Tomasz Umerle
The Institute of Literary Research
Polish Academy of Sciences
tomasz.umerle@ibl.waw.pl
ORCID: 0000-0002-7335-0568

Renata Rokicka
The Institute of Literary Research
Polish Academy of Sciences
University of Warsaw
rt.rokicka@uw.edu.pl
ORCID: 0000-0002-1246-1811

Patryk Hubar-Kołodziejczyk
University of Warsaw
p.hubar@uw.edu.pl
ORCID: 0000-0001-5582-2042

Cezary Rosiński
The Institute of Literary Research
Polish Academy of Sciences
cezary.rosinski@ibl.waw.pl
ORCID: 0000-0002-6136-7186

Katarzyna Jarzyńska
The Institute of Literary Research
Polish Academy of Sciences
University of Warsaw
k.jarzynska2@uw.edu.pl
ORCID: 0000-0001-8181-963X

POZNAŃSKIE STUDIA SLAWISTYCZNE
NR 30 (2026)
DOI: 10.14746/pss.2026.30.9

Data przesłania tekstu do redakcji: 14.12.2025
Data przyjęcia tekstu do publikacji: 6.05.2026

Analysing and Tracking Digital Transformation in the Humanities through Software Mentions Detection*

* This article was written as part of the research project "Making Software FAIR: A machine-assisted workflow for the research software lifecycle" (CHIST-ERA-22-ORD-08), co-funded by the Narodowe Centrum Nauki (National Science Center), project number 2022/04/Y/HS2/00182. For the purposes of Open Access, the author has applied a CC-BY licence to the Author Accepted Manuscript (AAM) resulting from this submission.

Abstract: Umerle Tomasz, Hubar-Kołodziejczyk Patryk, Jarzyńska Katarzyna, Rokicka Renata, Rosiński Cezary, *Analysing and Tracking Digital Transformation in the Humanities through Software Mentions Detection*, "Poznańskie Studia Slawistyczne" 30. Poznań 2026. Wydawnictwo "Poznańskie Studia Polonistyczne," Adam Mickiewicz University, Poznań, pp. 205–231. ISSN 2084-3011.

The article examines digital transformation in the humanities by analysing software mentions in scholarly publications using the Softcite tool. A longitudinal dataset comprising journals in traditional linguistics and literary studies (TLL) as well as digital humanities (DH) was compiled, and software mentions were manually validated on a per-document basis. The results reveal a clear upward trend in software use across both domains, with DH journals consistently exhibiting much higher levels of software saturation, while only selected linguistics journals show notable growth in TLL. The study reveals substantial false-positive rates, especially in DH, and confirms that abstracts rarely contain software references. It also demonstrates that automated software detection can support research infrastructures in the Social Sciences and Humanities (SSH), though effective implementation requires human validation and may be strengthened through integration with curated software catalogues such as the SSH Open Marketplace.

Keywords: software; digital humanities; research infrastructure; natural language processing (NLP); information extraction

1. Introduction

Contemporary, modern research infrastructures, such as discovery services for scientific output, are increasingly being powered by advanced tools that automate the processes of data creation. Due to digital transformation and its impact on scientific methods, scientific publications routinely contain references to software, including research software. Research communities interested in preserving knowledge about software use in science, and those research infrastructures dedicated to facilitating the discovery of scientific output, both recognise the need to detect, identify, and systematically document software use in publications. The SoFAIR project (financed through the CHIST-ERA initiative) is another initiative in this area, which aims to further the research and innovation around automated software detection in scientific publications by developing datasets and workflows, and validating the outputs of existing tools and solutions. This article examines the research and infrastructural potential of the Softcite[1] tool, which was designed to support automated software detection, and serves as a key component of the SoFAIR workflow[2]. To this end, the article aims to use Softcite both to trace the adoption of software in humanities scholarship, and to assess the potential uptake of the solution in domain infrastructures, especially scientific output aggregators (GoTriple) and software catalogues (SSH Open Marketplace). Through this method, we are seeking to answer the following research questions: 1) What does digital transformation look like in the humanities as seen through the lens of automated software detection? And 2) How can technological advancement in data extraction (i.e., extraction of software mentions) support existing research infrastructures in the humanities?

1 https://github.com/softcite
2 https://github.com/SoFairOA/documentation

2. Background and State-of-the-Art

Digital transformation (DT) is defined as "a process that aims to improve an entity by triggering significant changes to its properties through combinations of information, computing, communication, and connectivity technologies" (Vial, 2019). In the humanities, a clear sign of this process is the emergence and rapid development of digital humanities (DH), an interdisciplinary field of research that has become the subject of studies in its own right. In this article we aim to go beyond the well-researched area of the emergence of DH as an indication of DT. We will demonstrate and compare the DT processes within this subfield and across the more traditional humanities, with the aim of providing a more comprehensive account in the humanities. The hypothesis is that we should expect to find evidence of DT across the humanities owing to the pervasive influence of digital technologies on all aspects of research, even beyond DH itself. The question is to what extent have "traditional" and "DH" humanities been saturated with mentions, and for this purpose we have built a dataset of humanities publications spanning decades (primarily the timespan of 2001 to 2025) to capture relevant trends using automated detection of software mentions.

2.1. DH emergence and evolution

The DT in the humanities has been examined across numerous studies, which have attempted to define and contextualize the nature of DH as an emergent subfield within the broader field of the humanities. Bibliometric analyses, historical reconstructions, and disciplinary reflections illustrate how DH has evolved from early computational applications to a mature, interdisciplinary area of inquiry grounded in both technological innovation and critical reflection.

Sula and Hill (2019) have provided an empirical reconstruction of the early DH landscape through a systematic analysis of two foundational journals, *Computers and the Humanities* (CHum) (1966–2004) and *Literary and Linguistic Computing* (LLC) (1986–2004). Their study challenges the prevalent assumption that early humanities computing (HC) was narrowly focused on textual analysis. Although textual studies accounted for the majority of publications (59.3% in CHum, 72.4% in LLC),

a substantial amount of research also addressed technology (15%–16%), sound (8.9% in CHum), and multimedia (5%–6%). The authors concluded that early DH was characterized by greater thematic and disciplinary diversity than commonly assumed which they describe as a "big tent" encompassing multiple media and methodological traditions. Notably, the dominance of Anglophone institutions (primarily from the USA and UK) underscores the field's initial geographical concentration.

Yan, Li, and Liu (2024) analyzed publications from the *Digital Scholarship in the Humanities* (DSH) journal to trace the field's thematic and epistemological development. Their findings confirm that DH, which originated from corpus linguistics and early computational projects such as *Index Thomisticus* (circa 1950), has progressively expanded its scope. Initially centered on text encoding and linguistic analysis, the field evolved toward encompassing technology critique, digital infrastructure, and knowledge production studies. The journal's change in name from *Literary and Linguistic Computing* (LLC) to *Digital Scholarship in the Humanities* in 2015 symbolically marks this paradigmatic shift. The authors define DH as "the observation, description, and interpretation of human knowledge with the help of emerging digital technologies," emphasizing linguistics as a foundational methodological framework. While European and North American institutions, particularly in the UK and the USA, remain dominant, new research centers are emerging in Asia and South America, indicating the global diffusion of the field.

Finally, Salerno (2024) has revisited the ongoing debate over wheather DH constitutes a distinct academic discipline or an interdisciplinary practice. Tracing its intellectual lineage to the collaboration between Roberto Busa and IBM in 1949, Salerno frames DH as an outgrowth of *humanities computing* that gradually gained epistemological independence from computer science. The author argues that DH can be conceptualized as a new type of discipline, comparable to engineering, that integrates technological methodologies into humanities inquiry in ways intended to benefit both scholarly research and society at large. Although some scholars interpret DH as a Kuhnian revolution, Salerno questions whether the concept of "normal science" applies within the humanities context. The institutionalization of DH, manifested in dedicated academic programs (e.g., the first undergraduate curriculum in Pisa,

2002), specialized journals, and scholarly associations, nonetheless supports its recognition as a distinct and self-sustaining domain of research.

Spinaci, Colavizza, and Peroni (2022) presented a large-scale bibliometric study seeking to delineate the intellectual and institutional boundaries of DH. Their work addresses persistent challenges in defining the field bibliometrically and evaluating the coverage of DH publications in major citation databases. To operationalize their analysis, the authors compiled a comprehensive journal list categorized as "exclusively," "significantly," or "marginally" related to DH. The comparative assessment of major databases revealed that Crossref provides the best coverage of journals exclusively devoted to DH scholarship (3,989 indexed articles), whereas Dimensions offers the most extensive overall coverage across all categories (127,792 DH-related articles). Using Leveraging the Dimensions' dataset, they mapped the intellectual landscape of DH, uncovering strong connections with adjacent fields such as computational linguistics, natural language processing, and digital libraries. The core DH citation cluster was found to be centered around quantitative literary studies, including stylometry and authorship attribution. The study also demonstrates a strong overlap between publication venues and citation clusters, indicating that journals remain a key organizing principle in the intellectual structure of DH research.

The literature reviewed above points to two partly separate areas of research: studies of digital transformation in the humanities and studies focused on the automated detection of software mentions. This article connects these two areas by using software-mention detection not as an NLP task, but as a scientometric instrument for observing digital transformation in humanities publishing. The contribution of the article lies in testing whether automated software detection can help trace differences between traditional linguistics and literary studies journals and digital humanities journals, while also demonstrating the potential applicability of such data in research infrastructures.

2.2. Automated detection of software mentions

Automated detection of software mentions is a method commonly used to investigate how research software is applied across scientific domains. One recent initiative advancing this area is the SoFAIR project,

which aims to develop infrastructures for software mentions through the integration of GROBID (GeneRation Of BIbliographic Data) and the Softcite tool. As described by Du et al. (2021), the Softcite dataset was developed as the "goldstandard" corpus for software mentions. This dataset comprises 4,093 manually annotated software mentions extracted from the full text of 4,971 open-access academic articles in biomedicine and economics published between 2000 and 2010[3]. The utility of the dataset was demonstrated using a baseline conditional random field (CRF) model, which achieved strong sequence-labeling results, reporting an overall micro-average F-score of 0.81 (precision: 0.86; recall: 0.76). Softcite leverages GROBID (in development since 2008), which employs deep learning algorithms, statistical models, and rule-based methods to process PDF documents and extract structured metadata, including titles, authors, abstracts, affiliations, and citations (Lopez, 2009). Its outputs, available in both annotated PDF and TEI/XML formats, serve as input for downstream NLP pipelines such as software mention extraction.

Barbot et al. (2022, presented at the DH conference) developed a machine-learning-based named entity recognition (NER) model for identifying digital humanities tools. They compiled around 1.9 million sentences (from PLOS Sociology/Linguistics and DH conference abstracts) and seeded around 55,000 tool-name patterns from DH tool directories (TAPOR) and Wikidata. After manual annotation (via spaCy's Prodigy) and training, the final NER model achieved an F1score of approximately 0.91–0.92 in tool-name recognition. Applied to 470 DH publications (54,841 sentences), it automatically suggested 2,257 distinct tool names, revealing how DH scholarship frequently mentions a variety of specialized software.

To facilitate the large-scale macro-analysis of software usage and citation practices, Schindler et al. (2022) developed a supervised

3 Additionally, **SoMeSci** (Schindler *et al.*, 2021): A "5-star open data" knowledge graph with 3,756 software mentions in 1,367 PubMed Central articles, https://tinyurl.com/597ftbzx 30.04.2026. In SoMeSci, mentions are richly annotated with relations such as version, developer, URLs, and whether the mention reflects usage or creation of the software. It distinguishes software types (application, library, etc.) and provides entity linking for NER/relation extraction tasks. + SoFAIR dataset.

information extraction pipeline, which was trained on the comprehensive Software Mentions in Science (SoMeSci) gold standard corpus. This approach was applied to over 3.2 million scientific articles from PubMed Central (PMC), resulting in the creation of a software knowledge graph (SoftwareKG). This dataset comprises 11.8 million software mentions, structured into more than 300 million Resource Description Framework (RDF) triples. SoftwareKG was then utilized to provide extensive insights into the evolution of software usage and citation patterns over time across various scientific fields, journal ranks, and publication impact.

Finally, González-Guardia et al. (2023) introduced Softalias-KG, a novel knowledge graph that is explicitly focused on structuring software aliases. This knowledge graph was constructed by cleaning and transforming the clustered alias data generated during a recent large-scale analysis by the Chan Zuckerberg Initiative (CZI), which extracted software mentions from over 3.8 million papers in PubMed Central. Softalias-KG organizes over 50,000 unique aliases into more than 34,000 unique software application groups, and it is enriched with additional software entities and metadata imported from Wikidata. This demonstration highlights the utility of the knowledge graph (KG) as a reconciliation service, integrating it with the state-of-the-art Softcite NER model. In operation, Softcite detects candidate software mentions, which are then used to query Softalias-KG via SPARQL to retrieve the canonical name for the software application and the corresponding essential metadata, such as URLs derived from external sources like PyPI, Comprehensive R Archive Network (CRAN), or SciCrunch.

Both research on DT in the humanities and research on automated software detection remain emergent fields, to which this study offers an original contribution. We identify a gap in the studies of DT in the humanities beyond DH, as well as a specific gap in the research on automated software detection. The gap concerns the still limited scientometric application of automated software detection methods. Existing studies tend to focus either on tool extraction (i.e. building knowledge bases on software and tools) or on developing models for automating detection, thus focusing on the identification of mentions themselves rather than their distribution across scholarly infrastructures, particularly journals.

3. Methodology and Limitations

To capture long-term trends in the humanities as a whole, we built a dataset of articles and abstracts from journals that have been in publication for at least 20 years. The selection was made using DOAJ (Directory of Open Access Journals). The dataset was split into two main categories: "traditional" linguistics and literary studies (TLL), and DH journals. We focused on the full texts of publications, although we also processed a number of abstracts. In this context, "traditional" refers to journals whose mission and scope do not explicitly address digital culture in any way; this was manually verified by studying the journals' documentation that was available on their websites. This selection was intentional and aimed at examining the state of digital transformation in linguistics and literary studies. Therefore, journals explicitly labeled as digital humanities were excluded. Including DH journals could have distorted the assessment of the level of digital transformation in linguistics and literary studies journals. Due to the nature of automated software detection tools, we limited the data set to English-language publications. TLL includes two titles related to linguistics, two titles related to literary studies, and three titles that focus on both of these disciplines simultaneously. In the category of pure linguistics, we included the journals *Ibérica* and *Linguistic Discovery*. In the category of pure literary studies, we selected the journals *Colloquy: Text, Theory, Critique* and *Nineteenth-Century Gender Studies*. Journals that addressed both disciplines simultaneously and met our selection criteria were *British and American Studies*, the *International Journal of English Studies* (IJES), and *Facta Universitatis Series: Linguistics and Literature*. The analysis included 3,372 full-text articles in the field of linguistics and literary studies (Fig. 1) and 1,819 abstracts (Fig. 2). We compared these texts with journals from the field of DH, where we relied on a list prepared by Spinaci et al. (2022). The list contains the journals' metadata, which are as follows: a numerical identifier, electronic and print International Standard Serial Number (ISSN), the title, the URL, and a classification indicating the extent to which the journal is associated with digital humanities research. For the full-text analysis of DH journals, we chose *Digital Humanities Quarterly* (DHQ) and *Code4lib*, which allowed us to examine 1,536 full-text articles (Fig. 1). For

abstracts we used journals from all categories from the list (exclusively, significantly, or marginally DH); due to constraints related to content availability, we processed 16,621 abstracts (Fig. 2) from this list, relying on their availability in the Crossref service.

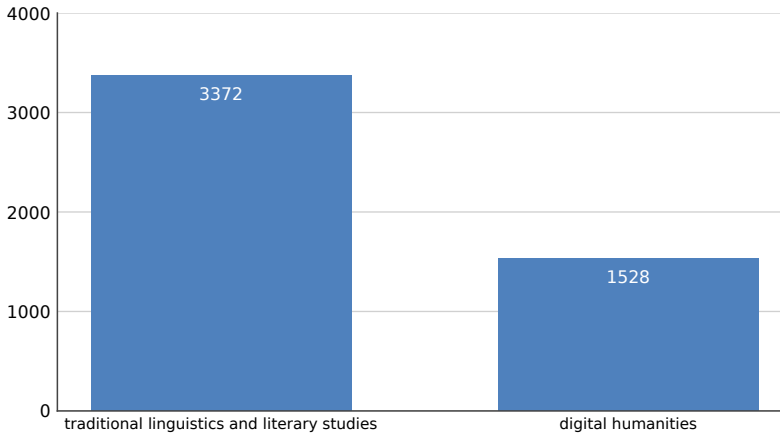

Fig. 1. Number of full text articles analyzed in the study

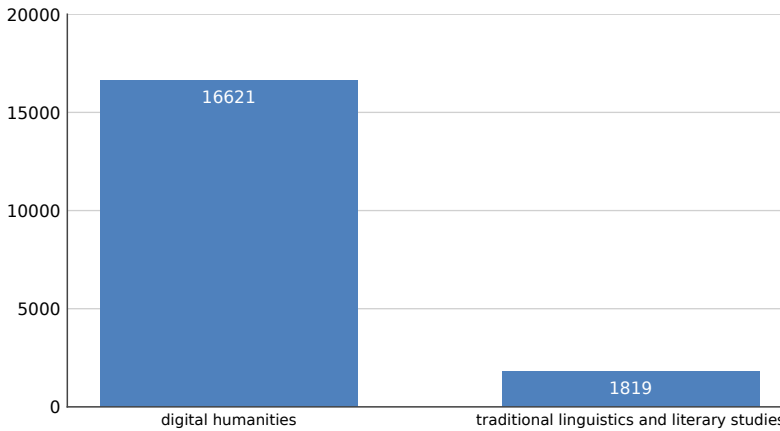

Fig. 2. Number of abstracts analyzed in the study

For the defined datasets, we used automated software detection based on the Softcite tool to extract data on software mentions[4]. However, we adopted a "mention per document" rule, meaning that we aimed only to identify the presence of software use in each article, regardless of how many times the software was mentioned in a single publication. In other words, we did not take into account either the location of the mention within the text or the Total number of occurrences. This approach allowed us to manually verify all outputs as either true positives (TPs) or false positives (FPs). True positives are instances correctly identified by the algorithm as belonging to a given class, whereas false positives occur when the algorithm incorrectly classifies an instance as belonging to that class. This approach has evident limitations: it does not account for potentially missed detections and instead relies solely on the output generated by the tool. This methodological decision partly stems from the SoFAIR project research that led to the creation of the SoFAIR dataset described by Pride et al. (2025a), in which state-of-the-art manual annotation for each software occurrence was performed for every software occurrencen to create another gold-standard dataset alongside Sofcite and SoMeSci. Such analysis is extremely time- consuming and exceeded the capacities of our research team. The research goal was more focused on understanding. Our approach differs from research focused on identifying and validating every individual software mention. For example, a single study may refer repeatedly to the same tool, which can lead to models achieving statistically strong performance in software-mention detection, while offering limited scientometric value for determining whether publications meaningfully rely on software.

The use of the Softcite tool in this study serves a scientometric, document-level purpose rather than an NLP benchmarking purpose. Recent SoFAIR research demonstrates that Softcite's performance varies substantially across corpora (Pride et al., 2025b), with lower results observed for heterogeneous datasets such as SoMeSci and SoFAIR than for its native SoftCite evaluation dataset. The SoFAIR dataset itself was designed precisely as a multidisciplinary gold-standard resource for

4 https://tinyurl.com/2s48xf43 30.04.2026.

evaluating and improving software-mention extraction tools, containing nearly 500 full-text papers and more than 9,000 annotated software mentions across 18 disciplines.

We therefore do not assume high recall for the present humanities corpus. Our manual validation estimates precision among Softcite-detected candidates – in the TLL subset, 528 of 1,141 detected candidates were confirmed as true positives – but it does not measure recall, because non-detected mentions were not exhaustively annotated. Consequently, the reported rates should be interpreted as conservative lower-bound estimates of articles containing explicit and formally detectable software mentions. Since our unit of analysis is the document rather than the individual occurrence, missed mentions affect the principal indicator only when all true software mentions in a given article remain undetected. We cannot exclude the possibility of period-specific variation in recall; therefore, the diachronic results should be interpreter as cautious relative indicators rather than precise measures of saturation.

In addition to full-text analysis, we also decided to include abstracts, for two reasons. It allowed us to process more publications and extend the scope of the research, and also offered opportunities to make up for the limitations of the "one mention per document" rule by including information concerning the location of the mention.

Finally, journals themselves played an important role in our study, as we assumed that understanding DT in the humanities requires attention to the "institutional" aspect and historical development of research communities organized around journals. The diachronic dimension of the analysis tracked the frequency of software mentions across the years of journal publication. This made it possible to observe temporal trends. Comparisons between TLL and DH journals provided insights into disciplinary differences, while within-journal trends highlighted the pace of methodological change within specific research communities.

4. DT in the Humanities: TLL vs DH

4.1. DT in TLL

Among the 2,844 articles from TLL, the first results showed 198 software mentions (Fig. 3). After this, the results were revised manually. These results were subsequently manually verified, which demonstrated that 71% of the detected mentions were true positives (TPs), while 29% were false positives (FPs) (Fig. 3). Only the true positives were included in the subsequent analysis.

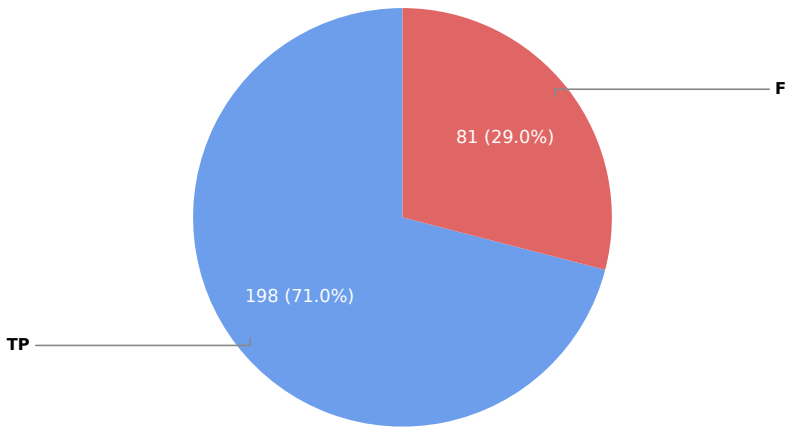

Fig. 3. True-positive and false-positive software mentions in TLL in Softcite (N = 279)

Among the seven TLL journals (Fig. 4), *Nineteenth-Century Gender Studies* published the largest number of articles during the examined period (586). However, it also had the lowest ratio of software mentions: only 1.02% (6) of its articles contained any software reference. *Linguistic Discovery* had the fewest articles for the examined period (150). Of these, 11.33% (17) contained software mentions. In the case of IJES, 15.71% of articles (58) contained software mentions among all those examined (369). Among

the 285 articles from *Iberica*, 27.01% (77) contained software. *Colloquy* published 463 articles, of which only 1.29% (6) contained software mentions.

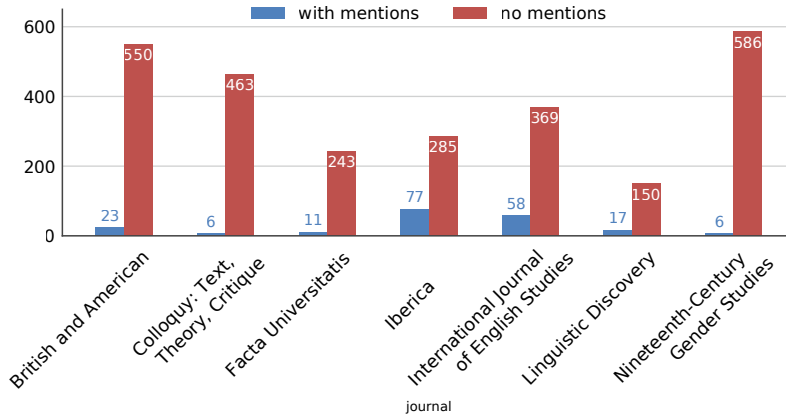

Fig. 4. Articles mentioning software in comparison to the total number of articles in the selected journals – Softcite

Of the 198 articles with software mentions, most came from *Iberica* (38.9%) and IJES (29.3%). The journal with the third largest number of software mentions in its articles was BAS (11.6%). The least number of software mentions found in articles was in *Colloquy* and *Nineteenth-Century Gender Studies* (3%).

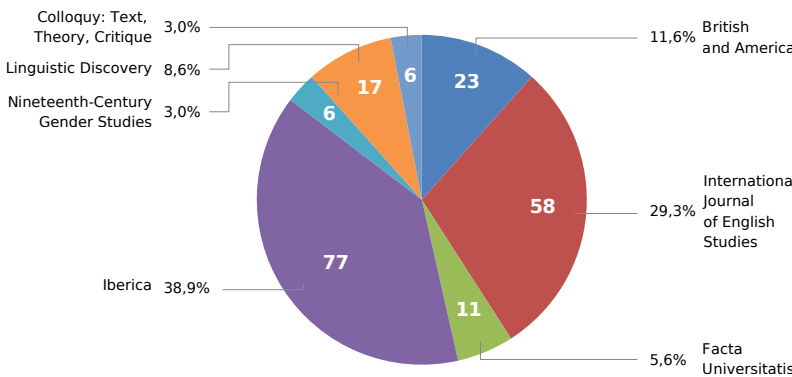

Fig. 5. Journal articles mentioning software – Softcite (N = 198)

Across the entire examined period (Fig. 6), beginning in the early 2000s, the data indicate a generally upward trajectory in software mentions, although the growth is neither linear nor uniform. During the first decade, the number of mentions increased slowly, with clear year-to-year fluctuations: some years exhibited temporary rises, while others returned to lower values. A series of modest peaks appeared around 2012–2013, reaching approximately 15 mentions, followed by another smaller increase in 2016–2017, when counts again approached roughly 14 mentions.

A more substantial increase became visible only from 2018 onward. From this point, software mentions rose more sharply, culminating in the highest values observed in the entire dataset in 2020–2021 (20 mentions). This intensified peak is possibly associated with the COVID-19 pandemic, during which reliance on computational tools, remote analysis, and digital research infrastructures grew markedly across the humanities.

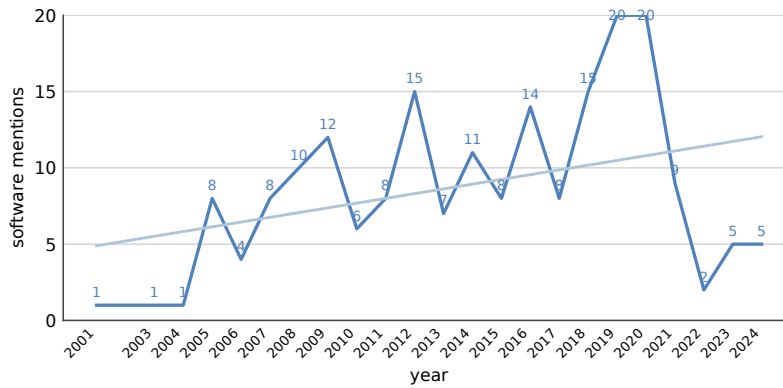

Fig. 6. Mentions of software in TLL journals from 2001 to 2025

The situation described above can be examined in terms of the contribution of each journal to the described increases and decreases for the number of software mentions across the analysed articles. The first peak, observed between 2008 and 2010, is driven primarily by software mentions in *IJES* and *Linguistic Discovery*, with some contributions from

Iberica. A second, more pronounced but shorter-lived peak occurred in 2012, indicating a temporary surge in software citation activity across these journals. In 2020, software mentions were concentrated mainly in *IJES*, *Nineteenth-Century Gender Studies*, and *Iberica*. During the period 2001–2007, *IJES*, *Linguistic Discovery*, and *Nineteenth-Century Gender Studies* recorded the fewest software mentions.

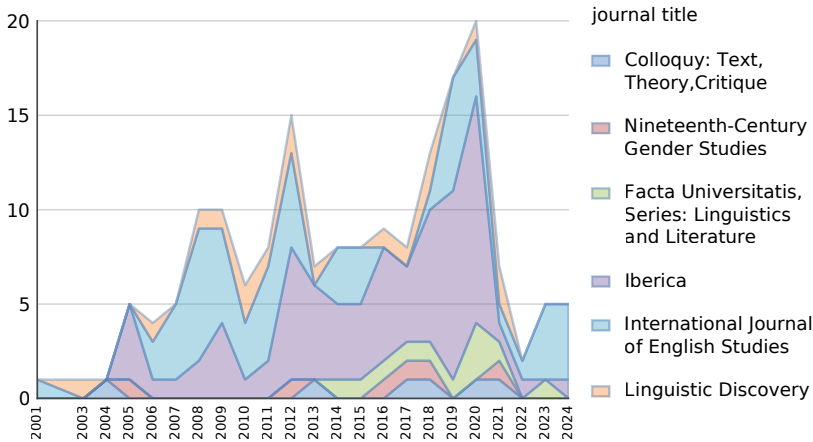

Fig. 7. Comparison of software mentions in selected journals from 2001 to 2025 – Softcite

4.2. DT in DH

Of the 1,227 DH full-text articles, 1,028 contained software mentions. The results show a nearly even distribution between true positives (50.8%) and false positives (49.2%) (Fig. 8), indicating a substantial proportion of incorrect detections within the dataset.

Both *Code4Lib* and *DHQ* display a general upward trajectory in software mentions, though with differing temporal dynamics. *Code4Lib* exhibits its first pronounced peak between 2013 and 2014, followed by a visible decline from 2019 to 2022 and another notable increase in 2023–2024. In contrast, *DHQ* demonstrates a steadier and more consistent increase without major declines; substantial growth become visible from 2016 onward, reaching its peak around 2020–2021.

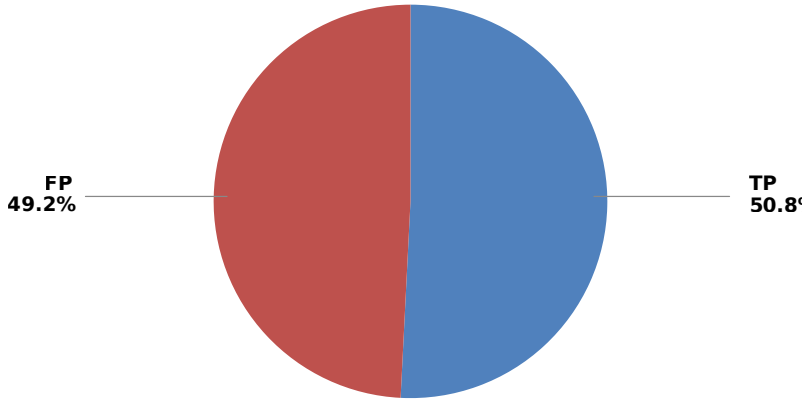

Fig. 8. True-positive and false-positive mentions in DH articles (N = 1028) – Softcite

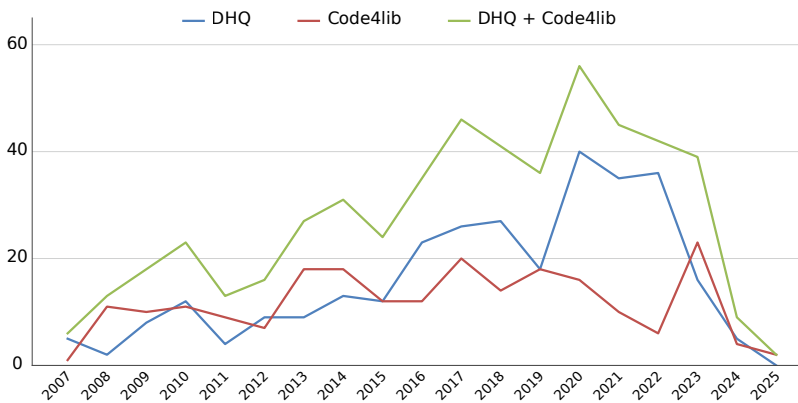

Fig. 9. TP software mentions in *Code4Lib* and DHQ for 2007–2025 – Softcite

4.3. DH vs TLL: A Comparison

Because the number of software mentions in several individual journals is small, yearly journal-level rates should be interpreted cautiously. For this reason, the article presents absolute journal-level counts while using aggregated mention rates for broader comparison between TLL and DH journals. To standardize the results, the mention rate was

calculated by dividing the number of software mentions in a given year by the total number of articles published that year. This normalized measure enables more objective comparisons across all articles of a given type with other disciplines and journals. The data show relatively regular fluctuations over time, with differences reaching 10%. Despite these variations, a clear upward trend remains visible, as both the lowest and highest values gradually increase throughout the observed period.

The mention rate across all DH full texts from 2007 to 2025 shows a generally upward trend, though with visible fluctuations. Between 2007 and 2016, the proportion oscillates between 20% and 40%, after which it consistently exceeds 40%, reaching nearly 60% in 2021.

In DH full texts, the percentage of software mentions is consistently high, often ranging from 30% to almost 60%. In contrast, TLL full texts display much lower levels throughout the entire period, generally remaining below 10%, with only occasional increases reaching around 15%. This contrast underscores the substantial differences in software referencing practices across the two corpora.

Particularly noteworthy are the results for abstracts, which were used to supplement the full-text analysis. The distribution of true positive (TP) and false positive (FP) mentions in abstracts for language and literature studies based on a dataset of 22 abstracts reveals that 86.4% of abstracts contain TP software mentions, while 13.6% do not. In the DH abstracts' dataset containing 271 texts, 56.1% are TPs and 43.9% are not. With

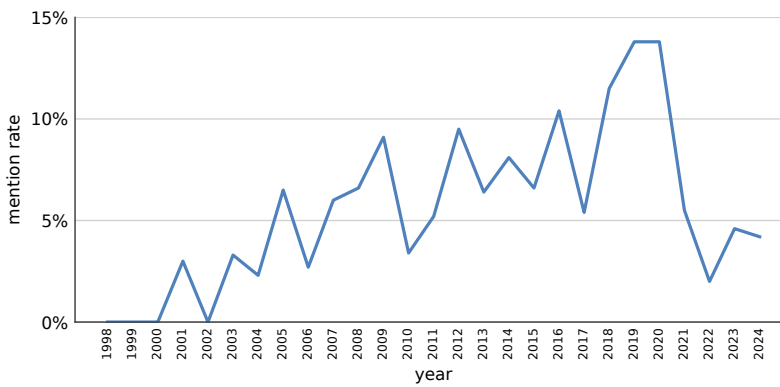

Fig. 10. Mention rate in TLL full texts – Softcite

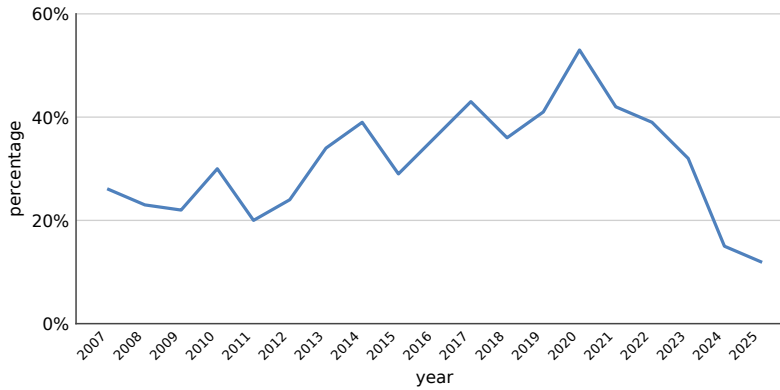

Fig. 11. Mention rate across all DH full texts – Softcite

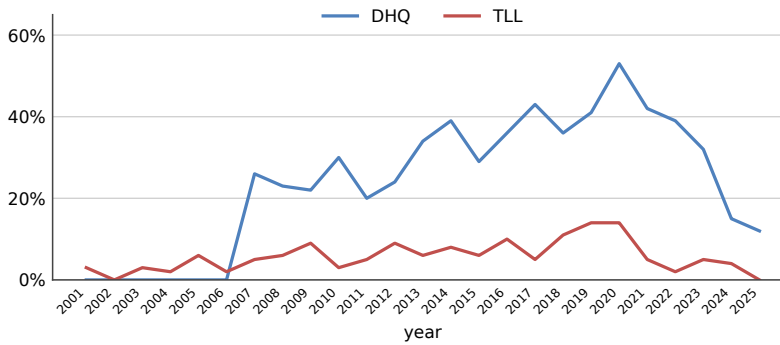

Fig. 12. Comparison of full-text software mentions for TLL and DH

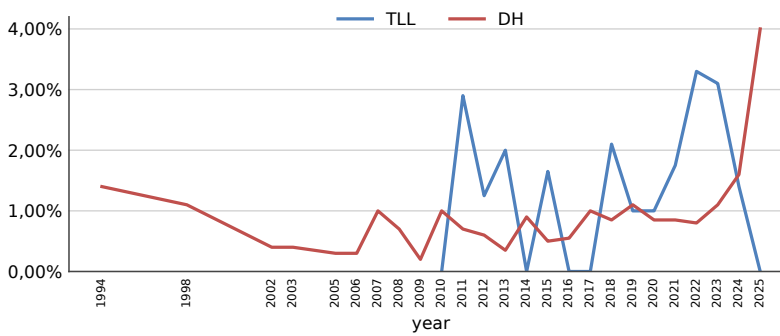

Fig. 13. Mention rates in TLL and DH abstracts

regard to mention rates in abstracts, TLL abstracts generally remained close to 0%, only periodically reaching values around 3%. In the case of DH abstracts, the rate oscillated around 1% and only rarely reached higher values. These results demonstrate that software is only infrequently referenced in abstracts.

5. The Uptake of Automated Software Detection in SSH Research Infrastructures

The SoFAIR project aims to embed automated software detection in scholarly infrastructures by deploying machine-learning models at the scholarly-content aggregator such as CORE. This approach enables repositories of research publications, including Hyper Articles en Ligne (HAL), to receive notifications about detected software mentions and validate them within repository infrastructures. Submitting authors may then review and confirm the detected software references. The CORE aggregator plays a central role in this workflow, as it \ integrates and deploys extended GROBID software – including enhanced Softcite ML models for extraction and disambiguation – directly within its ingestion pipeline. Once software mentions are extracted, CORE sends a validation request through its repository dashboard to the relevant institutional repositories. A repository, exemplified in the demonstrator case by HAL, may then accept the request and forward it to the manuscript's authors for validation. After authors confirm and validate the detected software mention, the repository issues an asset-registration request to Software Heritage (SWH). Software Heritage, a universal source code archive, permanently archives the newly identified and validated software asset and assigns it a persistent identifier (PID). This PID is subsequently returned to the repository. Finally, repositories adapt their systems to expose explicit links between the research manuscript and the software asset within their metadata feeds, ensuring interoperability.

Within the SSH context, this workflow could prove valuable for key services for content aggregation and tool cataloguing. We will analyse the potential application of these solutions through two infrastructures:

the European Open Science Cloud (EOSC) GoTriple service and the SSH Open Marketplace.

GoTriple is a multilingual discovery platform for SSH. It was developed within the *Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration* (TRIPLE) project and is now managed by OPERAS AISBL. GoTriple harvests metadata from over 1,400 sources – including academic databases, institutional repositories, data portals, etc. – and indexes millions of items, particularly publications, but also authors' profiles, project descriptions, and datasets. As an aggregator, GoTriple manages relationships between data providers (e.g., repositories, scientific databases) and end-users, including individual researchers. The platform enables researchers to create profiles r, organizing their publications, and curate personal research information. Current projects, such as LUMEN and GRAPHIA, aim to develop additional functionalities, including 1) linking publications with datasets and the software used in research – a feature requested by non-SSH communities and considered increasingly essential; 2) adapting and extending the OpenCitations databases through the development of an SSH Citation Index. In relation to SoFAIR's software-detection methods, GoTriple could incorporate Softcite directly into its indexing pipeline. More specifically, software mentions, together with PIDs, could be associated with corresponding GoTriple document entries. This would enable users to search GoTriple by software name or PID, thereby identifying publications that employ a given tool. Such integration would improve the *findability* of software and reinforce the FAIR principle that software should be treated as a research object equipped with unique identifiers. By exposing software mentions alongside standard metadata such as titles, authors, subjects, GoTriple can construct a richer semantic context for its data. This approach aligns with the development of an SSH knowledge graph as a future architectural foundation for GoTriple (see the GRAPHIA project). Such interoperability (FAIR principles I2/R1) would also allow other systems – including the SSH Open Marketplace and domain-specific catalogues – to consume and reuse the enriched metadata. As demonstrated in this study and in previous research, automated software detection requires human validation. In this respect, GoTriple's user-oriented functionalities could support

SoFAIR validation workflows. GoTriple users could confirm or correct software annotations associated with their papers.

The SSH Open Marketplace is a curated discovery portal for SSH. It pools and contextualizes resources that are important to SSH communities, and, especially, connects digital tools and services with other relevant, contextual resources: tool- and service-related training materials, datasets, and workflows. The platform was developed by the SSHOC (Social Sciences and Humanities Open Cloud) project under Horizon 2020 and is maintained by the major European SSH research infrastructures (Digital Research Infrastructure for the Arts and Humanities (DARIAH), Common Language Resources and Technology Infrastructure (CLARIN), Consortium of European Social Science Data Archives (CESSDA)) and their partners. The Marketplace fosters discovery; for example, a researcher can see which software tools are used in given datasets or workflows. Importantly, the Marketplace positions itself as an entry point for EOSC, ensuring that its curated resources conform to the community standards of FAIR metadata and interoperability. The SSH Open Marketplace targets a broad SSH audience: scholars in the humanities and social sciences, research data managers, librarians, and technical service providers. The platform's catalogue does not attempt to list every piece of software or dataset, but rather highlights selected, well-curated resources that have contextual relevance (for instance, tools cited in exemplary publications or used in training). This helps researchers locate methods, software, and data that are pertinent to specific research questions or workflows. The Marketplace also provides an API and data export for reuse, enabling the systematic analysis of its contents by other tools and infrastructures. Overall, the SSH Open Marketplace serves the SSH ecosystem by interlinking software and research content in a FAIR manner.

Within the context of SoFAIR, the SSH Open Marketplace could play a key role in enhancing the discoverability of software and metadata reuse in SSH research. Because it already catalogs tools and services, the Marketplace could ingest or link to software mentions detected in publications by the Softcite/SoFAIR pipeline. For example, when Softcite identifies a software tool in a scholarly article, that tool could be matched to an existing Marketplace entry (or trigger the creation of a new one),

with the publication being added as contextual evidence. By enriching its entries with detected software usage and persistent identifiers (such as DOIs or Software Heritage SoftWare Hash IDentifiers (SWHIDs)), the Marketplace would improve the *findability* (F1) and *interoperability* (I) of software across SSH literature. For instance, users could query the Marketplace using a software name or PID and retrieve associated publications, promoting reuse and credit for that software.

Moreover, the Marketplace's curation and contribution mechanisms could help validate SoftCite's output. The SoFAIR workflow includes aggregators, repositories, and software archives, but focuses less on community-organized software catalogues. Including such catalogues in automated software detection workflows could provide a new approach to the challenge of validating software mentions. In repositories, the authors depositing publications are notified about the detection of software, however, this does not necessarily mean that these authors are also responsible for developing the software, which can limit their motivation to validate mentions, especially where it demands work. Engaging software catalogues with curated tool entries could leverage the work of motivated curators of content who seek to enrich entries and monitor the impact of the software represented in the SSH Open Marketplace.

6. Limitations and Conclusions

The methodological approach adopted in this study was deliberately shaped by its scientometric rather than linguistic orientation. Unlike other studies in this area, which have aimed to identify all instances of software within the full text of a publication, our analysis operates at the document level. This means that the presence of a single detected software mention within an article is considered sufficient for our analysis, regardless of the number of times that particular tool appears elsewhere in the text. This design choice reflects the study's primary objective: to observe and quantify digital transformation processes in humanities scholarship rather than exhaustively map individual software occurrences. In this context, the unit of analysis is the publication itself,

which serves as an indicator of the broader integration of digital tools into research practices. While this approach inevitably limits the granularity of the data, potentially overlooking nuances such as the intensity of software use or the diversity of tools mentioned, it enables a more systematic, large-scale comparative perspective between traditional and digital humanities journals. By focusing on the presence rather than the frequency of software mentions, the study aims to capture a structural signal of digital adoption at the disciplinary level. This approach offers an aggregated view of how digital transformation manifests itself within the publication landscape of the humanities.

A further limitation concerns recall. The study validates the candidate mentions returned by Softcite but does not exhaustively annotate articles in which no candidate was detected. Consequently, the results should not be interpreted as a complete measurement of software use in humanities publications. They are better understood as conservative lower-bound estimates of explicit software visibility at the document level. This is sufficient for the comparative purpose of the article, but future work should include a stratified audit of articles for which Softcite returned no candidates across journals and periods in order to estimate false-positive rates more directly.

Taking into account these limitations, the following conclusions can be drawn from our study:

1. We can observe an upward trend in software mentions both in TLL and DH, with fluctuations along the way. While DH journals are naturally characterized by a higher density of software mentions, the linguistic journals have a sizable software mention rate (*Revista, Iberica* – 27% and IJES – 16%), despite not being explicitly positioned as DH journals or journals with a DT-related mission. DT in literary journals, as understood through automated software mention detection, appears to remain marginal.
2. This approach, which focuses on identifying articles with and without software mentions, sheds an alternative light on the performance of models. The manual validation of software mentions on a "per-document" basis reveals a sizable false positive detection rate. The variation in model performance in relationship to the article

saturation with mentions could be investigated further (Should the multiple, properly detected mentions in a single article, and of the same software, be valued the same?).

3. Softcite achieves higher performance for traditional journals than in the DH ones (approximately 70% in TLL, and around 50% in DH journals), which could be attributed to the more complex and technical terminology from DH journals.
4. Abstracts, in general, contain few software mentions. This is important for the research community, which considers abstracts a key vehicle for providing access to scientific knowledge: abstracts therefore offer limited information about software used in humanities research.
5. SSH research infrastructures could benefit from automated software detection; but however, such approaches require validation of detected software mentions, which may slow down adoption. It is therefore important to consider how research infrastructures can support automated software detection through existing software catalogue communities, such as the SSH Open Marketplace, and to further investigate incentives for engaging in the validation of the outputs generated by tools such as Softcite.

7. Research data

The data produced in the course of this study are publicly accessible via the Zenodo repository at: https://doi.org/10.5281/zenodo.17908205

References

Barbot, L., Fischer, F., Moranville, Y., Pozdniakov, I. (2019). *Which DH tools are actually used in research?* Weltliteratur, December 6. https://tinyurl.com/3fje8tp5 30.04.2026.

Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021b). *Softcite dataset: A dataset of software mentions in biomedical and economic research publications*. "Journal of the Association for Information Science and Technology", vol. 72, no. 7, pp. 870–884. https://doi.org/10.1002/asi.24454

González-Guardia, E., Lopez H., Garijo, D. (2023). *Softalias-KG: Reconciling Software Mentions in Scientific Literature*, In: *ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Greece: Athens*. https://tinyurl.com/2p2e47ds 30.04.2026.

Lopez, P. (2009). *GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications*. In: "Lecture Notes in Computer Science", pp. 473–474. https://doi.org/10.1007/978–3–642–04346–8_62

Pride, D., Cancellieri, M., Knoth, P., Rosiński, C., Umerle, T., Rudnicka, E., Monteil, A., Foppiano, L., Docekal, M., Scalbert, S., Romary, L., Oleksy, M. (2025a). SoFAIR *Project Dataset – An annotated dataset of software mentions in full text scientific articles* (Version 2). The Open University. https://doi.org/10.21954/ou.rd.30374830.v2

Pride, D., Guenci, M., Docekal, M., Peroni, S., Knoth, P. (2025b). *Identifying and classifying software mentions in full text scholarly documents*. In: *Proceedings of the 2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 261–264. IEEE. https://doi.org/10.1109/JCDL67857.2025.00041

Salerno, E. (2024). *Digital humanities: Mission accomplished? An analysis of scholarly literature*. "Cultures of Science", vol. 7, no. 1, pp. 34–48. https://doi.org/10.1177/20966083241234379

Schindler, D., Bensmann, F., Dietze, S., Krüger, F. (2021). *SoMeSci- A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles*. Preprint of CIKM 2021 Resource Paper, pp. 4574–4583. https://doi.org/10.1145/3459637.3482017

Spinaci, G., Colavizza, G., Peroni, S. (2022). *A map of Digital Humanities research across bibliographic data sources*. "Digital Scholarship in the Humanities", vol. 37, no. 4, pp. 1254–1268. https://doi.org/10.1093/llc/fqac016

Sula, C. A., Hill, H. V. (2019). *The early history of digital humanities: An analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004)*. "Digital Scholarship in the Humanities", vol. 34, no. 4, pp. 716–731. https://doi.org/10.1093/llc/fqz072

Vial, G. (2019). *Understanding digital transformation: A review and research agenda*. "The Journal of Strategic Information Systems", vol. 28, no. 2, pp. 118–144. https://doi.org/10.1016/j.jsis.2019.01.003

Yan, J., Li, Q., Liu, H. (2024). *Defining digital humanities and examining its relationship with linguistics through the lens of Digital Scholarship in the Humanities*. "Digital Scholarship in the Humanities", vol. 40, no. 1, pp. 354–380. https://doi.org/10.1093/llc/fqae075

TOMASZ UMERLE – PhD, assistant professor at the Institute of Literary Research of the Polish Academy of Sciences, and innovation and development specialist at the Poznań Supercomputing and Networking Center. OPERAS Innovation Lab manager. Principal investigator of the Polish contribution

to the "Making Software FAIR" project, funded in part by the Polish National Science Center.

- **CEZARY ROSIŃSKI** – PhD, assistant professor, documentation and data specialist at the Department of Current Bibliography, Institute of Literary Research of the Polish Academy of Sciences. Literary scholar and Python programmer. Co-author of the 'Polish Literary Bibliography'. His research interests include the relationship between space and literature, information architecture, data processing for humanities research, and Linked Open Data. Editor of the 'Forum of Poetics/Forum of Poetics'. Involved in international and national research projects.

- **PATRYK HUBAR-KOŁODZIEJCZYK** – PhD, digital humanities specialist at the Digital Humanities Center. Information architect and Python developer. Senior assistant at the Department of Informatology, Faculty of Journalism, Information and Bibliology, University of Warsaw. Collaborates with the Institute of Applied Linguistics at the University of Warsaw and WSB Merito University. Co-leader of the Artificial Intelligence Special Interest Group within the OPERAS-EU. His research interests include machine learning, natural language processing and Semantic Web technologies.

- **KATARZYNA JARZYŃSKA** – PhD student at the Doctoral School of Social Sciences at the University of Warsaw in the field of social communication and media studies. Librarian at the Joint Libraries of WFiS UW, IFiS PAN, and PTF. Research fellow in the "Making Software FAIR" project. Researcher in the NPRH grant project "Virtual Encyclopaedia of Print Culture in Pre-War Warsaw".

- **RENATA ROKICKA** – PhD student at the Doctoral School of Social Sciences at the University of Warsaw in the field of social communication and media studies. Librarian at the Joint Libraries of WFiS UW, IFiS PAN, and PTF. Research fellow in the "Making Software FAIR" project. Researcher in the NPRH grant "Virtual Encyclopaedia of Print Culture in Pre-War Warsaw".