic, as indicated by the title, is either a mix of at least two genres (here: politics and religion) or leans more towards a genre that is different to the form (here: politics).

Below is a record (T202404) from our sample with the assigned category of religion, illustrating the overlap between religion and politics and the difficulty

of evaluating which of the two genres is the most fitting description for such works.

> The young cobler of Glocester: or, Magna Charta [*sic*] discours'd of between a poor man and his wife. As also several high-church principles discussed, ... Together, with reflections upon several of the vices of the high-church clergy, by the cobler and his wife.

One of the annotators assigned a score of 1 on the basis that the subtitle "reflections upon several of the vices of the high-church clergy" indicates that the work is primarily religious in nature, and thus the most fitting genre label. The other annotator agreed that religion was a plausible genre, but thought the work was even more likely to have a political thrust, as evidenced by the title's references to Magna Carta and High Church principles (which often went hand in hand with support for the Tory party). Here, the annotators' motivations for assigning their respective scores are a testament to how their respective judgement is shaped by their own academic interest and domain knowledge.

The category with the lowest precision, history, provides interesting insights into how history writing was practised in the early modern period. On the face of it, a title containing phrases such as "a history of" should serve as a reasonably good predictor that the work in question belongs to the genre of history. However, during our annotation process, we discovered that works seemingly historical in nature could, in fact, sometimes be entirely fictional. Distinguishing between biographies of real and invented persons was particularly difficult, as early modern biographers tended to title their works "true" or "authentic" regardless of whether their subject ever existed, and fiction writers frequently mimicked the writing style and conventions of nonfiction genres (e.g. ESTC records T110215 and T173534). An additional challenge was in deciding on the cut-off point at which a work is considered historical. In the early modern period, many historiographers wrote works covering events up to the present, often with a view to intervening in contemporaneous political debates. Deciding whether a work describing recent events, such as a war that happened 20 years earlier, belonged to the genre of history or politics was not always straightforward, especially since the title itself often gives little insight into whether the work is primarily focused on describing the past or on using the past for present-day politics. One example of this overlap is record N9975 below:

> M[emoirs and observations] of the occurrences of Europe, Since the treaties of Nimeguen and Ryswick, with relation to the present treaty at Utrecht. Shewing, that it is of the last importance that England have a footing upon the continent.

This was initially labelled as history, but both annotators gave a score of 0.5 and suggested that politics was more fitting. The period covered by the work, as indicated by the title, is 1678 (Treaty of Nijmegen) to 1712, the year of publication. When the work was published, negotiations for the Treaty of Utrecht were still ongoing and the subtitle clearly indicates the author's contemporaneous political concerns.

The dividing line between religion and history—rather distinct categories in the twenty-first century—was also much more blurred in the early modern period. Biblical criticism as an academic subfield was still in its infancy, and writers of religious history freely mixed real events occurring in biblical times (e.g. the siege of Jerusalem in 70 CE) with events and figures only described in the Bible and for which historical veracity cannot be established (e.g. ESTC record W21041). Assigning such works to either history or religion was a judgement call for the annotators, since it was difficult to estimate the balance between primarily religious elements and primarily historical elements without consulting the work in question. These complexities explain why the precision rate for history is significantly lower than for all other categories. We sought to mitigate difficulties of assigning a genre in multiple ways, for example by discussing particularly complex cases, assigning a score of 0.5, and looking at a small number of original texts to see if a general rule for annotating such cases could be determined.

6 Patterns and biases of genres in the ESTC and ECCO

The new genre data introduced in this paper significantly extends the possibilities of using the ESTC and its subsets for quantitative analyses of early modern history. As established in section 2, the ESTC (alongside its subsets) has been the primary resource in attempts to chart the topic or genre composition of publishing outputs in early modern Britain. For the eighteenth century, which covers most of the editions in the ESTC, such work has only been possible for subsets or samples of the whole catalogue until now. Furthermore, there has been no framework that would have placed the entire catalogue—spanning more than 300 years—within the confines of a single concise and comprehensive classification scheme. Our categorisation covers 94 percent of all records in the ESTC. To the best of our knowledge, figure 4 is the most comprehensive depiction of the composition of early modern book publishing in the English-speaking world yet produced.

Our initial examination of figure 4 adds significant support to previous claims made about large-scale book publication trends and highlights unexpected developments that should be studied in more detail in future research. We can see the same relative decline of religion and rise of literature during the

eighteenth century as noted by earlier and more limited quantitative studies (Suarez 2009; Tolonen et al. 2021). Our results make it increasingly likely that these developments are robust and are not confined to particulars of any specific classification scheme or subset of ESTC editions analysed.

One development that warrants closer scrutiny is the rise in the proportion of political works during the decade of the English Civil War (1640s). Even after the end of the war, the proportion of political works does not return to the pre-war level. Since the relationship between the English Civil War and the emergence of a "public sphere" of political discourse in England has been discussed in qualitative research (Zaret 2000), this finding is potentially valuable if it can be regarded as reasonably accurate.[15] However, since the English Civil War decade is also a threshold that separates different segments of the ESTC (consisting of several preceding catalogues), data transitions happening in the 1640s should be analysed with special care, as they can also be linked to different biases and classification conventions of the different segments (Tiihonen 2020).

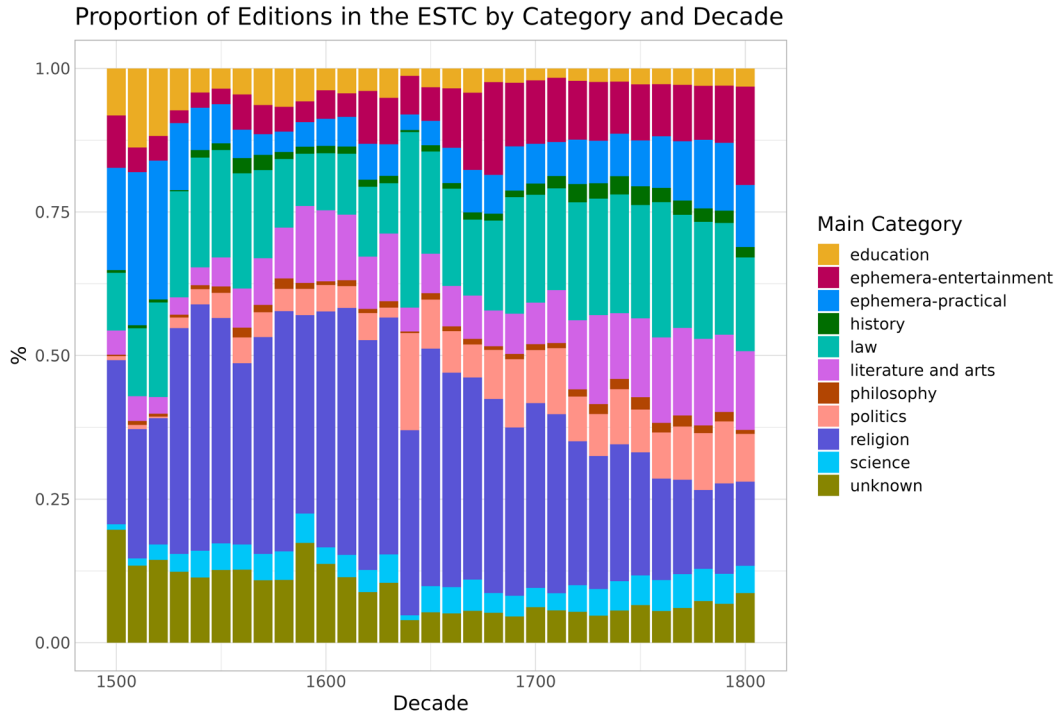

Figure 4: Main categories of editions in the ESTC 1500–1800

Our categorisation may also be useful for studies focusing primarily on other datasets. The ESTC is the default resource used to examine topic or genre-related biases in other collections of early modern books, such as ECCO (Tolonen,

15. The fact that politics was a major topic in the publications of the 1640s is in itself an extremely well-known phenomenon that has been extensively studied (Raymond 2003). However, here we can compare the decade of the English Civil War to roughly 150 years both before and after it at the scale of the entire ESTC.

Mäkelä, and Lahti 2022) or the early modern section of the HathiTrust Digital Library (Almelhem et al. 2023). For researchers working with other datasets, a more comprehensive classification of the ESTC renders it a more accurate comparison point in terms of genre and topic composition, and can thus help to reveal biases in the dataset being studied.

In figure 5, we compare the category composition of the ESTC to that of ECCO, the most comprehensive full-text collection of books printed in eighteenth-century Britain. As the figure demonstrates, the collection is not an accurate representation of all the surviving editions in the ESTC. Law texts and practical ephemera are underrepresented, whereas science, history and philosophy are overrepresented to a considerable degree. This comparison extends earlier quantitative work on the biases present in ECCO (Tolonen, Mäkelä, and Lahti 2022) that could not utilise a fully classified ESTC. We can now argue more confidently that ECCO is especially skewed towards texts related to Enlightenment categories—such as philosophy and science—and that it has a tendency to exclude administrative and practical documents even though they account for a significant proportion of early modern publications.

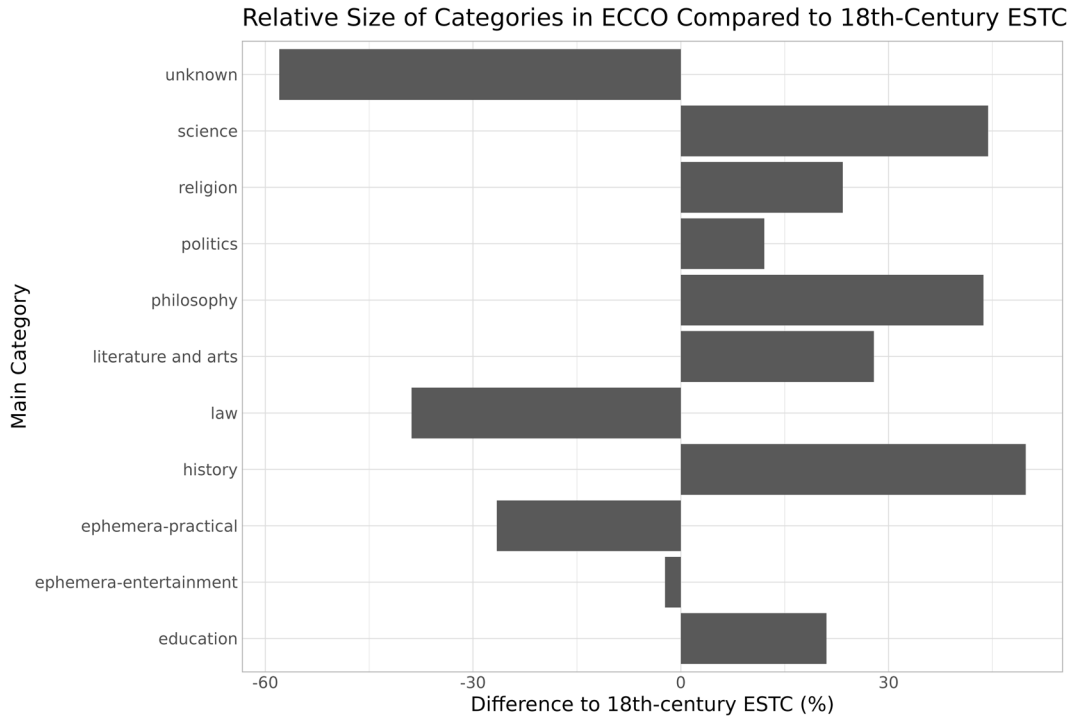

Figure 5: Relative "bias" of ECCO compared to the ESTC by main category

7 Conclusion

Even more important than the immediate findings are the yet-unexplored possibilities offered by the new data, as well as the lessons learned from the

classification process. Together, our workflow development and evaluation process demonstrate that a good understanding of the data can create a simple, eclectic and transparent workflow that produces relevant results even for a complex historical dataset like the ESTC.[16] A systematic but hands-on and humanities-oriented process of evaluating the results helped us to pinpoint the degree to which the data enrichment workflow was effective: manual evaluation allowed us not only to decide whether a category was ontologically correct, but also to consider the plausibility of any assigned category, which ended up being a very fruitful means of dealing with the various topical and thematic convolutions of early modern texts.

Our evaluation revealed that, in the ESTC, certain document types (e.g. politics and law) might benefit from a more complex classifier that leverages techniques from machine learning and computational linguistics, since keywords and titles can miss the essential elements relevant for classification. Despite the limitations discussed earlier, unsupervised and semi-supervised methods might have their use in discovering "genres" that are characterised by a complex configuration of linguistic, topical or even visual cues.[17] Thus, our results highlight the necessity for projects that study early modern genres and related questions from various angles. New techniques like multimodal learning might help to derive more holistic computational interpretations of genre in the early modern period, but as we demonstrate, classification workflows can also be driven mostly by domain expertise in (for example) the bibliographic material.

The process described here will be updated. Since it is already in use in various COMHIS research projects, it is reasonable to wait for some time (e.g. a year) to see if any suggestions for improvements emerge from experiences accumulated in real-life research. We can already name a few changes and expansions that could be implemented. In a multi-step workflow with many shifting parts, there are always individual elements like specific keywords which, in hindsight, could have been added or removed.[18] Larger or more targeted evaluation samples could help us learn more about the size and nature of category overlaps. One step to be taken in the near future is to enrich the pre-1700 ESTC with the keyword information from the Universal Short Title Catalogue (USTC). This

16. In the sense that various sources of information are used and the option deemed to be the best is used for classification.
17. For example, satirical illustrations may be a distinguishing feature between law and politics in early modern publications (for the political role of images, see e.g. Pierce 2008). To identify at scale which kind of images commonly appeared in which genres, it may be possible to build on a currently ongoing research strand in COMHIS, which uses deep learning algorithms to detect and classify visual elements in eighteenth-century books (Wang et al., forthcoming; for an introduction to this methodology, see Wang 2023).
18. For example, author one accepted the keyword "pirates" to map to the category literature, but no longer agrees with his previous reasoning, as it could also easily be related to political discussion too.

should significantly extend the coverage of records with a label before the eighteenth century. Another potential step is to create a version of the data in which multiple categories can exist for any given document. This would not be technically difficult to implement and, as our evaluation results demonstrate, category ambiguity was an important characteristic of early modern books published in the Anglosphere, making a strong case for a multi-label version. The data could also benefit from a formal approach to modelling genres. Scope notes for genres could improve rigour and cross-disciplinary communication, and an explicitly hierarchical genre model could be implemented for additional technical clarity. Combined with formal identifiers for the genres, this could make the data introduced in this article more compatible with Linked Open Data (LOD) standards. In the best-case scenario, genre data could be linked with other historical datasets in the future.

References

Almelhem, Ali, Murat Iyigun, Austin Kennedy, and Jared Rubin. 2023. "Enlightenment Ideals and Belief in Progress in the Run-Up to the Industrial Revolution: A Textual Analysis." *IZA Discussion Paper* No. 16674. http://dx.doi.org/10.2139/ssrn.4668604.

Amory, Hugh. 1983. "Review of *Eighteenth-Century British Books: An Author Union Catalogue* by F. J. G. Robinson, G. Averley, D. R. Esslemont, and P. J. Wallis and *Eighteenth-Century British Books: An Index to the Foreign and Provincial Imprints in the Author Union Catalogue* by *F. J. G. Robinson*, J. M. Robinson, and C. Wadham." *The Papers of the Bibliographical Society of America* 77 (1): 80–84. https://doi.org/10.1086/pbsa.77.1.24302881.

Baker, Keith Michael. 1964. "The Early History of the Term 'Social Science'." *Annals of Science* 20 (3): 211–26. https://doi.org/10.1080/00033796400203074.

Barr, Bernard. 1981. "Paying for the Eighteenth Century: The ECBB." *Library Review* 30 (4): 229–32. https://doi.org/10.1108/eb012729.

Corns, Thomas. 1986. "Publication and Politics, 1640–1661: An SPSS-based Account of the Thomason Collection of Civil War Tracts." *Literary and Linguistic Computing* 1 (2): 74–84. https://doi.org/10.1093/llc/1.2.74.

Feather, John. 1986. "British Publishing in the Eighteenth Century: A Preliminary Subject Analysis." *The Library* s6-VIII (1): 32–46. https://doi.org/10.1093/library/s6-VIII.1.32.

Fielding, David, and Shef Rogers. 2017. "Copyright Payments in Eighteenth-Century Britain, 1701–1800." *The Library* 18 (1): 3–44. https://doi.org/10.1093/library/18.1.3.

Gants, David. 2002. "A Quantitative Analysis of the London Book Trade 1614–1618." *Studies in Bibliography* 55 (January): 185–213. https://doi.org/10.1353/sib.2005.0004.

Gillings, Matthew, and Andrew Hardie. 2023. "The Interpretation of Topic Models for Scholarly Analysis: An Evaluation and Critique of Current Practice." *Digital Scholarship in the Humanities* 38 (2): 530–43. https://doi.org/10.1093/llc/fqac075.

Grajzl, Peter, and Peter Murrell. 2024. "A Macroscope of English Print Culture, 1530–1700, Applied to the Coevolution of Ideas on Religion, Science, and Institutions." *Social Science History* 48 (3): 489–519. https://doi.org/10.1017/ssh.2024.17.

Hill, Alexandra. 2018. *Lost Books and Printing in London, 1557–1640. An Analysis of the Stationers' Company Registers*. Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004349209.

Ijaz, Ali Zeeshan, Hege Roivainen, and Leo Lahti. 2019. "Analytical Edition Detection In Bibliographic Metadata." DataverseNL, V2. https://doi.org/10.34894/IWQ5BO.

Jenner, Mark. 2011. "London." In *The Oxford History of Popular Print Culture*, edited by Joad Raymond and Gary Kelly, 294–307. Oxford: Oxford

University Press. https://doi.org/10.1093/acprof:osobl/9780199287048.003.0023.

Kapoor, Sayash, and Arvind Narayanan. 2023. "Leakage and the Reproducibility Crisis in Machine-Learning-based Science." *Patterns* 4 (9). https://doi.org/10.1016/j.patter.2023.100804.

Lahti, Leo, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. 2019. "Bibliographic Data Science and the History of the Book (c. 1500–1800)." *Cataloging & Classification Quarterly* 57 (1):5–23. https://doi.org/10.1080/01639374.2018.1543747.

McHugh, Mary. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22 (3): 276–282. PMID: 23092060.

McKenzie, Donald. 2002. "Printing and Publishing 1557–1700: Constraints on the London Book Trades." In *The Cambridge History of the Book in Britain*, edited by John Barnard and Donald McKenzie, 4:553–67. Cambridge: Cambridge University Press. https://doi.org/10.1017/CHOL9780521661829.028.

Murphy, Sean. 2019. "Shakespeare and his Contemporaries: Designing a Genre Classification Scheme for Early English Books Online 1560–1640." *ICAME Journal* 43 (1): 59–82. https://doi.org/10.2478/icame-2019-0003.

Nicoll, Allardyce. 1959. *History of English Drama 1660–1900: Volume 6, A Short-title Alphabetical Catalogue of Plays*. Cambridge: Cambridge University Press.

Pierce, Helen. 2008. *Unseemly Pictures: Graphic Satire and Politics in Early Modern England*. New Haven: Yale University Press.

Raven, James. 2018. *What is the History of the Book*? Cambridge: Polity Press.

Raymond, Joad. 2003. *Pamphlets and Pamphleteering in Early Modern Britain*. Cambridge. Cambridge University Press.

Raymond, Joad. 2011. "The Development of the Book Trade In Britain." In *The Oxford History of Popular Print Culture*, edited by Joad Raymond and Gary Kelly, 59–75. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199287048.003.0006.

Revelle, William. 2024. "Psych: Procedures for Personality, and Psychological Research." The Comprehensive R Archive Network. Version 2.4.12. https://CRAN.R-project.org/package=psych.

Rosson, David, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. "Reception Reader: Exploring Text Reuse in Early Modern British Publications." *Journal of Open Humanities Data* 9 (1): 1–11. https://doi.org/10.5334/johd.101.

Ryan, Yann Ciarán, and Mikko Tolonen. 2024. "The Evolution of Scottish Enlightenment Publishing." *The Historical Journal* 67 (2): 223–55. https://doi.org/10.1017/S0018246X23000614.

Shadrova, Anna. 2021. "Topic Models Do Not Model Topics: Epistemological Remarks and Steps towards Best Practices." *Journal of Data Mining & Digital Humanities* (October). https://doi.org/10.46298/jdmdh.7595.

Simmons, Richard. 2002. "ABCs, Almanacs, Ballads, Chapbooks, Popular Piety and Textbooks." In *The Cambridge History of the Book in Britain,* edited by John Barnard and Donald McKenzie, 4:504–13. Cambridge: Cambridge University Press. https://doi.org/10.1017/CHOL9780521661829.025.

Snyder, Henry L. 1983. "Review of *Eighteenth-Century British Books: An Author Union Catalogue Extracted From the British Museum General Catalogue of Printed Books, the Catalogues of the Bodleian Library, and of the University Library, Cambridge*, by F. J. G. Robinson, G. Averley, D. R. Esslemont, and P. J. Wallis." *Eighteenth-Century Studies* 16 (3): 342–46. https://doi.org/10.2307/2738357.

Suarez, Michael. 2009. "Towards a Bibliometric Analysis of the Surviving Record, 1701–1800." In *The Cambridge History of the Book in Britain*, edited by Michael Suarez and Michael L. Turner, 5:37–65. Cambridge: Cambridge University Press. https://doi.org/10.1017/CHOL9780521810173.003.

Tiihonen, Iiro. 2020. "From Explosion to Implosion. A Quantitative Analysis of the English Civil War Print Production." Master's thesis, University of Helsinki. http://hdl.handle.net/10138/314293.

Tiihonen, Iiro, Leo Lahti, and Mikko Tolonen. 2024. "Print Culture and Economic Constraints: A Quantitative Analysis of Book Prices in Eighteenth-Century Britain." *Explorations in Economic History* 94 (October). https://doi.org/10.1016/j.eeh.2024.101614.

Tiihonen, Iiro, Yann Ryan, Lidia Pivovarova, Aatu Liimatta, Tanja Säily, and Mikko

Tolonen. 2022. "Distinguishing Discourses: A Data-Driven Analysis of Works and Publishing Networks of the Scottish Enlightenment." In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, edited by Karl Berglund, Matti La Mela, and Inge Zwart, 3232:120–34. CEUR Workshop Proceedings. Uppsala, Sweden: CEUR. http://ceur-ws.org/Vol-3232/#paper09.

Tolonen, Mikko, Mark J. Hill, Ali Zeeshan Ijaz, Ville Vaara, and Leo Lahti. 2021. "Examining the Early Modern Canon: *The English Short Title Catalogue* and Large-Scale Patterns of Cultural Production." In *Data Visualization in Enlightenment Literature and Culture*, edited by Ileana Baird, 63–119. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-54913-8_3.

Tolonen, Mikko, Eetu Mäkelä, and Leo Lahti. 2022. "The Anatomy of Eighteenth Century Collections Online (ECCO)." *Eighteenth-Century Studies* 56 (1): 95–123. https://doi.org/10.1353/ecs.2022.0060.

Wang, Ruilin. 2023. "Visual Element Extraction and Grouping from Eighteenth-Century Books Based on Deep-Learning Methods." Master's thesis, University of Helsinki. http://hdl.handle.net/10138/565060.

Wang, Ruilin, Lidia Pivovarova, Yann Ciarán Ryan, and Mikko Tolonen. Forthcoming. "Semi-Supervised Contrastive Training for Similar Image Identification in a Large Collection of Historical Books." In *Scandinavian Conference on Image Analysis (Reykjavík, Iceland, 23–25 June)*. Lecture Notes in Computer Science.

Watt, Tessa. 1990. "Publisher, Pedlar, Pot-Poet: The Changing Character of the Broadside Trade, 1550–1640." In *Spreading the Word. The Distribution Networks of Print 1550–1850*, edited by Robin Myers and Michael Harris, 61–82. Winchester: Saint Paul's Bibliographies.

Whitten, David O. 1978. "Democracy Returns to the Library: The Goldsmiths'-Kress Library of Economic Literature." *Journal of Economic Literature* 16 (3): 1004–6.

Zaret, David. 2000. *Origins of Democratic Culture: Printing, Petitions and the Public Sphere in Early-Modern England*. Princeton: Princeton University Press.

Zhang, Jinbin, Yann Ciarán Ryan, Iiro Rastas, Filip Ginter, Mikko Tolonen, and Rohit Babbar. 2022. "Detecting Sequential Genre Change in Eighteenth-Century Texts." In ***Proceedings of the Computational Humanities Research Conference 2022 (Antwerp, Belgium, 12–14 December)***, edited by Folgert Karsdorp, Alie Lassche, and Kristoffer Nielbo, 243–55. https://ceur-ws.org/Vol-3290/#short_paper2630.

Appendices

Appendix A—Supplementary data

Table 7: Full list of subcategories

Act	Adventure	Advertisement	Advice	Aesthetics
Agriculture	Almanac	Ancient history	Antiquarianism	Applied knowledge
Architecture	Astrology	Astronomy	Bible	Bill
Biography	Calendar	Catalogue	Catechism	Chemistry
Classics	Comedy	Commerce	Conduct of life	Cooking
Declaration	Devotional exercise	Dictionary	Doctrine	Fable

For children	Geography and nature	Handbook	Hobbies and games	Human mind
Hymn	Language	Linguistics	Logic	Manners
Mathematics	Medicine	Memoir	Metaphysics	Meteorology
Military applications	Moral	Music	Navigation	News
Novel	Opera	Painting	Pastoral letter	Periodical
Petition	Physics	Play	Poem	Political thought
Prayer book	Proceeding	Proclamation	Psalm	Report
Review	Satire	Satire poem	Sermon	Song
Speech	Technology	Theatre programme	Theoretical	Theory of knowledge
Travel	Trial	Utopia		

Appendix B—Technical details

Prominence score F of a main category for an edition **x** and main category **c** was calculated in the following way:

$$\mathbf{F(c,x)} = \sum_{k \in (K_x \cap K_c)} ln(Q(k))$$

Where **k** is a keyword or title n-gram in the ESTC, K_x is the set of keywords or title n-grams related to the edition **x**, and K_c is the set of keywords or title n-grams related to main category **c.** Function **Q** measures the number of times that the keyword or n-gram appears in the ESTC. In verbal form: the prominence score is a sum of natural logarithms of the prominences of keywords or n-grams that relate to an edition and one main category. Constructed in this manner, the equation favours the information provided by the most common keywords and title n-grams. The main category of the edition was chosen by picking the category **c** with the highest prominence score.

Appendix C—Exploratory statistical approaches

The primary explorative method used to find relevant keywords and n-grams for the classification workflow was to computationally detect those that were heavily associated with a main category in an already labelled subset of the data, and then analyse these keywords and n-grams manually to ascertain whether they had a direct and meaningful relation to that main category. If the connection between the keyword or n-gram and the main category was deemed plausible enough, it was typically included in the labelling process.

The first author still reserved the right to make any substance-based edits, for example by assigning the connection between an n-gram and main category wholly by hand or by adding n-grams or keywords that did not emerge through this process to the workflow.

In practice, this was implemented by analysing how often a keyword or n-gram manifested in a given main category (e.g. whether more than half of its appearances were concentrated in one main category). Another important measure was how disproportionately a keyword or n-gram manifested in a specific category compared to what we would have expected, if the topic or n-gram was equally distributed between different main categories. Since it was possible to implement these comparisons to all of the n-grams and topics analysed at once (combined with other filtering choices like how common the keyword or n-gram was in the data), it enabled the first author to restrict manual evaluation work to those keywords and n-grams that had at least a substantial correlational relationship to some main category, and hence a much greater likelihood of being useful for classification. This process required pre-labelled data to infer n-grams and topics with "overrepresentation" in a main category. In our case, such data was initially provided via manually labelled data. It became possible to repeat this investigation of correlations between topics and n-grams and main categories with a larger dataset when the data grew. The replication package of this article includes a script and an output table—similar to the tables that were then checked manually to select keywords and n-grams for the workflow—that demonstrate this approach.

This process is an example of utilising noisy correlations to find—rather than confirm—connections between main categories and n-grams/topics that have a meaningful interpretation. Manual confirmation of the included n-grams and topics was very important, as the reasons for specific correlations were occasionally very misleading, and it would have been questionable to project them to further classification. For example, the n-gram "our lord" was heavily overrepresented in the category ephemera-practical. This is understandable, as it is a jargon-like expression in the full title of early modern almanacs ("published in the year of our lord X"). But "our lord" has no substantial correlation to this main category. As it has a substantial correlation to categories like religion and politics, it is possible that mapping the bigram to ephemera-practical would lead to many out-of-sample documents in those categories being labelled as ephemera-practical.

This example highlights the more general motivations of our method. Given the high-dimensional and heterogeneous nature of our data, correlations as such are heavily influenced by random chance and the particulars of the specific subset of data used to detect them, making projections unreliable. As the possible explanations for correlations vastly outnumber those that tie a keyword or n-gram to a meaningful relationship with a main category, the

default approach was to require a historically grounded and robust explanation for a topic or keyword to be included. This does not mean that the keywords and n-grams included in the workflow are wholly unproblematic (we are not aware of such instances), but as we hope to demonstrate, sometimes correlation reflects genuine connections between the units of information that we use and genres as we conceptualise them.

An example of an included n-gram would be the bigram "performed at," which had a heavy statistical association with the category "arts." This expression was frequently used in the titles of plays (belonging to the category of arts[19]) published in London, and we are not aware of any serious convolutions that could lead to other publication types using this expression consistently. Still, there was an element of concession in including expressions like this in the workflow, as the connection between the bigram and the main category to which it maps is of a second-degree nature. A publication could address something being performed without necessarily being a publication that best belongs in the category arts, but to the best of our understanding, the expression was mostly used in the context of plays.

An even stronger example of an included n-gram is the bigram "a sermon." Sermons were an established text type in eighteenth-century Britain, and thousands of publications define themselves as such in their title. Even this bigram is not wholly unproblematic, as there are individual instances of political or philosophical essays and commentaries whose titles contain phrases like "in a response to a sermon preached at." Still, having "a sermon" in the title is an almost ideal piece of information for the workflow, as it maps directly to the classification scheme (sermons are assigned to the category "religion" by default).

19. In the first four steps of the workflow, arts and literature are still treated as separate categories, after which they are merged together.

TRANSFORMATIONS
A DARIAH Journal

The Sixth Generation of the Perseus Digital Library and a Workflow for Open Philology

Gregory Crane[1], James Tauber[2], Alison Babeu[1], Lisa Cerrato[1], Charles Pletcher[1], Clifford Wulfman[3], Sergiusz Kazmierski[4], Farnoosh Shamsian[5]

1 Tufts University
2 Signum University
3 Princeton University
4 Regensburg University
5 Leipzig University

Transformations, A DARIAH Journal
Volume 1, 2025
https://transformations.episciences.org

Dates
Received: 15/11/2024
Accepted: 03/04/2025
Published: 10/06/2025

DOI: 10.46298/transformations.14780

Abstract

This paper presents an overview of recent developments by the Perseus Digital Library in creating the Beyond Translation reading environment, a foundational component in the transition toward Perseus 6, built on the ATLAS (Aligned Text and Linguistic Annotation Server) architecture. It highlights the integration of diverse open data sources from multiple digital humanities projects, all brought together to support an innovative, richly layered digital reading experience. Following this, the paper details the key services offered by Beyond Translation, including text-translation alignments, advanced morpho-syntactic analysis, audio annotations and enhanced access to integrated reference resources such as commentaries and dictionaries. The paper concludes with a discussion of forthcoming enhancements and the future trajectory of the ATLAS architecture.

Keywords: multilingual corpora, text annotation, Linked Open Data, Named Entity Recognition (NER), cultural heritage

EPIsciences

To cite this article: G. Crane, J. Tauber, A. Babeu, L. Cerrato, C. Pletcher, C. Wulfman, S. Kazmierski, and F. Shamsian. (2025). The Sixth Generation of the Perseus Digital Library and a Workflow for Open Philology *Transformations. A DARIAH Journal*, vol. 1, p. 1-25, https://doi.org/10.46298/transformations.14780

Die sechste Generation der Perseus Digital Library: Ein Workflow für offene Philologie

Zusammenfassung

Dieser Beitrag gibt einen Überblick über die jüngsten Entwicklungen innerhalb der Perseus Digital Library, welche im Rahmen der Schaffung der Leseumgebung Beyond Translation erfolgt sind. Beyond Translation bildet im Kontext dieser Entwicklung eine grundlegende Komponente des Übergangs zu Perseus 6 und baut auf der ATLAS-Architektur (Aligned Text and Linguistic Annotation Server) auf. Der Beitrag beleuchtet insbesondere die Integration unterschiedlicher, offener Datenquellen, die aus zahlreichen Digital Humanities-Projekten zusammengeführt wurden, um eine innovative und vielschichtige, digitale Leseerfahrung möglich zu machen. Davon ausgehend werden die Schlüsselfunktionen von Beyond Translation erläutert und dabei das Alignment von Originaltexten und Übersetzungen, komplexe morpho-syntaktische Analysen, Audioannotationen und der erweiterte Zugriff auf integrierte Referenzressourcen wie Kommentare und Wörterbücher berücksichtigt. Abschließend werden in Arbeit befindliche und geplante Erweiterungen sowie Entwicklungsperspektiven der ATLAS-Architektur thematisiert.

Schlüsselwörter: Mehrsprachige Korpora, Textannotation, Verknüpfte Offene Daten, Erkennung benannter Entitäten (NER), Kulturelles Erbe

Gregory Crane was the lead author of this piece. James Tauber was primarily responsible for the new work on the ATLAS architecture performed in 2024.

Introduction

We report here on the workflow that we needed to develop in order to integrate the growing range of openly licensed, born-digital and, increasingly, machine-actionable publications. Our developmental work focused upon textual data for Ancient Greek, Latin, Old English, Classical Arabic and Classical Persian, but the challenges that we have had to address are relevant to sources in a wide range of languages, both ancient and modern. During the course of the early twenty-first century, major projects have emerged that support forms of publication that either extend well beyond, or in fact, have no counterparts in print culture. Systems such as INCEpTION and Perseids allow us to add exhaustive linguistic annotations to our corpora and develop treebanks. The Ugarit platform allows us to align, both manually and automatically, words and phrases in a source text with their equivalents in a translation. The Recogito system and the ToposText project have made it possible to align place names in source texts with gazetteers and then to generate customised maps. David Chamberlain individually developed Hypotactic, a site that allows readers to visualise 250,000 lines of metrically analysed Greek and Latin poetry and to download the underlying data under a Creative Commons licence. Efforts such as these have provided new classes of data for automatic analysis and provided new ways for readers to visualise machine-actionable annotations of source texts.

The work presented here seeks to complement these and other efforts by addressing the challenge of bringing these streams of data together. This work lays the foundation for research into how human readers can best take advantage of the increasingly rich classes of annotation that are already—or soon will be—available online. How do we read in an environment where the answers to far more questions are available than was thinkable in print culture? How do we critically assess automated systems that allow us to interact directly with sources in languages and from cultural contexts of which we may have little or no knowledge? How do we not only translate but also localise knowledge about one cultural context (e.g. European scholarship on Greco-Roman culture) for readers from very different contexts (e.g. Persian-speakers encountering the *Iliad* and the *Odyssey*, for whom Ferdowsi's *Book of Kings* shapes their understanding of epic poetry)? And, of course, what are the roles that increasingly intelligent Large Language Models can—and should—play in helping us engage with a human record that is far too large and complex for any of us to master?

Perseus 6 has been designed to be a publishing workflow that organises complementary data into an integrated reading environment.[1] This paper focuses on the ways in which we have organised the data and describes the current state of

1. The work that we describe here was developed as part of the Perseus on the Web: Preparing for the Next Thirty Years project, with support from the US National Endowment for the Humanities, the Data Intensive Studies Center and the Faculty of Arts and Sciences at

the ATLAS (Aligned Text and Linguistic Annotation Server) architecture. While this is the sixth version of the Perseus Digital Library, Perseus 6 represents a major step beyond its predecessors. Whereas Perseus 5 (described below) can represent and integrate digital versions of print editions (e.g. critical editions with interactive textual notes, links to lexicon and commentary entries), Perseus 6 was designed to bring together an expandable range of born-digital classes of annotation. An online ATLAS server with some initial functionality is now available and public services will expand throughout 2025.[2] Most of the ATLAS data is, however, now available on GitHub and that data will be the focus of this paper in its current version.

The purpose of this publication is to introduce people to the problems that we have addressed. Others may build on what we have done or, having seen our work, choose to develop completely different solutions. The work that we have done, however, addresses the need to bring together complementary datasets that are currently available under an open licence, but split across multiple repositories and systems.[3] Our hope is that more people will see what can be done when we bring together the growing wealth of information that open scholarly projects are producing. The work presented here is a work in progress and represents development as of November 2024.

Our goal was to create a workflow to organise, rather than create, textual data that had been produced by, and was available in, platforms that were open but separate. In the quarter of a century since Creative Commons licences emerged in the early 2000s,[4] multiple projects in a range of countries have developed robust workflows to produce one or more classes of open data. Projects such as Perseus, Perseids, PROIEL, GLAUx Trees, Daphne and Opera Graeca Adnotata (Celano 2024) have all published treebanks of Greek and Latin (Hudspeth, O'Connor, and Thompson 2024; Gorman 2020; Keersmaekers et al. 2019; Keersmaekers 2021; Keersmaekers and Van Hal 2022). Recogito allows users to associate place names in source texts with gazetteers such as Pleiades and enables automatic mapping (Simon et al. 2017; Barker, Palladino, and Gordin 2024). INCEpTION (Eckart de Castilho et al. 2018; Berti 2019) enables linguistic and named entity annotation as well as links to authority lists, but we need to turn to separate workflows (such as the Ugarit translation alignment editor [Palladino et al. 2023], which supports

Tufts University. See also Crane (2019, 2023); Crane, Babeu, et al. (2023); Crane, Shamsian, et al. (2023); and Crane et al. (2024).

2. Preliminary versions for most ATLAS services are available on an internal Tufts.edu server and are ready to be published on the public-facing https://atlas.perseus.tufts.edu/.
3. Data use and reuse often flows in both directions. While one goal of Beyond Translation and Perseus 6 has been to aggregate and reuse data produced by relevant projects, some of the projects listed in this paper, such as Perseids, Opera Graeca Adnotata, SEDES and Ugarit, have utilised data created by Perseus and its related partner project, Open Greek and Latin.
4. For more on the history of CC licences, see: https://creativecommons.org/timeline/.

word- and phrase-level alignments between source texts and translations). An entirely separate workflow emerged to produce and make available machine-actionable metrical analyses for more than 250,000 lines of Greek and Latin poetry (David Chamberlain's Hypotactic). The DICES project publishes metadata identifying and classifying direct speech in Greek and Latin epic (Forstall, Finkmann, and Verhelst 2022). Individual scholars have published projects such as SEDES, which identifies words in Greek epic that are in statistically surprising metrical positions (Sansom 2021; Sansom and Fifield 2023). The Ajax Multi-Commentary project uses the rich tradition of scholarship on Sophocles to show how we can aggregate and organise multiple commentaries (Romanello and Najem-Meyer 2024). The Homer Multitext project publishes new high-resolution images and diplomatic editions of the *Iliad*, as well as *scholia*, with links between transcriptions and images (Dué and Ebbott 2019; Smith and Blackwell 2023).[5]

Background

Before we outline the development of our own work over the past 40 years, we want to articulate some of the principles that have shaped this work and that are relevant to the most recent activities described in the rest of this paper. Sustainable integration of different categories of data has been a driving force behind the development of Perseus from the beginning. Planning for what is now the Perseus Digital Library began in the spring of 1985, with a substantial equipment grant from the Xerox Corporation and a planning grant from the Annenberg/CPB Project. Continuous development began in 1987 and has continued ever since.

The earliest versions of Perseus emphasised two classes of integration. **First**, one inspiration for Perseus was the realisation that a single system could display not only textual, but also visual information—a fundamentally radical idea in the early 1980s, when state-of-the-art computer terminals displayed monowidth (typewriter-style) ASCII characters without any formatting. As a graduate student, co-author Crane needed to move between two different Harvard libraries as he moved from philological to archaeological and art historical data, but even the best print publications had relatively few images—typically black and white and very low resolution in comparison with born-digital images. A major goal for Perseus was to combine, in a single digital space, textual data of various types with visual information such as images, maps, satellite photos and drawings (Crane 1996; Smith, Rydberg-Cox, and Crane 2000). To demonstrate such integration, we often used the description of Croesus' dedications at Delphi, showing how we could bring together maps, images and drawings of the site and objects much more effectively than was feasible with print. In the late 1990s, as

5. For more on the data used from these projects, and on collaboration between such projects and Perseus, see Crane, Shamsian, et al. (2023) and Crane (2023).

many other projects (the German Arachne project in particular) began to make digital images and metadata about the art and archaeological record available online, we shifted our focus to the textual record. In 2024, however, we began (as will be discussed below) to exploit the International Image Interoperability Framework to begin integrating the textual and material records once again.

Second, as early as the 1980s, we harnessed automatic analysis to create new links between previously separate classes of textual data. The Morpheus system (Crane 1991) is a rule-based morphological analyser for Classical Greek and Latin that was first developed in 1985 in the C programming language and is still in use today (Keersmaekers et al. 2019; Keersmaekers 2021; Keersmaekers and Van Hal 2022). Morphologically, Greek and Latin are much more complex than languages such as English, French and even German. In print culture, readers have often struggled to match inflected forms on a printed page (e.g. *ênenkas*, "you (sg) carried") with the relevant dictionary entry (e.g. *pherô*, "I carry"). Given an inflected Greek or Latin form, Morpheus provides every possible morphological analysis (e.g. *ênenkas* is second person singular aorist indicative active) and normalised dictionary form (e.g. *pherô*), matching its tables of stems, endings and combination rules. This system had two basic applications: (1) readers could click on a form and follow links to the morphological analysis and then to machine-readable Greek and (starting in the late 1990s) Latin dictionaries; (2) users could generate searches for those forms: for example, they could ask for *pherô*, "carry," and retrieve *fereis*, "you (sg.) carry," *eferon*, "I or they were carrying," *oisete*, "you (pl.) will carry," *enêngektai*, "she/he/it has been carried" and so on.

Third, from the earliest stages of planning we designed Perseus to be as sustainable as possible. The software development and scholarly explorations that led to Perseus began with work on the *Thesaurus Linguae Graecae*, a corpus of Greek literature then available on magnetic tape for third party development (now locked behind a proprietary paywall). This experience made it clear that content could, and should, be separate from the software for publication and analysis. The life cycle of software is short; while data, especially in fields such as Greco-Roman studies, has value that persists not only over years, but over centuries and millennia.

Our work began before the Text Encoding Initiative (TEI) would document guidelines on the use of textual markup, but we were already familiar with the principles behind the TEI. On a snowy night in 1985, the authors of DeRose (1990) presented the case for a generalised model of text content (with a talk that bore the same title as their classic paper: "What is text, really?") As a result, we adopted SGML (the predecessor to XML) and would later revise our markup to follow the TEI. The investment in TEI XML was onerous at first; tools for creating and validating SGML were still at an early stage of development and we had to throw out much of the laboriously added markup when we published our sources in

HyperCard. The investment would pay dividends over the years, however, as tools improved and the markup increasingly facilitated maintenance and updates. One of Perseus's first major digitisation projects, for example, the *Intermediate Liddell Scott Greek-English Lexicon* (based upon Liddell and Scott 1882), was completed in 1985 and is still in active use 40 years later. The current version is available on GitHub (https://github.com/helmadik/MiddleLiddell). The GitHub version has notably been enhanced by Helma Dik, a scholar who works with, but has never been part of, the Perseus project.[6] In our view, a sustainability strategy should—by choosing an open licence and making public statements—encourage others to take ownership of and enhance any digital scholarly product.

Perseus before Perseus 6.0

The first five versions of the Perseus Digital Library augmented data extracted from print sources (e.g. TEI XML transcriptions of editions and reference works, and catalogue entries on art objects and archaeological sites converted into metadata) with automatically generated annotations (such as the Morpheus output described above and named entity annotations classifying people, places and organisations, which were then linked these to authority lists such as the Pleiades Gazetteer).

These five versions reflect two parallel threads of development. On the one hand, the evolution of Perseus reflects how rapidly software has changed and how agile we have needed to be. The publication platform for Perseus 1.0 and Perseus 2.0—Apple's HyperCard—had its final release in 1998. Perseus 3.0, written primarily in the Perl programming language, appeared in its earliest form in 1995. It remained in use for almost a decade, giving way to Java-based Perseus 4 in 2004. Perseus 4, in turn, remains the most widely used version and has now been in use for more than two decades. Development of Perseus 4 ended in 2013, however, and the system now runs on virtual machines based primarily on a computing environment that is more than a decade old. We began work on a new platform, primarily written in Python, in 2018 and we continue to build upon this platform today.

On the other hand, whereas we have abandoned earlier code bases, we continue to use and enhance data that was collected decades ago—in one case (the so-called Middle Liddell Greek-English Lexicon based upon the *Intermediate Liddell Scott Greek-English Lexicon*) almost 40 years. Many of the older XML files have changelogs that go back to the late 1980s. While we hope that the longevity of software platforms will improve over time (and lessen the need to

6. For more than 15 years, Helma Dik has also led work at the University of Chicago that has made openly licensed materials from Perseus and other sources available through its Philologic system, with its own browsing and advanced searching capabilities.

engineer new systems), we are pleased (and honestly relieved) with the extent to which our data has proven to be sustainable. We have of course had to spend substantial amounts of time updating our content—we still have not updated every Perseus text so that it is compatible with the Canonical Text Services data model—but each change has reflected an upgrade that made our data better structured and more sustainable.

- 1992: Perseus 1.0 *Interactive Sources and Studies on Ancient Greece*, published by Yale University Press. This included a videodisc, CD ROM and print documentation. Production language: Apple's HyperCard.
- 1995: A web version of the Perseus Digital Library can be found at http://www.perseus.tufts.edu/. David A. Smith was the lead developer, and the primary programming language was Perl. As noted below, we were able to produce a web version of Perseus years before the CD ROM-based Perseus 2.0 could be published.
- 1997: Perseus 2.0: *Interactive Sources and Studies on Ancient Greece*, published by Yale University Press. This included five CD ROMs and print documentation. Production language: Apple's HyperCard. Perseus 2.0 built upon the software backend workflow and frontend developed for Perseus 1.0. The five CD ROMs allowed us to move on from the analogue videodisc of Perseus 1.0 and to create the first fully digital version.
- 2000: Platform Independent Perseus, published by Yale University Press. This contained the same content as Perseus 2.0, but was accessible both on Macintosh and Windows systems.
  - 2000: The Perseus Digital Library on the Web became sufficiently well-developed for us to consider it a full version of Perseus and describe it as Perseus 3.0.
- 2004: Perseus 4.0 (known as the Hopper). This second web version of the Perseus Digital Library is available at http://www.perseus.tufts.edu/hopper/. David Mimno was the lead developer and the primary programming language was Java. Active development on Perseus 4.0 ended in 2013, but the system continues to run on a suite of virtual machines, supported by Tufts University. Two decades later, Perseus 4.0 remains (as of November 2024) the most commonly used version of the Perseus Digital Library.
- 2018: Perseus 5.0 (known as the Scaife Viewer).[7] James Tauber was the lead developer. The primary development language was Python. Scaife was based upon the Canonical Text Services (CTS) data model. Scaife built upon the CapiTainS Software Suite and Guidelines for Citable Texts[8] and

7. The Scaife Viewer is named after Ross Scaife, a digital classics pioneer who embodied the spirit of collaboration and who set an early example in establishing open access and openly licensed data as the standards upon which digital classics now depends. The initial release of the Scaife Viewer was on March 15, 2018, the tenth anniversary of his premature passing on March 15, 2008.
8. The CapiTains suite of tools and guidelines were developed by Thibault Clérice, Bridget Almas and Matthew Munson. For an interesting discussion of their development and important lessons learned, see Almas and Clérice (2017).

allowed Perseus to publish new content. This currently includes 2,669 works in 3,776 editions and translations (1,941 in Greek and 631 in Latin), with 83.8 million words in all languages (40.6 million in Greek, 16.4 million in Latin). Brill adopted Scaife as the platform for all of its scholarly editions (Brill's Scholarly Editions). Users of Scaife who have access to these editions (most of which are available behind a paywall) will see the resemblance with Scaife in the page design. As of November 2024, two editions (*A Literary History of Medicine* and *The Pez Brothers' Correspondence*) are open access.[9] While Brill's content may be largely proprietary, the contributions its support team has made to the Scaife software platform are available on GitHub. In effect, Scaife represents a collaboration between Perseus, an academic project committed to open scholarship, and Brill (now De Gruyter Brill), a traditional publisher that relies upon proprietary control (although it has begun to support open access as a publication option).

Readers scanning the above list of Perseus versions will notice a chronological anomaly: the version of the Perseus Digital Library, which we later came to refer to as Perseus 3.0, was first published in 1995—two years before Yale would be able to publish Perseus 2.0. Our earlier investment in structured data (such as TEI SGML/XML) allowed David A. Smith to create an initial version of the Perseus Digital Library as an unfunded side project. While we had worked with Yale University Press to publish the first editions of Perseus and had, in so doing, hoped to help Yale and other university presses develop the infrastructure to develop digital projects, such a development was premature in the 1990s, and partnership with a publisher no longer made sense when the World Wide Web emerged in the early 1990s.

The Scaife Viewer was able to support core features of digital editions, including interactive textual notes, automatic dictionary lookup and integration of commentary and reference works. Brill, in particular, applied Scaife not only to traditional editions, but also to editions of so-called fragmentary works (e.g. Jacoby Online), which document surviving evidence of texts that have disappeared. Editions of fragmentary works extract quotations from surviving texts that describe, paraphrase or explicitly quote works that are otherwise lost. Fragmentary editions are thus meta-editions, with one edition of a fragmentary work building on dozens or hundreds of other editions.[10]

9. The Berlin-Brandenburg Academy of Sciences and Humanities has also experimented with Scaife and has published sections of its Greek edition and German translation of the *De Locis Affectis* by Galen on Scaife: https://scaife.perseus.org/library/urn:cts:greekLit:tlg0057.tlg057/.
10. For more on the state of the art for digital editions of fragmentary texts, beyond the approach taken by Scaife and Brill, see for example Berti (2021).

The Beyond Translation project

The Beyond Translation project[11] (2019–2023: Crane 2019, 2023; Crane, Babeu, et al. 2023; Crane, Shamsian, et al. 2023; Crane et al. 2024) allowed us to develop a prototype for Perseus 6. Our primary goal was to create a reading environment that could integrate and present a much wider range of born-digital annotations than had been feasible in Perseus 1 to 5. Treebanks, for example, contain not only part-of-speech tagging and regularised dictionary forms for each word in a corpus, but also syntactic role and dependency for each word in a sentence. We needed to be able to represent texts not only as texts with annotations, but also graphically as trees. We also needed to be able to represent multiple layers of annotations associated with a text. The following examples illustrate several categories of data that we wanted to represent and that required us to integrate data from standoff markup with one or more textual sources.

First, traditional commentaries constitute one well-known document class that requires standoff markup. Print commentaries can present text on the upper part of the page and commentary on the bottom, but the text and commentary are nevertheless separate—albeit parallel—documents.[12] The basic principle is that commentaries quote and comment upon phrases, words and (often) morphemes or other subsets of words. Perseus 4 simply used citations to link texts and commentary: for example, a reader looking at Thucydides book 1, chapter 33, section 2 would see any comments that were associated with that chunk of text. With Perseus 6, we now can link from spans in the source text to a commentary.

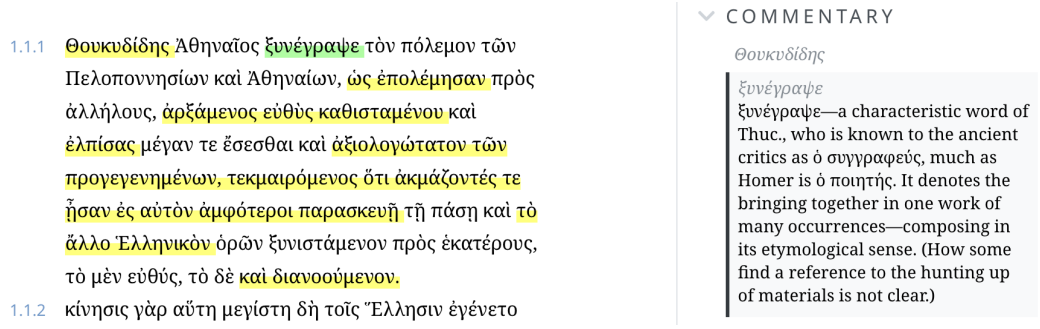

Figure 1: A commentary on the opening of Thucydides' *History of the Peloponnesian War*

Beyond Translation. Thucydides. *History of the Peloponnesian War*. Commentary: https://preview.scaife-viewer.eldarion.com/thucydides-commentary

In figure 1, yellow highlighting identifies words that have comments. The reader has selected *xunegrapse*, "he composed/wrote," and the system displays

11. https://beyond-translation.perseus.org/
12. Perseus team member Sarah Abowitz recently explored the content and structure of print commentaries and how they might be improved for digital library users, see Abowitz, Crane, and Babeu (2024).

the comment on this particular word. The ability to associate particular spans of a text with data can be applied to many classes of annotation, not just traditional commentary.

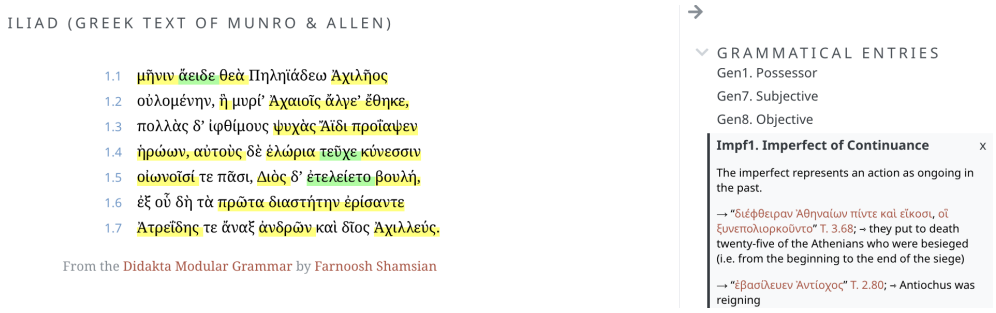

Figure 2: Links to a language-specific (Ancient Greek) grammar

Beyond Translation, *Iliad*. Didakta Modular Grammar by Farnoosh Shamsian: https://beyond-translation.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7?mode=grammatical-entries

Figure 2 illustrates links from individual words to explanations in a machine-readable grammar. The reading environment uses highlighting to identify words with grammatical annotations and then displays just those grammatical features that appear in the selected passage. When readers select an annotated word, they see the grammatical explanation on the right, while all words with the same grammatical feature have an additional layer of highlighting: for example, *aeide*, "sing!"*, teuche*, "it was fashioning," and *eteleieto*, "it was brought to completion," have all been tagged as instances of the imperfect of continuance. Links from spans to external datasets allow us to support an open-ended range of annotation classes, of which the following provide some examples.

Second, we consider word- and phrase-level alignments between source texts and translations. We can generate these alignments between Greek or Latin source texts and English translations with reasonable accuracy (ca. 80%), but we can also manually create alignments between source texts and translations. Further, we can generate born-digital alignments designed to illustrate the literal meanings of a source text. And of course, in a digital reading environment, we can offer both literal and literary translations so that readers can choose which they wish to see.

In figure 3, Maryam Foradi manually aligned words and phrases taken from Persian poetry by Hafez with the 1891 English translation by Henry Wilberforce Clarke. In the figure above, the reader has moused over "Saki" and sees that name highlighted in the Persian original. More importantly, however, the reader can see that the words in light red have no equivalent in the original Persian. By showing what the translator has added, the annotator has revealed to readers with no knowledge of Persian that a layer of religious allegory has

been imposed upon the original (Palladino, Foradi, and Yousef 2021; Foradi et al. 2019). Translation alignment can be revealing in itself, but becomes much more powerful when readers can go beyond the translation and explore the form and function of each word in a source text.

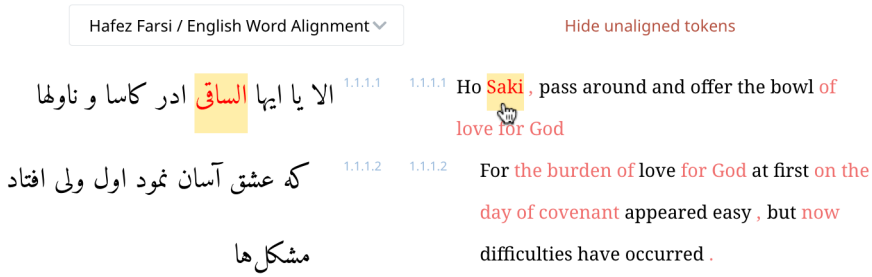

Figure 3: Manual alignment of a Persian poem by Hafez with a nineteenth-century English translation

Beyond Translation. *Divan*. Hafez Farsi/English word alignment: https://beyond-translation.perseus.org/reader/urn:cts:farsiLit:hafez.divan.perseus-far1-hemis:1.1?mode=alignments&rs=urn%3Acite2%3Ascaife-viewer%3Aalignment.v1%3Ahafez-farsi-english-word-alignment-temp

This leads to the second class of data: linguistic annotations. The Perseus team had begun planning to develop rich linguistic annotation in 2001. More than one million words each of Greek and Latin are available in manually treebanked form, while machine learning allows us to produce automatically generated treebanks for any online Greek or Latin corpus.

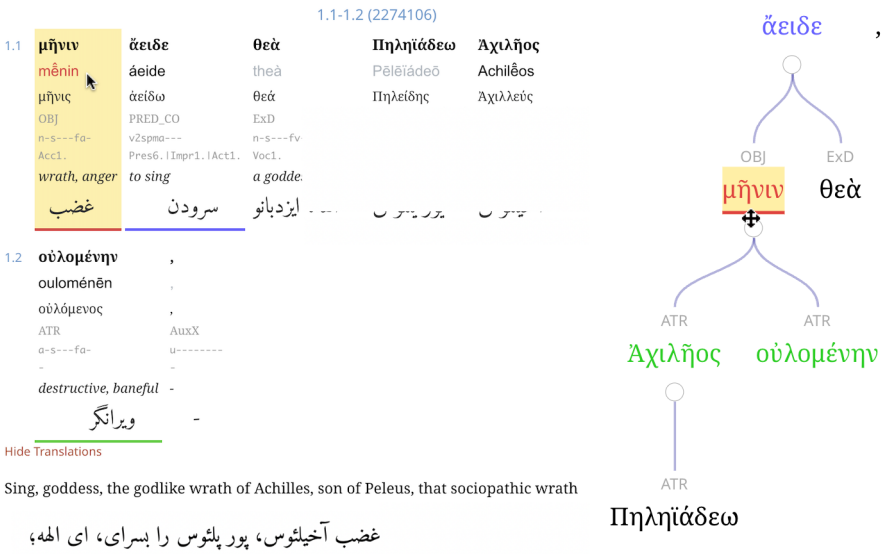

Figure 4: The opening of the Homeric *Iliad* with linguistic annotations

Beyond Translation. *Iliad*. Greek text from Munro & Allen. Treebank by Gregory Crane from Ancient Greek Dependency Treebank: https://beyond-translation.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7?mode=syntax-trees&collectionUrn=urn%3Acite2%3Abeyond-translation%3Atext_annotation_collection.atlas_v1%3Ail_gregorycrane_gAGDT

Figure 4 above (*left*) shows annotations for the *Iliad* that address the needs of readers who wish to push beyond translation alignments and explore the form and function of each word.There are seven layers of annotation for each word in the opening of the Homeric *Iliad*.

1. Transliteration
2. Regularised dictionary form
3. Syntactic role (using a dependency grammar for cross-lingual analysis)
4. Part-of-speech tagging
5. A (potentially language-specific) grammatical tag
6. An English gloss
7. A Persian gloss

These seven layers are drawn from different sources and, for all practical purposes, can only be managed as a series of linked datasets. Layer 1 was automatically generated—unlike languages such as Arabic and Persian, Greek regularly includes short vowels and can be automatically transliterated. Layers 2, 3 and 4 are derived from the Perseus Dependency Treebank. Layer 5 is based on a separate stream of annotations that link words in the text to a language-specific Greek grammar to shed light on grammatical features that cannot be inferred from a dependency grammar that was designed to represent shared features across multiple languages. Layers 6 and 7 are English and Persian glosses. Persian is included because collaborator Farnoosh Shamsian has completed a dissertation on localising the study of Ancient Greek in Persian.

Translations of the Greek sentence into English and Persian are visible below. The goal is to support an open-ended number of modern languages and to localise the platform as a whole (e.g. replace the English with Persian or some other language).

Third, the Greek quoted above is in poetic form and that poetic form is not easy for those studying Greek to decipher. Here we can draw upon yet another class of annotation: machine-actionable metrical analyses and recorded readings so that audiences can hear poetry as poetry and learn how to read poetry in metrical form.

The figure below presents a metrical analysis for the opening lines of the *Iliad*, with darker grey representing long syllables, lighter grey shorter syllables and dark vertical bars delineating breaks between the six metrical units of the hexameter. To the right, a button allows readers to listen to a recording of the lines being read. Together, the diagram and sound recording make it possible for readers with no Greek to learn how to read Greek poetry (this is now a regular assignment for students who do not know Greek). When readers combine the metre and recording with the aligned translation and linguistic annotation, they are able to engage directly with the Greek in ways that were not previously feasible.

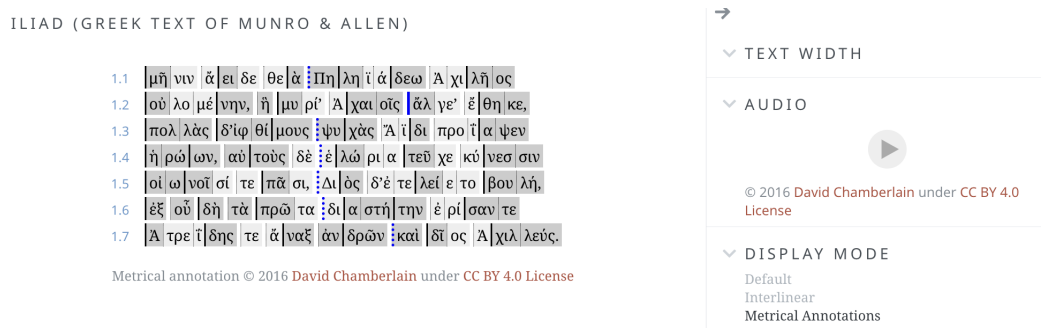

Figure 5: Metrical analysis and recording of the opening of the *Iliad*

Beyond Translation. *Iliad*. Greek text from Munro & Allen. Metrical annotations by David Chamberlain, imported from Hypotactic: https://beyond-translation.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7?mode=metrical

Note that the metrical annotations do not correspond to word breaks: metrical units often begin and end in the middle of words. Thus, this class of annotation (along with other forms such as analysis of the morphemes within words) requires an ability to go beyond the token level and annotate chunks of a word.

Other examples discussed elsewhere include the use of named entity annotation to generate maps and social networks and links at the word and character level between transcriptions and images of inscriptions, papyri or manuscripts.

The Beyond Translation project implemented initial frontends for born-digital annotations such as those listed above. In so doing, Beyond Translation also created an initial backend that imported different types of data from different projects into a coherent backend. In late 2023, a new National Endowment for the Humanities (NEH)-funded project, Perseus on the Web: Preparing for the Next Thirty Years, reviewed and reorganised the backend in general and the particular formats that we used to make data from multiple sources interoperable.

Perseus 6, the ATLAS architecture and the CTS Data Model

While Scaife addressed a number of core needs, James Tauber and his collaborators (in particular Jacob Wegner) felt that they needed an approach to complement the Scaife architecture. With funding from Mellon and then from the NEH, they began developing ATLAS, the Aligned Text and Linguistic Annotation Server architecture.

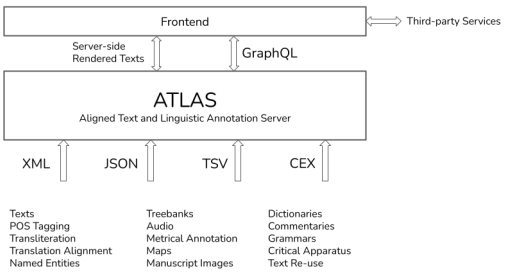

Figure 6: The ATLAS architecture

The CTS data model allows us to integrate data between CTS-compliant TEI XML and a wider range of data in ATLAS. It also allows us to identify chunks of data with a URN (Uniform Resource Name)[13] such as the following:

```
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7
```

- cts—This defines the CTS protocol. The data that follows in the URN is unique within this CTS protocol.
- greekLit—This is the name space that Perseus and the Open Greek and Latin project use to define Ancient Greek literature.
- tlg0012—This describes a text group. In most cases, a CTS text group will correspond to an author (e.g. Plato or Vergil), but the more general term text group allows us to also deal with cases (such as the New Testament) where we need to be able to address a collection of works by multiple authors as a single unit (tlg0031). Likewise, we use tlg0012 to describe the *Iliad* and the *Odyssey*, whether or not we view these to be the works of a single individual.
- tlg001—This describes the particular work and, in this case, tlg0012.tlg001 designates the *Iliad*.
- perseus-grc2—This designates the particular edition of the *Iliad* that we are citing (in this case, the 1908–1920 Munro & Allen *Oxford Classical Text of the Homeric Epics*).
- 1.1-1.7—This defines a text range that extends from line 1 through to line 7 of book 1 of the *Iliad*.

The following sections provide examples for some, but not all, of the annotation classes that we manage in ATLAS. Readers can follow the repositories available on GitHub for both the tagging pipeline and data preparation for more information.

13. https://en.wikipedia.org/wiki/Uniform_Resource_Name

Scaife texts in ATLAS

While ATLAS helps us integrate data from many different sources, it also provides us with a simpler way to integrate basic textual data. We have completed an initial conversion into ATLAS of all of the texts that are now available in Scaife.

The CapiTainS CTS software library has explicit guidelines on how to structure XML markup, how to organise XML files in a directory and how to represent metadata. These requirements help develop more interoperable collections and can thus simplify search, analysis and visualisation. Nevertheless, the barrier to entry is substantial. ATLAS allows us to add texts that are not CapiTainS-compliant. ATLAS only requires a flat, tab-separated values (TSV) format that includes a unique reference and a chunk of text. We can convert TEI XML into this format, producing a TSV file that contains every book/chapter/section of a particular edition of Thucydides' *History of the Peloponnesian War* or every line from a particular edition of Sophocles' *Antigone*. Thus, the TEI XML version of Henry Stuart Jones' edition of Thucydides (available in Scaife) is in the following format:

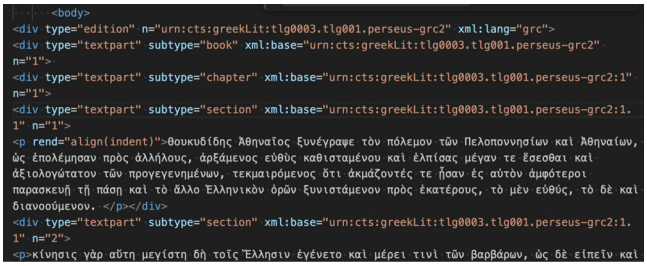

Figure 7: Opening of Thucydides in TEI XML

TEI XML file for Thucydides. GitHub. Perseus Digital Library: https://github.com/PerseusDL/canonical-greekLit/blob/master/data/tlg0003/tlg001/tlg0003.tlg001.perseus-grc2.xml

For Perseus 6, we create a TSV file with the name tlg0003.tlg001.perseus-grc2.tsv and then store the lines version of this file for annotation.

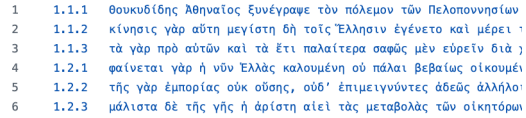

Figure 8: ATLAS TSV version of the opening of Thucydides

TSV file for Thucydides. GitHub. Scaife Viewer: https://github.com/scaife-viewer/tagging-shard-06/blob/main/tlg0003/tlg001/tlg0003.tlg001.perseus-grc2.tsv

Because the TSV file becomes the basis for further computation, we can add new texts much more easily by converting them into this TSV identifier+text format. The more heavily structured format of the CapiTainS framework does

promote greater consistency, but the TSV identifier+text format is useful for many purposes—indeed, more useful than the XML.

Morpho-syntactic analysis

Our goal is to provide multiple layers of linguistic annotation.

1. Curated annotations produced by human annotators. More than two million words of Greek and Latin have been manually treebanked and made available on GitHub. In at least one case, we have different manual treebanks for the same passage (e.g. both the PROIEL project and Vanessa Gorman have annotated the same passages in Herodotus) and will need to help people compare different linguistic annotations of comparable authority. For now, however, our goal is to determine whether we have curated annotations for given passages and offer these first.

2. Automatically generated treebanks with some curation. The GLAUx project has published 20 million words of treebanked Greek, including manually and automatically generated data. The accuracy for the automated work, especially lemmatisation and part-of-speech tagging, is very good. GLAUx includes a number of texts that are not available in the Perseus/Open Greek and Latin collection (particularly short and fragmentary works), but Perseus contains 40 million words of Greek and 16 million words of Latin.

3. Automatically generated treebanks (currently using spaCy) with the best available models. We use the greCy and LatinCy spaCy pipelines to generate default morpho-syntactic analyses for all texts in Perseus. The spaCy treebank data provides a default layer upon which we can rely in cases where we do not have curated treebank data or hybrid human/machine treebanks like GLAUx.

The spaCy analyses of Open Greek and Latin texts are more bulky (ca. 5 gigabytes) than GitHub prefers to support in a single repository. We have broken the data up into multiple smaller repositories (the information about how to find data on particular authors is here). For now, we store this data in the CoNLL-U format.

```
1.1.1   0       Θουκυδίδης      PROPN   Ne      Case=Nom|Gender=Masc|Number=Sing         Θουκυδίδης      nsubj   2
1.1.1   1       Ἀθηναῖος        ADJ     A-      Case=Nom|Degree=Pos|Gender=Masc|Number=Sing      Ἀθηναῖος        amod    0
1.1.1   2       ξυνέγραψε       VERB    V-      Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act      συγγράφω        ROOT    2
1.1.1   3       τὸν     DET     S-      Case=Acc|Definite=Def|Gender=Masc|Number=Sing|PronType=Dem       ὁ       det     4
1.1.1   4       πόλεμον NOUN    Nb      Case=Acc|Gender=Masc|Number=Sing         πόλεμος obj     2
1.1.1   5       τῶν     DET     S-      Case=Gen|Definite=Def|Gender=Masc|Number=Plur|PronType=Dem       ὁ       det     6
1.1.1   6       Πελοποννησίων   ADJ     A-      Case=Gen|Degree=Pos|Gender=Masc|Number=Plur      Πελοποννήσιος   nmod    4
1.1.1   7       καὶ     CCONJ   C-      καί     cc      6
1.1.1   8       Ἀθηναίων        ADJ     A-      Case=Gen|Degree=Pos|Gender=Masc|Number=Plur      Ἀθηναῖος        conj    6
1.1.1   9       ,       PUNCT   Z               ,       dep     2
1.1.1   10      ὡς      SCONJ   G-              ὡς      mark    11
1.1.1   11      ἐπολέμησαν      VERB    V-      Aspect=Perf|Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin|Voice=Act      πολεμέω advcl   9
1.1.1   12      πρὸς    ADP     R-              πρός    case    13
1.1.1   13      ἀλλήλους        PRON    Pc      Case=Acc|Gender=Masc|Number=Plur|PronType=Rcp    ἀλλήλων obl     11
```

Figure 9: Linguistic analysis of the opening of Thucydides using the greCy pipeline

The figure above shows output for the opening of Thucydides.

Dictionaries

We have added most of the dictionaries available in Perseus to ATLAS and plan to add more (notably the *Intermediate Liddell Scott Greek-English Lexicon* and Slater's *Lexicon to Pindar*). The format is JSON and captures the structure and inline formatting in the TEI XML. An entry from Cunliffe's *Lexicon of the Homeric Dialect* follows below:


```
{"headword": "ἀγηνορίη", "data": {"content": "<p>-ης, ἡ</p> <p>[ἀγήνωρ.]</p>"},
"senses":

[{"label": "1", "urn": "urn:cite2:exploreHomer:senses.atlas_v1:1.117",
"definition": "Courage, spirit",

"citations":
[{"urn": "urn:cite2:scholarlyEditions:citations.v1:1.117_1",
"data": {"ref": "Il. 12.46", "quote": null, "urn":
"urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:12.46"}},
{"urn": "urn:cite2:scholarlyEditions:citations.v1:1.117_2",
"data": {"ref": "Il. 22.457", "quote": null, "urn":
"urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:22.457"}}]},

{"label": "2", "urn": "urn:cite2:exploreHomer:senses.atlas_v1:1.118",
"definition": "The quality in excess or with arrogance.",

"citations": [], "children":
[{"label": "", "urn": "urn:cite2:exploreHomer:senses.atlas_v1:1.119",

"definition": "In pl.", "citations":
[{"urn": "urn:cite2:scholarlyEditions:citations.v1:1.119_1", "data": {"ref": "Il.
9.700", "quote": null, "urn":
"urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:9.700"}}]}]}], "urn":
"urn:cite2:exploreHomer:entries.atlas_v1:1.60"}
```


Figure 10: An entry from Cunliffe's *Lexicon of the Homeric Dialect*

Scaife Viewer. GitHub: https://github.com/scaife-viewer/atlas-data-prep/tree/main/test-data/dictionaries/cunliffe-1-lex

Textual notes

Textual notes documenting different versions of a source constitute a core element for any critical edition. The unofficial but de facto development partnership with Brill (which uses Scaife for its scholarly editions) has provided us with a solution to that problem. We can and will include textual notes wherever these are available.

For now, we store textual notes as a special type of commentary. Figure 11 below modifies the first line of "The Passionate Shepherd to His Love" by Christopher Marlowe. The default line is: "Come live with mee and be my love." The annotation above it indicates that there is a note associated with the phrase "live with me" and that this note, in this case, is a textual variant. The underlying data uses the TEI XML guidelines for textual editions to encode the variants. The Marlowe edition was actually produced in the late 1990s and we were able to import TEI annotations that were more than 25 years old, illustrating the initial stability and sustainability of that particular work and of this aspect of TEI XML.

```
{
  "references": [
    "urn:cts:engLit:mds822-32.tpsthl-1599.pdl-eng:1.1"
  ],
  "commentary": "<span>If thou wilt live</span>",
  "fragment": "live with mee",
  "ve_refs": [
    "1.1.t2",
    "1.1.t3",
    "1.1.t4"
  ],
  "idx": "1",
  "urn": "urn:cite2:scaife-viewer:commentary.v1:commentary2",
  "witnesses": [
    {
      "value": "Rs",
      "label": "MS Rodenbach"
    }
  ]
},
```


Figure 11: A textual note for a poem by Christopher Marlowe

Beyond Translation.

Text alignments

Another category of data and services provided is text alignments.[14]


```
{
  "urn": "urn:cite2:scaife-viewer:alignment-record.v1:iliad-word-alignment-parrish-998078bc3bab42978b47fa8e8b852cae_3",
  "relations": [
    [
      "urn:cts:greekLit:tlg0012.tlg001.parrish-eng1:1.1.t4",
      "urn:cts:greekLit:tlg0012.tlg001.parrish-eng1:1.1.t5"
    ],
    [
      "urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1.t1"
    ]
  ]
},
```


Figure 12: A textual alignment between two words in an English translation and one word in a Greek edition of the *Iliad*

In figure 12, each alignment between one text and another is a unique annotation with a unique, citable identifier. The format allows for many-to-many alignments. In the example above, two tokens in Amelia Parrish's translation of the first book of the *Iliad* are aligned with one token of Greek. An individual could create alignments showing how different translations analysed the same word or the individual could align a pre-existing translation with the source text. They could also create a born-digital translation designed for alignment. We use the term text alignments because the same method could be used to align different editions of a Greek text as well as a Greek text with an English or Persian translation.

Syntax trees (treebanks)

Figure 13 shows part of a treebank represented as JSON. The tagset for this treebank is based on the Perseus Dependency Treebank. Our goal is to begin

14. While only a brief example is given here, further discussion of text alignment and how it has been implemented in Beyond Translation and the beginning of Perseus 6 can be found here: https://pdldatajournal.pubpub.org/pub/knjho2r7/release/1. See also Crane, Shamsian, et al. (2023).

using the tagset from the Universal Dependencies (UD) framework. That transition can largely be done automatically but will require some curation. The UD data, however, can be represented by the JSON that we have chosen so that we can manage treebanks in multiple formats.


```
[
  {
    "urn": "urn:cite2:beyond-translation:syntaxTree.atlas_v1:tlg0008-tlg001-grc-1",
    "treebank_id": "1",
    "words": [
      {
        "id": 1,
        "value": "φύλαρχος",
        "head_id": 79,
        "relation": "SBJ",
        "lemma": "φύλαρχος",
        "tag": "n-s---mn-"
      },
      {
        "id": 2,
        "value": "δ'",
        "head_id": 79,
        "relation": "AuxY",
        "lemma": "δέ",
        "tag": "d--------"
```


Figure 13: A treebank represented as JSON

Syntax Tree. ATLAS. Available on GitHub: https://github.com/scaife-viewer/atlas-data-prep/blob/main/test-data/annotations/syntax-trees/gorman_syntax_trees_017_tlg0008.tlg001.perseus-grc1.json

Audio annotations

The GitHub repository with audio annotations is available here.

urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1	https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_1.mp4
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.2	https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_2.mp4
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.3	https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_3.mp4
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.4	https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_4.mp4
urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.5	https://storage.googleapis.com/explorehomer-prod-media/tlg0012.tlg001/audio/1/line_5.mp4

Figure 14: Lines of the *Iliad* signed to recorded performances of each line

At present, we support alignments of text chunks to particular MP4 files. Each line in the TSV file points to a line of the *Iliad*, with the first field pointing to the text of a particular edition and the second to a recorded reading stored in a server.

Attributions/credits

Arguably the most important challenge we face is to preserve and aggregate fine-grained credits for born-digital annotations.[15] Credits are easily represented for articles, editions, and even individual commentary notes. In a true digital library, however, a researcher may contribute one or more machine-actionable annotations: for example, they may provide the syntactic structure of a

15. The repository for annotation and credit information is available here: https://github.com/scaife-viewer/atlas-data-prep/tree/main/test-data/annotations/attributions.

sentence, identify which Antonius is which in a given context, create alignments between a Greek word and its translations in half a dozen passages or may publish metrical analyses for 250,000 lines of Greek and Latin poetry.

In the original workflow for the Perseus Dependency Treebank, for example, two independent annotators analysed each sentence while a senior editor reviewed and adjudicated places where the annotators differed. A final corrected version was published, with three credits for each sentence. When another project, run by a former Perseus project member and long-time collaborator, adapted the treebank to a new format, the credits were initially lost. There was no ill intention. There was simply no existing framework to preserve credits in the new format and the project needed to change format. Likewise, another major treebank project includes both automatically generated and manually curated treebank data. This second project indicates whether a sentence was automatically or manually produced, but the treebank does not identify the annotator. Because the manually produced treebanks are all available on GitHub, it is possible to identify who did what, but this second treebank project did not do so. Its focus was to aggregate existing treebanks and produce new ones as quickly as possible. Both treebank projects were working with limited time and probably intended to go back to resolve the credits issue, but did not find that practical.

In Beyond Translation, we preserved all the credit information that we received. We now have an initial framework by which to represent credits from multiple projects and to make it possible for contributors to develop portfolios showing their contributions across projects.


```
[
  {
    "role": "Annotator",
    "person": {
      "name": "Alex Lessie"
    },
    "organization": {
      "name": "University of Pennsylvania, Philadelphia, PA, USA"
    },
    "data": {
      "references": [
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277120",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277121",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277122",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277123",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277124",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277125",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277126",
        "urn:cite2:exploreHomer:syntaxTree.v1:syntaxTree-tlg0012-tlg001-perseus-grc2-2277127",
[
  {
    "person": {
      "name": "Farnoosh Shamsian"
    },
    "role": "Annotator",
    "organization": {
      "name": "Universität Leipzig: Leipzig, Sachsen, DE"
    },
    "data": {
      "references": [
        "urn:cite2:scaife-viewer:alignment.v1:crito-greek-english-word-alignment-7b34509f15734bd7a20b873aeb08eaa1",
        "urn:cite2:scaife-viewer:alignment.v1:crito-greek-farsi-word-alignment-tr1-7b34509f15734bd7a20b873aeb08eaa1",
        "urn:cite2:scaife-viewer:alignment.v1:crito-greek-farsi-word-alignment-tr2-7b34509f15734bd7a20b873aeb08eaa1",
```


Figure 15: Sample attribution data for treebanks and translation alignments

Treebank alignment: Alex Lessie. Text alignment: Farnoosh Shamsian.

Figure 15 shows credits for treebanking individual sentences in the *Iliad* (above) and for aligning particular words and/or phrases between the Greek text of the *Crito* and an English translation, and between a Greek text and a Persian translation. We can now begin to aggregate credits as seen in figure 16.

Translator	Farshid Rahimi	3
Translator	Mahdi Shojaian	3
Annotator	Alex Lessie, University of Pennsylvania, Philadelphia, PA, USA	2,081
Annotator	Daniel Lim Libatique, College of the Holy Cross, Worcester, MA, USA	293
Annotator	Florin Leonte, Central European University of Budapest, Hungary	89
Annotator	Jack Mitchell, Tufts University, Medford, MA, USA	2,410

Figure 16: An initial report showing what contributions individuals have made to the ATLAS server

For now the data is relatively simple. In figure 16 we can see that Farshid Rahimi and Mahdi Shojaian each contributed three translated sentences. Alex Lessie helped treebank 2,081 sentences. However, the goal is to provide richer information (e.g. allow viewers to see the translated or treebanked sentences) and also to show where annotators have contributed to more than one project (e.g. treebanks and translations), but the above is a first step in that direction.

Next steps

As the ATLAS backend takes shape, our focus will shift to the next stage of work:

1. We need to build out the services available on the evolving ATLAS server: https://atlas.perseus.tufts.edu/.
2. We need to refine the ATLAS data already available on GitHub.
3. We need to add frontend support, developed in Beyond Translation (and discussed above) into the Scaife architecture. We will implement Perseus 6 by adding the ATLAS backend and Beyond Translation UI widgets to the earlier Scaife architecture.

Conclusions and ongoing work

This paper provides information about the first release of the ATLAS architecture and representative data.

References

Abowitz, Sarah, Gregory Crane, and Alison Babeu. 2024. "Bridging the Understanding Gap: Helping Readers Engage Directly with Foreign-Language Sources More Easily." Paper presented at the 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL *'24*) (Hong Kong, China, 16–20 December). Zenodo. Version v1. https://zenodo.org/records/14278721.

Almas, Bridget, and Thibault Clérice. 2017. "Continuous Integration and Unit Testing of Digital Editions." *Digital Humanities Quarterly* 11 (4). https://hal.science/hal-01709868/.

Barker, Elton, Chiara Palladino, and Shai Gordin. 2024. "Digital Approaches to Investigating Space and Place in Classical Studies." *The Classical Review* 74 (1): 1–19.https://doi.org/10.1017/S0009840X23002858.

Berti, Monica. 2019. "Named Entity Annotation for Ancient Greek with INCEpTION." In *Proceedings of the CLARIN Annual Conference 2019*, edited by Kiril Simov and Maria Eskevich, 1-4. Leipzig, Germany: CLARIN.

Berti, Monica. 2021. *Digital Editions of Historical Fragmentary Texts*. Digital Classics Books 5. Heidelberg: Propylaeum.https://doi.org/10.11588/propylaeum.898.

Celano, Giuseppe G. A. 2024. "Opera Graeca Adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek." ArXiv.https://arxiv.org/abs/2404.00739.

Crane, Gregory. 1991. "Generating and Parsing Classical Greek." *Literary and Linguistic Computing* 6 (4): 243–45. https://doi.org/10.1093/llc/6.4.243.

Crane, Gregory. 1996. "Building a Digital Library: The Perseus Project as a Case Study in the Humanities." In *DL '96: Proceedings of the First ACM International Conference on Digital Libraries (Bethesda, USA, 20–23 March)*, 3–10. New York: Association for Computing Machinery. https://dl.acm.org/doi/pdf/10.1145/226931.226932.

Crane, Gregory. 2019. "Beyond Translation: Language Hacking and Philology." *Harvard Data Science Review* 1 (2). https://doi.org/10.1162/99608f92.282ad764.

Crane, Gregory. 2023. "Beyond Translation: A Reading Environment for the Next Generation Perseus Digital Library." *Perseus Journal of Data Preservation and Sustainability*. https://pdldatajournal.pubpub.org/pub/el65xygp/release/1.

Crane, Gregory, Alison Babeu, Lisa Cerrato, Farnoosh Shamsian, James Tauber, and Jacob Wegner. 2023. "Perseus 6.0: Towards a Next Generation Reading Environment for Born-Digital Editions and Corpora." In *Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (Santa Fe, USA, 26–30 June)*, 297–98. IEEE. https://doi.org/10.1109/JCDL57899.2023.00066.

Crane, Gregory, Alison Babeu, Lisa Cerrato, Farnoosh Shamsian, Jacob Wegner, James Tauber, and Charles Pletcher. 2024. "Beyond Translation: Translation as Gateway Rather Than Endpoint." White paper on a concluded project. Tufts University.

Crane, Gregory, Farnoosh Shamsian, Lisa Cerrato, Alison Babeu, Amelia Parrish, Carolina Penagos, James Tauber, and Jake Wegner. 2023. "Beyond Translation: Engaging with Foreign Languages in a Digital Library." *International Journal of Digital Libraries* 14 (March): 163–176. https://doi.org/10.1007/s00799-023-00349-2.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What Is Text, Really?" *Journal of Computing in Higher Education* 1 (2): 3–26. https://doi.org/10.1007/BF02941632.

Dué, Casey, and Mary Ebbott. 2019. "The Homer Multitext within the History of Access to Homeric Epic." In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 239–56. Berlin, Boston: De Gruyter Saur. https://doi.org/10.1515/9783110599572-014.

Eckart de Castilho, Richard, Jan-Christoph Klie, Naveen Kumar, Beto Boullosa, and Iryna Gurevych. 2018. "Linking Text and Knowledge Using the INCEpTION Annotation Platform." In *2018 IEEE 14th International Conference on E-Science (e-Science) (Amsterdam, Netherlands, 29 October to 01 November)*, 327–28. IEEE. https://doi.org/10.1109/eScience.2018.00077.

Foradi, Maryam, Jan Kaßel, Johannes Pein, and Gregory R. Crane. 2019. "Multi-Modal Citizen Science: From Disambiguation to Transcription of Classical Literature." In *HT '19: Proceedings of the 30th ACM Conference on Hypertext and Social Media (Hof, Germany, 17–20 September)*, 49–53. New York: Association for Computing Machinery. https://doi.org/10.1145/3342220.3343667.

Forstall, Christopher W., Simone Finkmann, and Berenice Verhelst. 2022. "Towards a Linked Open Data Resource for Direct Speech Acts in Greek and Latin Epic." *Digital Scholarship in the Humanities* 37 (4): 972–81. https://doi.org/10.1093/llc/fqac006.

Gorman, Vanessa B. 2020. "Dependency Treebanks of Ancient Greek Prose." *Journal of Open Humanities Data* 6 (1): 1. https://doi.org/10.5334/johd.13.

Hudspeth, Marisa, Brendan O'Connor, and Laure Thompson. 2024. "Latin Treebanks in Review: An Evaluation of Morphological Tagging Across Time." In *Proceedings of the*

1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024) (Bangkok, Thailand and online, August), 203–18. Kerrville, Texas: Association for Computational Linguistics. https://aclanthology.org/2024.ml4al-1.21.

Keersmaekers, Alek. 2021. "The GLAUx Corpus: Methodological Issues in Designing a Long-Term, Diverse, Multi-Layered Corpus of Ancient Greek." In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021 (online, August)*, edited by Nina Tahmasebi, Adam Jatowt, Yang Xu, Simon Hengchen, Syrielle Montariol, and Haim Dubossarsky, 39–50. Kerrville, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.lchange-1.6.

Keersmaekers, Alek, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. "Creating, Enriching and Valorizing Treebanks of Ancient Greek." In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 109–17. Kerrville, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7812.

Keersmaekers, Alek, and Toon Van Hal. 2022. "Creating a Large-Scale Diachronic Corpus Resource: Automated Parsing in the Greek Papyri (and Beyond)." *Natural Language Engineering* 30 (5): 1035–1064. https://doi.org/10.1017/S1351324923000384.

Liddell, Henry George, and Robert Scott. 1882. *An Intermediate Greek-English lexicon: Founded Upon the Seventh Edition of Liddell and Scott's Greek-English Lexicon*. Oxford: Oxford University Press.

Palladino, Chiara, Maryam Foradi, and Tariq Yousef. 2021. "Translation Alignment for Historical Language Learning: A Case Study." *Digital Humanities Quarterly* 15 (3). http://www.digitalhumanities.org/dhq/vol/15/3/000563/000563.html.

Palladino, Chiara, Farnoosh Shamsian, Tariq Yousef, David J. Wright, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2023. "Translation Alignment for Ancient Greek: Annotation Guidelines and Gold Standards." *Journal of Open Humanities Data* 9 (1): 22. https://doi.org/10.5334/johd.131.

Romanello, Matteo, and Sven Najem-Meyer. 2024. "A Named Entity-Annotated Corpus of 19th Century Classical Commentaries." *Journal of Open Humanities Data* 10 (1): 1. https://doi.org/10.5334/johd.150.

Sansom, Stephen A. 2021. "*Sedes* as Style in Greek Hexameter: A Computational Approach." *TAPA (Society for Classical Studies)* 151 (2): 439–67. https://doi.org/10.1353/apa.2021.0017.

Sansom, Stephen A., and David Fifield. 2023. "SEDES: Metrical Position in Greek Hexameter." *Digital Humanities Quarterly* 17 (2). https://digitalhumanities.org/dhq/vol/17/2/000675/000675.html.

Simon, Rainer, Elton Barker, Leif Isaksen, and Pau De Soto Cañamares. 2017. "Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2." *Journal of Map & Geography Libraries* 13 (1): 111–32. https://doi.org/10.1080/15420353.2017.1307303.

Smith, Neel, and Christopher Blackwell. 2023. "Analytical Developments for the *Homer Multitext*: Palaeography, Orthography, Morphology, Prosody, Semantics." *International Journal on Digital Libraries* 24 (3): 179–84. https://doi.org/10.1007/s00799-023-00380-3.

Smith, David A., Jeffrey A. Rydberg-Cox, and Gregory R. Crane. 2000. "The Perseus Project: A Digital Library for the Humanities." *Literary and Linguistic Computing* 15: 15–26. https://doi.org/10.1093/llc/15.1.15.

Websites cited

Ajax Multi-commentary Project: https://github.com/mromanello/ajax-multi-commentary.

Ancient Greek and Latin Perseus Dependency Treebanks: https://perseusdl.github.io/treebank_data/.

Arachne: https://arachne.dainst.org/.

Brill's Scholarly Editions: https://scholarlyeditions.brill.com/.

CapiTainS Software Suite and Guidelines for Citable Texts: http://capitains.org/.

Daphne: https://github.com/francescomambrini/Daphne.

De Gruyter Brill: https://degruyterbrill.com/.

DICES (Digital Initiative for Classics: Epic Speeches): https://www.dices.uni-rostock.de/en/about-dices/.

GLAUx Trees: https://glaux.be/.

Homer Multitext: https://www.homermultitext.org/.

Hypotactic: https://hypotactic.com/.

INCEpTION: https://inception-project.github.io/releases/32.1/docs/user-guide.html.

International Image Interoperability Framework: https://iiif.io/.

Jacoby Online: https://scholarlyeditions.brill.com/bnjo/.

A Literary History of Medicine: https://scholarlyeditions.brill.com/lhom/.

Opera Graeca Adnotata: https://github.com/OperaGraecaAdnotata/OGA.

Pedalion Trees. GitHub.https://github.com/perseids-publications/pedalion-trees.

Perseids Treebanking Environment: https://github.com/perseids-publications.

Perseus 4.0 (the Hopper): https://www.perseus.tufts.edu/hopper/.

Perseus 5.0 (the Scaife Viewer): https://scaife.perseus.org/; source code: https://github.com/scaife-viewer.

Perseus—Beyond Translation: https://beyond-translation.perseus.org/.

Perseus 6.0—Scaife ATLAS Server: https://atlas.perseus.tufts.edu/.

Perseus under Philologic: https://perseus.uchicago.edu/.

Pleiades: https://pleiades.stoa.org/.

PROIEL: https://www.hf.uio.no/ifikk/english/research/projects/proiel/.

Recogito: https://recogito.pelagios.org/.

SEDES: https://github.com/sasansom/sedes.

Text Encoding Initiative (TEI): https://tei-c.org/.

The Pez Brothers' Correspondence: https://scholarlyeditions.brill.com/pez/.

Ugarit translation alignment editor: https://ugarit.ialigner.com/.

Universal Dependencies framework: https://universaldependencies.org/.

Introduction to Workflows – Part 2

Anne Baillot, Françoise Gouzi, Toma Tasovac

In the first part of our introduction to the inaugural issue of *Transformations*, we underlined how workflows, although key to the research process, still lack proper recognition. With this second series of articles dedicated to workflows, *Transformations* continues its contribution to changing that situation.

In the section on metadata-based workflows, Emiliano Degl'Innocenti, Francesco Pinna, Alessia Spadi, and Federica Spinelli examine the role of infrastructures in stabilising and ensuring the reusability of research workflows. Their article, **Supporting executable scientific workflows in a clustered Infrastructure: DARIAH-IT and H2IOSC**, introduces AEON (dAriah sErvice Oriented iNfrastructure), an upcoming platform designed to create, execute and manage scientific workflows based on a variety of rich metadata.

Infrastructures also play an important role in facilitating text-based workflows, as shown in **Fluffy Publication Workflow: Preserving Humanities Research Data with the TextGrid Repository** by Ubbo Veentjer, Stefan Buddenbohm, José Calvo Tello, Stefan E. Funk, Ralf Klammer, Nanette Rißler-Pipka, Alex Steckel, Lukas Weimer, George Dogaru, and Mathias Göbel. The authors describe the improvements to the TextGrid repository both in its operational procedures and through researchers' involvement in usability initiatives that guide its ongoing development.

Both papers reflect on how the DARIAH community contributes to the implementation of the FAIR principles while providing the infrastructural backbone needed to make workflows truly reusable. They also highlight the inherent complexity of solutions designed to encompass a range of research settings and questions, and offer concrete examples on the potentials and pitfalls of standardisation.

With the contribution by Kévin Roger, we open a new chapter: one dedicated to musicology. In **The Music Encoding Initiative: Its Generation Workflows**, Roger systematically addresses a range of steps, tools, and processes relevant to digital musicology workflows, providing an overview of the current methods and weighing their advantages and limitations. It serves as a unique entry point for those interested in working with the Music Encoding Initiative, and an illuminating perspective for non-specialists seeking to learn more about the challenges in the field.

While this first issue of *Transformations* cannot capture every aspect of digital scholarship in which workflows are pivotal or increasingly becoming a more valued asset, it still reveals the breadth of their relevance. We hope to have opened new avenues of inquiry, and made the contribution of workflows to a healthy research landscape more tangible—along with the underlying infrastructural challenges they entail. We also hope that future publications will build on this foundation, deepening reflection on workflows and further improving access to the rich array of resources and insights they bring to light.