Received: 15 January 2021 | Revised: 30 April 2021 | Accepted: 26 May 2021

DOI: 10.1002/asi.24533

RESEARCH ARTICLE

JASIST WILEY

Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape

Jan Luhmann | **Manuel Burghardt**

Computational Humanities Group, Leipzig University, Leipzig, Germany

Correspondence
Manuel Burghardt, Computational Humanities Group, Leipzig University Augustusplatz 10, 04109 Leipzig, Germany.
Email: burghardt@informatik.uni-leipzig.de

Abstract

Although digital humanities (DH) has received a lot of attention in recent years, its status as "a discipline in its own right" (Schreibman et al., A companion to digital humanities (pp. xxiii–xxvii). Blackwell; 2004) and its position in the overall academic landscape are still being negotiated. While there are countless essays and opinion pieces that debate the status of DH, little research has been dedicated to exploring the field in a systematic and empirical way (Poole, Journal of Documentation; 2017:73). This study aims to contribute to the existing research gap by comparing articles published over the past three decades in three established English-language DH journals (Computers and the Humanities, Literary and Linguistic Computing, Digital Humanities Quarterly) with research articles from journals in 15 other academic disciplines (corpus size: 34,041 articles; 299 million tokens). As a method of analysis, we use latent Dirichlet allocation topic modeling, combined with recent approaches that aggregate topic models by means of hierarchical agglomerative clustering. Our findings indicate that DH is simultaneously a discipline in its own right and a highly interdisciplinary field, with many connecting factors to neighboring disciplines—first and foremost, computational linguistics, and information science. Detailed descriptive analyses shed some light on the diachronic development of DH and also highlight topics that are characteristic for DH.

1 | INTRODUCTION

The rise of digital information technology and the accompanying "computational turn" has fundamentally changed the way we do research (Berry, 2011). Since the humanities do not have a distinguished tradition of using computer-based research methods, current developments in this academic branch are referred to as *digital humanities* (DH). Although Schreibman et al. (2004) suggested that we "consider digital humanities as a discipline in its own right," its purpose and status have continued to be the subject of numerous debates and its very nature is still being negotiated. The difficulty in defining DH arises from the disciplinary diversity and a set of rather heterogeneous scholarly practices within the field (Svensson, 2010). One popular attempt at embracing this interdisciplinarity and heterogeneity, is to conceptualize DH as an inclusive "big tent."[1] However, Terras (2013) challenges this concept as it gives a false idea about the actual, limited institutional and structural openness of DH, while it also shies away from defining what fits under the tent and what does not, blurring and obfuscating the actual disciplinary boundaries.

Along these lines, Roth (2019) identifies three currently existing subfields as distinct and non-overlapping communities of DH: "digitized humanities," which deals with the creation, use, and analysis of digitized resources; "numerical humanities," also known as "computational humanities," which approaches humanities research questions through computational models; and "humanities of the digital," which is devoted to the study of digital phenomena such as online communication from a humanities perspective. Extending Roth's taxonomy, Burghardt (2020) has added "public humanities," which are concerned with scholarly communication in the humanities, digital publishing, and electronic learning.

From a historic perspective, McCarty (2015) notes that DH has only developed self-awareness as a field or "discipline" in its own right since the 2000s. He also points out that in terms of its standing with related disciplines it has "a center all over the disciplinary map and a circumference that is at best uncertain" (p. 79). A continuing part of the debate about the status of DH is its relationship to the field of information science (Burghardt & Luhmann, 2021), with which it shares an interest in information management, data modeling, and libraries. Robinson et al. (2015) have also discussed the possibility of a joint future between the two. In contrast, Gladney (2012) considers DH an unnecessary invention from the perspective of information science.

The lack of clarity about what actually constitutes DH—beyond being an encounter of some sort between the humanities and the digital—and what sets it apart from similar fields have led some to deny it is or ever will become a discipline in its own right (Alvarado, 2012; Cordell, 2016). Even more critical voices are inclined to dismiss DH as mere hype or an empty buzzword that is only used to generate attention and raise funding for ill-conceived humanities research (Brennan, 2017; Fish, 2018). Still others see DH as a movement (Adams & Gunn, 2013; Holm et al., 2015), pioneering a scientific or computational turn in the humanities (G. Hall, 2012), while Pannapacker (2012) notes that "it won't be long until the digital humanities are, quite simply, 'the humanities'" (p. 233).

This brief survey of views on DH is not intended to be exhaustive, but rather provides a rough overview of the major debates that have surrounded DH in recent years. In fact, debating the status of DH has evolved into a genre of its own (Kirschenbaum, 2013), as with some notable anthologies such as "Defining Digital Humanities" (Terras et al., 2013), "Interdisciplining Digital Humanities" (J. T. Klein, 2015), and multiple volumes of the "Debates in the Digital Humanities" series (Gold, 2012; Gold & Klein, 2016; L. F. Klein & Gold, 2019).

1.1 | Goals of this study

One point raised by Poole (2017) is that the discussion of the identity of DH so far relies largely on anecdotal evidence, while "the field would benefit from exploring itself more empirically" (p. 107). This study aims at contributing to Poole's suggestion by systematically assessing how DH is situated in the academic landscape. We will do so by comparing research articles published over the past three decades in three established DH journals with research articles published in journals from 15 other academic disciplines.

As a method of analysis we use topic modeling (latent Dirichlet allocation [LDA]; as proposed by Blei et al., 2003), combined with recent approaches that aggregate topic models using hierarchical agglomerative clustering (Vega-Carrasco et al., 2020). Topic modeling allows us to find patterns of word usage, commonly referred to as topics, among the research articles, which can be interpreted as research topics, concepts, or aspects of academic discourse. Topic modeling is widely used for studying corpora of academic (for some examples, see Goldstone & Underwood, 2012, 2014; Griffiths & Steyvers, 2004; McFarland et al., 2013), enabling us to determine diachronic trends among research topics (D. Hall et al., 2008) and to calculate topical similarity between articles and disciplines.

2 | RELATED WORK

In this section we will provide an overview of existing approaches to the scientometric analysis of DH and its academic and disciplinary status. First to be mentioned are Sula and Hill (2019), who have presented a quantitative analysis of research articles from "Computers and the Humanities" (*CHum*) and "Literary and Linguistic Computing" (*LLC*) from 1966 to 2004. More concretely, they investigated the diachronic development of different types of media objects that are studied in the papers as well as the authors' locations and discipline affiliations.

Gao et al. (2017), who are interested in disciplinary relationships to English studies, history, information science, computational linguistics and natural language processing, and statistics, conducted a study of co-citation networks and citation patterns in *CHum*, *LLC* and "Digital Humanities Quarterly" (*DHQ*) articles from 1966 to 2016. In a similar vein, Tang et al. (2017) analyzed intellectual cohesion over time, based on co-authorship, co-citation, and article keywords. Another interesting approach is provided by Weingart and Eichmann-Kalwara (2017) and by Weingart (2016), as they measured trends in papers accepted for the international annual Digital Humanities conference. They considered author affiliations to disciplines and locations, co-authorship, representation, and diversity (see also Eichmann-Kalwara et al., 2018), as well as research topics that are based on the topic keywords manually assigned by article authors.

Although topic modeling has become very popular in DH and in particular in computational literary studies (see, e.g., Meeks & Weingart, 2012; Robertson, 2016; Schöch, 2017), it has been used very little for analyzing DH in a scientometric way. This is rather surprising as topic modeling has already been used with great success to analyze the content of academic publications for other disciplines (see, e.g., Goldstone & Underwood, 2014; Griffiths & Steyvers, 2004; D. Hall et al., 2008; Riddell, 2014). One of the few examples in DH can be found with Meeks (2011), who used topic modeling on a rather small scale of 50 texts and 20 topics to visualize the discourse in articles that discuss the status of DH. This work has been extended by Callaway et al. (2020), who analyzed 334 articles of the same genre, using a topic model of 55 topics. They related the identified topics to metadata on authors' gender identity and disciplinary affiliation, revealing definitional aspects in the articles including methodological paradigms (e.g., coding, distant reading) and community values, as well as reflections on the lack of diversity in the DH community. Earlier, Sula (2013) developed a conceptual model of the relation between DH and library and information science, which identified five topics (using LDA topic modeling) in 86 abstracts of DH-related research in the library, information science and technology abstracts (LISTA) database.

Puschmann and Bastos (2015) took yet another approach by studying two academic online platforms that are related to DH and analyzing topics among user posts, using co-word analysis and topic modeling. They showed how each platform tends toward certain sub-areas of DH and disciplinary affiliations.

This review of related work reveals that a large-scale study of DH journals in comparison with other academic journals is still missing, which is the main motivation for this research. In the following section we will describe the methodological details of the corpus creation and data analysis approach.

3 | METHODS

To examine how DH relates to other disciplines and how it positions itself in the overall academic landscape, we compared established journals from DH with further journals from 15 other scientific disciplines. Journal articles are an important means of scholarly communication and for this reason they are often used as subjects of scientometric studies (Paul & Girju, 2009; Sula & Hill, 2019). As a minor caveat, however, it has to be pointed out that they are not always coextensive with a "discipline" and disciplinary boundaries do not necessarily align with journal boundaries. Thus, whenever we make assumptions about DH as a discipline, these are guided and possibly biased by the medium of journal articles and our selection of journals. The actual comparison of disciplines and journals is based on shared topics, which we compute by using topic modeling in combination with an agglomerative hierarchical clustering procedure (Vega-Carrasco et al., 2020).

3.1 | Corpus creation and sampling

3.1.1 | Selection of academic disciplines and journals

As we want to investigate the role and position of DH on the stage of academic disciplines, the core of our corpus consists of three DH journals that are well-established in the community and that have been around for some time. The journal with the longest history is *Computers and the Humanities* (*CHum*), which has been around since 1966 and was renamed *Language Resources and Evaluation* (*LRE*) in 2005. *Literary and Linguistic Computing* (*LLC*) was founded in 1986 and was renamed *Digital Scholarship in the Humanities* (*DSH*) in 2015. *LLC/DSH* is published by Oxford University Press on behalf of the European Association for DH (EADH) and the Alliance of Digital Humanities Organizations (ADHO). *Digital Humanities Quarterly* (*DHQ*) is the youngest DH journal in our corpus, being founded in 2007 by the Association for Computers and the Humanities (ACH). To allow for a diachronic perspective, we selected articles published in the above journals over a period of three decades; specifically, 1990–2019. In addition to the subcorpus of DH articles, we included further subcorpora of 15 other disciplines that are related to DH in some way, resulting in a total of 34,041 articles with over 299 million tokens (for a detailed overview see Table 1). We let an existing study by Sula and Hill (2019) guide our selection of disciplines, as it provides a comprehensive overview of authors who published in *CHum* and *LLC* along with their affiliations to core disciplines.

For each of the chosen disciplines we selected English-language, peer-reviewed journals with regular issues from 1990 to 2019, which publish original research articles on a broad range of topics. By researching the history of each journal as well as checking bibliometrical

TABLE 1 Overview of disciplines and selected journals

Discipline	Journal	Publisher	Decade	Number of articles	Number of tokens	Sampling proportion in discipline-decade stratum	Number of articles overall	Number of tokens overall	Mean number of tokens per article (±SD)
Applied computer science	Artificial Intelligence	Elsevier	1990	285	3,181,682	25.00%	744	8,650,403	11,627 (±4,991)
			2000	339	3,828,459	25.00%			
			2010	120	1,640,262	25.00%			
	IEEE Transactions on Software Engineering	IEEE	1990	131	1,567,353	25.00%	291	3,636,601	12,497 (±4,085)
			2000	78	916,275	25.00%			
			2010	82	1,152,973	25.00%			
	IEEE/ACM Transactions on Networking	IEEE	1990	48	474,211	12.50%	199	2,197,209	11,041 (±2,833)
			2000	70	754,769	12.50%			
			2010	81	968,229	12.50%			
	Science of Computer Programming	Elsevier	1990	144	1,486,396	37.50%	572	6,289,476	10,996 (±4,208)
			2000	271	2,869,946	37.50%			
			2010	157	1,933,134	37.50%			
Art	Art Journal	Routledge	1990	412	1,768,127	33.33%	839	4,115,126	4,905 (±2,637)
			2000	292	1,475,695	33.33%			
			2010	135	871,304	33.33%			
	The Journal of Aesthetics and Art Criticism	Wiley-Blackwell	1990	298	1,967,108	33.33%	815	5,910,454	7,252 (±3,188)
			2000	284	2,203,994	33.33%			
			2010	233	1,739,352	33.33%			
	Visual Arts Research	University of Illinois Press	1990	149	831,833	33.33%	444	2,340,459	5,271 (±2,367)
			2000	163	958,826	33.33%			
			2010	132	549,800	33.33%			
Classical studies	The American Journal of Philology	Johns Hopkins University Press	1990	274	1,883,813	50.00%	595	5,170,403	8,690 (±4,291)
			2000	218	2,066,509	50.00%			
			2010	103	1,220,081	50.00%			
	The Classical Quarterly	Cambridge University Press	1990	498	3,131,373	50.00%	1,392	8,893,435	6,389 (±4,278)
			2000	542	3,410,261	50.00%			
			2010	352	2,351,801	50.00%			
Computational linguistics	Computational Linguistics	MIT Press	1990	215	1,928,721	62.50%	656	6,681,793	10,186 (±5,338)
			2000	199	2,055,768	62.50%			
			2010	242	2,697,304	62.50%			
		Elsevier	1990	61	540,857	12.50%	143	1,272,761	8,900 (±2,863)

TABLE 1 (Continued)

Discipline	Journal	Publisher	Decade	Number of articles	Number of tokens	Sampling proportion in discipline-decade stratum	Number of articles overall	Number of tokens overall	Mean number of tokens per article (±SD)
	Computer Speech & Language		2000	54	474,567	12.50%			
			2010	28	257,337	12.50%			
	Machine Translation	Springer	1990	112	925,316	25.00%	290	2,630,862	9,072 (±4,114)
			2000	94	900,335	25.00%			
			2010	84	805,211	25.00%			
Digital humanities	Computers and the Humanities/Language Resources and Evaluation	Springer	1990	330	1,747,961	50.00%	909	5,683,944	6,253 (±2,997)
			2000	237	1,284,925	45.00%			
			2010	342	2,651,058	33.33%			
	Digital Humanities Quarterly	Alliance of Digital Humanities Organizations	1990	0	0	0.00%	385	2,779,115	7,218 (±3,259)
			2000	49	342,752	10.00%			
			2010	336	2,436,363	33.33%			
	Literary and Linguistic Computing/Digital Scholarship in the Humanities	Oxford University Press	1990	287	1,338,981	50.00%	1,254	6,624,759	5,283 (±2,615)
			2000	393	2,001,329	45.00%			
			2010	574	3,284,449	33.33%			
History	Journal of Contemporary History	Sage Publications	1990	271	2,464,026	37.50%	777	7,789,832	10,026 (±3,090)
			2000	253	2,452,560	37.50%			
			2010	253	2,873,246	37.50%			
	Journal of World History	University of Hawaii Press	1990	96	914,695	25.00%	348	3,814,847	10,962 (±3,612)
			2000	116	1,354,098	25.00%			
			2010	136	1,546,054	25.00%			
	The Historical Journal	Cambridge University Press	1990	368	4,251,616	37.50%	972	11,614,908	11,949 (±3,305)
			2000	387	4,752,435	37.50%			
			2010	217	2,610,857	37.50%			
Information science	JASIST	Wiley-Blackwell	1990	253	1,601,670	50.00%	3,087	21,400,368	6,932 (±3,162)
			2000	1,605	10,512,349	50.00%			
			2010	1,229	9,286,349	50.00%			
	Journal of Documentation	Emerald Group Publishing	1990	173	1,119,768	50.00%	922	6,737,621	7,308 (±2,755)
			2000	327	2,260,000	50.00%			
			2010	422	3,357,853	50.00%			

(Continues)

TABLE 1 (Continued)

Discipline	Journal	Publisher	Decade	Number of articles	Number of tokens	Sampling proportion in discipline-decade stratum	Number of articles overall	Number of tokens overall	Mean number of tokens per article (±SD)
Linguistics	Language	Linguistic Society of America	1990	124	1,511,073	33.33%	390	4,714,007	12,087 (±5,409)
			2000	182	2,140,424	33.33%			
			2010	84	1,062,510	33.33%			
Linguistics	Linguistic Inquiry	MIT Press	1990	306	2,409,660	33.33%	750	6,190,414	8,254 (±5,375)
			2000	310	2,588,178	33.33%			
			2010	134	1,192,576	33.33%			
	Natural Language & Linguistic Theory	Springer	1990	153	1,967,398	33.33%	431	5,721,620	13,275 (±4,760)
			2000	138	1,877,389	33.33%			
			2010	140	1,876,833	33.33%			
Literary studies	Journal of Narrative Theory	Eastern Michigan University	1990	13	108,423	6.25%	259	2,326,106	8,981 (±2,821)
			2000	161	1,450,661	6.25%			
			2010	85	767,022	6.25%			
	Modern Fiction Studies	Johns Hopkins University Press	1990	279	2,456,073	46.88%	735	7,066,726	9,615 (±2,929)
			2000	302	3,089,770	46.88%			
			2010	154	1,520,883	46.88%			
	New Literary History	Johns Hopkins University Press	1990	452	3,652,913	46.88%	1,008	8,533,845	8,466 (±3,188)
			2000	371	3,203,998	46.88%			
			2010	185	1,676,934	46.88%			
Mathematics	Advances in Mathematics	Elsevier	1990	32	286,705	12.50%	231	2,448,891	10,601 (±4,503)
			2000	44	429,291	12.50%			
			2010	155	1,732,895	12.50%			
	Annals of Mathematics	Princeton University	1990	282	2,908,003	43.75%	880	9,667,338	10,986 (±4,738)
			2000	385	4,275,391	43.75%			
			2010	213	2,483,944	43.75%			
	Journal of the American Mathematical Society	American Mathematical Society	1990	223	2,366,627	43.75%	552	6,139,542	11,122 (±4,607)
			2000	231	2,549,738	43.75%			
			2010	98	1,223,177	43.75%			
Musicology	Ethnomusicology	University of Illinois Press	1990	131	1,120,965	25.00%	403	3,988,194	9,896 (±4,379)
			2000	157	1,708,681	25.00%			
			2010	115	1,158,548	25.00%			
	Popular Music	Cambridge University Press	1990	198	1,586,551	25.00%	553	4,874,583	8,815 (±3,504)

TABLE 1 (Continued)

Discipline	Journal	Publisher	Decade	Number of articles	Number of tokens	Sampling proportion in discipline-decade stratum	Number of articles overall	Number of tokens overall	Mean number of tokens per article (±SD)
			2000	244	2,129,560	25.00%			
			2010	111	1,158,472	25.00%			
	The Journal of Musicology	University of California Press	1990	216	1,964,969	50.00%	444	4,481,535	10,094 (±4,355)
			2000	134	1,392,629	50.00%			
			2010	94	1,123,937	50.00%			
Philosophy	Philosophical Topics	University of Arkansas Press	1990	188	2,070,764	33.33%	524	5,691,598	10,862 (±3,730)
			2000	160	1,794,465	33.33%			
			2010	176	1,826,369	33.33%			
	Philosophy and Phenomenological Research	International Phenomenological Society	1990	393	2,904,100	33.33%	1,057	9,479,760	8,969 (±4,468)
			2000	423	4,006,881	33.33%			
			2010	241	2,568,779	33.33%			
	The Journal of Philosophy	Journal of Philosophy, Inc.	1990	222	1,902,382	33.33%	593	5,386,218	9,083 (±3,637)
			2000	227	2,130,914	33.33%			
			2010	144	1,352,922	33.33%			
Political science	American Journal of Political Science	Wiley-Blackwell	1990	462	4,299,884	62.50%	1,536	16,661,043	10,847 (±2,913)
			2000	499	5,682,871	62.50%			
			2010	575	6,678,288	62.50%			
	Political Science Quarterly	Academy of Political Science	1990	242	2,400,567	37.50%	565	5,878,427	10,404 (±2,909)
			2000	218	2,326,021	37.50%			
			2010	105	1,151,839	37.50%			
Sociology	American Sociological Review	Sage Publications	1990	409	4,120,679	50.00%	998	11,923,130	11,947 (±4,314)
			2000	289	3,667,216	50.00%			
			2010	300	4,135,235	50.00%			
	Sociology	Sage Publications	1990	304	2,416,913	50.00%	1,114	9,028,347	8,104 (±1,849)
			2000	410	3,369,687	50.00%			
			2010	400	3,241,747	50.00%			
Statistics	Journal of the American Statistical Association	American Statistical Association	1990	647	4,727,507	50.00%	1,836	15,930,005	8,676 (±3,336)
			2000	640	5,727,853	50.00%			
			2010	549	5,474,645	50.00%			
	The American Statistician	Taylor & Francis	1990	368	1,329,057	50.00%	925	3,846,699	4,159 (±2,213)
			2000	355	1,551,436	50.00%			

(Continues)

TABLE 1 (Continued)

Discipline	Journal	Publisher	Decade	Number of articles	Number of tokens	Sampling proportion in discipline-decade stratum	Number of articles overall	Number of tokens overall	Mean number of tokens per article (±SD)
			2010	202	966,206	50.00%			
Theoretical computer science	Information and Computation	Elsevier	1990	244	2,425,925	50.00%	787	8,601,518	10,930 (±4,359)
			2000	347	3,885,918	50.00%			
			2010	196	2,289,675	50.00%			
	Theoretical Computer Science	Elsevier	1990	686	5,735,299	50.00%	1,836	16,490,531	8,982 (±4,138)
			2000	740	6,759,564	50.00%			
			2010	410	3,995,668	50.00%			

statistics on Google Scholar[2] and Scimago,[3] we ensured we selected only those journals that can be considered well-established and central in their respective fields.

3.1.2 | Data acquisition

The majority of articles in our corpus were gathered from JSTOR, who—upon our request—kindly provided us with the articles as plain text files that were extracted from PDF sources via optical character recognition (OCR).[4] Articles of non-humanities disciplines were mostly retrieved via the CrossRef[5] text mining service as PDF files.

Additionally, IEEE computer science articles were downloaded as PDFs via the bulk download function of IEEE Xplore,[6] *Computational Linguistics* articles via MIT Press,[7] and *LLC/DSH* articles via Oxford University Press.[8] We used GROBID[9] to convert each PDF file into XML, from which the article text was extracted. *DHQ* articles were downloaded from their official website as native XML files.[10] To balance the length of articles between disciplines, we only retained articles with a length between 1,000 and 20,000 tokens. We tried to automatically remove non-research documents, which we identified by generic titles, such as "Editorial," "Book Review," "Introduction," and "News and Notes." We also removed those articles whose authors were not explicitly recorded in the metadata.

3.1.3 | Preprocessing and data cleaning

All articles were tokenized, tagged with part-of-speech (POS) tags, lemmatized, and lowercased using spaCy (Honnibal et al., 2020). Regarding lemmatization, some manual corrections were made for a couple of domain-specific plural-only nouns (humanities, linguistics, data, media, etc.). Judging by manual inspection, the texts gathered from JSTOR contained some noise from erroneous OCR, mostly resulting from word hyphenation in original sources. To clean up these errors, we concatenated any bigram including a hyphen in those instances where the bigram occurs less frequently in the corpus than the concatenated unigram. For example, the bigram hyphenation occurs much less frequently than the unigram hyphenation; therefore each occurrence of hyphenation is replaced with hyphenation. Bigrams without hyphens present (e.g., discipline, information) were left untouched because these cannot be resolved unambiguously. We will see later that the most common word particles resulting from this circumstance (-tion, -ity, -ing, -sion, etc.) are filtered out by LDA, by being grouped into a separate topic which we manually exclude from our analysis. Other than that, the OCR noise did not have any discernible negative impact on the quality of the topics we analyzed.

3.1.4 | Feature selection and phrase concatenation

We chose to perform topic modeling exclusively on nouns, proper nouns, and noun phrases. This provides noise reduction, since all function words, which are generally not relevant for identifying topics, are eliminated. As Martin and Johnson (2015) have shown, a nouns-only approach leads to better interpretable topics while also speeding up the LDA training process. Our decision is further motivated by research on automatic terminology extraction from academic texts, which shows that technical terms referring to concepts or entities are mostly single nouns and noun phrases that adhere to certain POS patterns cf. (cf. Justeson & Katz, 1995; Manning & Schütze, 1999, pp. 153–157; Lang et al., 2018). In order to extract noun phrases, we concatenated 2-, 3-, and 4-grams that occurred at least 50 times and showed a high association value (≥0.3), measured by *normalized pointwise mutual information* (NPMI; Bouma, 2009). The candidates were then filtered by POS patterns, as originally proposed by Justeson and Katz (1995).

To further reduce noise, terms which include fewer than three alphabetic characters or any punctuation, apart from apostrophes and hyphens, as well as terms on a manually defined stop list of highly frequent words were removed.[11] Finally, we limited the vocabulary being used for the topic modeling to the 20,000 terms with the highest average tf-idf scores term frequency-inverse document frequency, as proposed by Salton & McGill, 1986).

3.2 | Sampling

As can be seen in Table 1, the number of articles per discipline, journal and decade varies greatly. To ensure that each discipline and decade has the same impact on topics and that the influence of journals does not change diachronically, each run of topic modeling is performed using a stratified sample of articles as training data.

We decided to draw a fixed number of 500 articles per discipline and decade, implying moderate random oversampling for some disciplines, and random undersampling for others. With 3 decades, 16 disciplines, and 500 articles per discipline-decade stratum, each sample from the corpus consists of 24,000 articles. Within each discipline-decade stratum, articles are further stratified by journal to keep the proportions of journals associated with any given discipline constant across the three decades. We set a fixed percentage for each journal, which define how many of the 500 articles were to be drawn from it (see Table 1 for details). This means that should we later observe diachronic trends in the occurrence of topics, we can rule out the possibility that a change in journal weighting was the cause. Per discipline, the shares are distributed equally among journals. Exceptions are made for journals with substantially fewer articles than necessary in order to avoid excessive oversampling. For these journals we manually set a lower percentage, roughly based on the number of articles present per decade. All assigned journal percentages are displayed in Table 1.[12] Another exception is the DH subcorpus, where *DHQ* was not published until 2007. Here, the journal percentages change between decades: in the 1990s, the sample is evenly split between *CHum/LRE* and *LLC/DSH*; in the 2000s, 10% are sampled from *DHQ*, 45% from *CHum/LRE*, and 45% from *LLC/DSH*; and in the 2010s, the distribution is even among the three journals.

All topic models and statistics based on probability distributions of the final aggregate topic model were calculated on the basis of corpus samples according to this scheme. For evaluation purposes, 10 articles per discipline-decade stratum (in total 480 articles) were randomly picked and fixed as testing data for perplexity calculation, as described in the following section.

3.3 | Topic modeling

Topic modeling is an unsupervised machine-learning approach for discovering latent hidden semantic structures (topics) in large collections of documents, based on patterns of word co-occurrences. For our corpus analysis, we use the LDA approach, which is one of the most popular topic modeling techniques (as proposed by Blei et al., 2003). The implementation of LDA used here is the standard model of the natural language processing toolkit MALLET (McCallum, 2002), which uses Gibbs sampling (see Steyvers & Griffiths, 2007; Yao et al., 2009) to train the model and also features a method for the optimization of asymmetric *alpha* and symmetric *beta* Dirichlet priors, based on Wallach, Mimno, and McCallum (2009).

3.3.1 | Evaluation metrics

In the context of corpus analysis, topic modeling methods have been criticized for producing non-replicable results, which are additionally heavily influenced by manually—and oftentimes arbitrarily—set parameters (Schmidt, 2012). Mohr and Bogdanov (2013) and Blei (2012) therefore point

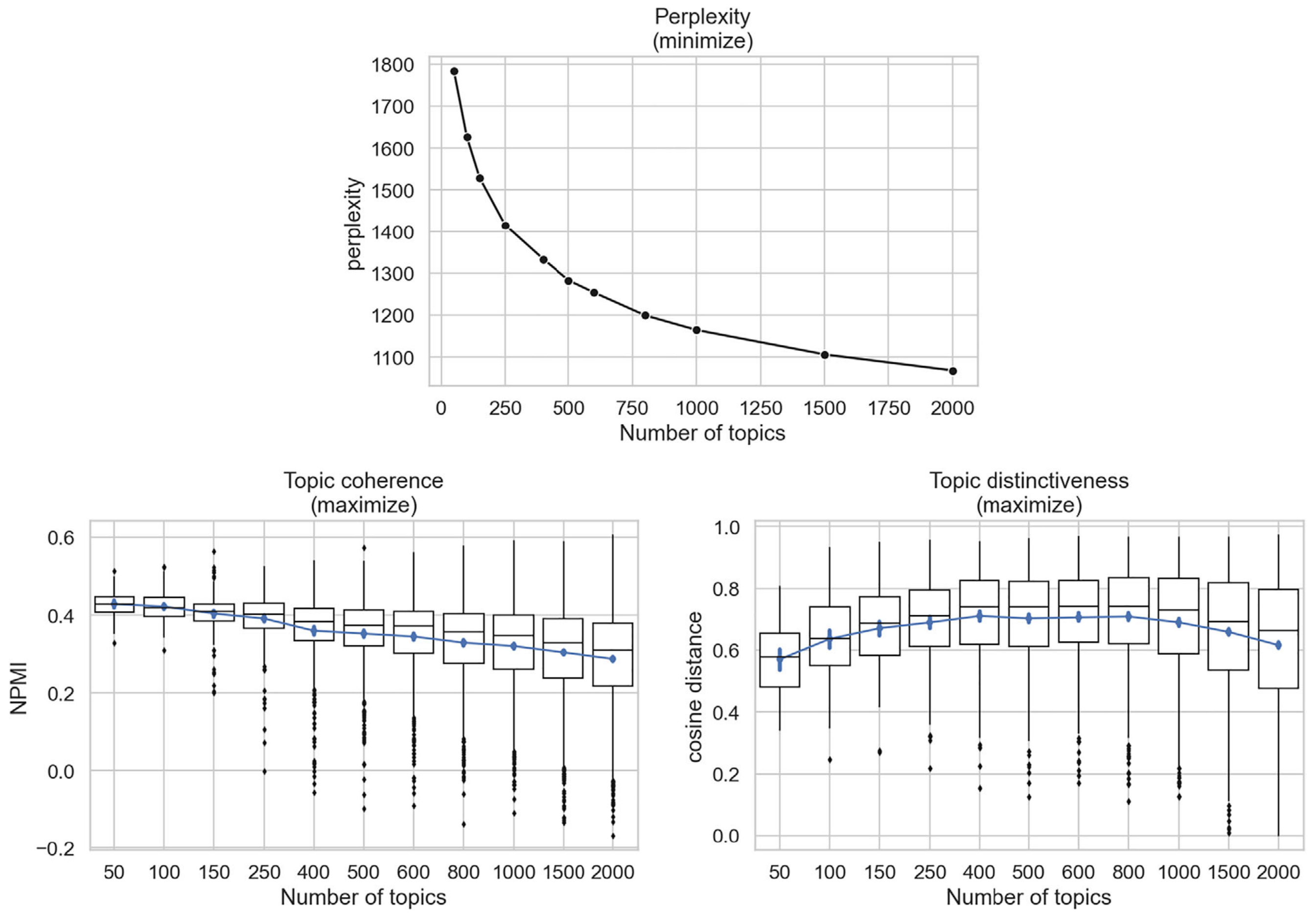


FIGURE 1 Perplexity, topic coherence, and topic distinctiveness scores for models with varying K, with blue points showing mean and standard deviation [Color figure can be viewed at wileyonlinelibrary.com]

to the increased responsibility on and required caution of the researcher in the process of modeling and interpreting the topics. We addressed these issues by evaluating models using established metrics to justify our parameter settings. The evaluation metrics we used in this study include *perplexity* (see Blei et al., 2003; Wallach, Murray, et al., 2009), *topic coherence* (see Blair et al., 2020; Lau et al., 2014) and *topic distinctiveness* (see Vega-Carrasco et al., 2020). Overall, our evaluation process is closely related to that of Vega-Carrasco et al. (2020) and Burghardt and Luhmann (2021).

3.3.2 | Estimating the number of topics

Using the evaluation measures, we examined the effect of different numbers of topics (K; in the range 50–2,000) on the resulting topic model. Each model is fitted for 1,000 iterations, which proves sufficient for the log-likelihood values to converge. In Figure 1 we can see that a reasonable trade-off between favorable evaluation scores is achieved with a setting of K in the range 400–800. After examining the resulting topic models qualitatively, we decided to use the model with $K = 800$, as it offered the most details, especially in topics that may be relevant to DH.

3.4 | Topic aggregation

The idea of ensuring the reliability of LDA results by aligning topics from multiple models was probably first mentioned by Steyvers and Griffiths (2007). In recent years, this idea has been further developed, with Blair et al. (2020), Blair et al. (2016), Rieger et al. (2020), and Vega-Carrasco et al. (2020) introducing approaches to aggregate multiple topic models using clustering techniques. We have taken up this idea of topic aggregation to increase the reliability of our model.

In the first step, 10 LDA models were obtained using the previously determined hyperparameters and differing random states, each fitted on a new corpus sample. This gave us a total of 10 times 800, or 8,000 topics. These are then clustered using hierarchical agglomerative clustering, based on the cosine distance between their term probability distributions.

3.4.1 | Hierarchical agglomerative clustering

With this clustering method, each topic is initially grouped within its own cluster. Then, the two clusters nearest to each other are merged into a new cluster, with the whole process being repeated iteratively. The distance of clusters is measured by cosine distance. Also, the average linkage method is used, where the distance between two clusters is defined by the average of the pairwise distances of all topics in the clusters. We specified several distance thresholds (0.25, 0.35, and 0.45), at which the fusion of clusters stops.

Following Vega-Carrasco et al. (2020), the resulting topic clusters were then filtered by size; that is, the number of original topics that were grouped within each cluster. The cluster size is interpreted to reflect the stability or recurrence of a topic in a similar form across multiple LDA runs. For this purpose, a threshold for the minimum cluster size between 1 (no filtering) and 10 (corresponding to the number of original models) was considered. Appropriate settings of the two parameters, distance threshold and minimum cluster size, were evaluated using the same evaluation metrics as for estimating the number of topics.

3.4.2 | Selection of clustering parameters

Figure 2 presents the results of the evaluation of the aggregate topic models. The colors of the connected points represent different distance thresholds; the x-axis indicates the threshold of minimum cluster size. The horizontal black lines mark the values of the original models from which the aggregate models were created, with the dotted lines indicating the standard deviation.

Especially in crucial areas, the differences in the values of topic coherence and topic distinctiveness are very small and might deviate if the experiment were to be repeated. We decided to set a cosine distance threshold of 0.35 and a minimum cluster size of 6 for the final aggregate topic model; however, we would like to emphasize

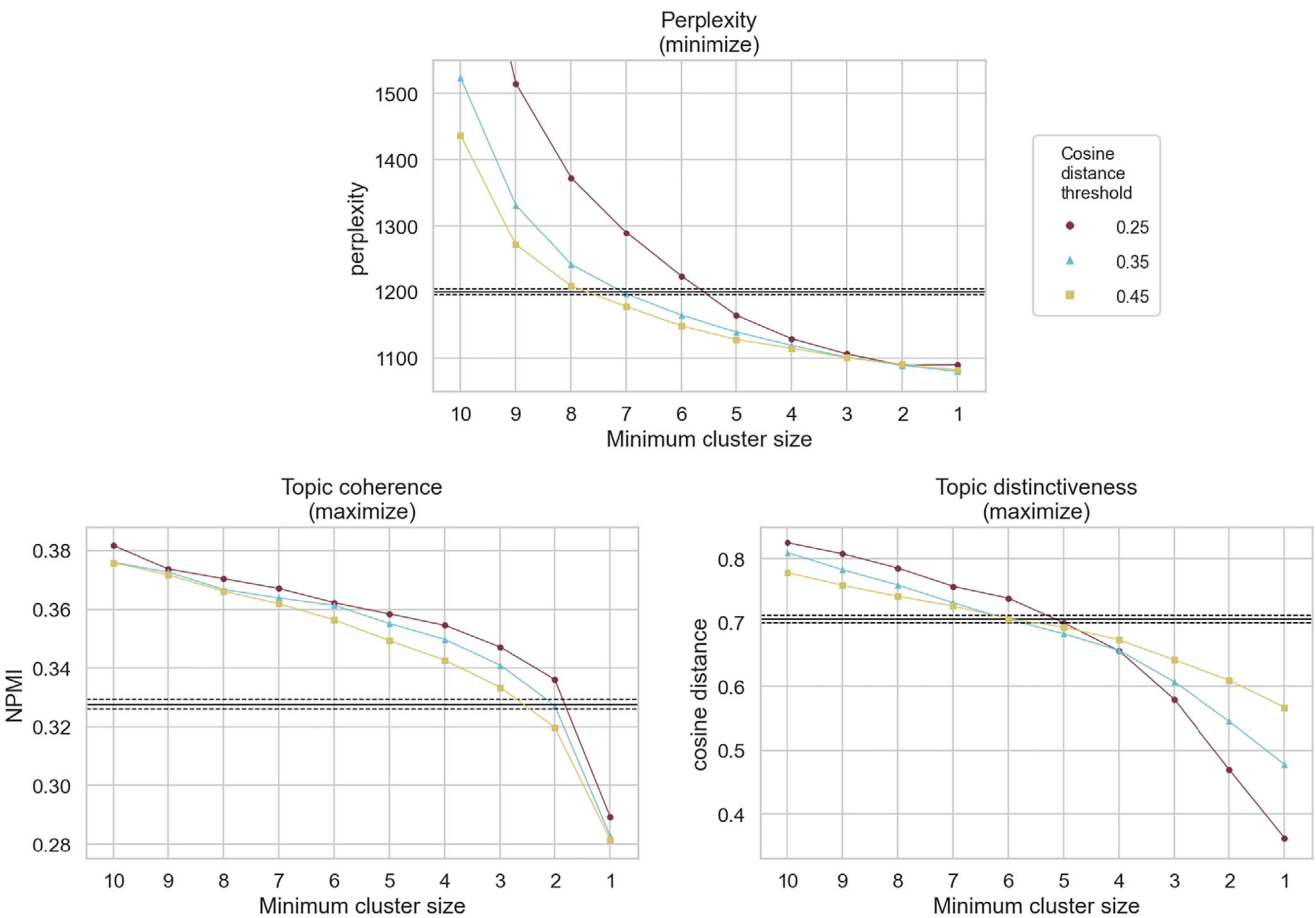


FIGURE 2 Evaluation values of aggregate topic models, using different minimum cluster sizes and cosine distance thresholds. Mean values are shown for topic coherence and topic distinctiveness. Black horizontal lines show the mean values of the original topic models, with dotted lines indicating standard deviation. Very high perplexity values have been truncated for better visibility of relevant values [Color figure can be viewed at wileyonlinelibrary.com]

that some neighboring candidates of threshold combinations would have been an equally viable choice. The final aggregated topic model based on the selected thresholds consists of 625 topic clusters, which are available online for further investigation.[13] The term probability distribution of each topic cluster was computed as the centroid of the term probabilities of the topics found in the cluster. Based on these centroids, the document topic probability distributions were re-inferred using the "iterated pseudo-count" method by Wallach, Murray, et al. (2009). Taking a closer look at the remaining topics, it becomes obvious that the topic aggregation method not only ensures higher reliability, but also filters out most low-coherence topics. We further optimized the aggregated topic model by manually excluding 10 topic clusters (see Table 2) that were obviously not related to aspects of academic discourse, but rather capture the occurrence of non-English passages, of personal names and month names, and of word particles presumably caused by OCR processing.

3.5 | Digital humanities in the academic landscape

In this section, the relationships of topics, articles, disciplines, and journals will be considered and visualized from different perspectives. Each of these approaches is reductive in that it can only focus on individual aspects of the relationships within the model. Taken together, however, they will hopefully shed some light on the status of DH in the overall landscape of academic disciplines.

3.6 | Distances of discipline centroids

First, we wanted to gain an overview of how the academic disciplines relate to each other, based on the corpus articles and the topic model. The topic probability distribution of each article can be also understood as a topical embedding or a vector representation in a latent space, consisting of 615 dimensions, corresponding to the previously defined topics. This way, the distance between the vectors of two articles can then be interpreted as the degree of similarity between the articles in terms of shared topics. To represent a group of articles—for example, grouped by discipline, journal, or decade, or a combination thereof—we computed the centroid of all vectors within this group. This centroid then also revealed the mean probability of topics for this group.

Figure 3 shows a heatmap of cosine distances between disciplines based on discipline centroids. To ensure a consistent influence of journals and decades on each centroid, these were calculated based on a stratified corpus sample, with the particularity that here we oversample so that each article occurs at least once in the sample. The ordering of disciplines follows a dendrogram obtained using average linkage clustering on cosine distances.

Judging by the dendrogram, disciplines can be clustered largely into the following three groups:

1. Math and statistics, and theoretical and applied computer science; we note that statistics appears to be rather distinct.
2. Linguistics, computational linguistics, DH, and information science; we note that among these, linguistics stands out as being more distant.
3. Humanities, arts, and social sciences; here we note an especially high similarity between sociology and political science, as well as between history, literary studies, and art history and theory.

TABLE 2 Overview of manually removed topics

Topic no.	Top terms	Reason for exclusion
54	cation nition speci ned erent schema nite rst viewpoint…	OCR noise
55	tion ing ment con ity sion cal com ture cation ter tie ence…	OCR noise
37	van dutch netherlands jan amsterdam van_der het dat een zijn…	Non-English terms
58	que french qui est paris par jean de_la une dans pour pierre…	Non-English terms
188	der den ich nicht von ist ein german mit hat sie eine sich…	Non-English terms
189	der von berlin germany zur vienna die friedrich munich…	Non-English terms
351	est non qua quod cum latin qui sed quam esse passage aut…	Non-English terms
366	che del della florence non venice rome italy giovanni con…	Non-English terms
204	march june may october september january july april november…	Months
262	david john robert michael richard paul peter james cambridge…	Personal names

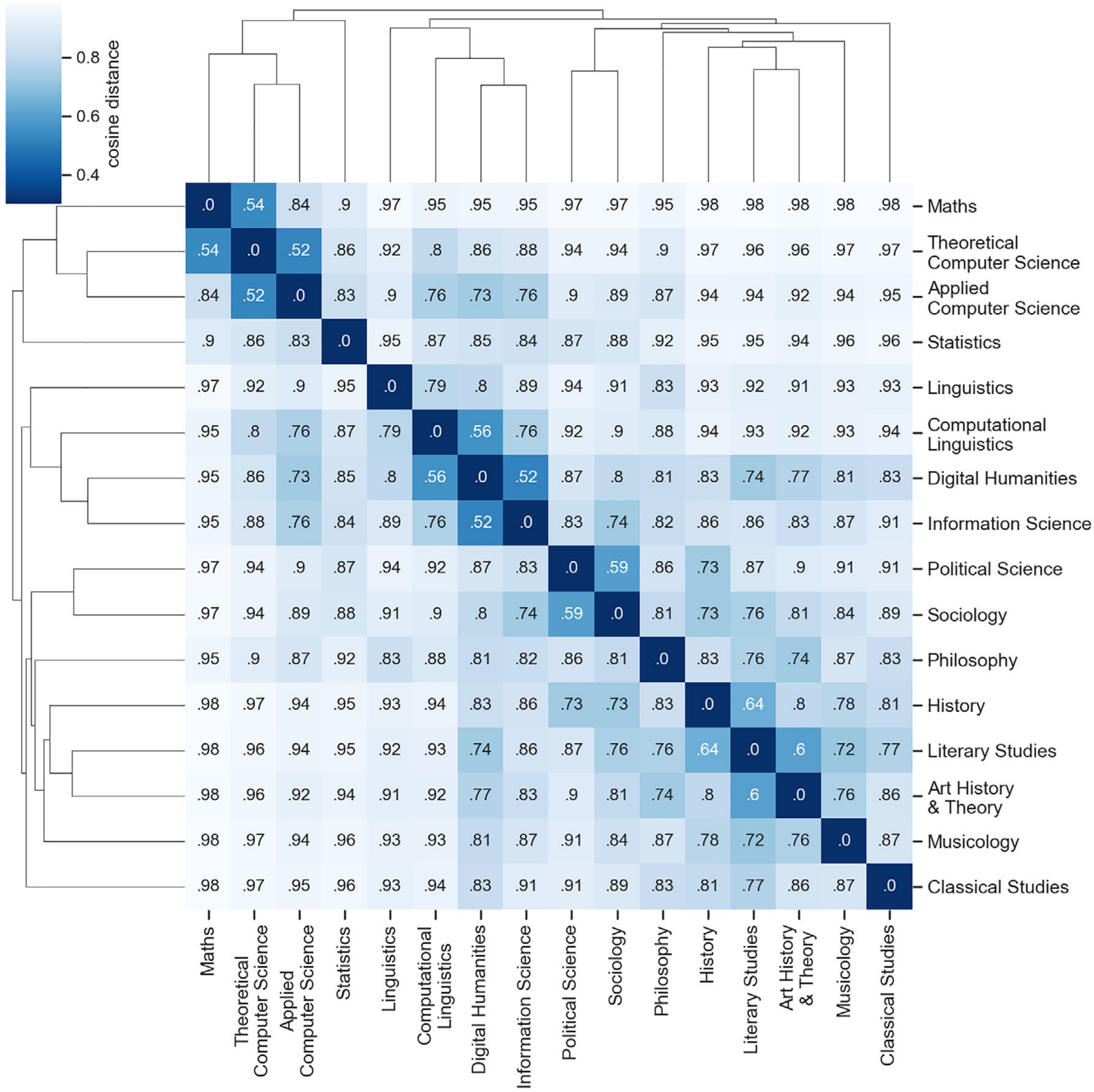


FIGURE 3 Heatmap of cosine distance between topic distribution centroids of disciplines, based on a stratified corpus sample. The dendrogram of discipline centroids uses average linkage [Color figure can be viewed at wileyonlinelibrary.com]

The overall most distinct disciplines seem to be statistics, linguistics, and philosophy. DH shows a strikingly high similarity to both computational linguistics and information science, and also moderate similarities to applied computer science, literary studies, and art history and theory. Interestingly, computational linguistics shows a much higher similarity to DH than to linguistics. The cosine distance between DH and information science matches exactly the cosine distance between theoretical and applied computer science. DH also has the highest average similarity to all other disciplines, which implies that it appears to be central to all disciplines in the latent space. However, any conclusions that can be drawn from this approach must be at a rather superficial level, as it assumes each discipline to be a static entity represented by a single vector and does not take into account the dispersion of topics among disciplines, let alone among individual journals.

3.7 | Projection of inter-article distances

A more substantial overview can be provided on the level of inter-article distances (see Figure 4). To achieve this, we used UMAP (McInnes et al., 2018) as a technique for dimension reduction. Based on the cosine distance measure, the topical vector representations of articles are projected to two-dimensional vectors. In this projection, the distances of local neighborhoods among the original vectors are approximately preserved. The UMAP technique

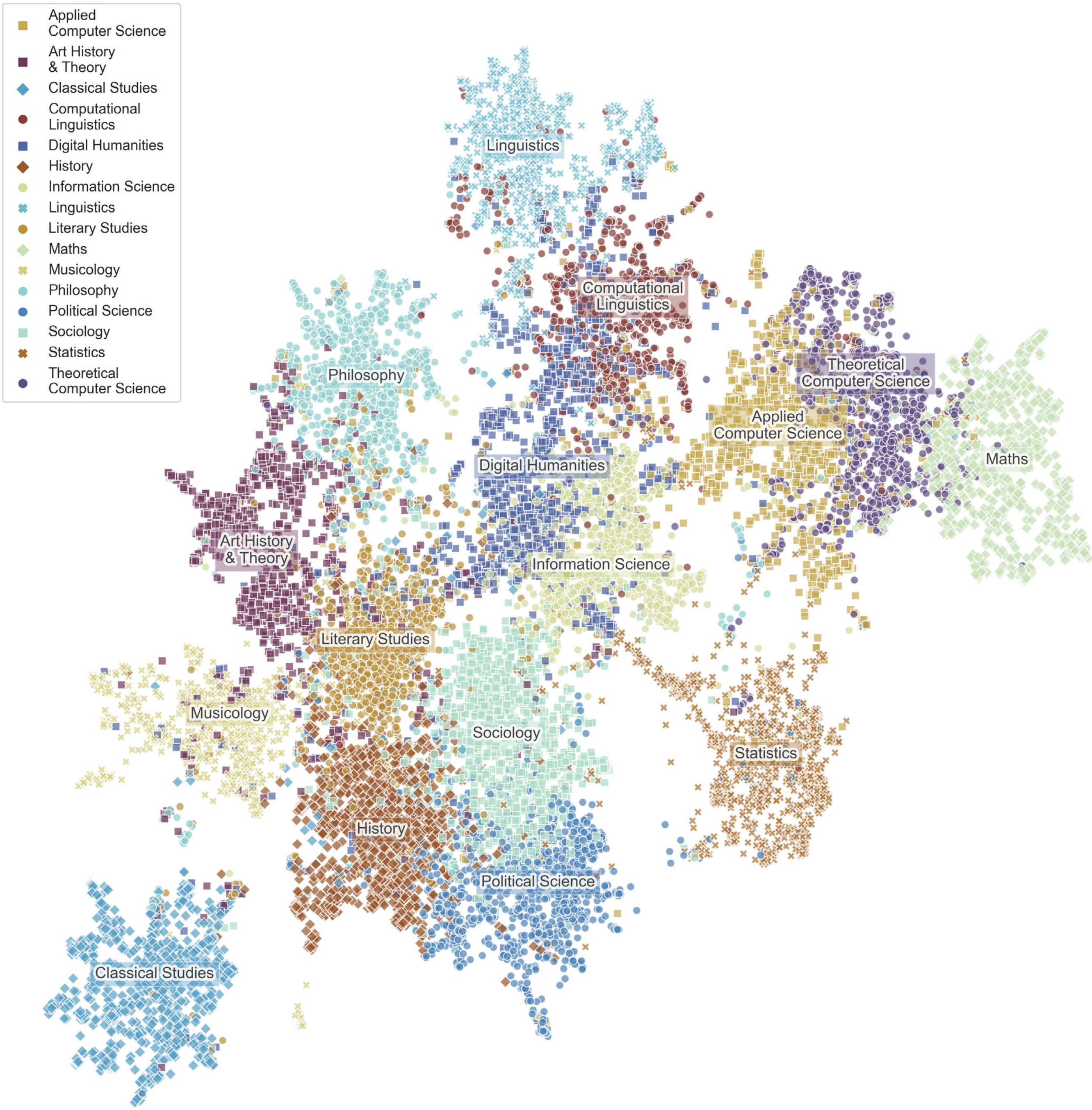


FIGURE 4 UMAP projection of a stratified sample of corpus articles, where duplicate articles (caused by oversampling) are removed. Each point represents one article, where color and shape indicate discipline and the distance between points indicates the cosine distance between articles' topic probability distributions [Color figure can be viewed at wileyonlinelibrary.com]

is non-deterministic and relies on manually set parameters that influence the outcome. The parameter *n_neighbors* controls how many nearest neighbors of each article are considered to be part of its local neighborhood. The lower the value, the more fine-grained the projection representing the underlying data structures, while a higher value allows for a more global representation. We experimented with settings in the range of 20–200, finding that the representations did not differ fundamentally. However, a higher value seemed to better capture the overall relationships between disciplines (see Figure 3) and also brought more stability among multiple UMAP runs using different random states; therefore we set *n_neighbors* to 200.

Figure 4 displays each article as a point in a latent space, with color and shape indicating its assigned discipline. The previously noted observations of clusters and

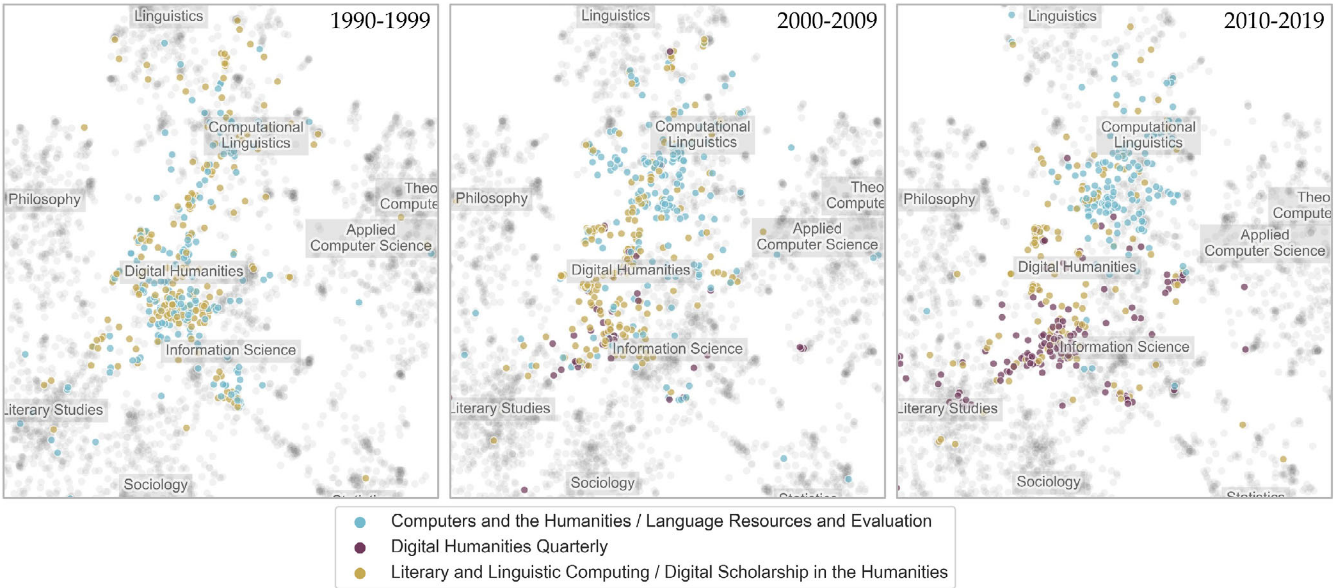
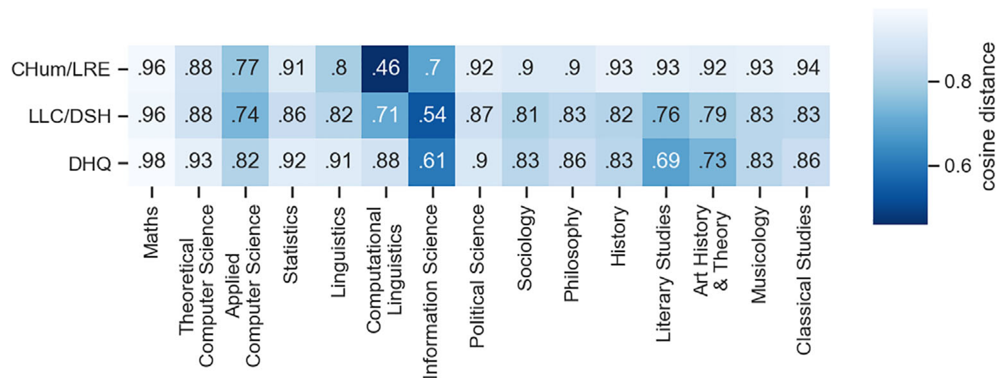


FIGURE 5 Detail of DH articles by journal and decade, using the same UMAP projection as in Figure 4 [Color figure can be viewed at wileyonlinelibrary.com]

FIGURE 6 Cosine distance of centroids of DH journals and disciplines, based on articles published between 2007 and 2019 [Color figure can be viewed at wileyonlinelibrary.com]

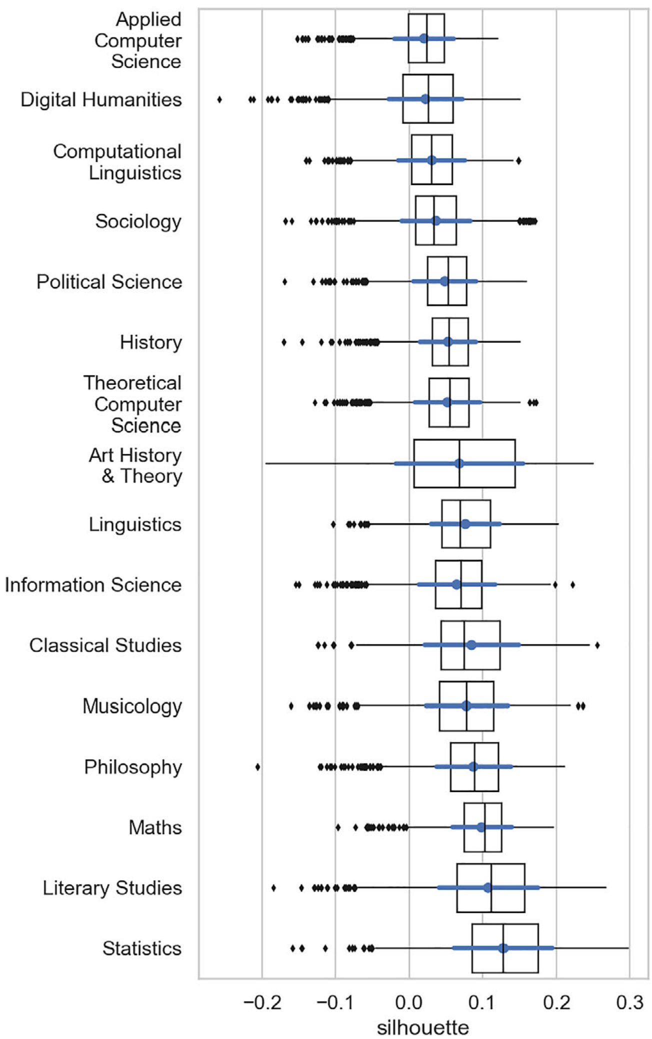


nearest neighbors between disciplines (see Figure 3) are essentially confirmed here. Interestingly, each discipline appears as a clearly demarcated group of articles in a core area, with areas of gradual transition to adjoining disciplines and few outliers in other areas of the projection. DH articles manifest themselves alongside information science articles and reach far into the areas of computational linguistics and linguistics. We can observe some kind of transitional area between DH and three other disciplines; specifically, literary studies, applied computer science, and statistics. There is a slight variation in the extent to which the individual disciplines appear more as a cohesive unit or are more scattered. DH articles seem to appear slightly more scattered than, for instance, musicology, math, or linguistics articles. However, this observation has to be taken with a word of caution, since it is based on a single UMAP representation. An approach for measuring this dispersion or scatteredness of disciplines is presented in a subsequent section, using silhouette values.

3.7.1 | Detail of digital humanities articles

Figure 5 provides more detail on DH articles in the wider academic landscape and is divided into three subplots, one for each decade of publication. All article points are recolored to gray, while points of DH articles are colored by journal. Each of these plots is created using the previously fitted UMAP model.

It becomes evident from this detail view that the space occupied by each journal varies greatly between decades. In the 1990s, *CHum/LRE* (then still named Computers and the Humanities) and *LLC/DSH* (then still named Literary and Linguistic Computing) occupy much of the same space, which can be seen as a core area of



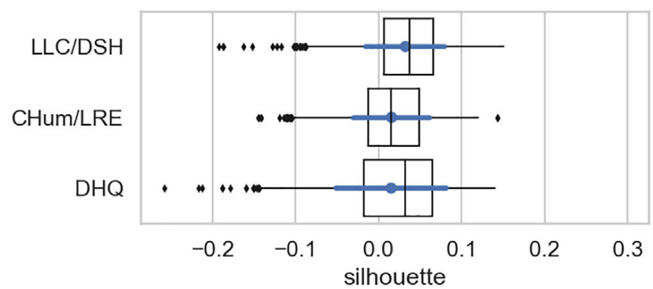


FIGURE 7 Left: Silhouette values per article, grouped by discipline; Right: Silhouette values per article, grouped by DH journals [Color figure can be viewed at wileyonlinelibrary.com]

DH at the fringes of information science, but is also sporadically scattered in the direction of computational linguistics and literary studies. In the 2000s (when both journals were renamed and *DHQ* enters the scene), we see *CHum/LRE* articles shift toward computational linguistics. The few *DHQ* articles manifest themselves primarily at a passage toward literary studies. *LLC/DSH* articles appear mostly between those two poles. In the 2010s, this divide seems to be even more distinguished, with *CHum/LRE* appearing almost exclusively in the computational linguistics area, while *DHQ* articles appear more scattered across literary studies and applied computer science, with a distinguishable cluster toward information science.

Reverting to the visualization approach shown in Figure 3, the proximity of each journal to other disciplines can also be measured by the cosine distance between the centroids of the article groups. For each of the three journals and each discipline, we compute a centroid of the mean topic probabilities based on a stratified sample. Because *DHQ* was first published in 2007, we restricted the articles considered for calculating these centroids here to a publication period of 2007–2019, to ensure a fair comparison. The heatmap in Figure 6 illustrates that *CHum/LRE* has a very close proximity to computational linguistics. What may be surprising is how distant the journal is from the traditional humanities, arts, and social sciences. *LLC/DSH* and *DHQ* each show the highest similarity to information science and are in general closer to the humanities; for the case of *DHQ* this is particularly true for literary studies and art history and theory. Of all three DH journals, *LLC/DSH* shows the highest similarity to applied computer science. Strikingly, *DHQ* appears very distant to computational linguistics. Overall, Figure 6 confirms the observations from Figure 5, which suggests that because the journals have

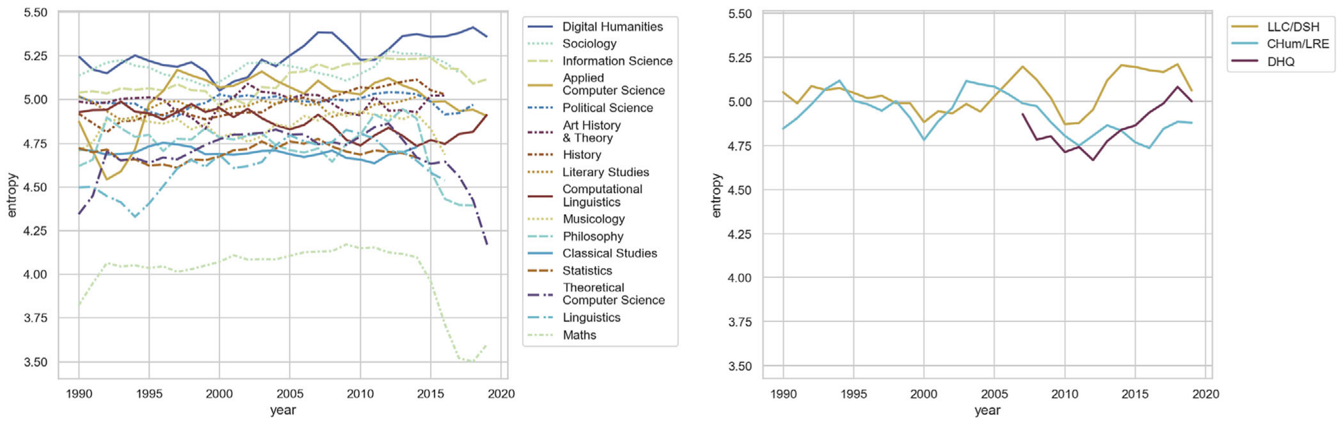


FIGURE 8 Left: Topic entropy per discipline and year (based on a stratified corpus sample); Right: Topic entropy per DH journal and year [Color figure can be viewed at wileyonlinelibrary.com]

developed very differently from 1990 to 2019, the differentiation of topics by journal is advisable.

3.8 | Silhouette

In this section we will introduce silhouette values (Rousseeuw, 1987) as a means to reliably measure the cohesiveness and dispersion of disciplines. Silhouette values are typically used for evaluating clustering approaches. For our use case, silhouettes are intended to work in the following way. If, for each article, we assume its assigned discipline to be its cluster label, the silhouette measures how well an article fits with the other articles in its own discipline compared with how well it would fit if it were assigned to another discipline nearest to it. The silhouette value ranges from −1 to 1. A negative value indicates a higher similarity of an article to articles of another discipline than to those of its own discipline; that is, it appears misplaced within its discipline. A positive value indicates a higher similarity to the articles of its own discipline than to any other; that is, it fits well within its discipline.

The left of Figure 7 shows that some of the lowest silhouette values are actually to be found among DH articles. Regarding mean and median values, DH is one of the lowest scoring disciplines, together with applied computer science, computational linguistics, sociology, political science, history, and theoretical computer science. The low silhouette values of these disciplines indicate substantial overlap of many of their articles (also see Figures 3 and 4). Broken down by journal (see the right side of Figure 7), it becomes apparent that some of the lowest silhouette values of DH are to be found among *DHQ* articles, while on average *LLC/DSH* articles score highest, indicating that this journal publishes more articles that describe topics characteristic of DH.

3.9 | Topic entropy

Another measure that seems apt to capture the dispersion of disciplines is the conditional topic entropy, originally proposed by D. Hall et al. (2008) to measure the "diversity of ideas" in computational linguistics conference proceedings. In short, it can be defined as the Shannon entropy (Shannon, 1948) computed on topic probabilities. In this study, we measured entropy values for the topic probability distribution of each discipline in each year, based on the stratified corpus samples. To ensure a reliable result, we took mean values of 10 iterations of calculations with resampling in each run.

The left side of Figure 8 presents the entropy values per discipline over time. Of all disciplines, DH (solid blue line) achieves the highest values, indicating a high diversity of topics. The right side of Figure 8 shows topic entropy per DH journal and suggests that the increase in overall topic entropy can be attributed to a slight increase of entropy in *LLC/DSH* articles, but mainly to the presence of a third journal: *DHQ*. The significantly lowest topic entropy is achieved by math, which shows the limits of our model's ability to capture discursive complexity. We would argue that these values certainly do not reflect a low topical diversity, but rather the nature in which ideas and concepts are communicated in math articles, which may be in a more formalized language, with a less diverse vocabulary, and in particular non-verbally, via mathematical formulas.

3.10 | Diagnostic topics

Thus far we have considered the topic model mostly as a numerical representation of our corpus, without dealing with the semantic quality of individual topics.

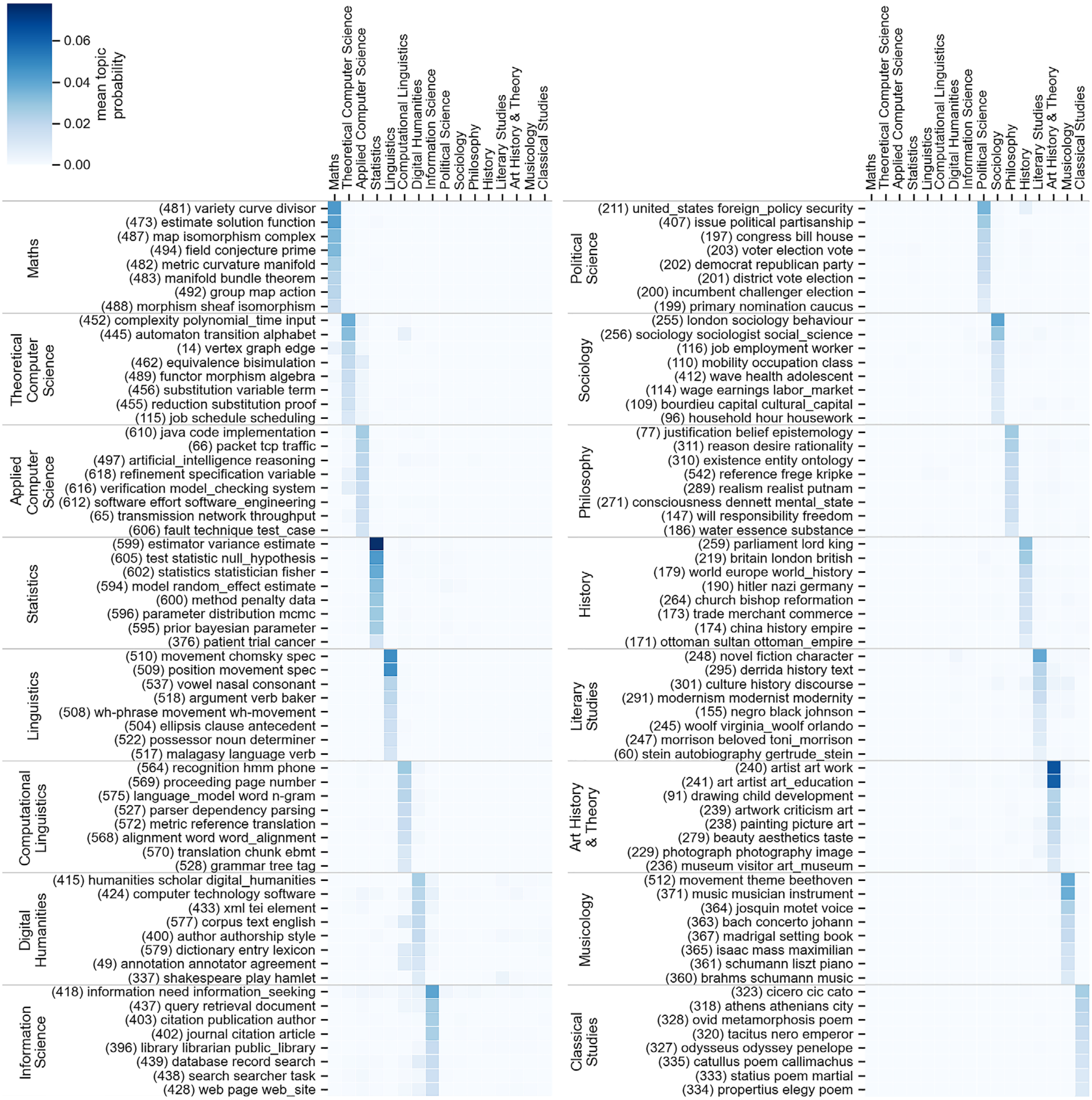


FIGURE 9 Eight diagnostic topics per discipline, where color indicates the mean probability for each discipline; disciplines are ordered according to Figure 3, topics are ordered by mean probability for each discipline [Color figure can be viewed at wileyonlinelibrary.com]

In a first step, we now want to identify the most characteristic and distinguishing topics for each discipline; that is, those topics that mainly caused the disciplines to manifest themselves as more or less cohesive clusters in Figure 4. Griffiths and Steyvers (2004) refer to these most characteristic topics of a given group of articles as *diagnostic topics*. They define the corresponding weighting as the ratio of the mean topic probability for articles of a relevant group to the sum of mean topic probabilities across other groups. If we consider disciplines as groups, the diagnostic weighting of topics is calculated by dividing the topic probability centroid of a relevant discipline by the sum of centroids of all other disciplines.

For Figure 9 the eight most diagnostic topics have been selected per discipline. The heatmap colors indicate the mean probability of a topic for each discipline. Disciplines are ordered by their linkage retrieved for Figure 3. Topics are binned by the discipline for which they have been picked as diagnostic and are ordered within those discipline bins by their mean topic probability.

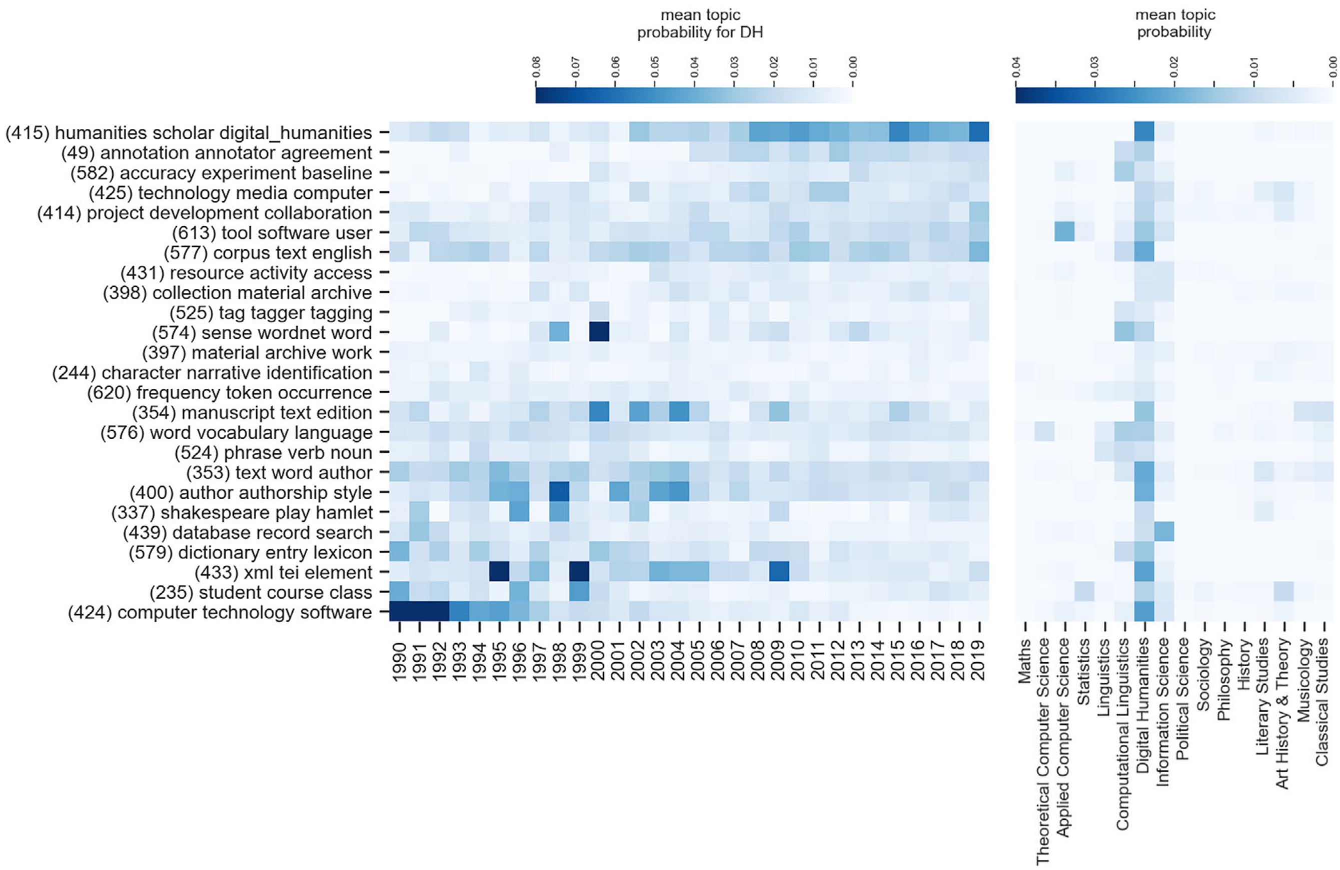


FIGURE 10 Left: 25 most-diagnostic DH topics and their probability for each year, order of topics by slope of regression line; Right: Mean probability of the same topics for each discipline [Color figure can be viewed at wileyonlinelibrary.com]

FIGURE 11 Top 20 terms of topic 424

```
computer technology software machine computing program system

application processing information ibm user tool technique

development cd-rom device hardware screen problem
```

Each topic is displayed as its topic number (defined by the hierarchical agglomerative clustering linkage during topic aggregation) and its three most probable terms.

From the distribution of probabilities across disciplines, it appears that most diagnostic topics are nearly unique to their respective disciplines. The most evident overlaps are observed in the diagnostic DH topics 577, 579, and 49, which also have a high probability among computational linguistics articles. Some redundancies are also present; for instance, topics 509 and 510 (linguistics), which both seem related to syntactic movement, or topics 402 and 403 (information science), which both seem related to citation analysis. In the diagnostic topics of literary studies, musicology, and classical studies, we noticed a high proportion of personal names of authors, philosophers, composers, and historical figures among the topic terms, while the topics among other disciplines seem to be dominated mostly by terminological vocabulary. Three disciplines have topics exhibited that appear to reflect self-referential discourse, including statistics with topic 602, sociology with topics 255 and 256, and DH with topic 415.

A more in-depth look at DH is provided in Figure 10. The displayed topics are the 25 most-diagnostic topics of DH, with their probability per year shown diachronically. Following Griffiths and Steyvers (2004), we calculated a linear regression line on these per-year probabilities. The topics are ordered by the slope of their respective regression lines, with the topics at the top exhibiting the highest diachronic increase, and the topics at the bottom humanities scholar digital_humanities discipline tool research

scholarship project humanities_computing humanist work field digital

computing method study technology visualization archive text

FIGURE 12 Top 20 terms of topic 415

the highest diachronic decrease. On the right, for each topic we see their mean probability for each discipline over all time periods, presented in the same way as in Figure 9.

Figure 10 (left) reveals that topic 424 is the most declining topic. While it is the most prominent topic for 1990 to 1992, it has almost disappeared by 2013. It is also rather unique to DH, judging by the very low probabilities among other disciplines. From its top 20 terms (see Figure 11), we see that this topic represents a broad reflective discourse on the emergence of (personal) computers and their usage. Interestingly, this topic is almost non-existent in computer science articles.

At the other end of the spectrum, topic 415 is found to have the highest diachronic increase in probability, with particularly high probabilities in the 2008–2019 period, where (except for 2009) it appears as the most prominent topic. This topic represents a reflective discourse on DH as an academic field and the use of computing in humanities in general (see Figure 12).

Furthermore, the topics about text encoding (433), authorship and stylometry (400), and word sense disambiguation (574) are the ones that appear most prominently at isolated points in time. These sporadic appearances can be traced back to journals' thematic special issues. The diachronically most consistent topics seem to be about scholarly collaboration (414), tool usage (613), text corpora (577), and text analysis (353, 576). Regarding the uniqueness of topics for DH, about half of the displayed topics show equal or even increased probabilities to other disciplines. The topics about training in machine learning (582) and about word sense disambiguation (574) show a higher probability toward computational linguistics, while the topic about tool usage (613) indicates a higher probability toward applied computer science. The topic about databases (439) shows a higher probability toward information science than to DH. Surprisingly, the latter topic does not even appear in articles of applied computer science.

All in all, the topics that appear to be most unique for DH are topics about DH itself (415), annotation (49), (scholarly) collaboration (414), text corpora (577), text editions (354), text analysis (353), authorship and stylometry (400), dictionaries (579), text encoding (433), and the emergence of computing (424).

4 | CONCLUSION AND FUTURE DIRECTIONS

In this study we have used an aggregated topic model to analyze a large corpus of academic journals. In doing so, we were able to sketch out an academic landscape of existing disciplines, where shared topics are a basic means to assess relations between different disciplines. This experimental setup enabled us to investigate the role of DH, as we were able to observe how the articles of three established DH journals are distributed across the disciplinary map. By using a corpus that spans three decades of publications we were also able to identify diachronic patterns and trends. While our outline of the statistical measures and visualizations used above has explained some of the most prominent patterns in our data, we will add some higher-level conclusions at this point.

The overall research question underlying this study is concerned with the status of DH—is it a discipline in its own right (Schreibman et al., 2004) or is it rather a cross-disciplinary endeavor that brings digital information technology to existing humanities disciplines (McCarty, 2015)? Our data shows that actually both seem to be the case (see Figure 4). DH is just as clearly demarcated as its own cluster as are other disciplines. At the same time, the low silhouette values of DH and its central position in the academic map (reflected by the high average similarity to all other disciplines) indicates a considerable interdisciplinary orientation.

A distinguished proximity can be found to the fields of computational linguistics and information science (see Figure 3). In the case of computational linguistics, this is presumably due to a shared canon of methods in the analysis of textual data, which continue to be the predominant object of study in DH. The relationship between DH and information science has been the subject of lively discussion before, ranging from kinder conceptions of a joint future (Robinson et al., 2015) to a more hostile atmosphere (Gladney, 2012), where the more established information science rejects DH as an intruder into the traditional realm. We refer to Burghardt and Luhmann (2021) here, where the relationship between DH and information science is discussed in more detail. They show that although DH and information science have a lot of methodological overlap there are also many uncontested areas, which can be assigned relatively clearly to the respective disciplines.

Despite its proximity to computational linguistics and information science, it is interesting to note that DH is only loosely connected to applied computer science. This observation could be interpreted as an underpinning of the characterization of the more traditional "humanities computing" and the more recent "computational humanities," who both claim relationships to computer science. Another observation to be made here is that DH and information science have the exact same cosine distance as can be found between applied computer science and theoretical computer science (see Figure 3). Taking a look at the UMAP projection (see Figure 4) it even looks like they form two larger clusters of their own. As for the case of applied and theoretical computer science, this is to be expected, as they are widely considered two parts of the same discipline, which is computer science. By analogy, DH and information science could likewise be seen as two halves of a superordinate discipline that has yet to be named.

While we were able to gain some data-driven insights into the disciplinary status of DH, we also want to highlight some of the methodological and conceptual limitations of our approach. Although we have selected a comprehensive corpus of journals, it has to be assumed that the genre of journal articles is biased in a certain way when it comes to objectively representing the scholarly communication of different disciplines. As for the case of DH, much of the academic discourse takes place at conferences, most notably the international DH conference, which is an annual event with several hundred contributions each year. One could argue that the barrier to publish research as an extended abstract or a short paper at one of the regular DH conferences is much lower than to publish it as a full-blown journal article, resulting in a more diverse scholarly communication in DH. In addition, journals typically give a certain thematic direction, thus further limiting the diversity of research (see Figure 5). In a follow-up study we want to investigate in more detail how the actual topics are distributed among the three DH journals, but also among different DH conferences.[14] As one of the reviewers of this paper suggested, the author names of DH abstracts could also be used as seeds to identify DH publications in journals that are not explicitly branded as DH (for instance *PLos One* or *JASIST*) and thus create a "diasporic corpus" of DH. We also plan a follow-up study in which we accompany the topic modeling approach with further methods, such as co-citation analysis (as used in Gao et al., 2017), to investigate how well the method of topic modeling is actually suited to determine disciplinary boundaries of academic fields. Finally, a major limitation of our study can be seen in its focus on DH as an evolving discipline, while the other disciplines are treated as static entities. It would be worthwhile to investigate the dynamics of disciplines such as computational linguistics and information science to see how their relationship to DH develops through the course of time.

ACKNOWLEDGEMENTS

Our great thanks go to the reviewers of this article, whose comments were extremely helpful in improving the text and who also gave us exciting suggestions for follow-up studies. Open access funding enabled and organized by Projekt DEAL.

ORCID

Jan Luhmann https://orcid.org/0000-0002-7099-4659
Manuel Burghardt https://orcid.org/0000-0003-1354-9089

ENDNOTES

1. *Big tent DH* was coined as the official theme of the international DH 2011 conference at Stanford University. https://dh2011.stanford.edu/. Accessed April 30, 2021.
2. https://scholar.google.de/citations?view_op=top_venues. Accessed April 30, 2021.
3. https://www.scimagojr.com/. Accessed April 30, 2021.
4. For technical details of JSTOR's processing, see https://www.jstor.org/dfr/about/technical-specifications. Accessed April 30, 2021.
5. https://www.crossref.org/education/retrieve-metadata/rest-api/text-and-data-mining-for-researchers/. Accessed April 30, 2021.
6. Retrieved from https://ieeexplore.ieee.org/
7. https://mitpressjournals.org/loi/coli/. Accessed April 30, 2021.
8. https://academic.oup.com/journals/pages/help/third_party_data_mining/. Accessed April 30, 2021.
9. https://github.com/kermitt2/grobid. Accessed April 30, 2021.
10. https://digitalhumanities.org/dhq/. Accessed April 30, 2021.
11. This stop list includes: way, example, case, part, result, use, time, one, point, introduction, section, note, state, table, tab, figure, fig, kind, other hand, same time.
12. Although they have a large number of articles, the journals *Artificial Intelligence* (applied computer science) and *Machine Translation* (computational linguistics) are also assigned a low percentage, because they are thematically very specific and therefore, in our view, should not constitute too large of a part of the disciplines' samples.
13. https://doi.org/10.5281/zenodo.4728006
14. For an overview of DH conferences see "The Index of Digital Humanities Conferences," which documents almost 500 conferences for a time span of 60 years. https://dh-abstracts.library.cmu.edu/. Accessed April 30, 2021.

REFERENCES

Adams, J. L., & Gunn, K. B. (2013). Keeping up with … digital humanities. Keeping Up With…. http://www.ala.org/acrl/publications/keeping_up_with/digital_humanities. Accessed April 12, 2021.

Alvarado, R. C. (2012). The digital humanities situation. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 50–55). University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0005

Berry, D. M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, *12*, 1–22.

Blair, S. J., Bi, Y., & Mulvenna, M. D. (2016). Increasing topic coherence by aggregating topic models. In F. Lehner & N. Fteimi (Eds.), *Knowledge science, engineering and management—KSEM 2016* (pp. 69–81). Springer. https://doi.org/10.1007/978-3-319-47650-6_6

Blair, S. J., Bi, Y., & Mulvenna, M. D. (2020). Aggregated topic models for increasing social media topic coherence. *Applied Intelligence*, *50*(1), 138–156. https://doi.org/10.1007/s10489-019-01438-z

Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, *2*(1), 8–11.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, R. E. de Castilho, & M. Stede (Eds.), *Proceedings of GSCL* (pp. 31–40). Gunter Narr Verlag.

Brennan, T. (2017). The digital-humanities bust. *The Chronicle of Higher Education*, *64*(8). https://www.chronicle.com/article/the-digital-humanities-bust Accessed April 12, 2021.

Burghardt, M. (2020). Theorie und digital humanities – eine bestandsaufnahme. AG digital humanities theorie. https://dhtheorien.hypotheses.org/680. Accessed April 12, 2021

Burghardt, M., & Luhmann, J. (2021). Same same, but different? On the relation of information science and the digital humanities. A scientometric comparison of Academic Journals using LDA and hierarchical clustering. In T. Schmidt & C. Wolff (Eds.), *Information between data and knowledge. Information science and its neighbors from data science to digital humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)—Regensburg, Germany, 8–10 March 2021* (pp. 173–199). Werner Hülsbusch.

Callaway, E., Stone, H., Turner, J., & Halstrom, A. (2020). The push and pull of digital humanities: Topic modeling the "what is digital humanities?" genre. *Digital Humanities Quarterly*, *14*(1).

Cordell, R. (2016). How not to teach digital humanities. In M. K. Gold & L. F. Klein (Eds.), *Debates in the digital humanities 2016* (pp. 459–474). University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb.39

Eichmann-Kalwara, N., Jorgensen, J., & Weingart, S. (2018). Representation at digital humanities conferences (2000–2015). In E. Losh & J. Wernimont (Eds.), *Bodies of information: Intersectional feminism and digital humanities* (pp. 72–92). University of Minnesota Press. https://doi.org/10.5749/j.ctv9hj9r9.9

Fish, S. (2018). Stop trying to sell the humanities. *The Chronicle of Higher Education*, *64*(36). https://www.chronicle.com/article/stop-trying-to-sell-the-humanities Accessed April 12, 2021.

Gao, J., Duke-Williams, O., Mahony, S., Bold, M. R., & Nyhan, J. (2017). The intellectual structure of digital humanities: An author co-citation analysis. In R. Lewis, C. Raynor, D. Forest, M. Sinatra, & S. Sinclair (Eds.), *12th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2017. Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO).

Gladney, H. M. (2012). Long-term digital preservation: A digital humanities topic? *Historical Social Research/Historische Sozialforschung*, *37*(3), 201–217. https://doi.org/10.12759/hsr.37.2012.3.201-217

Gold, M. K. (2012). *Debates in the digital humanities*. University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.001.0001

Gold, M. K., & Klein, L. F. (2016). *Debates in the digital humanities 2016*. University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb

Goldstone, A., & Underwood, T. (2012). What can topic models of pmla teach us about the history of literary scholarship. *Journal of Digital Humanities*, *2*(1), 39–48.

Goldstone, A., & Underwood, T. (2014). The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, *45*(3), 359–384. https://doi.org/10.1353/nlh.2014.0025

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*, 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 363–371). Association for Computational Linguistics.

Hall, G. (2012). There are no digital humanities. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 133–136). University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0016

Holm, P., Jarrick, A., & Scott, D. (2015). *The digital humanities. Humanities world report 2015* (pp. 64–83). Springer. https://doi.org/10.1057/9781137500281_4

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in python*. Zenodo. https://doi.org/10.5281/zenodo.1212303

Justeson, J., & Katz, S. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27. https://doi.org/10.1017/S1351324900000048

Kirschenbaum, M. G. (2013). What is digital humanities and what's it doing in english departments? In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining digital humanities: A reader* (pp. 211–220). Ashgate Publishing. https://doi.org/10.5749/minnesota/9780816677948.003.0001

Klein, J. T. (2015). *Interdisciplining digital humanities: Boundary work in an emerging field*. University of Michigan Press. https://doi.org/10.3998/dh.12869322.0001.001

Klein, L. F., & Gold, M. K. (2019). *Debates in the digital humanities 2019*. University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk

Lang, C., Schneider, R., & Suchowolec, K. (2018). Extracting specialized terminology from linguistic corpora. In E. Fuß, M. Konopka, B. Trawiński, & U. H. Waßner (Eds.), *Grammar and corpora 2016* (pp. 425–434). Heidelberg University Publishing. https://doi.org/10.17885/heiup.361.509

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In S. Wintner, S. Goldwater, & S. Riezler (Eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530–539). Association for Computational Linguistics.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martin, F., & Johnson, M. (2015). More efficient topic modelling through a noun only approach. In B. Hachey & K. Webster (Eds.), *Proceedings of the Australasian Language Technology Association Workshop, ALTA 2015, December 8–9, 2015, Parramatta, Australia* (pp. 111–115). Association for Computational Linguistics.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu. Accessed April 12, 2021.

McCarty, W. (2015). Becoming interdisciplinary. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A new companion to digital humanities* (pp. 69–83). Wiley. https://doi.org/10.1002/9781118680605.ch5

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, *41*(6), 607–625. https://doi.org/10.1016/j.poetic.2013.06.004

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

Meeks, E. (2011). Comprehending the digital humanities. Digital Humanities Specialist. https://dhs.stanford.edu/comprehending-the-digital-humanities/. Accessed April 12, 2021.

Meeks, E., & Weingart, S. (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, *2*(1), 1–6.

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, *41*(6), 545–569. https://doi.org/10.1016/j.poetic.2013.10.001

Pannapacker, W. (2012). Digital humanities triumphant? In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 233–234). University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0022

Paul, M. J., & Girju, R. (2009). Topic modeling of research fields: An interdisciplinary perspective. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, & N. Nikolov (Eds.), *Recent advances in natural language processing, RANLP 2009, 14–16 September, 2009, Borovets, Bulgaria* (pp. 337–342). Association for Computational Linguistics.

Poole, A. H. (2017). The conceptual ecology of digital humanities. *Journal of Documentation*, *73*, 91–122. https://doi.org/10.1108/JD-05-2016-0065

Puschmann, C., & Bastos, M. (2015). How digital are the digital humanities? An analysis of two scholarly blogging platforms. *PLoS One*, *10*(2), 1–15. https://doi.org/10.1371/journal.pone.0115035

Riddell, A. B. (2014). *How to read 22,198 journal articles: Studying the history of german studies with topic models. Distant readings: Topologies of German culture in the long nineteenth century* (pp. 91–114). Boydell & Brewer.

Rieger, J., Koppers, L., Jentsch, C., & Rahnenführer, J. (2020). Improving reliability of latent dirichlet allocation by assessing its stability using clustering techniques on replicated runs. *arXiv Preprint*.

Robertson, S. (2016). The differences between digital history and digital humanities. In M. K. Gold & L. F. Klein (Eds.), *Debates in the digital humanities 2016*. University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb.28

Robinson, L., Priego, E., & Bawden, D. (2015). Library and information science and digital humanities: Two disciplines, joint future? In F. Pehar, C. Schlögl, & C. Wolff (Eds.), *Re:Inventing information science in the networked society. Proceedings of the 14th International Symposium on Information Science, ISI 2015, May 19–21, Zadar, Croatia* (pp. 44–54). Werner Hülsbusch. https://doi.org/10.5281/zenodo.17969

Roth, C. (2019). Digital, digitized, and numerical humanities. *Digital Scholarship in the Humanities*, *34*(3), 616–632. https://doi.org/10.1093/llc/fqy057

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc. https://doi.org/10.5555/576628

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, *2*(1), 49–65.

Schöch, C. (2017). Topic modeling genre: An exploration of French classical and enlightenment drama. *Digital Humanities Quarterly*, *11*(2).

Schreibman, S., Siemens, R., & Unsworth, J. (2004). The digital humanities and humanities computing: An introduction. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities*. Blackwell. https://doi.org/10.1002/9780470999875

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 427–448). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203936399

Sula, C. A. (2013). Digital humanities and libraries: A conceptual model. *Journal of Library Administration*, *53*(1), 10–26. https://doi.org/10.1080/01930826.2013.756680

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*, *34*, i190–i206. https://doi.org/10.1093/llc/fqz072

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

Tang, M.-C., Cheng, Y. J., & Chen, K. H. (2017). A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses. *Scientometrics*, *113*(2), 985–1008. https://doi.org/10.1007/s11192-017-2496-6

Terras, M. (2013). Peering inside the big tent. In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining digital humanities: A reader* (pp. 263–270). Ashgate Publishing.

Terras, M., Nyhan, J., & Vanhoutte, E. (Eds.). (2013). *Defining digital humanities: A reader*. Ashgate Publishing. https://doi.org/10.4324/9781315576251

Vega-Carrasco, M., O'Sullivan, J., Prior, R., Manolopoulou, I., & Musolesi, M. (2020). Modelling grocery retail topic distributions: Evaluation, interpretability and stability. *arXiv Preprint*.

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22: 23rd Annual Conference on Neural Information Processing Systems. Proceedings of a meeting held 7–10 December 2009, Vancouver, British Columbia, Canada* (pp. 1973–1981). Curran Associates, Inc.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. M. (2009, 2009). Evaluation methods for topic models. In A. P. Danyluk, L. Bottou, & M. L. Littman (Eds.), *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14–18* (pp. 1105–1112). ACM. https://doi.org/10.1145/1553374.1553515

Weingart, S. (2016). Submissions to DH2017. http://scottbot.net/submissions-to-dh2017-pt-1/. Accessed April 12, 2021.

Weingart, S., & Eichmann-Kalwara, N. (2017). What's under the big tent?: A study of ADHO conference abstracts. *Digital Studies/Le Champ numérique*, *7*(1), 6. https://doi.org/10.16995/dscn.284

Yao, L., Mimno, D. M., & McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. In J. F. Elder, IV, F. Fogelman-Soulié, P. A. Flach, & M. J. Zaki (Eds.), *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28–July 1, 2009* (pp. 937–946). ACM. https://doi.org/10.1145/1557019.1557121

How to cite this article: Luhmann, J., & Burghardt, M. (2021). Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology*, 1–24. https://doi.org/10.1002/asi.24533