PROJECT TEAM

Zoe G. LeBlanc (PI)
School of Information Sciences,
University of Illinois Urbana-Champaign

Jeri E. Wieringa
Center for Digital Humanities
Princeton University

Coding DH Project

Exploring DH Coding Communities and Practices on GitHub

LEARN MORE

Introducing *Coding DH*

Since 2008, GitHub has become vital for both software companies and scholars. Although scholars are a small part of the user base, GitHub's academic significance has grown, hosting everything from large research projects to syllabi and datasets. In Digital Humanities (DH), there's been a focus on using GitHub for version control and collaborative coding. But despite its importance, the role of GitHub in DH research and teaching remains under explored. The *Coding DH* project investigates GitHub's influence on knowledge production and community formation in DH, focusing on:

1. **Identifying DH on GitHub**
2. **Studying Global DH Coding Communities**
3. **Evolution of DH Coding Practices**

Contextualizing *Coding DH*

Studying platforms like GitHub is common in Mining Software Repositories (MSR), a Computer Science subfield. However, MSR research mainly enhances software engineering practices rather than examining coding communities. Unlike studies using archived data collection projects like GHTorrent, we employ custom code to work directly with GitHub APIs.

Beyond CS, there are two previous studies that are relevant to *Coding DH*, which explored coding practices in Library Science and Journalism, respectively. But these studies rely on a far smaller sample size (~1000 and ~100). To date, the only study of DH and GitHub was in 2016, when Lisa Spiro and Sean Morey Smith surveyed DH scholars on their coding practices. *Coding DH* builds on this existing work, as well as DH scholarship on Twitter and citation communities.

Challenges & Limitations of GitHub Data

- **API Rate Limits:** GitHub imposes rate limits on API usage, making data collection a slow process. Compiling this dataset has taken over two years using the GitHub Search API and General APIs.
- **Metadata Version History:** Version history does not apply to metadata—data such as bios, descriptions, or topics are not archived via the GitHub API, meaning only the most current version is available.
- **Data Access:** GitHub prevents API access to certain activities, such as Projects or Teams, as well as private repositories, so our data analysis will always represent a lower bound of activity.
- **Interpreting Code Work:** Not all interactions represent the same level of engagement or labor, so it is essential to consider both what data is missing and what the existing data represents. We also do not argue that coding is required to be a DH scholar, but instead want to explore those that are engaging with DH on GitHub.

Protecting Users' Data Privacy

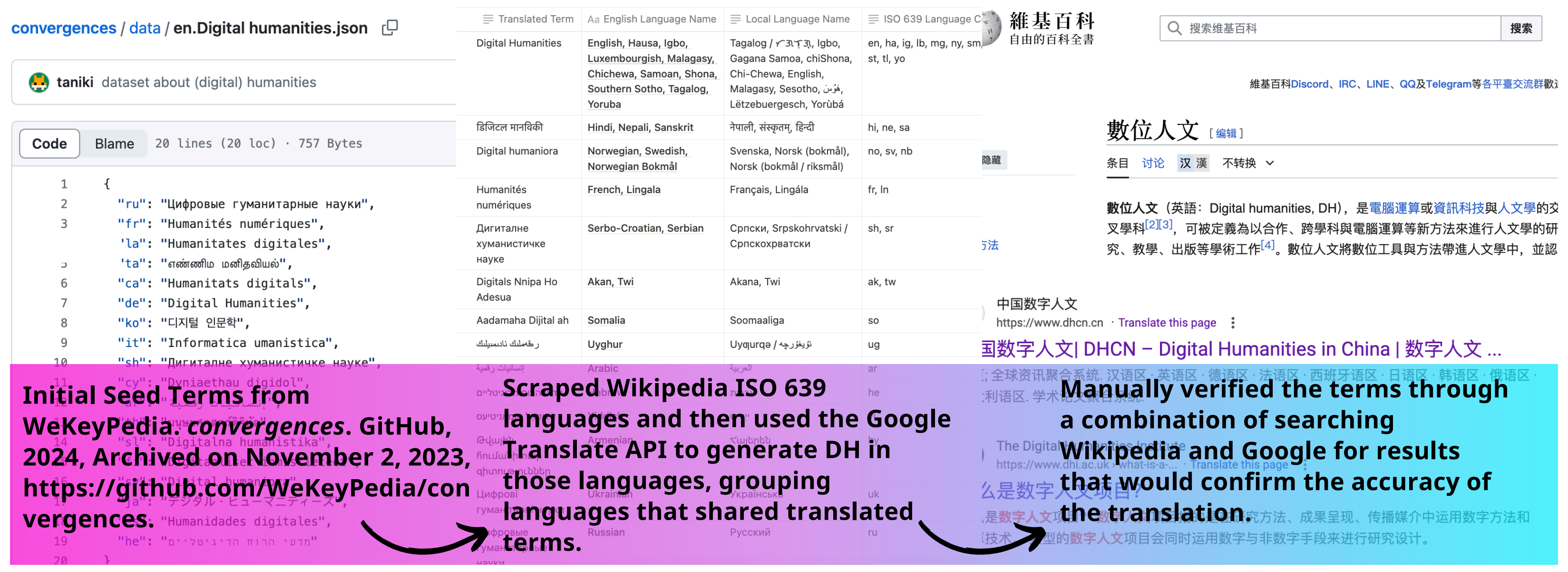
GitHub is primarily designed for sharing code, so users do not typically expect their interactions, bios, and affiliations to be studied. While the data used in this study is technically public, it exists in a gray area between public and personal information. GitHub's terms of service permits researchers to use public, non-personal information for research purposes if the resulting publications are open access. However, *Coding DH* is committed to prioritizing users' data privacy, and therefore will **not release any identifiable user data**, in either aggregate or individual form, unless we have explicit user consent or the user is extremely active on the platform.

Our goal is to **balance user privacy with recognizing those contributing to DH code work**, which often isn't as visible as citations or grants.

Searching for DH Activity on GitHub

Identifying DH Globally

To find DH activity on GitHub, we initially used the GitHub Search API, which returns results for users, repositories, organizations, topics, issues, gists, and even code, as long as it includes the search query.



convergences / data / en.Digital humanities.json

taniki dataset about (digital) humanities

Code Blame 20 lines (20 loc) · 757 Bytes

```
{
  "ru": "Цифровые гуманитарные науки",
  "fr": "Humanités numériques",
  'la': "Humanitates digitales",
  'ta': "எண்ணிம மனிதவியல்",
  "ca": "Humanitats digitals",
  "de": "Digital Humanities",
  "ko": "디지털 인문학",
  "it": "Informatica umanistica",
  "sh": "Дигиталне хуманистичке науке",
  "cy": "Dyniaethau digidol",
  "sl": "Digitalna humanistika",
  "ja": "デジタル・ヒューマニティーズ",
  "Humanidades digitales",
  "he": "מדעי הרוח דיגיטליים"
}
```

Translated Term	English Language Name	Local Language Name	ISO 639 Language C
Digital Humanities	English, Hausa, Igbo, Luxembourgish, Malagasy, Chichewa, Samoan, Shona, Southern Sotho, Tagalog, Yoruba	Tagalog / ᜆᜄᜎᜓᜄ᜔, Igbo, Gagana Samoa, chiShona, Chi-Chewa, English, Malagasy, Sesotho, هَوُسَ, Lëtzebuergesch, Yorùbá	en, ha, ig, lb, mg, ny, sm, st, tl, yo
डिजिटल मानविकी	Hindi, Nepali, Sanskrit	नेपाली, संस्कृतम्, हिन्दी	hi, ne, sa
Digital humaniora	Norwegian, Swedish, Norwegian Bokmål	Svenska, Norsk (bokmål), Norsk (bokmål / riksmål)	no, sv, nb
Humanités numériques	French, Lingala	Français, Lingála	fr, ln
Дигиталне хуманистичке науке	Serbo-Croatian, Serbian	Српски, Srpskohrvatski / Српскохрватски	sh, sr
Digitals Nnipa Ho Adesua	Akan, Twi	Akana, Twi	ak, tw
Aadamaha Dijital ah	Somalia	Soomaaliga	so
رەقەملىك ئادىمىيلار	Uyghur	Uyqurqə / ئۇيغۇرچە	ug
إنسانيات رقمية	Arabic	العربية	ar
			he
	Armenian		hy
Цифрові гуманітарні	Ukrainian	Українська	uk
Цифровые гуманитарные науки	Russian	Русский	ru

維基百科 自由的百科全書

搜索維基百科 搜索

維基百科Discord、IRC、LINE、QQ及Telegram等各平臺交流群歡迎

數位人文 [编辑]

条目 讨论 汉 漢 不转换

數位人文（英語：Digital humanities, DH），是電腦運算或資訊科技與人文學的交叉學科[2][3]，可被定義為以合作、跨學科與電腦運算等新方法來進行人文學的研究、教學、出版等學術工作[4]。數位人文將數位工具與方法帶進人文學中，並認

中国数字人文
https://www.dhcn.cn · Translate this page

国数字人文| DHCN – Digital Humanities in China | 数字人文 ...

The Digital Humanities Institute
https://www.dhi.ac.uk › what-is-a-... · Translate this page

么是数字人文项目？

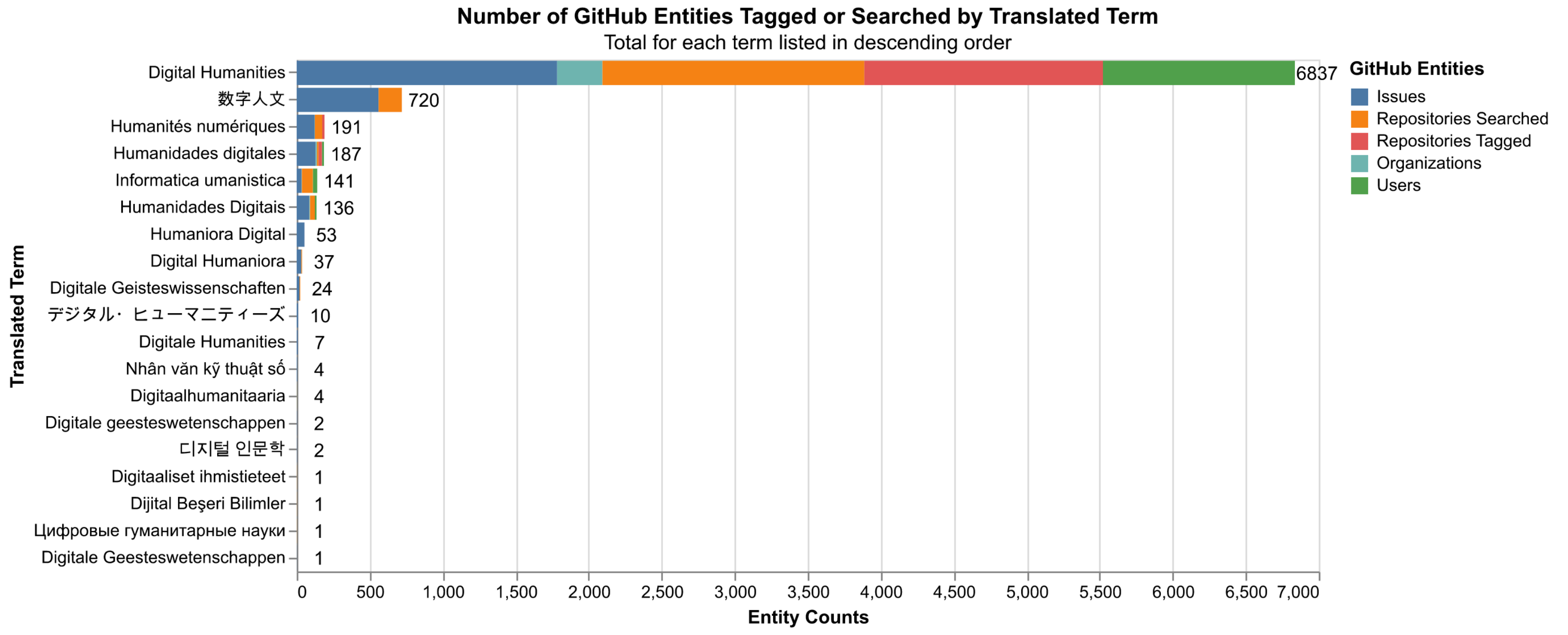
Initial Seed Terms from WeKeyPedia. *convergences*. GitHub, 2024, Archived on November 2, 2023, https://github.com/WeKeyPedia/convergences.

Scraped Wikipedia ISO 639 languages and then used the Google Translate API to generate DH in those languages, grouping languages that shared translated terms.

Manually verified the terms through a combination of searching Wikipedia and Google for results that would confirm the accuracy of the translation.

Since GitHub is a global platform, we not only searched for DH in English, but also across all **183** ISO-639 languages. To generate these terms, we used a combination of existing DH translations and automated translation that we verified manually, giving us a total of **106** unique terms for DH that correlate to **123** languages.

Digital Humanities Dominant?



Number of GitHub Entities Tagged or Searched by Translated Term
Total for each term listed in descending order

Our initial search yielded results for only **18** of 106 terms, corresponding to **30** potential languages. While Chinese, French, Spanish, Italian, and Portuguese are well represented, 'Digital Humanities' is the dominant term. This raises the question of whether GitHub is mainly used by English-speaking scholars. However, given GitHub's over 100 million active users globally, this is unlikely.

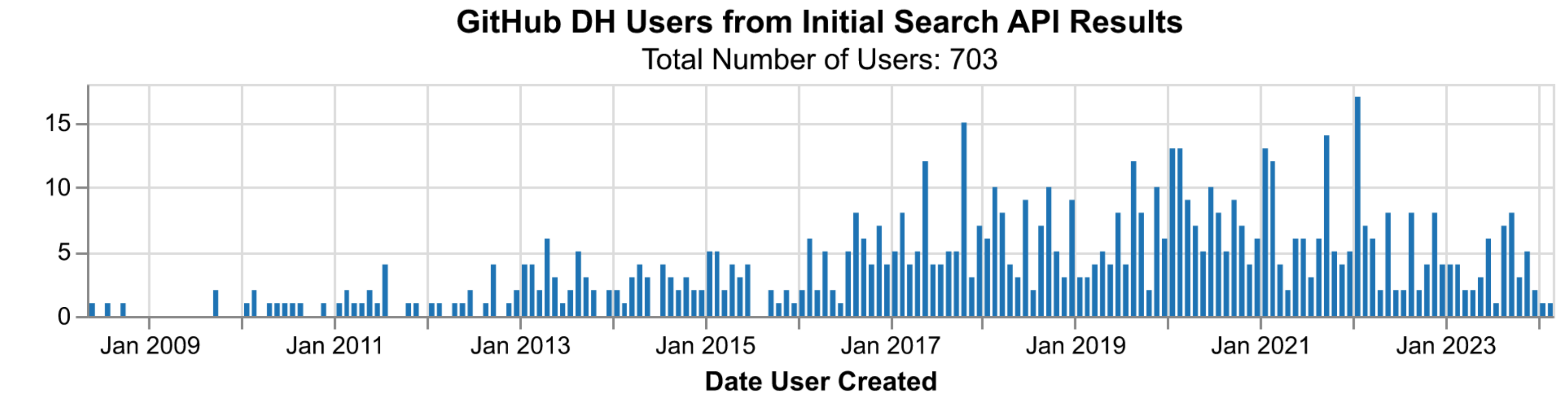
Instead, the dominance of the term "Digital Humanities" likely reflects the platform's limitations. For example, **GitHub encodes all top-level repository and user names in ASCII, which does not support character-based languages**. This likely leads to the use of the term DH even in projects where English is not the primary language.

Additionally, the platform has evolved over time, which presents challenges for data collection. For example, topics were introduced in early 2017 and are less consistently used in older repositories. Many users and repositories also have minimal descriptive text or do not explicitly use the term DH despite being well-known within the DH community.

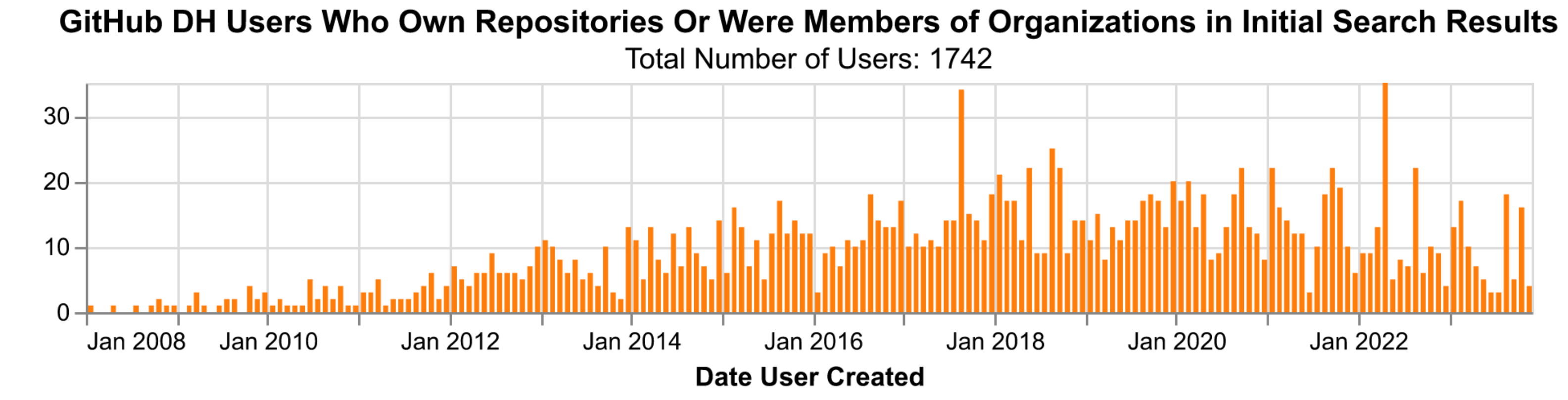
While our initial search uncovered some DH-related activity on GitHub, it also revealed the limitations of the search approach, which relies on users' explicit engagement with DH and the consistency of GitHub's data.

Expanding & Exploring DH Coding Communities

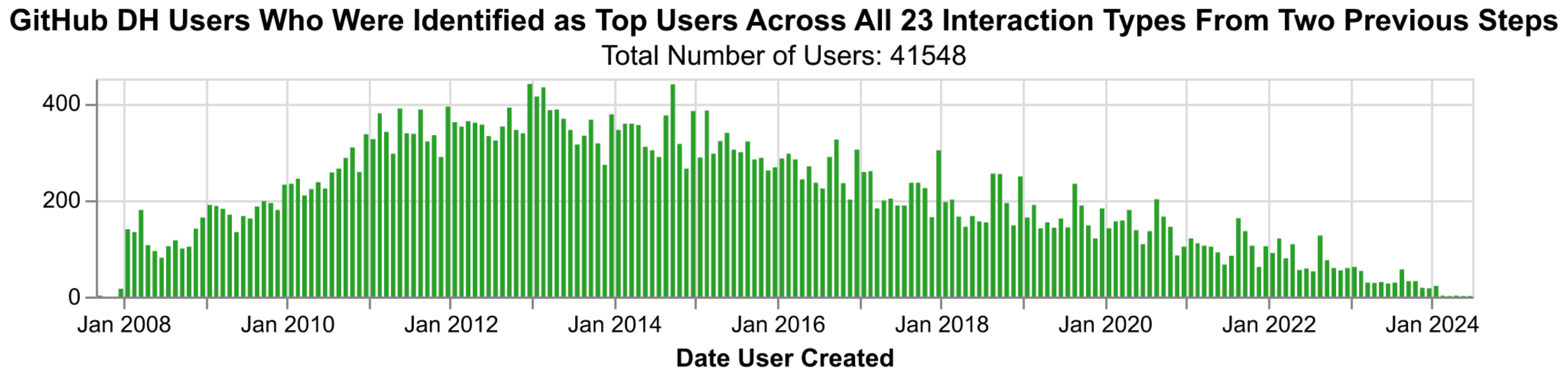
To get a more comprehensive view of DH users on GitHub, we **added two more steps to our data collection process**. First, we added any user who owned a repository or was a member of an organization that was identified as part of DH in our initial search. Second, we then also scraped data on **23** different interaction types between users and other GitHub entities, and added those users whose activity levels were in the top 25% of activities, for a total of **~43993** unique DH users.



GitHub DH Users from Initial Search API Results
Total Number of Users: 703

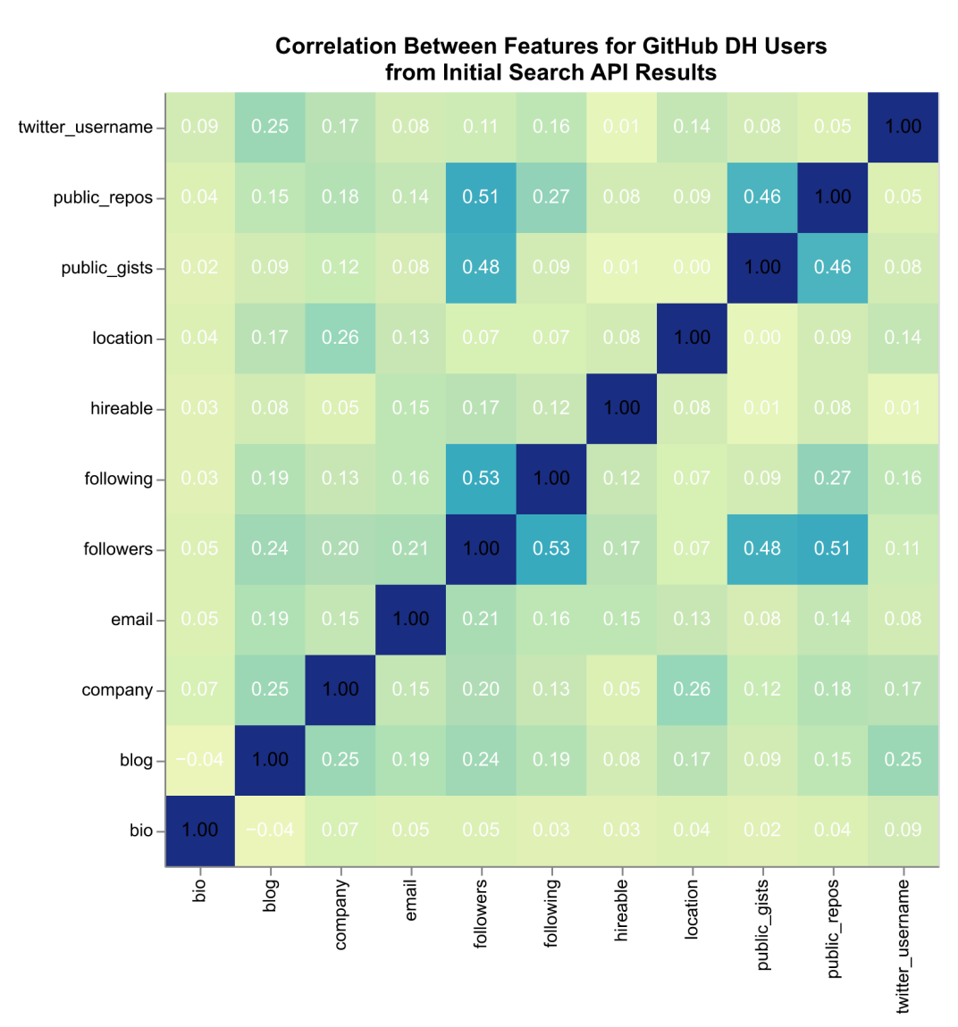


GitHub DH Users Who Own Repositories Or Were Members of Organizations in Initial Search Results
Total Number of Users: 1742

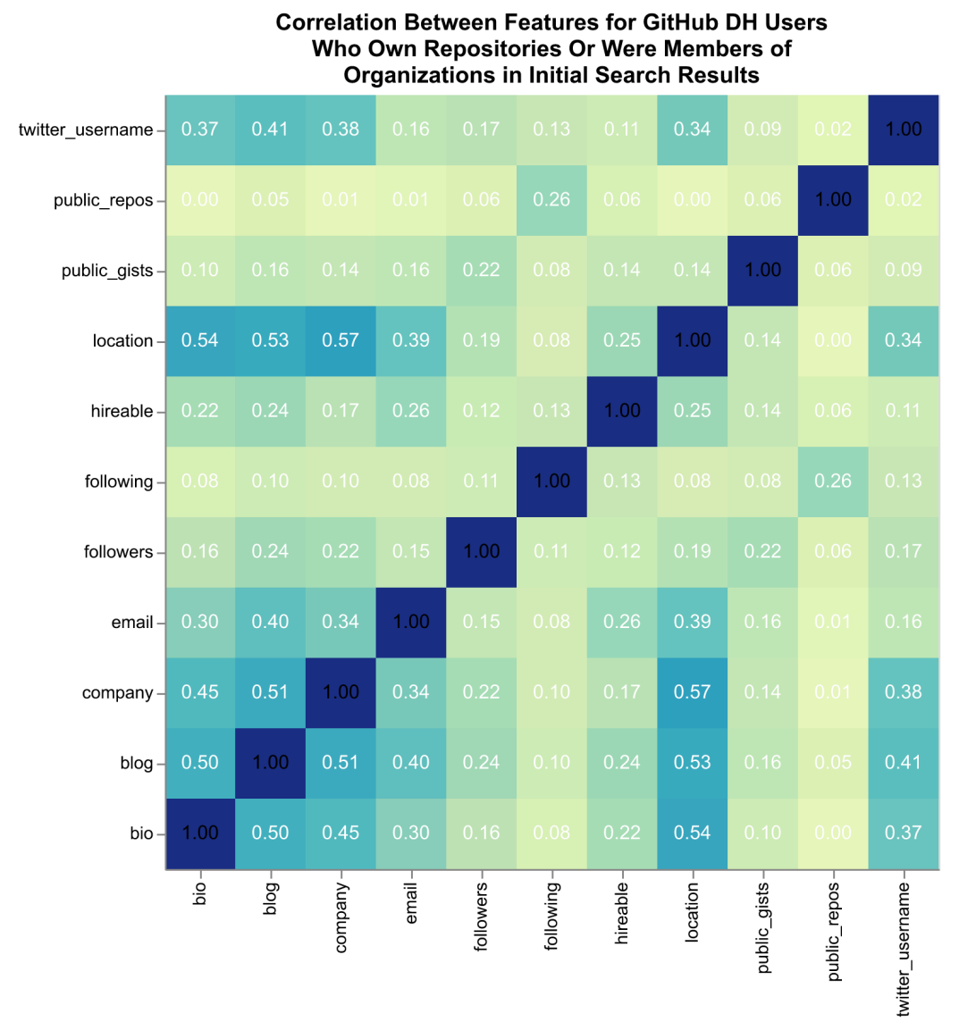


GitHub DH Users Who Were Identified as Top Users Across All 23 Interaction Types From Two Previous Steps
Total Number of Users: 41548

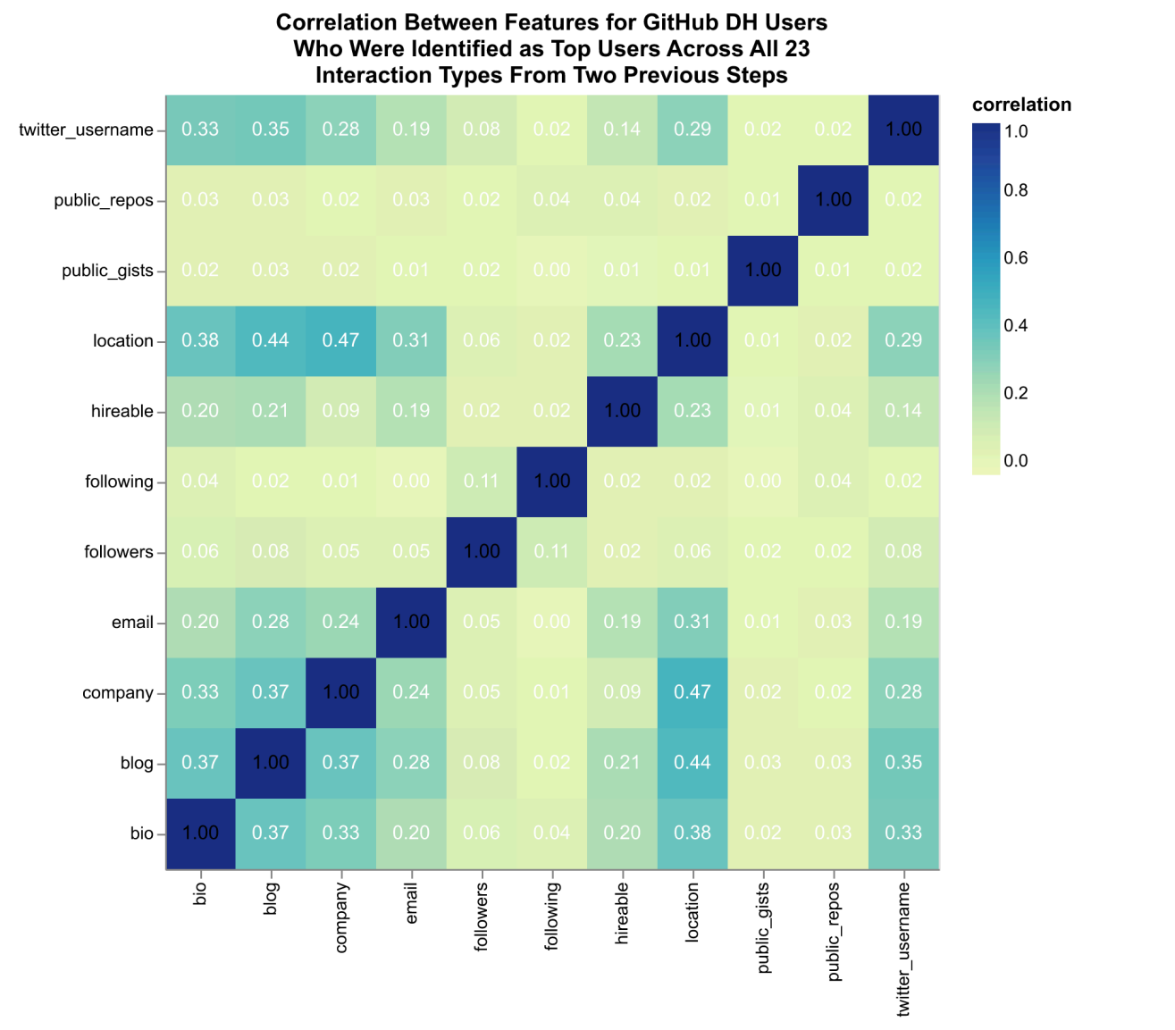
These graph show that while few GitHub users explicitly mention DH in their name or bio, many users are involved with DH repositories or organizations, and even more engage across them. But there is also a noticeable decline in the third stage. This trend is partly due to the graph reflecting the date when users were created, however, it also raises questions about who gets to be a coder in DH and whether we are providing that opportunity to the next generation of scholars.



Correlation Between Features for GitHub DH Users from Initial Search API Results



Correlation Between Features for GitHub DH Users Who Own Repositories Or Were Members of Organizations in Initial Search Results



Correlation Between Features for GitHub DH Users Who Were Identified as Top Users Across All 23 Interaction Types From Two Previous Steps

Exploring metadata for each data collection stage shows that DH users from our initial search have weaker correlations across categories, indicating more heterogeneity among these users compared to later stages. While these patterns may be due to sample size, future *Coding DH* research will explore if these trends also correspond to linguistic communities.

Ultimately, by detailing our data collection process and initial results, we argue that we need to view coding as more than just best practices in DH, and instead understand how these trends are shaping the future of DH.