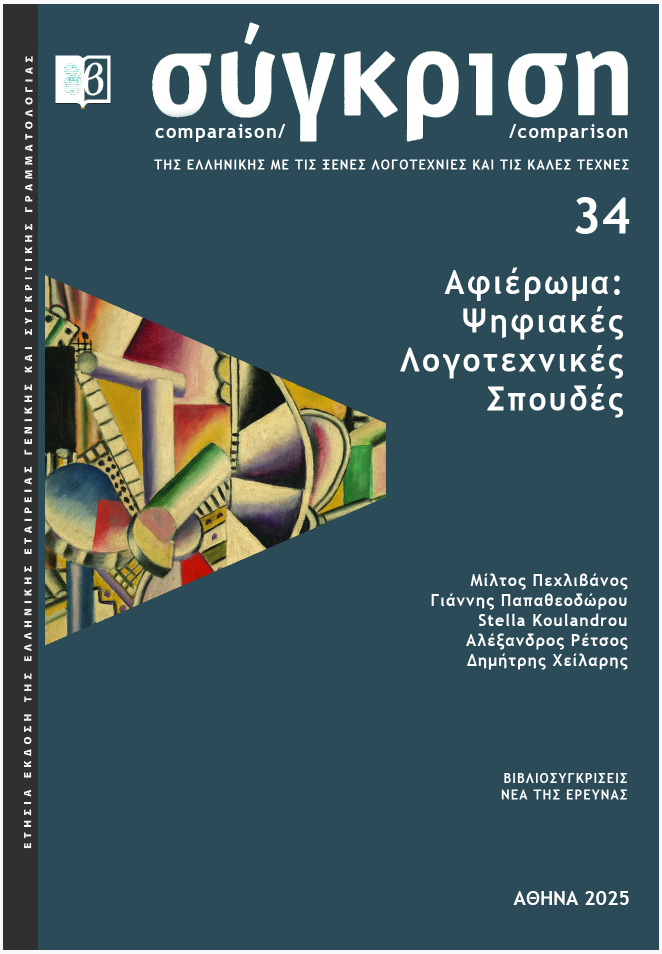
Σύγκριση/Comparaison/Comparison

Αρ. 34 (2025)

Ψηφιακές Λογοτεχνικές Σπουδές



Enhancing Digital Humanities Research Through the Analysis of Software Repositories

Emmanouil S. Rigas, Maria Papoutsoglou, Georgia M. Kapitsaki, Vasileios Vasileiadis, Aikaterini Tiktopoulou

Βιβλιογραφική αναφορά:

Rigas, E. S., Papoutsoglou, M., Kapitsaki, G. M., Vasileiadis, V., & Tiktopoulou, A. (2026). Enhancing Digital Humanities Research Through the Analysis of Software Repositories. *Σύγκριση/Comparaison/Comparison*, (34), 176–194. ανακτήθηκε από https://ejournals.epublishing.ekt.gr/index.php/sygkrisi/article/view/43712

EMMANOUIL S. RIGAS,[1] MARIA PAPOUTSOGLOU,[2] GEORGIA M. KAPITSAKI,[3] VASILEIOS VASILEIADIS,[4] AIKATERINI TIKTOPOULOU[4]

[1]School of Medicine, Aristotle University of Thessaloniki, [2]School of Informatics, Aristotle University of Thessaloniki, [3]Department of Computer Science, University of Cyprus, [4]School of Philology, Aristotle University of Thessaloniki

Enhancing Digital Humanities Research Through the Analysis of Software Repositories

1. Introduction

Digital humanities (DH) is an academic field in the intersection of computer science and the humanities,[1,2] encompassing a wide range of methods for the analysis, preservation, and dissemination of cultural and historical data (Berry, 2012). In recent years, this field has expanded rapidly. This expansion is driven by advances in digitization, data science, and web technologies among others. By integrating computational tools with traditional humanities scholarship, DH enables new modes of inquiry, such as large-scale text mining, geospatial analysis, and network visualization, which were previously impractical or impossible (Schreibman, 2015).

One of the central goals of DH is to make cultural heritage more accessible and analyzable through digital means. Projects involving the digitization of archives, the development of online platforms for scholarly collaboration, and the application of artificial intelligence to textual and visual data are now common (Kirschenbaum, 2016; Svensson, 2010). At the same time, DH raises important questions about the ethics of data curation, the politics of representation, and the epistemological implications of algorithmic analysis in humanistic inquiry (Klein and Gold, 2016).

In recent years, open-source platforms such as GitHub[3] have become central to the development and dissemination of tools, datasets, and collaborative projects in the Digital Humanities sector (Spiro and Smith, 2016). GitHub facilitates software versioning and code sharing and can also reflect evolving priorities, methodologies, and communities within the field. By examining repositories labeled or tagged as Digital Humanities, one can trace patterns in tool development, adoption of programming languages, institutional affiliations, and levels of community engagement.

In this paper, we aim to analyze GitHub repositories related to digital humanities. To drive our research, we devised the following Research Questions (RQs):

- RQ1: How have digital humanities related repositories evolved over the years? This question examines how the number of repositories related to digital humanities has changed over time and whether these repositories are actively maintained. Understanding temporal trends helps assess the growing interest and sustained engagement of the community in the development of software tools for the digital humanities.

[1] https://whatisdigitalhumanities.com
[2] https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2023
[3] https://github.com

- RQ2: What are the common licensing schemes used in digital humanities related repositories? In this RQ, we analyze the open-source licenses adopted by repository creators. Licensing choices reflect how developers wish to share, reuse, or protect their work, and can indicate levels of openness, community involvement, or academic versus commercial orientation in this multidisciplinary field.
- RQ3: Which programming languages are most commonly used in digital humanities related repositories? This question explores the main programming languages employed in digital humanities software projects. Identifying language preferences helps us understand the technical landscape of the field and the types of tools or frameworks that are most accessible or relevant for scholars and developers working in this area.
- RQ4: What topics are most frequently discussed by developers contributing to digital humanities repositories on GitHub? This RQ applies topic modeling on repository descriptions to uncover thematic patterns in the development of digital humanities software. By doing so, we aim to reveal the intersection of software engineering practices with scholarly goals and educational needs, offering insights into how computational thinking supports research and pedagogy in the humanities.

To the best of our knowledge, this is the first large-scale empirical study that systematically analyzes Digital Humanities repositories hosted on GitHub. While prior work has examined individual tools, platforms, or theoretical aspects of DH software development, little attention has been given to understanding the broader software ecosystem and its evolution over time. By combining statistical analysis with topic modeling, our work contributes a data-driven overview of how DH projects are developed, shared, and sustained in open-source environments. This perspective provides valuable insights into the technological foundations of the field, highlighting both current practices and emerging trends that shape digital scholarship.

The rest of the paper is structured as follows: Section 2 presents related works and outlines the novelty of this paper. Section 3 describes the data collection and pre-processing steps. Section 4 presents the results of the empirical analysis of the data. Section 5 discusses threats to validity and presents implications to practitioners. Finally, Section 6 concludes and presents ideas for future improvements.

2. Related work

Digital Humanities is an emerging field which has drawn the attention of the research community recently. As can be observed in Figure 1 searching the Scopus.[4] library using the keyword *digital humanities*, publications exist already from the year 2000. However, a steeper increase is observed from the year 2010 onward, while from the year 2022 approximately 1250 documents are published every year indicating an established trend towards the expansion of the wide digital humanities sector.

[4] https://www.elsevier.com/products/scopus.

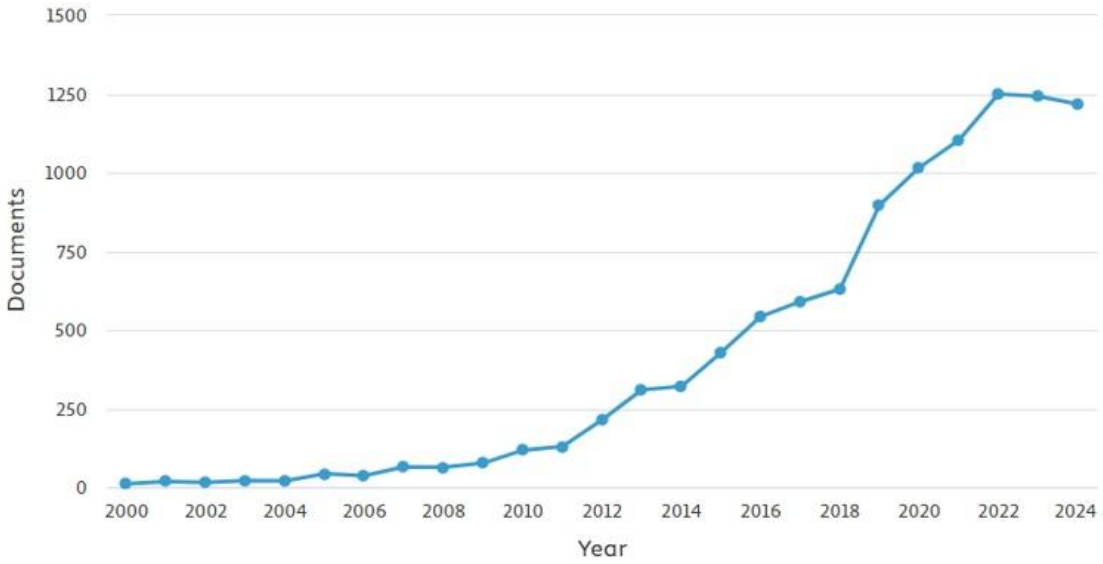


Figure 1 Documents published according to Scopus using the keyword digital humanities.

The works published in the DH field are diverse, but given that this is a new scientific field, many target to introduce the readers in DH and what this field brings out. For example, according to Berry (2012) the shift towards DH has not only introduced new tools and methods but also challenged traditional epistemologies and ontologies, as seen in the growing tension between "close" and "distant" reading practices. The emergence of the digital humanities reflects this broader transformation, as the field moves from its roots in "humanities computing" (Svensson, 2016) toward a more expansive and contested domain that blends computational approaches with critical and interpretive scholarship. Internal debates continue to shape the field's identity, focusing on questions of inclusion, technical proficiency, and its limited engagement with issues of race, gender, and social justice (Liu, 2013). Efforts to map this evolving landscape have led to the development of typologies and paradigmatic models that describe DH as encompassing diverse roles: technology as tool, object of study, expressive medium, laboratory, and activist space (Svensson, 2010). At the same time, scholars such as Jones (2013) argue that DH must be understood in the context of broader cultural shifts, where digital networks have become ubiquitous and integrated into everyday life, influencing both the form and substance of contemporary humanities scholarship.

Recent research highlights a critical gap in the preservation of scholarly code and its associated content hosted on Git hosting platforms (GHPs) such as GitHub, GitLab, Bitbucket, and SourceForge. While traditional archiving efforts preserve PDFs of scholarly articles, the source code, issue threads, and collaborative discussions referenced within them remain largely unpreserved, threatening the reproducibility and longevity of research outputs (Escamilla et al., 2022). In an educational context, GitHub is also being used as a dynamic learning tool. The DHClassHub project, implemented in a digital humanities course, demonstrates how GitHub can foster collaborative, student-driven learning. Students not only share and debug code together but also contribute to a growing, sustainable archive of course-related development, showing the platform's value beyond short-term academic use (Beshero-Bondar, and Parker, 2017). Finally, the analysis of GitHub repositories to gain insights in the activity of software developers has been used in

other domains such as in the work of Rigas et al. (2023) where the authors studied GitHub trends for the emerging field of Electric Vehicles.

Building on these perspectives, recent work highlights how digital tools, particularly conversational agents, bridge digital humanities and computer science. Tassios et al. (2025) evaluate Large Language Models (LLMs) in chatbots for migrants, addressing low-resource languages such as Greek and stressing the need for domain-specific adaptation for social inclusion. Chlasta et al. (2022) present *MyMigrationBot*, a multilingual chatbot on Facebook supporting migrants in Europe, combining cloud deployment, socio-psychological instruments, and platform integration. Lelis et al. (2020) introduce the NADINE-bot, designed to answer administrative questions for asylum seekers via a two-step text similarity approach. Collectively, these studies show how infrastructures such as LLMs, APIs, and knowledge bases are used to meet cultural and social needs, illustrating DH's overlap with software engineering. Our work follows this line by examining GitHub repositories as a lens into this interface, revealing how software engineering communities contribute to DH practices.

In parallel, research increasingly leverages GitHub and related platforms as large-scale data sources for studying human, social, and economic phenomena. Wachs (2023) used developer geolocation data from GitHub to trace post-2021 migration and potential brain drain of Russian software developers, underscoring the societal value of repository data. Feng et al. (2025) analyzed millions of Stack Overflow posts to build a taxonomy of software tasks, linking them to salaries and job requirements and revealing evolving work patterns. Mészáros et al. (2024) examined library adoption dynamics on Stack Overflow, highlighting innovation and sustainability in programming ecosystems. Juhász et al. (2024) used programming language data from open-source repositories to extend economic complexity research to the digital economy, showing its relevance for GDP, inequality, and national diversification.

These works collectively demonstrate that open-source ecosystems such as GitHub are not only technical infrastructures but also rich empirical laboratories for investigating human behavior, knowledge transfer, and economic transformation. By situating our study of digital humanities repositories within this broader context, we emphasize the dual role of GitHub: as a collaborative coding platform central to software engineering practices and as a data source enabling new insights into multidisciplinary domains. Our approach contributes to this growing line of inquiry by focusing on the intersection of digital humanities and software engineering, revealing how repository activity reflects both scholarly engagement and broader trends in technological adoption, education, and cultural production.

To the best of our knowledge, this is the first study to leverage GitHub software repositories for an empirical analysis aimed at uncovering trends in software development within the digital humanities field.

3. Methodological process

Our data collection and analysis pipeline was implemented using the R programming language.[5] To retrieve the data, we used the official GitHub API and the *httr*

[5] https://www.r-project.org

package,[6] querying repositories with the keyword "*digital humanities*"". Although general, we opted to use only this term for constructing our dataset, as it is currently the most representative expression of the field and reflects the emerging technological orientation in the humanities. The data collection process yielded a total of 1,581 repositories, which served as the basis for subsequent analysis.

After collecting the raw data, we extracted key metadata fields for each repository to guide our empirical analysis framework. Table 1 presents an overview of the variables used to address each research question and the type of analytical method employed. To provide an overview of the process, Figure 2 illustrates the methodological workflow adopted in this study.

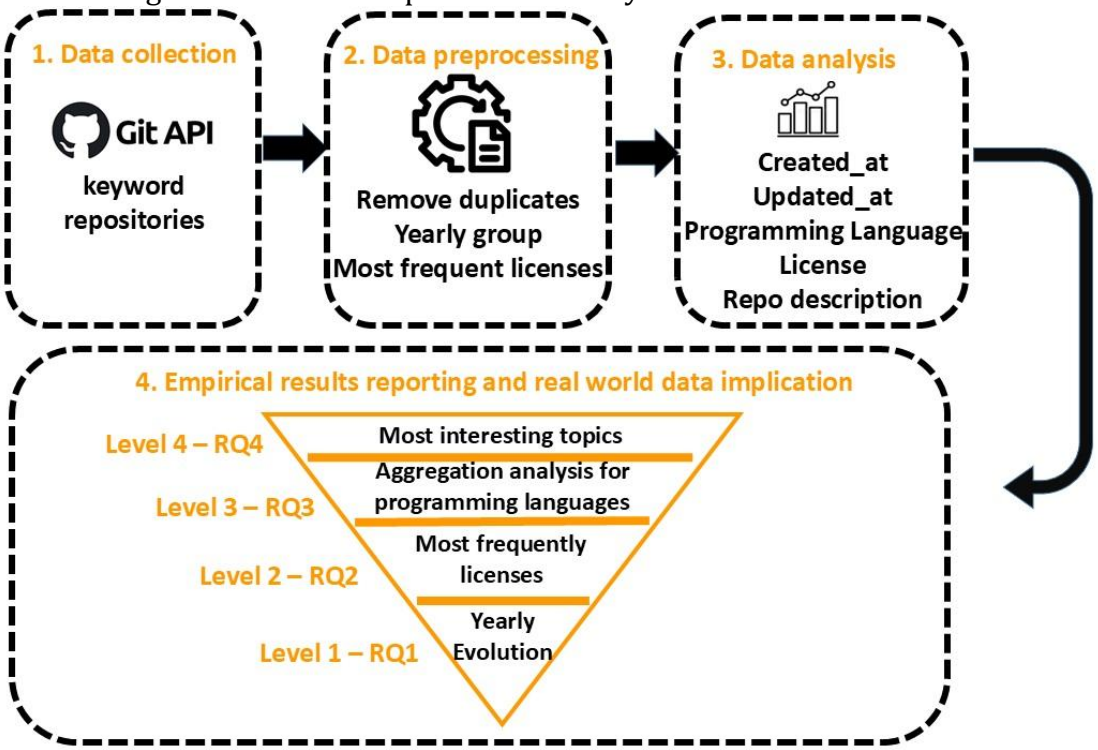


Figure 2 Data collection, pre-processing and analysis pipeline.

Table 1 Mapping of Research Questions to Variables and Methods

Research Question	Created_at	Updated_at	License	Language	Description	Method
RQ1: Evolution over time	X	X				Descriptive statistics
RQ2: Licensing schemes			X			Frequency analysis
RQ3: Programming languages				X		Frequency & aggregate metrics
RQ4: topics discussed					X	Topic modelling (LDA)

[6] https://cran.r-project.org/package=httr

As shown in Figure 2, the pipeline consists of four main stages. In the first stage, data were collected using the GitHub API by querying repositories with the keyword "digital humanities", ensuring the systematic retrieval of relevant projects. The second stage involved data preprocessing, including duplicate removal, yearly grouping, and the identification of the most frequent licenses. The third stage focused on data analysis, where key metadata fields, such as creation and update timestamps, programming language, license, and repository description, were analyzed to address the research questions. Finally, the fourth stage corresponds to empirical results reporting and interpretation of real-world data implications. The inverted pyramid highlights the hierarchical organization of the research questions, beginning with broad yearly trends (RQ1) and license distributions (RQ2), followed by programming language aggregation (RQ3), and culminating in topic modeling of repository descriptions (RQ4). This structured approach enabled both fine-grained insights and thematic interpretations, linking software development practices to the evolving role of digital humanities.

To explore RQ4, we applied topic modeling on the repository *description* field to uncover latent themes. The analysis was conducted using the *quanteda*[7] and *topicmodels*[8] packages in R. Preprocessing steps included converting text to lowercase, removing punctuation, numbers, and English stopwords, and applying stemming. Tokenization was performed using bigrams (n = 2) to improve semantic clarity. A document-feature matrix was created and transformed into a document-term matrix compatible with topic modeling. We employed Latent Dirichlet Allocation (LDA) using Gibbs sampling with k=5 topics and a fixed seed (1234) for reproducibility. From the resulting model, we extracted the term-topic matrix (φ), the document-topic matrix (θ), and the most representative keywords for each topic. We also computed the average topic proportions across documents and used the *LDAvis*[9] package for interactive visualization. This pipeline enabled the identification and interpretation of thematic patterns in digital humanities software development, offering insight into the intersection of computational and scholarly practices.

4. Empirical Results and Discussion

In order to answer all RQs, we relied on the data returned that contain various information on each repository (e.g., main programming language, number of forks). The results of each RQ are presented and discussed in the remaining of this section.

RQ1: How have digital humanities related repositories evolved over the years?
A first interesting insight (see Figure 3) is related to the growth trend in the number of GitHub repositories created under the "*digital humanities*" keyword. The plot in Figure 3 shows how interest in this field has evolved over the years.

Starting from 2009, we observe a very low number of repository creations, with only 1 repository that year. This low level of activity continues in the next years, as 2 repositories were created in 2010, 3 in 2011, and 10 in 2012. These numbers show a limited engagement with digital humanities in open-source

[7] https://cran.r-project.org/package=quanteda
[8] https://cran.r-project.org/package=topicmodels
[9] https://cran.r-project.org/package=LDAvis

software development at that time. However, from 2012 onward, we can observe a steady increase. The number of repositories rises to 17 in 2013, 34 in 2014, and 63 in 2015. This trend shows a growing awareness and integration of digital humanities into digital research and software projects. The trend increases with a faster rate especially in years 2016 (70 repositories) and 2017, where a sharp rise to 130 repositories exists.

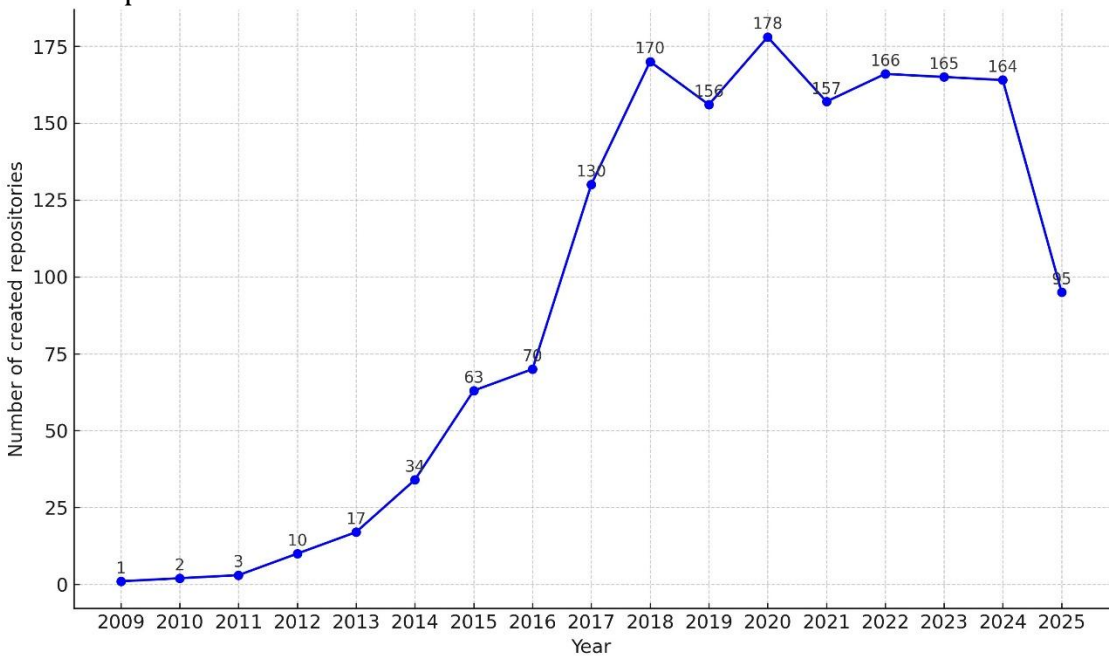


Figure 3 Yearly creation of repositories related to the keyword digital humanities on GitHub (2009–2025).

From 2018 to 2024, the number of newly created repositories remains consistently high with some ups and downs being observed, indicating a permanent community interest. This shows that the field of digital humanities has matured in relation to software development, and the community is actively contributing to the GitHub ecosystem.

In 2025, although the dataset contains data of the first half of the year, 95 repositories have already been created. If this rate continues through the year, approximately 228 repositories will be created, significantly surpassing all previous years. This projection supports the conclusion that interest in digital humanities remains strong and is likely to continue growing.

RQ2: Which are the usual licensing schemes in digital humanities related repositories?

The dataset provides a detailed breakdown of GitHub repository license usage across different years, categorized by the type of account (i.e., individual users and organizations). The license usage data offers meaningful insights into the evolution of open-source practices, preferences, and possibly underlying motivations for license choices.

For individual users, the data visualization (see Figure 4) reveals that the MIT License has remained the most dominant choice throughout the years. This is probably due to its simplicity and permissive nature, which enables developers to reuse, modify, and distribute code with minimal restrictions. The upward trend in its adoption reflects a broader cultural shift among developers towards openness

and collaboration, especially among smaller teams, or individual contributors. In recent years (approximately after 2021), we observe a more noticeable diversification of license usage. Licenses such as the Creative Commons Zero v1.0 Universal (CC0) and the GNU General Public License v3.0 (GPLv3) have gained traction. This trend may indicate an increasing awareness among users regarding intellectual property rights, as well as the implications of license terms in domains such as machine learning and data sharing.

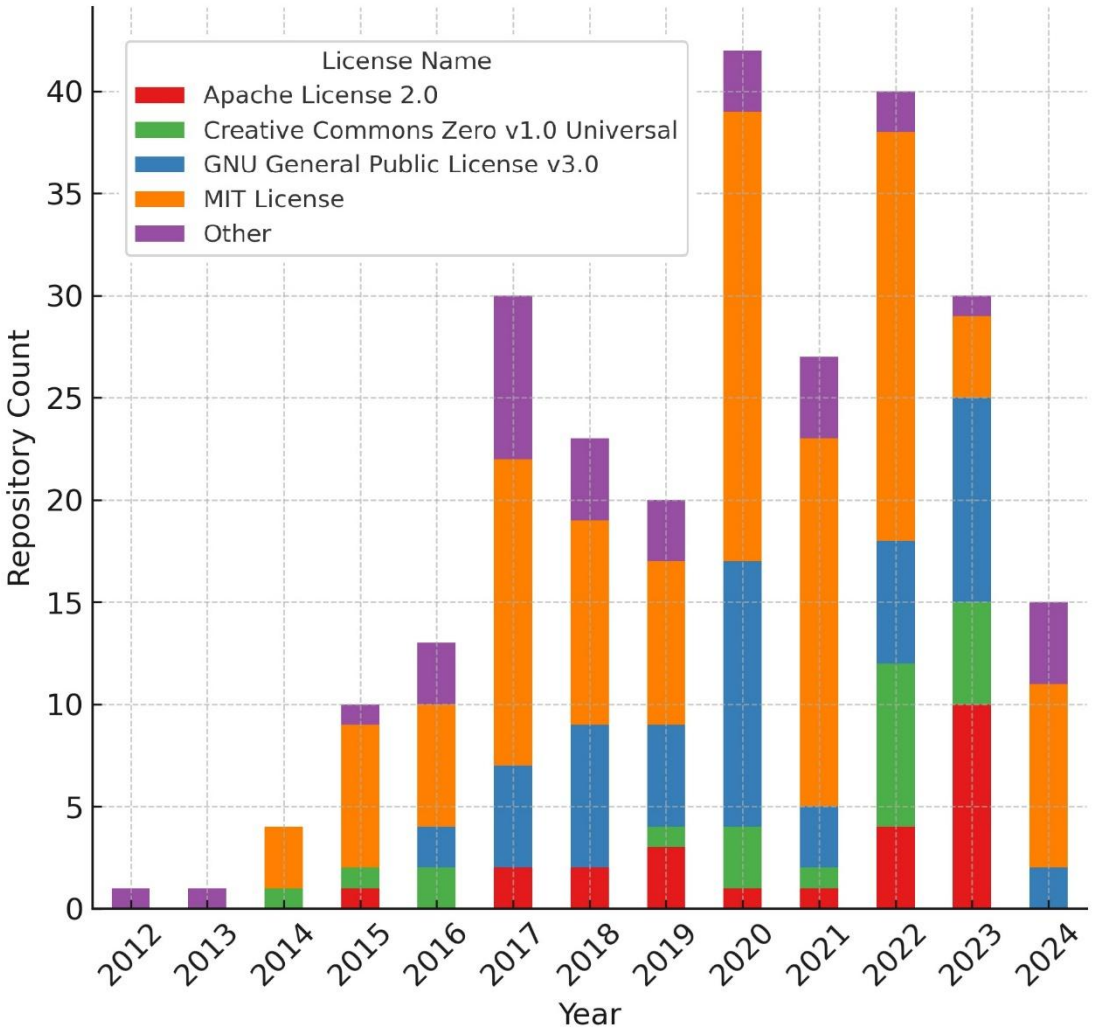


Figure 4 Yearly Preference of Top 5 Licenses for Users.

The inclusion of CC0, in particular, is notable in recent years. This license is often associated with datasets and content intended for public domain usage, which is an increasingly common practice in AI-related development, where data openness is critical for training large language models (LLMs) and similar systems. This change possibly reflects a broader alignment with the open data movement and the needs of developers working in generative AI or commercial data-intensive applications.

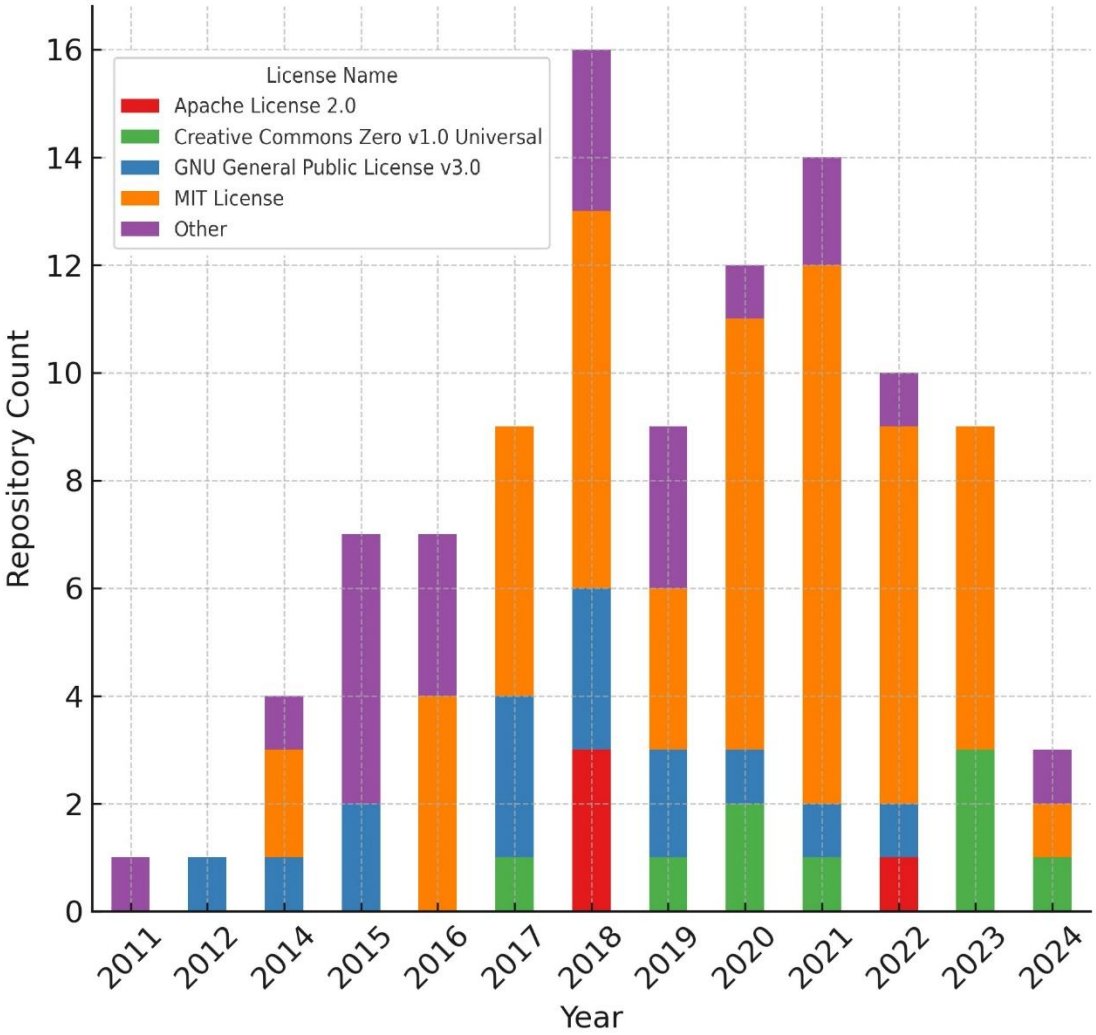


Figure 5 Yearly Preference of Top 5 Licenses for Organizations

In contrast, Figure 5 presents a more structured and conservative licensing landscape for organizations. Historically, the GNU General Public License v3.0 and the Apache License 2.0 have been the main choices. These licenses fit enterprise-level software development, offering strong protection for intellectual property and more elaborate conditions around distribution and patent rights. However, recent years show a relative decline in the adoption of the Apache License 2.0. This shift may be due to the fact that many organizations reevaluate the complexity of the Apache license, or due to a broader transition towards more permissive or community-friendly licenses, such as the MIT License, which appears to have gained modest ground in the organizational space. Another emerging trend is the incremental presence of CC0 in organizational repositories, which may suggest that some institutions are releasing datasets or models in the public domain, possibly in response to the growing demand for transparency and open access in AI research and digital humanities.

These evolving patterns underscore the dynamic nature of license selection and the broader interplay between technological evolution, legal frameworks, and community norms. As organizations navigate compliance requirements and developers seek frictionless collaboration, the license landscape continues to adapt,

potentially influenced by the demands of AI/LLM development, open science, and changing commercial expectations.

RQ3: Which are the most popular programming languages?
The programming languages identified in the GitHub repositories are summarized in Table 2. From these data, we observe that a variety of programming environments are used, showing the interdisciplinary and diverse tools the digital humanities community is using.

Table 2 Top programming languages of repositories.

Programming Language	#repos	M watchers	M forks	M days creation -> update
HTML	303	2.73	1.35	495.67
Jupyter Notebook	276	1.69	0.86	375.53
Python	170	1.22	0.42	474.29
JavaScript	103	3.37	0.77	506.51
CSS	64	1.36	0.84	604.13
R	46	0.67	0.50	473.97
Ruby	20	0.65	0.30	731.05
TeX	18	0.67	0.00	112.77
PHP	16	0.88	0.56	1453.07
SCSS	15	5.00	2.53	685.29

The most commonly used programming language is HTML. Strictly speaking, HTML is not a programming language, but a markup language used to structure and present content on the web, similar in this respect to XML. Its prominence nevertheless reflects a characteristic feature of digital humanities projects: the strong emphasis on web-based presentation, accessibility, and dissemination of scholarly outputs. The frequent appearance of HTML suggests that many repositories include front-end or presentation-layer components for sharing processed texts, visualizations, digital editions, or online exhibitions. Therefore, rather than interpreting HTML as evidence of programming activity in the narrow sense, we interpret it as an indicator of the importance of web delivery and human-readable digital presentation in digital humanities work.

Another frequently appearing entry is Jupyter Notebook, which may not represent a standalone programming language, but rather a development environment. However, Jupyter Notebooks are mostly used in combination with Python. GitHub identifies the dominant file type in a repository to infer the main programming language, which explains why repositories that use Python within Jupyter environments may be classified as “Jupyter Notebook”. Languages such as Python and R are particularly noteworthy due to their strong association with data analysis, statistical modeling, and natural language processing, tasks that are central to many digital humanities tasks. For example, Python's extensive library ecosystem (e.g. NLTK, spaCy, Pandas) makes it a powerful and efficient option for text mining and machine learning tasks applied to historical documents, literature corpora or digitized archives. Similarly, R is well-regarded for statistical computing and is

commonly used in projects that involve textual or quantitative analysis of cultural datasets.

The relatively high mean durations from creation to last update (often exceeding 400 days) across most languages suggest that many repositories are not simply experimental or abandoned. Instead, they reflect sustained development and usage, which may be indicative of collaborative academic projects or publicly funded initiatives. This trend supports the idea that digital humanities software development is long-term and oriented toward producing reusable, citable research outputs. Finally, metrics such as the average number of watchers and forks also provide valuable information. Although the overall values are modest, higher averages for languages such as Python, JavaScript, and SCSS may indicate broader community interest or reusability of specific tools. Forks, in particular, are indicative of reuse and adaptation, suggesting that some repositories serve as templates or foundations for derivative works.

In summary, the table highlights the diverse technological practices in the digital humanities domain. It reveals the field's blend of programming for analysis (e.g., Python, R), presentation (e.g., HTML, CSS), and platform customization (e.g., Jupyter, TeX). Understanding these patterns offers valuable information about the nature of scholarly coding and the practical tools scholars choose to build, analyze, and share digital cultural resources.

RQ4: What topics are most frequently discussed by developers contributing to digital humanities repositories on GitHub?

To explore patterns of technological focus and scholarly interest in digital humanities software development, we applied Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in repository descriptions. We extracted five distinct topics based on bigrams to improve interpretability. Table 3 presents the most frequent terms associated with each topic.

Table 3 Top 7 bigrams for each identified topic based on LDA modeling (k=5).

Topic	**Top 7 Bigrams**
Topic 1	digit_human, intro_digit, human_class, human_confer, human_summer, summer_school, relat_digit
Topic 2	human_digit, digit_knowledg, cours_digit, univers_bologna, repositori_contain, introduct_digit, github_repositori
Topic 3	digit_human, human_project, human_research, text_analysi, research_project, program_digit, human_tool
Topic 4	digit_human, final_project, project_digit, human_lab, tool_digit, topic_digit, ma_digit
Topic 5	digit_human, human_cours, human_univers, repositori_digit, human_research, method_digit, human_workshop

Topic 1 appears to be centered around *summer schools and training programs*, with bigrams such as *digit_human*, *intro_digit*, and *summer_school* suggesting an educational event focused on digital humanities technologies. Its high prevalence across documents, as seen in Figure 6, shows that such programs are a key driver for GitHub repository creation in the field. This is further supported by the topic proportion scores, with Topic 1 accounting for approximately 19.97 % of the total topic distribution across the corpus. Topics 2 to 5 form a cluster of thematically

related topics that seem to be related to *university courses and semester-long assignments*. For instance, Topic 2 includes bigrams such as *univers_bologna*, *cours_digit*, and *github_repositori*, indicating institutional course settings. Topic 3 emphasizes analytical tasks such as *text_analysi* and *research_project*, suggesting student work involving computational text methods, or potentially work within a research project. Topic 4 appears to focus on *final course projects*, with *final_project* and *ma_digit*, which likely refers to Master's level digital humanities programs. Finally, Topic 5 reflects *hands-on workshops* embedded in academic training.

These topics collectively illustrate a strong *educational orientation* in digital humanities software repositories, spanning both structured university curricula and extracurricular training. Importantly, this also reflects a broader trend toward *reskilling and upskilling* in the humanities through computational approaches. GitHub, in this context, serves not only as a development platform, but also as an educational tool for acquiring digital competencies in text analysis, tool creation, and research collaboration.

This educational-technological alignment also highlights the disciplinary intersection between software engineering and digital humanities. The fact that repositories are a product of course-related work implies that the training ecosystem in digital humanities is increasingly leveraging industry-standard software practices and tools.

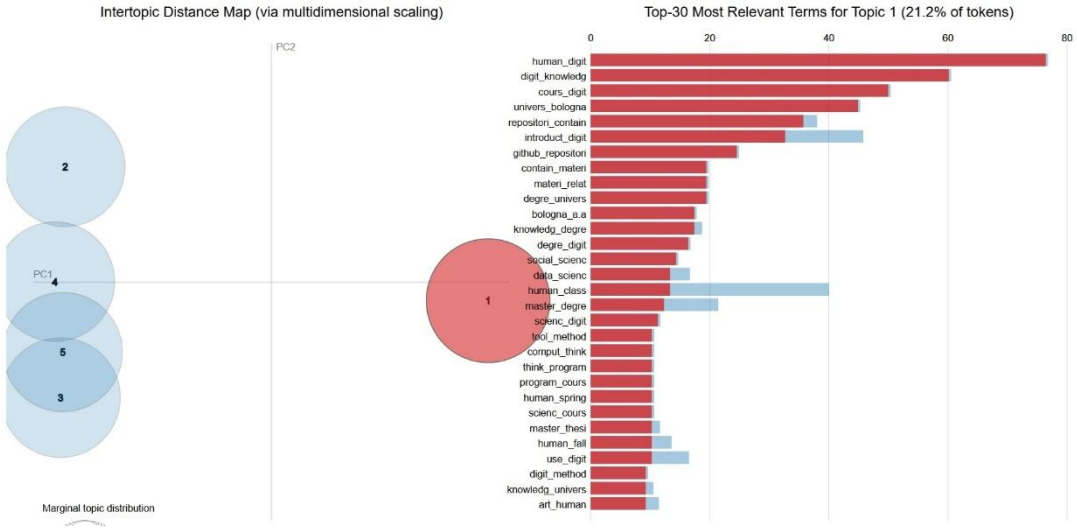


Figure 6 LDAvis intertopic distance map showing the semantic relationships and sizes of the 5 discovered topics.

5. Implications for Practitioners

The findings of this study have several implications for different practitioner groups. For researchers, the analysis of GitHub repositories provides a valuable perspective on the technological foundations of digital humanities, enabling them to track emerging trends, methodological innovations, and collaborative practices. By systematically examining programming languages, licensing choices, and thematic orientations, researchers can situate their work in relation to broader developments and identify potential areas for interdisciplinary collaboration. This approach not only contributes to the mapping of the digital humanities landscape but also provides an empirical basis for assessing the evolution of tools and practices in a field where scholarly and technical contributions increasingly intersect.

For students, the repositories studied highlight the growing use of GitHub as a learning and training environment in the humanities. Many projects reflect course assignments, workshops, or summer schools, illustrating how open-source platforms facilitate hands-on experience with computational methods. This has direct implications for lifelong learning, as students are exposed early to industry-standard practices such as version control, collaborative coding, and open licensing. Such exposure strengthens their ability to navigate both academic and professional contexts, equipping them with skills that are transferable across disciplines.

For the software engineering community, the study demonstrates how practices from computer science are being adapted and reused in a multidisciplinary setting. The digital humanities provide a testing ground for methods such as topic modeling, statistical analysis, and collaborative tool development, but framed in ways that engage with cultural data and humanistic inquiry. This interaction not only enriches the DH community but also feeds back into software engineering, highlighting how technical infrastructures can support diverse forms of knowledge production and scholarly communication.

Finally, the results also carry implications for the labor market. As more cultural institutions, research centers, and companies recognize the importance of data-driven methods in the humanities, demand is growing for professionals who can combine technical expertise with domain-specific knowledge. The patterns observed in programming languages and repository structures illustrate the types of skills that are valued in this emerging landscape. Employers can benefit from these insights when designing job profiles or training programs, while individuals entering the labor market can use them to better align their competencies with evolving expectations. The blending of technical and humanistic expertise underscores the importance of multidisciplinary training and signals new opportunities for career development at the intersection of culture, technology, and data science.

6. Threats to Validity

As with any empirical study, our work is subject to several threats to validity. Regarding construct validity, one possible limitation is the reliance on a single keyword, "*digital humanities*", for the data collection process. While this choice may exclude projects that are relevant but not explicitly labeled with this term, we argue that it is the most widely recognized expression of the field and therefore provides a representative sample. Another issue is related to the text used for topic modeling. Repository descriptions may vary in length and quality, but they generally contain the most enriched textual information available in GitHub repositories. For this reason, they were selected as the best available proxy to capture thematic patterns. Additional metadata such as programming languages, licenses, and activity timestamps were directly obtained from GitHub's API, which ensures consistency but depends on how repository owners define their projects.

Turning to internal validity, our results may be influenced by methodological decisions such as the preprocessing steps applied to the raw data and the parameters of the topic modeling. Choices like duplicate removal, grouping by year, and setting the number of topics (k=5) could affect the distribution of results. To mitigate these risks, we followed established practices in computational text analysis, reported all parameter settings, and ensured reproducibility by fixing random seeds.

External validity concerns the extent to which our findings can be generalized beyond the present dataset. Our analysis is limited to GitHub repositories tagged with "*digital humanities*", and results may differ if alternative keywords or other platforms such as GitLab or Bitbucket were included. Moreover, the strong educational orientation revealed in the topic modeling suggests that many repositories stem from courses, workshops, or training programs. While this reflects an authentic trend in the field, it may not capture the full scope of long-term research infrastructures or institutional projects outside GitHub.

Finally, in terms of conclusion validity, our interpretations depend on the robustness of the statistical and probabilistic techniques applied. For descriptive statistics on licenses and programming languages, frequency analysis is straightforward but may underrepresent categories with few repositories. For topic modeling, results may vary depending on parameterization or preprocessing decisions. To strengthen validity, we used widely adopted R packages (*quanteda, topicmodels, LDAvis*), reported reproducible parameters, and interpreted outcomes conservatively within the scope of the dataset.

7. Conclusion and Future Work

This study explored GitHub repositories related to digital humanities through an empirical lens, examining their temporal evolution, licensing preferences, programming languages, and thematic content via topic modeling. The results reveal that digital humanities repositories are not only growing steadily in number but are also shaped significantly by educational contexts such as university courses, workshops, research projects and summer schools.

The findings from topic modeling underscore the central role of education in the development of these repositories. Repositories are often created as part of structured academic activities, including final projects, semester-long assignments, and intensive summer schools. This educational focus reveals how digital competencies, especially in programming, data analysis, and collaborative software development, are being increasingly integrated into humanities training. Such evidence provides critical insight into the shifting landscape of humanities scholarship, where digital fluency is emerging as a key skill.

Importantly, this has direct implications for labor market analytics and reskilling research. The presence of open-source software artifacts developed during humanities education points to the increasing intersection between traditionally non-technical fields and software engineering practices. Online platforms like GitHub offer a repository of code, but also a lens into evolving professional competencies. This approach can be scaled across domains to identify where reskilling and upskilling trends are materializing, thus offering a novel method for tracking emergent interdisciplinary skill sets in the digital labor market.

In terms of future work, we plan to expand our dataset by incorporating a broader set of keywords beyond "digital humanities" to capture adjacent topics such as "cultural analytics", "computational history", "digital philology", "Scholarly Digital Editions" and "Text/Language Corpora". We aim to integrate additional web data sources such as academic-focused Q&A platforms (e.g., Stack Exchange or Humanities Commons) to validate and complement our GitHub-based findings. We also want to implement contributor-level network analysis to study collaboration patterns, identify key actors in digital humanities development, and measure

interdisciplinary connectivity. We also aim to enhance the topic modeling component using dynamic topic models or BERTopic to capture evolution and contextual nuance over time. Finally, we want to explore repository impact metrics (e.g., stars, forks, citation counts) to assess the visibility and scholarly reach of educationally created software.

Overall, our study contributes to the growing field of computational social science by showing how public software repositories can serve as empirical evidence for educational innovation and technological adaptation within the humanities. As digital scholarship expands, so does the need to track, understand, and support the digital transitions that occur across disciplines.

References

Berry, D. M. (2012). Introduction: Understanding the digital humanities. In *Understanding digital humanities* (pp. 1-20). London: Palgrave Macmillan UK.

Beshero-Bondar, E. E., & Parker, R. J. (2017). A GitHub Garage for a Digital Humanities Course. In *New Directions for Computing Education: Embedding Computing Across Disciplines* (pp. 259-276). Cham: Springer International Publishing.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Chlasta, K., Sochaczewski, P., Grabowska, I., & Jastrzębowska, A. (2022). MyMigrationBot: a cloud-based facebook social chatbot for migrant populations. *arXiv preprint arXiv:2208.13005*.

Escamilla, E., Klein, M., Cooper, T., Rampin, V., Weigle, M. C., & Nelson, M. L. (2022, September). The rise of GitHub in scholarly publications. In *International Conference on Theory and Practice of Digital Libraries* (pp. 187-200). Cham: Springer International Publishing.

Feng, X., Wachs, J., Daniotti, S., & Neffke, F. (2025). The building blocks of software work explain coding careers and language popularity. *arXiv preprint arXiv:2504.03581*.

Jones, S. E. (2013). *The emergence of the digital humanities* (p. 224). Taylor & Francis.

Juhász, S., Wachs, J., Kaminski, J., & Hidalgo, C. A. (2024). The software complexity of nations. *arXiv preprint arXiv:2407.13880*.

Kirschenbaum, M. G. (2016). What is digital humanities and what's it doing in English departments?. In *Defining digital humanities* (pp. 195-203). Routledge.

Klein, L. F., & Gold, M. K. (2016). *Debates in the Digital Humanities 2016*. University of Minnesota Press.

Lelis, A., Vretos, N., & Daras, P. (2020, July). Nadine-bot: An open domain migrant integration administrative agent. In *2020 IEEE international conference on Multimedia & Expo Workshops (ICMEW)* (pp. 1-6). IEEE Computer Society.

Liu, A. (2013). The meaning of the digital humanities. *pmla*, *128*(2), 409-423.

Mészáros, G., & Wachs, J. (2024). The dynamics of innovation in open source software ecosystems. *arXiv preprint arXiv:2411.14894*.

Rigas, E. S., Papoutsoglou, M., Kapitsaki, G. M., & Bassiliades, N. (2023, October). Mining software repositories to identify electric vehicle trends: the case of GitHUB. In *2023 10th International Conference on Behavioural and Social Computing (BESC)* (pp. 1-6). IEEE.

Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2015). *A new companion to digital humanities*. John Wiley & Sons.

Spiro, L. M., & Smith, S. M. (2016, July). Evaluating GitHub as a Platform of Knowledge for the Humanities. In *DH* (pp. 688-690).

Svensson, P. (2010). The landscape of digital humanities. *Digital humanities quarterly*, *4*(1).

Svensson, P. (2016). Humanities computing as digital humanities. In *Defining digital humanities* (pp. 159-186). Routledge.

Tassios, A., Tegos, S., Bouas, C., Manousaridis, K., Papoutsoglou, M., Kaltsa, M., Meditskos, G. (2025). LLM Performance in Low-Resource Languages: Selecting an

Optimal Model for Migrant Integration Support in Greek. *Future Internet*, *17*(6), 235.

Wachs, J. (2023). Digital traces of brain drain: developers during the Russian invasion of Ukraine. *EPJ Data Science*, *12*(1), 14.

Περίληψη

Εμμανουήλ Σ. Ρήγας, Μαρία Παπουτσόγλου, Γεωργία Μ. Καπιτσάκη, Βασίλειος Βασιλειάδης, Αικατερίνη Τικτοπούλου

Ενισχύοντας την έρευνα στις Ψηφιακές Ανθρωπιστικές Επιστήμες μέσω της ανάλυσης αποθετηρίων λογισμικού

Οι Ψηφιακές Ανθρωπιστικές Επιστήμες (ΨΑΕ) είναι ένας διεπιστημονικός τομέας που αναπτύσσεται δυναμικά και βρίσκεται στη διασταύρωση της επιστήμης των υπολογιστών και των ανθρωπιστικών επιστημών. Χαρακτηρίζεται από την ταχεία ανάπτυξη εργαλείων, εφαρμογών και επιστημονικών συνεισφορών την τελευταία δεκαετία. Ενώ ένα σημαντικό μέρος της βιβλιογραφίας έχει εξετάσει θεωρητικές προοπτικές, μεθοδολογικές προκλήσεις και μελέτες περιπτώσεων στον τομέα, λιγότερη προσοχή έχει δοθεί στη συστηματική ανάλυση των υποδομών λογισμικού που στηρίζουν μεγάλο μέρος των ΨΑΕ. Σε αυτήν την εργασία, αντιμετωπίζουμε αυτό το κενό διερευνώντας το τοπίο του λογισμικού Ψηφιακών Ανθρωπιστικών Επιστημών που φιλοξενείται στο GitHub, την πιο ευρέως χρησιμοποιούμενη πλατφόρμα για συνεργατική ανάπτυξη κώδικα. Συγκεκριμένα, συλλέξαμε και εξετάσαμε αποθετήρια με την ετικέτα “Ψηφιακές Ανθρωπιστικές Επιστήμες” και διεξήγαμε μια εμπειρική ανάλυση για να καταγράψουμε τις τάσεις στην εξέλιξή τους με την πάροδο του χρόνου, τους τύπους αδειών ανοιχτού κώδικα που υιοθετούνται συχνότερα, το εύρος των γλωσσών προγραμματισμού που χρησιμοποιούνται και τους θεματικούς προσανατολισμούς των έργων που προσδιορίζονται μέσω της μοντελοποίησης θεμάτων. Τα ευρήματα αποκαλύπτουν μια σταθερή και συνεπή αύξηση στον αριθμό των αποθετηρίων, υπογραμμίζοντας την αυξανόμενη εξάρτηση από υπολογιστικές μεθόδους σε διάφορους επιστημονικούς κλάδους των ανθρωπιστικών επιστημών, συμπεριλαμβανομένων των λογοτεχνικών σπουδών. Επιπλέον, τα αποτελέσματα δείχνουν ότι πολλά από αυτά τα έργα διαμορφώνονται από και συνδέονται με εκπαιδευτικά πλαίσια, υπογραμμίζοντας τη σημασία των Ψηφιακών Ανθρωπιστικών Επιστημών όχι μόνο ως ερευνητικού πεδίου αλλά και ως παιδαγωγικού τομέα που ενισχύει τον ψηφιακό γραμματισμό και τις τεχνικές δεξιότητες στην κοινότητα των ανθρωπιστικών επιστημών.

Λέξεις κλειδιά

Ψηφιακές Ανθρωπιστικές Επιστήμες, Αποθετήρια Λογισμικού, GitHub, Μοντελοποίηση Θεμάτων, Ανάλυση Δεδομένων.

Abstract

Emmanouil S. Rigas. Maria Papoutsoglou, Georgia M. Kapitsaki, Vasileios Vasileiadis, Aikaterini Tiktopoulou

Enhancing Digital Humanities Research Through the Analysis of Software Repositories

Digital Humanities is an increasingly dynamic and interdisciplinary field that lies at the intersection of computer science and the humanities, characterized by a rapid proliferation of tools, applications, and scholarly contributions over the past decade. While a substantial body of literature has examined theoretical perspectives, methodological challenges, and case studies in the domain, less attention has been devoted to systematically analyzing the software infrastructures that underpin much of this work. In this paper, we address this gap by investigating the landscape of Digital Humanities software hosted on GitHub, the most widely used platform for collaborative code development. Specifically, we collected and examined repositories tagged with the *Digital Humanities* keyword and conducted an empirical analysis to capture trends in their evolution over time, the types of open-source licenses most frequently adopted, the range of programming languages employed, and the thematic orientations of projects identified through topic modeling. The findings reveal a steady and consistent increase in the number of repositories, highlighting the growing reliance on computational methods across diverse humanities disciplines, including literary studies. Moreover, the results indicate that many of these projects are shaped by and connected to educational contexts, underscoring the importance of Digital Humanities not only as a research field but also as a pedagogical domain that fosters digital literacy and technical skills in the humanities community.

Keywords

Digital Humanities, Software Repositories, GitHub, Topic Modeling, Data Analysis.