11-26-2021

Exploring the Digital Humanities Research Agenda: A Text Mining Approach

Soohyung Joo
University of Kentucky, soohyung.joo@uky.edu

Jennifer Hootman
University of Kentucky, jlhootman@uky.edu

Marie Katsurai
Doshisha University, Japan

Exploring the Digital Humanities Research Agenda: A Text Mining Approach

Digital Object Identifier (DOI)
https://doi.org/10.1108/JD-03-2021-0066

Notes/Citation Information
Published in *Journal of Documentation*.

Exploring the Digital Humanities Research Agenda: A Text Mining Approach

Abstract

Purpose: This study aims to explore knowledge structure and research trends in the domain of digital humanities in the recent decade. The study identified prevailing topics, and then, analyzed trends of such topics over time in the digital humanities field.

Design/methodology/approach: Research bibliographic data in the area of digital humanities were collected from scholarly databases. Multiple text mining techniques were employed to identify prevailing research topics and trends, such as keyword co-occurrences, bigram analysis, structural topic models and biterm topic models.

Findings: Term-level analysis revealed that cultural heritage, geographic information, semantic web, linked data, and digital media were among the most popular topics in the recent decade. Structural topic models identified that linked open data, text mining, semantic web and ontology, text digitization, and social network analysis received increased attention in the digital humanities field.

Originality: This study applied existent text mining techniques to understand the research domain in DH. The study collected a large set of bibliographic text, representing the area of DH from multiple academic databases, and explored research trends based on structural topic models.

Introduction and Background

Despite the long-contested and quickly evolving nature of what is now commonly referred to as digital humanities (DH), most Humanists agree upon an early pioneer in humanities computing, the Italian Jesuit priest, Father Roberto Busa. Father Busa's work in applying computing to his humanist objective makes for as good of an origin story as any. In 1949, Busa set out to create an index of all the words in St. Thomas Aquinas' (and related authors) works, totaling an impressive 11 million medieval Latin words. To accomplish this, Busa sought the support of IBM's CEO, Thomas Watson, eventually creating a punch-card lemmatized concordance (Hockey, 2004). Over the next five decades, the field took on a variety of labels including Humanist Informatics, Literary and Linguistic computing, and Humanities Computing (Nyhan & Flinn, 2016). In 2004, fifty-five years after he started his groundbreaking, monumental work, Busa wrote the foreword to *A Companion to Digital Humanities* which formally introduced the term "digital humanities" (Brandeis Library, 2012).

In the sixteen years since its introduction, the term, "digital humanities," has been widely adopted. Even though we have a broad acceptance of the term, its definition and boundaries remain contested. The varied ways of defining digital humanities is evidenced in Matthew K. Gold's *Debates in Digital Humanities* (2012). For instance, John Unsworth stated that digital humanities is "using computational tools to do the work of the humanities" (Gold, 2012, p. 67). Julia Flanders explained that "digital humanities is a critical investigation and practice of the methods of humanities research in the digital medium" (Gold, 2012 p. 69). Ernesto Priego, with a slightly different take, defined digital humanities as "the scholarly study and use of computers and computer culture to illuminate the human record" (Gold, 2012 p. 69). Looking ahead, Ed Finn, expounded on the future of DH and said,

> I think digital humanities, like social media, is an idea that will increasingly become invisible as new methods and platforms move from being widely used to being ubiquitous. For now, digital

> humanities defines the overlap between humanities research and digital tools. But the humanities are the study of cultural life, and our cultural life will soon be inextricably bound up with digital media. (Gold, 2012, p. 68)

Finally, Matthew K. Gold shares his explanation of DH as "both a field with a discernable set of academic lineages, practices, and methodologies and a vague umbrella term used to describe the application of digital technology to traditional humanistic inquiry." Importantly, he adds that "what sets DH apart from many other humanities fields is its methodological commitment to building things as a way of knowing" (Gold, 2012, p. 68-69).

However, one chooses to define DH, there are hallmarks that are part of the nature of the field. These hallmarks include the application of technology to a research question(s); collaboration between disciplines, services, programs, and departments; iterative nature of projects requiring an attitude of experimentation and problem-solving; critically questioning the role and impact of the technology; creating space to ask new questions; and bringing new ways of exploring old data.

From the early days of humanities computing to DH today, the object of the field's research agenda has shifted in a number of directions. Though text and its analysis remains a popular object of research, it is no longer the focus of the conversation. Other objects of focus and methodologies have been pushing the DH research agenda in more recent years particularly as a more diverse group in academe become involved in DH projects, research, and teaching. For instance, academic librarians have been documenting their own role in DH collaborations with campus faculty. They have been providing critical reflections on how they support DH work and build their library's capacity to engage in DH partnerships on campus (Edmond et al. 2020; Hartsell-Gundy 2015; Logsdon et al. 2017; Siemens et al., 2010 and 2011). In 2015, Hartsell-Gundy et al. edited a volume that has become a touchstone resource for subject-specialist librarians. Their work sought to provide librarians with a sense of what is digital humanities and what kind of campus relationships are typically needed. The text also illuminated avenues and examples of collaboration in DH initiatives for subject librarians. On the heels of this edited volume came an article from Bello et al. (2017) detailing a capacity-building DH activity among librarians from two different divisions in the library. This was an example of "learn by doing" in which the librarians employed DH analysis tools to investigate collection development trends. Further, the authors advocated for librarians to leverage DH tools and applications in their own research and not solely engage in a supportive role on DH projects.

Logsdon et al. (2017) tackled the rising issue of a librarian's labor in DH projects. Looking at the structural inequality in academic labor, the authors advocate for making a librarian's expertise as a discourse mediator and affective labor more explicitly known. Although there are many disparate topics within the professional library and information science literature centering on digital humanities, another area of focus is sustainability. Edmond's et al. (2020) research findings hold that sustainability considerations and planning for DH projects should not only address data and technology but also the people involved - the user community, communications, and knowledge management.

Today, there are many different voices in the DH literature sharing their expertise, experiences, new findings, new questions, and new resources. These voices range from disciplinary faculty to information technologists, to programmers, to GIS and data specialists, to students (both graduate and undergraduate), to scholars working in museums and archives, and to academic librarians. Each voice brings something new to the field of digital humanities. And while one area in the DH literature may be focusing on textual analysis, another may be interrogating labor issues, artificial intelligence, or machine learning. The

common thread always being the humanist bringing computational power to bear on their questions and applying analysis to their findings.

Related Research

As DH has emerged as a distinct academic field, researchers have tried to explore research topics and knowledge structure in the field. Traditional bibliometrics studies have been most widely conducted in the investigation of the DH research domain. As an early effort, Wang and Inaba (2009) investigated the emergence of digital humanities between 2005 and 2008 based on the analysis of DH research publications (e.g., *Digital Humanities Quarterly* and *Literary and Linguistic Computing*). They made one of the early findings observing the evidence of expansion of DH research. Following these initial studies, most of the DH domain analyses have been made in the last five years. For instance, Gao et al. (2017) conducted author co-citation analysis using VOSviewer, which is a software tool designed to automatically analyze bibliographic records (Van Eck & Waltman, 2019). They looked into the co-citation relationships among core journals in DH such as *Computers and the Humanities*, *Digital Humanities Quarterly*, and *Literature and Linguistic Learning*. It is one of the typical, exemplar studies that examined scholarly communications using traditional bibliometrics analysis. Similarly, Wang (2018) also relied on VOSviewer to analyze bibliographic data related to DH, collected from Web of Science. The findings confirmed the exponential growth of DH research in the recent decade. Wang further explored most productive institutions, key authors, representative journals, and keyword co-occurrences in the DH field. Tang, Cheng, and Chen (2017) employed multiple bibliometrics techniques to explore the field of DH, such as co-authorship analysis, co-citation analysis, and bibliographic coupling. Their findings detected ten groups of authors with distinct specialties, ranging from general interests, digital infrastructure, author attribution, digital libraries, and to others. Chen and Tang (2019) specifically focused on the development of DH research in Taiwan based on co-authorship analysis. They observed that the recent popularity of collaborative efforts and adoption of computational methods in DH while earlier work tended to involve cultural heritage studies. Most recently, Su (2020) also used VOSviewer to investigate international-level research contributions in the field of DH. The countries located in the center of a collaborative relationship map include the United States, Germany, and England, revealing those three countries are leading the field of DH research.

Another line of research has investigated sub-areas, elements, and characteristics in the field of DH. Poole (2017) critically synthesized prior work in DH and identified the conceptual ecology of digital humanities by reviewing fundamental issues in DH. Poole claimed that DH represents a new emerging current of interdisciplinary intellectual research activities, which respond to various research problems and issues in humanities. Poole and Garwood (2018) further focused on the nature of interdisciplinarity and collaboration in the DH field empirically. They identified the benefits of collaborative work in DH as avoiding redundancy, exploding disciplinary silos, and more ambitious and larger-scale research. Kaplan (2015) conceptually defined three concentric areas of big data in DH, including big cultural data, digital culture, and digital experience, and discussed research challenges in each area. Lee and Wang's study (2018) revealed the nature of cross-disciplinarity in DH, which encompasses Computer Science, History, Chinese, Humanities, and Geography, among others. Recently, Münster and Terras (2020) focused on digital visualization techniques applied in humanities studies. They introduced the concept of visual digital humanities, which describes the approaches in DH consuming and producing pictorial information to respond to various research questions in humanities.

The field of DH has a close relationship with librarianship and library and information science (LIS). Several studies have investigated the elements of DH from the librarianship perspectives. Siemens et al. (2010; 2011) emphasized the collaborative nature of DH and libraries. Collaboration within digital project teams encompasses librarians, academics, students, computer programmers, and other individuals. Kamada (2010) explored the elements of DH related to libraries, for example, metadata, digital archives, XML, and e-books. Kamada emphasized the advantage of computer-assisted research in DH, especially the application of text mining. Koltay's study (2016) also points to the importance of digital data-intensive research in DH. Sula (2013) defined a conceptual model that represents the relationship between DH and librarianship. Sula identified five areas of research topics in the library context, including arts & humanities librarianship, digital infrastructure, knowledge production and collaboration, digital scholarship, and research communities. Bakkalbasi et al. (2015) specifically focused on the skills relevant to DH librarianship, such as digitization, citation and resource management software, metadata, and web skills, among others. Green (2014) analyzed five empirical cases in which academic libraries collaborated with faculty in DH research studies. Green argued the importance of librarians' support in the area of text encoding. Padilla (2016) examined types and characteristics of humanities data in libraries in regards to digital scholarship initiatives. Padilla investigated the existent data stewardship practices and identified three types of humanities data collection models based on data availability, data accessibility, and content. Data management is another area where academic libraries potentially collaborate with humanities researchers (Poole and Garwood, 2020). Data management supports robust DH project work and information professionals including librarians help with data management planning in the humanities field.

These prior studies have greatly contributed to the understanding of the DH domain. Particularly, traditional bibliometrics have been widely employed, such as co-author analysis, citation analysis, and keyword relationships to understand the research domain of DH (e.g., Gao et al., 2017; Wang, 2018; Tang, Cheng, & Chen 2017; Su, 2020). Additionally, those in library science fields, too, have made contributions to the DH research discipline. However, little research has applied latent Dirichlet allocation (LDA) topic models in the investigation of the DH field, especially for trend analysis of DH topics in the recent decade. As an attempt to holistically explore the entire sphere of DH, focusing on topic trends, this study employed existent text mining techniques including bigram analysis, keyword co-occurrence network, structural topic modeling, and bi-term topic modeling.

Methods

Data collection

For this study, we collected bibliographic records from multiple scholarly databases in May 2020: (a) Academic Search Complete, (b) Library Literature & Information Science Full Text (H.W. Wilson), and (c) Library, Information Science & Technology Abstracts (LISTA) via EBSCOhost and (d) Scopus. For those databases, we used the query of "digital humanities" for the title, abstract or keyword fields to ensure high recall rates. We refined the search results limited to research articles in English published from 2010 to 2020. Then, we removed any redundant records as well as the records that do not have any abstract information. After removing them, we ended up with 2,717 records. From the collected bibliographic records, two sets of text corpora were constructed: (a) the "abstracts" corpus and (b) the "subject terms" corpus. We deleted any copyright information from the abstract text using regular expression, which is not directly relevant to the content of an article. Since the data was collected in May 2020, relatively, there are a smaller number of records in 2020.

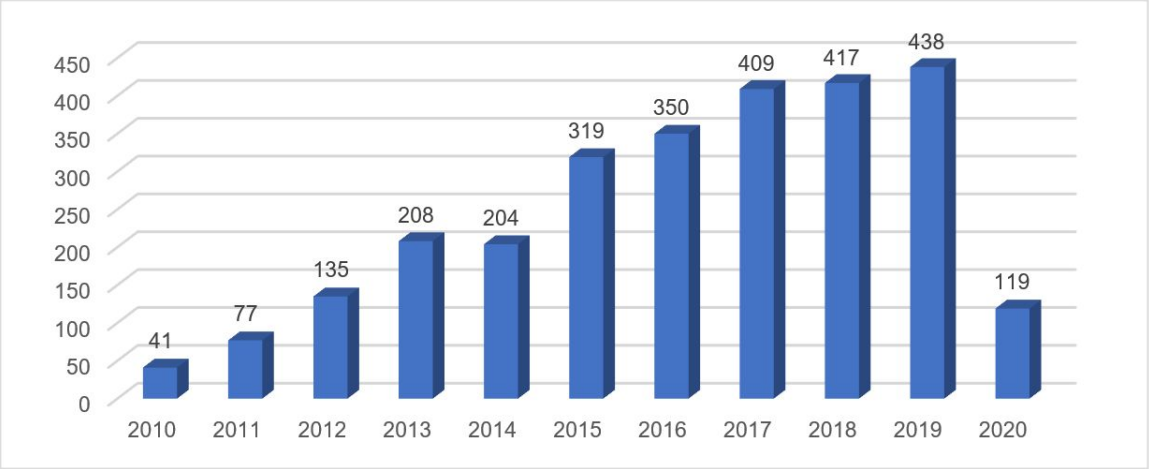

Figure 1. Number of articles by year

Data analysis

Multiple text mining procedures were carried out to explore topics and trends in the domain of DH research. The collected data underwent text preprocessing including tokenization and stopword removal. The following methods were applied: (a) term frequency analysis and bi-gram analysis were conducted to identity key concepts in the domain of digital humanities. We also made a network map of keyword co-occurrences. To examine the trends of key concepts, we looked into the changes of frequent bigrams over three sequential time periods, (i.e., 2010-2014, 2015-2017, and 2018-2020); (b) Latent Dirichlet Allocation (LDA) topic modeling was used to explore topics and concepts underlying the abstracts corpus. LDA is an unsupervised machine learning technique that detects hidden latent topics or themes from a set of unstructured text (Blei et al., 2013; Blei, 2012). More importantly, we traced the topic probabilities over time and identified hot and cold topics respectively; and (c) We employed bi-term topic modeling for the subject terms corpus. Because the subject terms are likely to be short text for each record, the conventional LDA method was not applicable. Thus, bi-term topic modeling, which resolves the problem of data sparsity in short text (Yan et al., 2013), was used alternatively. The bi-term topic modeling considers the whole corpus as a mixture of topics where each bigram is drawn from a specific topic independently. It can capture multiple topic gradients in a short text while keeping the correlation between words (Yan et al., 2013). Most prior studies employed co-word analysis (Wang and Inaba, 2019) or relied on existent software tools dedicated to traditional bibliometrics analysis (Gao et al. 2017; Wang 2018; Su 2020). Less research applied text mining and machine learning techniques to explore research topics and their trends in DH. This study employed LDA topic modeling to uncover prevailing topics and themes in the domain of DH research.

Results

Investigation of the DH domain based on term-level analysis

First, we investigated most frequent terms from both corpora. Table 1 presents 30 most frequent terms from the abstract and subject term corpora respectively. Not surprisingly, “digit” and “human” were the top two terms in both corpora. Also, the term “data” and “research” among the top across the corpora. For

the abstract corpus, we observed 14,392 unique terms and 226,235 tokens. In the abstract terms, the terms representing research articles appeared often among the top words, such as “research,” “studi,” “project,” “paper,” “work,” and “articl,” because the abstracts are likely to be a summary of each article. Also, the terms like “analysi,” “tool,” and “method” are related to research analysis and methods employed in the articles. Some terms in the abstract corpus reveal areas or disciplines in digital humanities, for example, “cultur,” “histori,” “librari,” and “archiv.” For the subject terms corpus, we observed 4,594 unique terms and 41,111 tokens. The subject terms exhibited more cohesive vocabularies to represent topics or areas of research in digital humanities. Computation and data analysis related terms ranked highly, such as “data,” “comput,” “system,” “analysi,” “visual,” “network,” “technolog,” and others. Also, we found that “librari" is highly ranked at 8th.

Table 1. Most frequent stemmed terms in each corpus

Abstract terms				Subject terms			
Rank	Term	Frequency	Percent	Rank	Term	Frequency	Percent
1	digit	5361	2.37%	1	digit	2886	7.02%
2	human	3574	1.58%	2	human	2201	5.35%
3	research	2865	1.27%	3	data	784	1.91%
4	data	2035	0.90%	4	comput	724	1.76%
5	studi	1651	0.73%	5	inform	669	1.63%
6	project	1644	0.73%	6	research	657	1.60%
7	develop	1267	0.56%	7	system	553	1.35%
8	paper	1232	0.54%	8	librari	520	1.26%
9	work	1197	0.53%	9	histori	508	1.24%
10	articl	1179	0.52%	10	analysi	408	0.99%
11	inform	1156	0.51%	11	cultur	373	0.91%
12	text	1151	0.51%	12	visual	369	0.90%
13	present	1055	0.47%	13	languag	362	0.88%
14	cultur	1048	0.46%	14	social	344	0.84%
15	analysi	1007	0.45%	15	semant	328	0.80%
16	tool	1000	0.44%	16	network	308	0.75%
17	method	988	0.44%	17	technolog	306	0.74%
18	collect	984	0.43%	18	scienc	294	0.72%
19	model	944	0.42%	19	archiv	263	0.64%
20	scholar	931	0.41%	20	histor	260	0.63%
21	technolog	923	0.41%	21	process	259	0.63%
22	histori	914	0.40%	22	model	251	0.61%
23	librari	894	0.40%	23	text	248	0.60%
24	histor	883	0.39%	24	learn	246	0.60%
25	approach	875	0.39%	25	knowledg	241	0.59%
26	provid	814	0.36%	26	web	239	0.58%
27	discuss	799	0.35%	27	educ	234	0.57%
28	comput	778	0.34%	28	studi	219	0.53%
29	archiv	769	0.34%	29	heritag	218	0.53%
30	practic	757	0.33%	30	manag	187	0.45%

To better identify topical terms in digital humanities, we extracted bigram terms from both the abstracts and subject terms (Table 2). Top bigrams reveal key topics, issues, and methods in the domain of DH. The bigrams extracted from the abstracts well represent sub-areas in DH, such as cultural heritage, digital scholarship, social media, digital libraries, linked open data, historical research, and others. Also, there were several bigrams indicating research methods and approaches, such as case study, big data, digital tools and technologies, and computational methods. The bigrams of subject terms exhibit more specific topics or methods. Some of these were geographic information, semantic web, natural language processing, culture heritage, and information retrieval. In addition, we observed bigrams related to recent computational techniques such as big data, text mining, machine learning, and artificial intelligence.

Table 2. Most frequent bigrams

Abstracts			Subject terms		
Rank	Bigram	Frequency	Rank	Bigram	Frequency
1	digital humanities	2244	1	digital humanities	1690
2	cultural heritage	220	2	humanities digital	222
3	humanities research	193	3	digital libraries	188
4	case study	167	4	information systems	141
5	paper presents	124	5	geographic information	110
6	humanities dh	116	6	semantic web	105
7	digital scholarship	108	6	natural language	105
8	big data	104	8	language processing	104
9	humanities scholars	101	9	cultural heritage	90
10	social sciences	94	10	humanities computing	88
10	social media	94	11	information retrieval	83
12	field digital	91	12	data mining	82
13	digital tools	90	13	humanities research	78
14	article discusses	89	14	cultural heritages	74
15	research data	78	15	big data	73
16	digital technologies	77	16	processing systems	72
17	humanities projects	76	17	linked data	62
18	case studies	75	18	history digital	61
19	digital libraries	74	19	character recognition	59
20	linked data	73	20	text mining	56
20	digital media	73	21	network analysis	56
22	humanities project	72	22	heritage digital	54
23	open data	70	23	linked open	53
24	use digital	69	23	open data	53
25	computer science	68	25	data visualization	51
25	research project	68	26	machine learning	50
27	historical research	67	27	user interfaces	47
28	paper present	65	27	artificial intelligence	47
29	design methodology	64	29	social sciences	46

30	methodology approach	63	29	social networking	46
30	computational methods	63	31	social media	45
30	originality value	63	31	united states	45
30	digital resources	63	33	humanities historical	44

As shown in Figure 2, we visualized the co-occurrence relationships between key terms. The network diagram places "digital" and "humanities" in the center of the network. The three terms, "history," "data," and "research," are also located in the central area of the map. Other keyword terms observed in the near-central area include "computer," "systems," and "information."

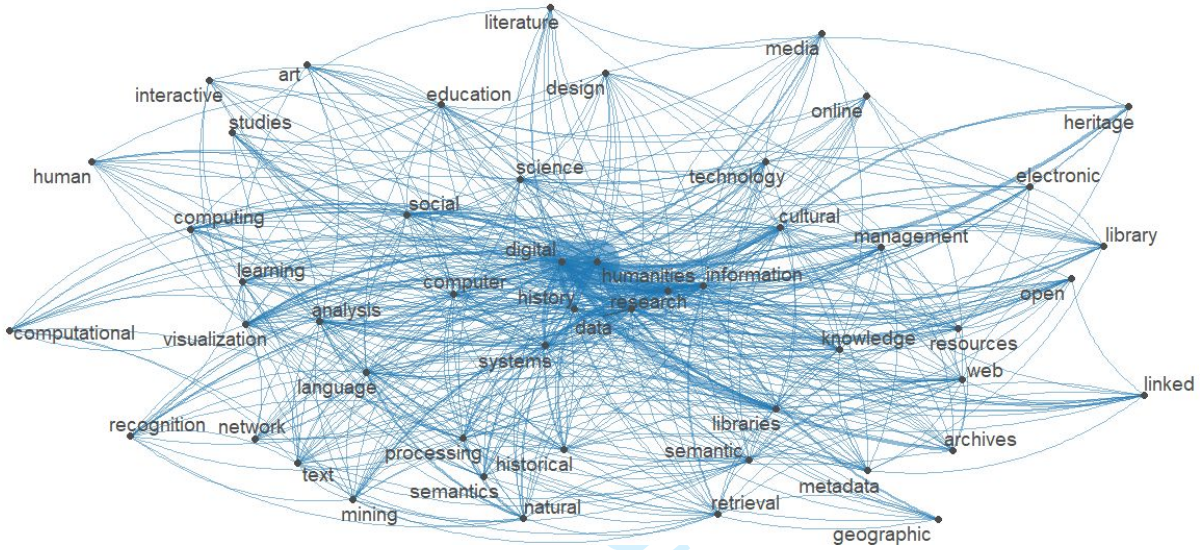

Figure 2. A network visualization of keyword co-occurrences

Next, we investigated the trends of bigrams over time. We divided the whole abstract dataset into three successive time periods (i.e., 2010–2014, 2015–2017, and 2018–2020) and then extracted bigrams that increased compared with the previous period. To evaluate the degree of increase, the score of each bigram was calculated using the size of documents in a particular period.

$$Score = DF_{cur} - DF_{prev} * \frac{\#doc_{cur}}{\#doc_{prev}}$$

DF_{cur} and DF_{prev} are the document frequencies of the target bigram at the current or previous time period, respectively, and $\#doc_{cur}$ and $\#doc_{prev}$ are the total number of documents at the current or previous time period, respectively. For the period 2010–2014, we show the original rank based on document frequency. We found that computational methods have received an increased attention recently, such as machine learning, natural language processing, topic modeling, and data science.

Table 3. Analysis of bigram trends (document frequency)

2010–2014 (665 documents)		2015–2017 (1078 documents)			2018–2020 (974 documents)		
Bigram	Rank	Bigram	Rank	Score	Bigram	Rank	Score

digital humanities	1	humanities dh	1	25.93	machine learning	1	17.96
humanities research	2	digital scholarship	2	23.03	cultural heritage	2	17.44
article discusses	3	case study	3	21.61	natural language	3	16.83
cultural heritage	4	humanities projects	4	19.79	also mentions	4	13.00
case study	5	paper presents	5	18.58	data available	5	12.10
field digital	6	humanities project	6	18.03	one hand	6	11.58
humanities scholars	7	language processing	7	15.76	language processing	7	10.83
paper present	8	text analysis	8	15.14	studies digital	8	10.58
paper presents	9	distant reading	9	14.14	topic modeling	8	10.58
social media	10	textual data	10	14.00	virtual reality	10	10.39
digital technology	11	research data	11	13.90	widely used	10	10.39
use digital	12	research projects	12	13.52	paper argues	12	10.29
big data	13	digital humanities	13	13.24	allow us	13	10.19
paper describes	14	purpose paper	14	13.14	second part	14	10.10
digital libraries	15	recent years	15	12.89	wide range	15	9.87
linked data	16	open data	16	12.65	article focuses	16	9.36
article presents	16	computational methods	17	12.52	data science	17	9.19
arts humanities	16	using digital	18	12.52	american studies	18	9.10
research project	16	natural language	18	12.52	archives museums	18	9.10
digital library	16	digital humanists	20	12.27	humanities dh	20	8.53
new media	16	supporting digital	21	11.00	spatial humanities	21	8.19
digital resources	22	purpose purpose	22	10.38	results show	22	8.16
historical research	22	article describe	23	10.00	topic modelling	23	8.10
computer science	22	domain experts	24	9.38	libraries museums	23	8.10
within digital	22	network analysis	25	9.27	practical implications	25	7.77
digital media	22	project management	26	9.00	social sciences	26	7.41
research questions	27	makes possible	26	9.00	large amounts	27	7.39
higher education	28	two case	26	9.00	neural networks	28	7.29
digital archives	28	results show	29	8.76	new perspectives	29	7.29
humanities social	28	text mining	30	8.65	open access	30	7.25

Topic trends analysis

We conducted topic modeling to detect prevailing topics, themes, and concepts in the domain of DH. Two different topic modeling approaches were applied for the two text datasets respectively: LDA topic modeling for the abstracts corpus and biterm topic modeling for the keywords corpus.

First, LDA topic modeling was carried out with the abstracts corpus. To determine the optimal number of topics, we used the "griffith2004" metrics, proposed by Griffiths and Steyvers (2004), included in the R "ldatuning" package (Nikita & Chaney, 2020). Using these metrics, we set the topic number (k) as 65. Figure 3 presents the topics by probability rank, and Table 4 presents the topic model results. Among all 65 extracted topics, we interpreted the top 49 topics. Lower ranked topics below 50th were likely to

exhibit less coherent top probable topic terms. Therefore, those were excluded from the analysis. To better interpret the topic model results, we looked up actual abstracts that contain topical terms. T20 indicates DH research in general. Popular research topics/themes or methods include open linked data (T35), text mining (T29), visualization (T9), and semantic web and ontology (T31). DH librarianship is another distinct topic highly ranked (T7). There are several topics related to literature, history or cultural heritage related topics, including T15, T36, T30, T4, T44, T39, and others.

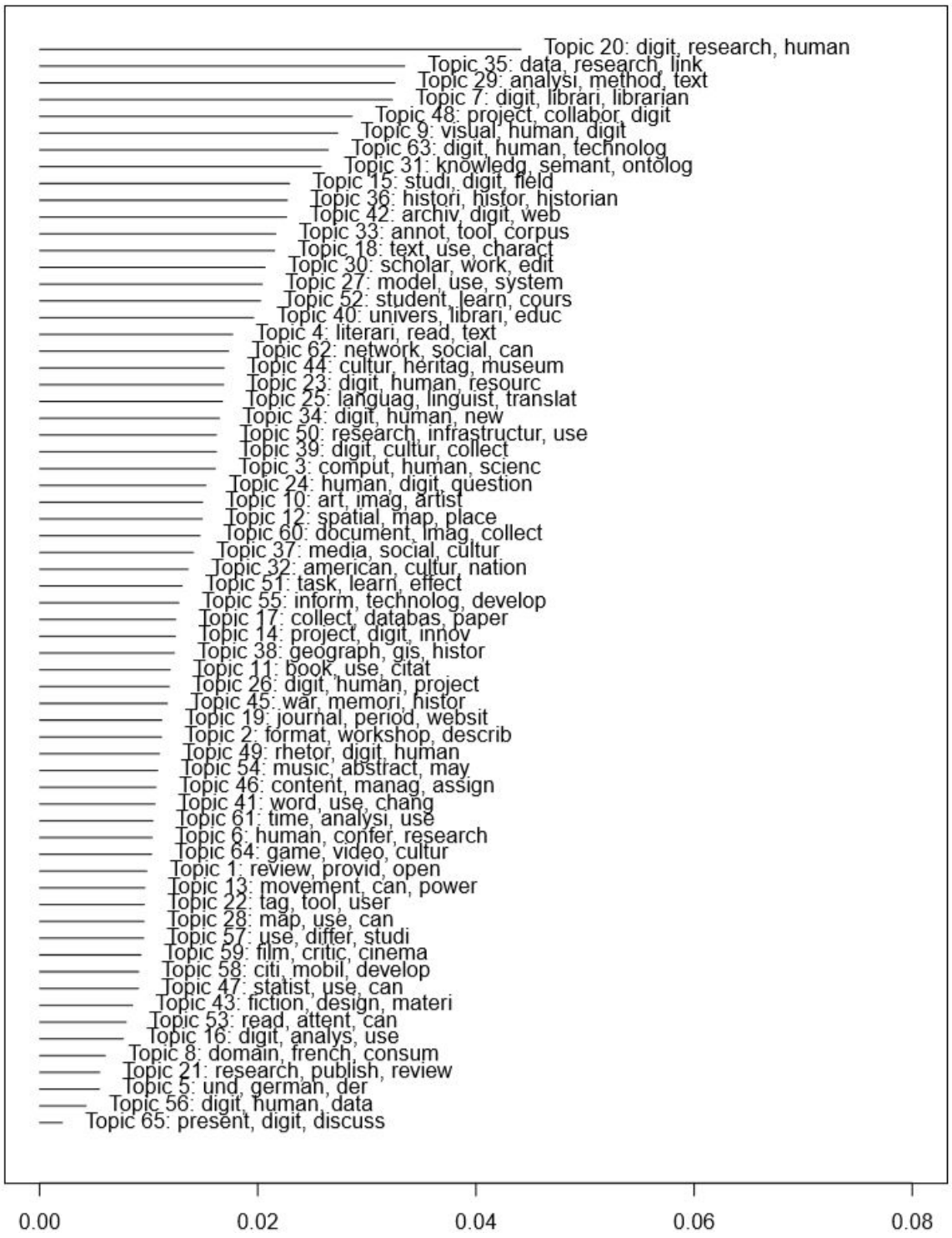

Figure 3. Estimated probabilities of topics

Table 4. Topic modeling results (sorted by probability)

Topic	Topic label	Most probable stemmed terms
T20	DH research	digit, research, human, librari, articl, includ, discuss, univers, develop, inform
T35	Linked open data	data, research, link, use, open, human, paper, big, applic, service
T29	Text mining	analysi, method, text, use, tool, techniqu, studi, corpus, mine, methodolog
T7	DH librarianship	digit, librari, librarian, human, scholar, scholarship, role, academ, research, support
T48	Collaboration work	project, collabor, digit, work, practic, team, initi, research, studi, manag
T9	Visualization	visual, human, digit, research, present, issu, uncertainti, within, support, design
T63	DH technology	digit, human, technolog, practic, work, within, critic, way, engag, object
T31	Semantic web and ontology	knowledg, semant, ontolog, link, use, web, paper, base, present, entiti
T15	Literature	studi, digit, field, human, literatur, will, approach, disciplin, practic, archaeolog
T36	History	histori, histor, historian, scienc, digit, articl, studi, research, oral, new
T42	Digital archives and preservation	archiv, digit, web, preserv, record, materi, sourc, document, collect, access
T33	Text annotations	annot, tool, corpus, text, digit, human, user, use, process, paper
T18	Text digitization and OCR	text, use, charact, histor, name, digit, document, ocr, ancient, can
T30	Medieval manuscripts	scholar, work, edit, manuscript, digit, mediev, new, author, articl, scholarship
T27	System model	model, use, system, approach, human, differ, propos, implement, inform, present
T52	Learning and education	student, learn, cours, teach, use, educ, digit, experi, human, undergradu
T40	University and libraries	univers, librari, educ, digit, collect, public, institut, school, faculti, partnership
T4	Literary criticism	literari, read, text, literatur, studi, close, digit, critic, method, distant
T62	Social network analysis	network, social, can, differ, relationship, studi, charact, use, graph, structur
T44	Cultural heritage	cultur, heritag, museum, object, digit, use, inform, technolog, collect, system
T23	DH resources	digit, human, resourc, impact, research, studi, use, develop, field, technolog
T25	Linguistics	languag, linguist, translat, semant, one, use, context, corpus, natur, present
T34	Diversity (Black and women)	digit, human, new, black, articl, practic, studi, women, chang, research
T50	Infrastructure	research, infrastructur, use, paper, digit, human, scholar, twitter, practic, scienc
T39	Cultural heritage and community engagement	digit, cultur, collect, user, heritag, engag, communiti, differ, interact, content
T3	Computational science	comput, human, scienc, copyright, develop, work, use, field, new, research

T24	Questions and answers	human, digit, question, author, method, use, articl, answer, shakespear, comput
T10	Arts	art, imag, artist, digit, work, collect, use, histori, experi, artwork
T12	Maps and spatial	spatial, map, place, space, histor, geograph, use, represent, narrat, deep
T60	Document and image collections	document, imag, collect, data, qualiti, histor, newspap, digit, extract, use
T37	Social media	media, social, cultur, digit, platform, technolog, new, emerg, use, product
T32	African American	american, cultur, nation, space, articl, african, world, urban, steampunk, map
T51	Crowdsourcing	task, learn, effect, use, particip, crowdsourc, studi, perform, result, motiv
T55	Information technology	inform, technolog, develop, knowledg, digit, infrastructur, research, polici, organ, human
T17	Collections and databases	collect, databas, paper, digit, set, use, brows, imag, process, can
T14	Digitization projects	project, digit, innov, univers, lab, develop, use, digitis, will, transcrib
T38	Geographic data and GIS	geograph, gis, histor, data, geographi, map, inform, use, spatial, develop
T11	Books and publishing	book, use, citat, digit, inform, librari, publish, cite, resourc, imag
T26	DH project	digit, human, project, process, focus, research, also, design, platform, differ
T45	History: war, politics	war, memori, histor, cultur, polit, use, world, period, past, state
T19	Journals	journal, period, websit, digit, index, issu, articl, studi, perform, theatr
T2	Metadata	format, workshop, describ, metadata, process, tool, digit, can, human, tei
T49	Rhetoric and essays	rhetor, digit, human, univers, practic, essay, theori, press, modern, paper
T54	Music and musicology	music, abstract, may, copyright, copi, user, publish, email, articl, musicolog
T46	Content management	content, manag, assign, present, student, event, servic, issu, support, resourc
T41	Language analysis	word, use, chang, studi, result, languag, mean, predict, differ, set
T61	Time analysis (history)	time, analysi, use, histor, method, chang, can, quantit, studi, word
T6	Conference papers	human, confer, research, digit, new, scienc, paper, address, comput, intern
T64	Video games and virtual reality	game, video, cultur, realiti, use, virtual, play, studi, effect, educ

Further, we traced topic trends in the DH research domain and identified eight hot topics and eight cold topics, which showed increasing or decreasing patterns over time. As shown in Figure 4, hot topics include open linked data (T35), text mining (T29), semantic web (T31), history (T36), text digitization (T18), social network analysis (T62), and maps (T12), among others. On the contrary, descending trend topics include DH librarianship (T7), archives/preservation (T42), medieval manuscripts (T30), universities and libraries (T40), infrastructure (T50), cultural heritage (T39), arts (T10), social media (T37), and some others.

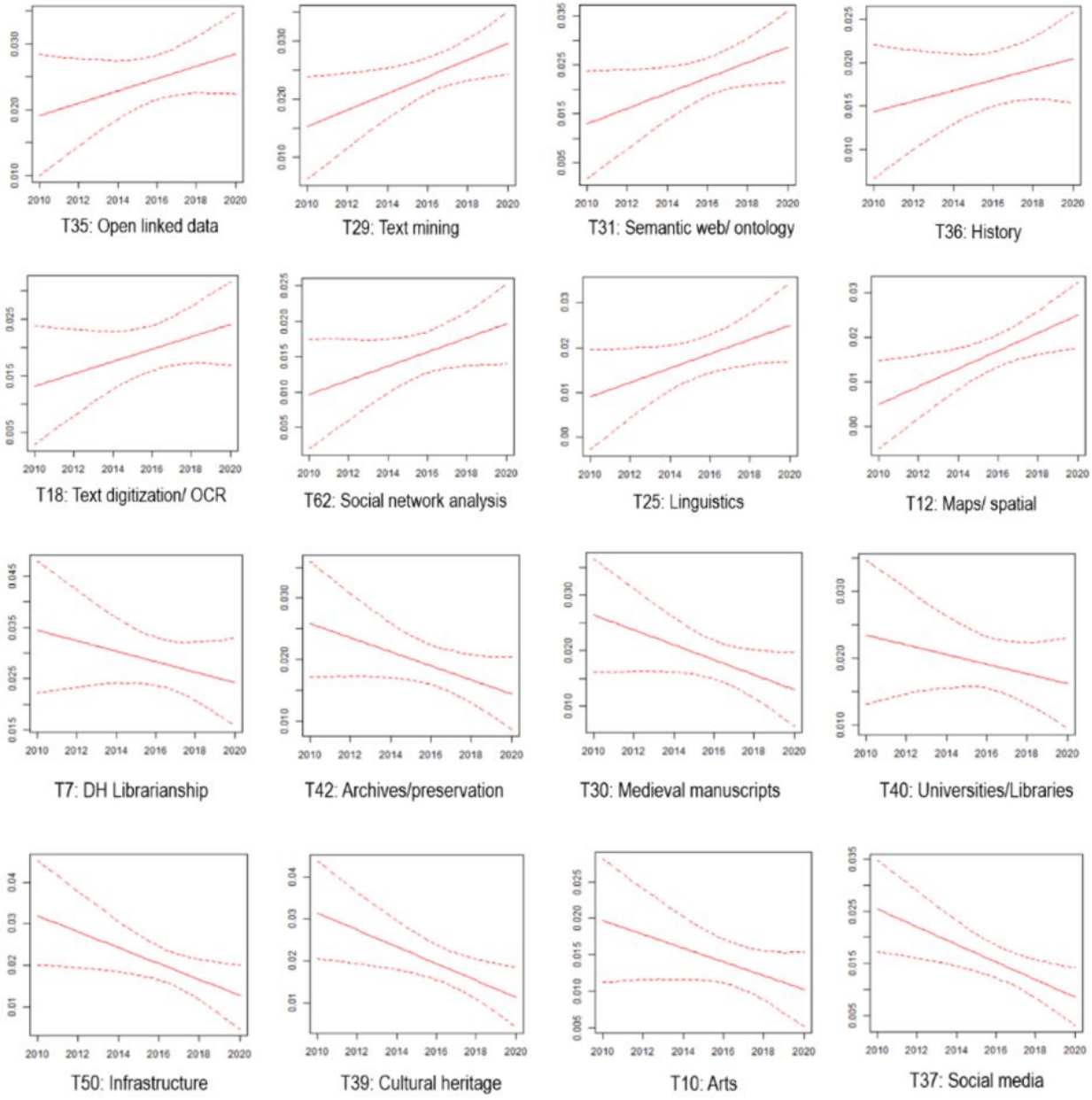

Figure 4. Select hot and cold topics

Finally, we analyzed the topics from the subject terms corpus using bi-term topic modeling. Bi-term topic modeling allowed us to identify prevalent topics represented in the subject terms (Figure 5). We observed an overlap with the topics from the abstracts corpus, such as linked open data, semantic web, natural language processing, and cultural heritage. Additionally, there are distinct topics related to maps, geographic information, computer graphics, image recognition, user interface, and visualization. Social media and virtual reality made up distinctive topics as well. Methods-related topics included computational linguistics, network analysis, machine learning, and artificial intelligence. Further, a separate topic included libraries in DH.

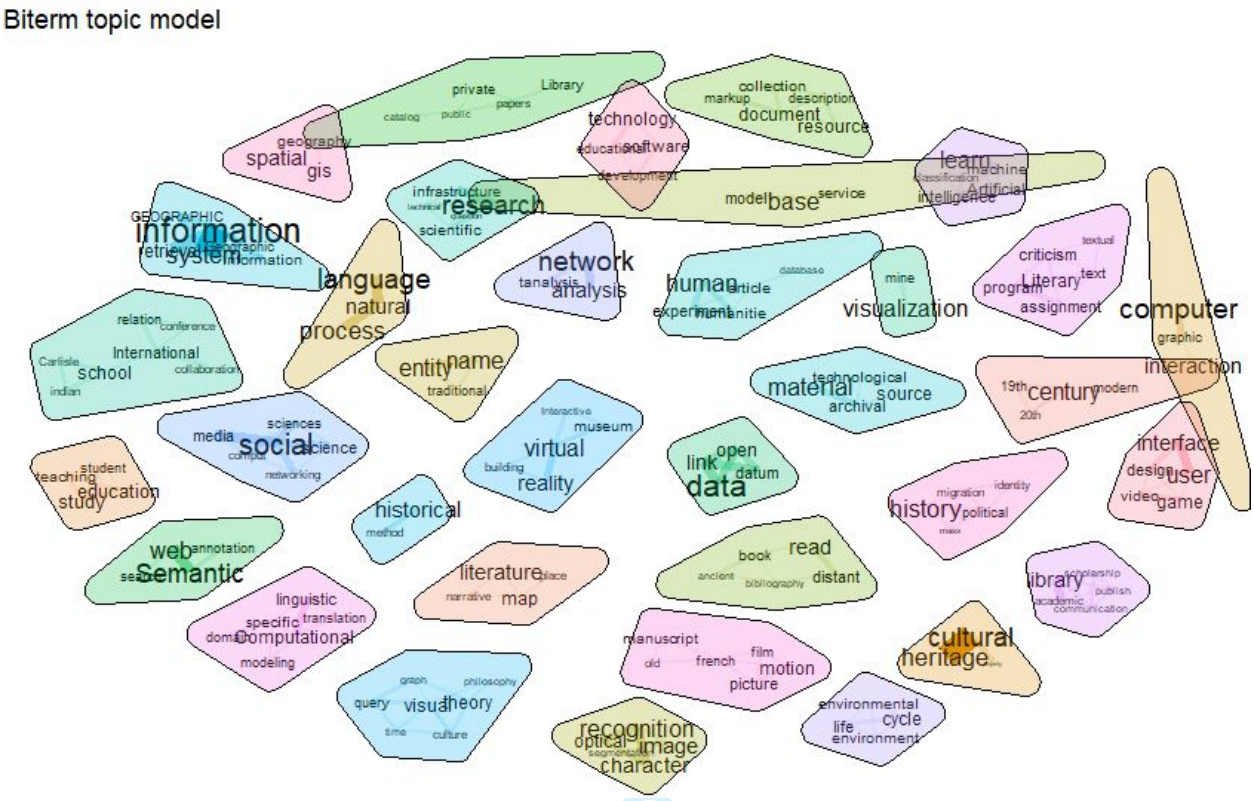

Figure 5. Bi-term topic model based on the subject terms

Discussion and conclusion

This study explored topics and topic trends in the domain of DH in the recent decade. We collected DH related research records comprehensively from multiple scholarly databases and created two distinct corpora (i.e., the abstracts corpus and the subject terms corpus). To explore different dimensions of the collected data, we relied on multiple text mining techniques. Specifically, we first looked at most frequent terms and bigrams from the collected text to examine key concepts in DH. Then, topic modeling was applied to both the text corpora to identify what the prevailing topics are in DH. More importantly, we further investigated the trends in topics over time and identified recently popular topics. Although we borrowed existent textual analysis methods, these methods have not been widely applied in analyses of the DH research domain.

The term frequency analysis revealed different dimensions of key concepts in DH. The findings from the term-level analysis reaffirm the key constructs and popular subject terms that were found in previous studies (Chen and Tang, 2019; Wang and Inaba, 2009; Wang, 2018; Tang, Cheng, and Chen 2017; Su 2020). The top terms revealed popular topics in DH such as culture, history, language, archives, and heritage. These terms represent primary subjects in humanities, and, thus, make their way into digital humanities research. Another important dimension observed from the top terms involves methods in DH. In particular, the subject terms showed several recent computational data analysis methods adopted in DH, such as “comput,” “visual,” “network,” and “technolog.” One of the underlying assumptions and

characteristics in DH lies in that computational methods have enabled humanities researchers to produce and develop new insights and arguments (Dobson, 2019). This basic term frequency analysis explicitly describes the adoption of computational methods in the domain of DH. Another notable observation is that DH is closely associated with librarianship. Some terms are relevant to the library context directly or indirectly, for example, "librari" and "archiv." There have been research efforts to define the roles of libraries in DH, and DH librarianship has become a distinct area in the library field (Sula, 2013; Poremski, 2017). Libraries are one of the critical entities that participate in DH communities. Librarians have provided assistance or, in many cases, have been directly involved in different types of DH initiatives, such as research projects which include digitization, preservation, text encoding, and cultural heritage (Cunningham, 2010; Sula, 2013; Green, 2014; Poremski, 2017). According to Siemens et al. (2011), DH remains closely associated with diverse entities, such as libraries/librarians, archives, historical organizations, and cultural heritage institutions. The term-level analysis provides additional evidence that these diverse stakeholders have been collaborators in the DH research domain.

The bigram analysis provided a more explicit set of terms that defines the key concepts in DH. Similar to the findings of term frequency analysis, the bigrams highlight popular topics in DH. Cultural heritage is one of the top ranked bigrams, revealing that it is a critical area in DH. There have been many digitization initiatives in collaboration with humanists and libraries to build digital cultural heritage collections, and it is considered one of the primary practices in DH (Tomasi, 2018). Digital libraries have long been associated with digital humanities (Siemens et al., 2011; Sula, 2013). The bigram "digital libraries" is ranked 3rd in the subject terms corpus, which implies close association between DH and libraries' efforts in digitization. Other noteworthy areas in DH according to the most frequent bigrams include geographic information, semantic web, linked data, and digital media. Particularly, DH has adopted geo-spatial techniques and utilized GIS and map data (Bodenhamer, 2013), which was reflected in the analysis of the top bigrams. Also, semantic web and linked data have been considered essential technical tools to organize and present knowledge that is produced and consumed in DH (Hyvönen, 2020). The bigram analysis also implied popular research methods. Interestingly, artificial intelligence appeared among the top thirty keyword bigrams. Artificial intelligence requires large scale data and increased computational capacity. It implies that DH has benefited from the most recent computational analysis and tools. In addition, the topic of user interfaces appeared as a distinct area of research according to the bigram analysis results. As digitized products are delivered via electronic tools, user interface and human-computer interaction emerged as one of the distinct subjects in DH research (Siemens et al., 2016). Then, we examined the changes of frequent bigrams over three sequential periods of time (i.e., 2010-14, 2015-17, and 2018-20). In the most recent period (2018-2020), machine learning emerged as the most frequently used bigram in documents. That is, machine learning has been a computational method that received increased attention in DH. Various machine learning techniques have been adopted in recent DH projects for different data analysis techniques (e.g., Fiorucci et al., 2020). In addition, natural language processing (NLP) and topic modeling have been used as popular tools in DH in recent years (e.g., Navarro-Colorado, 2018; De Luca, 2019). From the trends of bigrams over the three periods, we observed emerging sub-areas and topics in DH. Machine learning and NLP has received increased attention in recent years. Virtual reality is ranked among the top 10 bigram phrases in the period of 2018-2020. Machine learning or virtual reality related terms were not frequently observed in term-level analysis in earlier studies (Wang and Inaba 2009). The bigram trend analysis of this study reveals that machine learning and virtual reality have emerged as popular topics in DH most recently.

Topic modeling provided a structured, comprehensive portrayal of prevailing topics and concepts in DH. The LDA topic modeling extracted sixty-five topics, and we examined the top forty-nine that generated the most coherent terms that could be interpreted as topics. These topics mostly cover the findings from

existing literature (Poole 2017; Wang 2018; Tang, Cheng, and Chen 2017; Su 2020), such as digitization, crowdsourcing, archives, text encoding, visualization, and librarianship, amongst others. Interestingly, "linked open data" appeared as one of the highly probable topics from the abstracts corpus. Linked open data is closely associated with organization of cultural heritage resources (Zeng & Chen, 2018). Similarly, the semantic web concept has been applied to enhance the searchability of cultural heritage or digital resources. Knowledge organization tools for digital archives, such as linked open data and semantic web, were observed as distinct topics in DH. The topic modeling results also exhibit specific humanities disciplines, such as literature, history, linguistics, arts, African American studies, rhetoric, and music. This implies that the DH movement has impacted a variety of fields in the humanities, not limited to cultural heritage. Another important group of topics is methods used in DH. The results highlight that DH benefits from recent computational methods, such as text mining, visualization, social network analysis, and computational sciences. Also, several topics appeared as representing digital technologies (e.g., text digitization, OCR, and semantic web). Again, the essential component in DH is the adoption of digital technologies and computational methods. The roles of libraries are represented in the extracted topics. For example, T7 directly indicates DH librarianship. Some topics such as linked open data, digital archives and preservation, maps, image collections, and metadata are highly relevant to library services and practices (Bakkalbasi et al., 2015; Kamada, 2010; Sula, 2013). Academic librarians have expertise, knowledge, and skills relevant to DH. As a result, librarians/libraries often support, participate as collaborators, and are actively involved in DH initiatives and activities (Cunningham, 2010; Kamada, 2010; Millson-Martula & Gunn, 2017). The bi-term topic model extracted from the subject terms provided a more specific set. Most of the extracted topics from the bi-term topic modeling overlap with those from the LDA topic modeling, including linked open data, semantic web, cultural heritage, and OCR. In the bi-term topic model, natural language processing appeared explicitly as a separate topic. Interestingly, virtual reality is also detected as a distinct topic, which reveals the application of virtual reality techniques in presenting digital artifacts and digital cultural heritage (e.g., Colegrove & Mikel, 2018). The LDA results extracted more diverse topics and sub-areas in comparison with previous studies. Several prior studies that defined DH constructs tended to create a higher-level conceptual map, a conceptual ecology, or a set of categories (Kaplan 2015; Poole 2017; Su 2020). The machine learning technique enabled us to understand deeper levels of latent topics and interpreted forty-nine specific topics derived from the LDA topic models. Several of the topics that we extracted exhibited more specific concepts, such as war and politics, video games and virtual reality, African American, medieval manuscript, OCR, and several others.

One of the contributions of this study is our attempt to trace the changes of topic popularity over time in the past decade. Prior studies made efforts to identify research areas and topics in DH (Wang and Ibana, 2009; Wang, 2018; Tang, Cheng, and Chen 2017; Su 2020; etc.). However, to the best of our knowledge, few studies investigated the trends of research topics in the domain of DH. For example, Wang's study (2018) identified popular keywords based on co-occurrence analysis, but it did not investigate any trend pattern. This study examined the trends of topics identifying those that were hot and cold. The topics that received increased attention are likely to involve recent technologies and methods, such as linked open data, text mining, semantic web and ontology, text digitization, and social network analysis. This implies that, increasingly, the DH field has continued to employ digital technologies in humanities research and projects. In addition, the topic of maps showed an upward pattern, revealing increased use of spatial data in DH projects. The topics of history and linguistics were categorized as hot topics and these fields can benefit from text mining and digital maps and spatial analysis. The results imply that recent trends in DH reflect more extended, diverse topics in DH, especially involving the application of computational methods including text mining, linked open data, and digital spatial data, beyond traditional digitization

projects of cultural heritage. On the contrary, it seems that traditional digital archiving and preservation has received decreased attention in the domain. Even though digital archiving is considered a fundamental area (Poole 2017) in DH, its research is on a downward trend in most recent years. This is partly because other topics were emerging recently as shown above. As archiving and preservation are showing a declining trend, related topics, such as DH librarianship and cultural heritage, also exhibited a relative decrease of topic probability most recently.

Although academic librarians and libraries continue to play an active role in supporting and collaborating with DH researchers (Bakkalbasi et al., 2015; Edmond et al., 2020; Hartsell-Gundy 2015; Kamada, 2010), other topics have emerged overtaking previously trending topics. This has resulted in a relative downward trend in DH librarianship. This downward trend in the topic DH librarianship could be due to a variety of reasons. Two most likely reasons include the broad nature of librarians' work and the ubiquity of DH in the profession. The roles of libraries and their outputs are not limited to research publications. Librarians develop digital infrastructures to support DH projects, provide consultative services for DH researchers, manage digital artifacts, provide DH-related workshops and tutorials, and build digital archives (e.g., oral histories, image galleries). Further, perhaps Ed Finn's projections for the future of DH are manifesting at least partially in the field of librarianship. Indeed, DH has become ubiquitous in academic libraries steadily maturing and evolving while more attention is drawn towards a variety of computational methods and data.

References

Bakkalbasi, N., Jaggars, D., and Rockenbach, B. (2015), "Re-skilling for the digital humanities: Measuring skills, engagement, and learning", *Library management,* Vol 36 No. 3, pp.208-214.

Bello, L., Dickerson, M., Hogarth, M., and Sanders, A. (2017), "Librarians doing DH: A team and project-based approach to digital humanities in the library", *Collaborative Librarianship,* Vol 9 No. 2, pp.97-103.

Blei, D.M., Ng, A.Y., and Jordan, M.J. (2003), "Latent Dirichlet allocation", *Journal of Machine Learning Research,* Vol. 3, pp.993-1022.

Blei, D.M. (2012), "Probabilistic Topic Models." *Communications of the ACM* Vol. 55 No. 4, pp.77-84.

Bodenhamer, D.J., Harris, T.M., and Corrigan, J. (2013), "Deep mapping and the spatial humanities", *International Journal of Humanities and Arts Computing,* Vol. 7 No. 1-2, pp.170-175.

Brandeis Library. (2012), "What's "digital humanities" and how did it get here?", available at https://blogs.brandeis.edu/library/2012/10/09/whats-digital-humanities-and-how-did-it-get-here/ (accessed 20 March 2021)

Chen, K.H. and Tang, M.C. (2019), "A Bibliographic Analysis of Scholarly Publication in the Emerging Field of Digital Humanities in Taiwan", In Wong, S. R., Li, H., & Chou, M. (Ed.), *Digital Humanities and Scholarly Research Trends in the Asia-Pacific*, IGI Global, pp.140-157.

Colegrove, P.T. and Mikel, M. (2018), "Radical inclusion: immersive 360-degree video capture, dissemination, and use of emerging technology in support of traditional archival roles at a university library", *Proceedings of the Association for Information Science and Technology,* Vol. 55 No. 1, pp.779-780.

Cunningham, L. (2010), "The librarian as digital humanist: the collaborative role of the research library in digital humanities projects", *Faculty of Information Quarterly,* Vol. 2 No. 2. pp.1-11

De Luca, E.W., Fallucchi, F., Ligi, A., and Tarquini, M. (2019), "A research toolbox: a complete suite for analysis in digital humanities" In: Garoufallou E., Fallucchi F., William De Luca E. (eds) *Metadata and Semantic Research. MTSR 2019. Communications in Computer and Information Science*, vol 1057. Springer, Cham., pp.385-397.

Dobson, J. E. (2019). *Critical digital humanities: the search for a methodology.* University of Illinois Press, Champaign, IL, USA.

Edmond, J. and Morselli, F. (2020), "Sustainability of digital humanities projects as a publication and documentation challenge", *Journal of Documentation,* Vol. 76 No. 5, pp.1019-1031.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., and James, S. (2020), "Machine Learning for Cultural Heritage: A Survey", *Pattern Recognition Letters,* Vol. 133, pp.102-108.

Gao, J., Duke-Williams, O., Mahony, S., Ramdarshan Bold, M., & Nyhan, J. (2017). "The Intellectual Structure of Digital Humanities: An Author Co-Citation Analysis", available at: https://discovery.ucl.ac.uk/id/eprint/10052270/1/Nyhan%20The%20Intellectual%20Structure%20of%20Digital%20Humanities%20083.pdf (accessed 20 March 2021)

Gold, M. (2012), *Debates in the digital humanities*. University of Minnesota Press, Minneapolis, Minnesota, USA.

Green, H.E. (2014). "Facilitating communities of practice in digital humanities: Librarian collaborations for research and training in text encoding", *The Library Quarterly,* Vol. 84 No. 2, pp.219-234.

Griffiths, T.L. and Steyvers, M. (2004), "Finding scientific topics", *Proceedings of the National academy of Sciences,* Vol. 101(suppl 1), pp.5228-5235.

Hartsell-Gundy, A., Braunstein, L., Golomb, L. and Langan, K. (2015), *Digital humanities in the library: Challenges and opportunities for subject specialists*. Association of College and Research Libraries, Chicago, USA.

Hockey, S. (2004). "The History of Humanities Computing", In Schreibman, S., Siemens, R., & Unsworth, J. (Eds.), *A Companion to Digital Humanities*, Blackwell Publishing. pp.3-19

Hyvönen, E. (2020), "Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery", *Semantic Web,* Vol. 11 No. 1, pp.187-193.

Jessop, M. (2008), "The inhibition of geographical information in digital humanities scholarship", *Literary and Linguistic Computing,* Vol. 23 No. 1, pp.39-50.

Kaplan, F. (2015), "A map for big data research in digital humanities", *Frontiers in Digital Humanities,* Vol 2 No 1.

Kamada, H. (2010), "Digital humanities: Roles for libraries?", *College & Research Libraries News,* Vol. 71 No. 9, pp.484-485.

Koltay, T. (2016), "Library and information science and the digital humanities", *Journal of Documentation,* Vol. 72 No.4, pp.781-792.

Lee, H.L. and Wang, S. (2018), "Investigating digital humanities: a domain analysis of conference proceedings published in Taiwan, 2009-2016", *Journal of Library & Information Studies,* Vol. 16 No. 2, pp.1-23.

Logsdon, A., Mars, A., and Tompkins, H. (2017), "Claiming expertise from betwixt and between: Digital humanities librarians, emotional labor, and genre theory", *College & Undergraduate Libraries*, Vol. 24 No. 2-4, pp. 155-170.

Millson-Martula, C. and Gunn, K. (2017), "The digital humanities: Implications for librarians, libraries, and librarianship", *College & Undergraduate Libraries,* Vol. 24, pp.135-139.

Münster, S. and Terras, M. (2020), "The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures", *Digital Scholarship in the Humanities,* Vol. 35 No. 2, pp.366-389.

Navarro-Colorado, B. (2018), "On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry", *Frontiers in Digital Humanities,* Vol. 5 No. 15.

Nikita, M. and Chaney, N. (2020), "ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters", available at: https://cran.r-project.org/web/packages/ldatuning/index.html (accessed 20 March 2021)

Nyhan, J. and Flinn, A. (2016), *Computation and the Humanities Towards an Oral History of Digital Humanities*. Springer, Cham.

Padilla, T. (2016), "Humanities data in the library: Integrity, form, access", *D-Lib Magazine,* Vol. 22 No. 3.

Poole, A.H. (2017), "The conceptual ecology of digital humanities", *Journal of Documentation*, Vol. 73 No. 1, pp.91-122.

Poole, A.H. and Garwood, D.A. (2018), "Interdisciplinary scholarly collaboration in data-intensive, public-funded, international digital humanities project work", *Library & Information Science Research,* Vol. 40 No. 3-4, pp.184-193.

Poole, A. H., & Garwood, D. A. (2020). "Digging into data management in public-funded, international research in digital humanities", *Journal of the Association for Information Science and Technology,* Vol. 71 No. 1, pp. 84-97.

Poremski, M.D. (2017), "Evaluating the landscape of digital humanities librarianship", *College & Undergraduate Libraries,* Vol. 24 No. 2-4, pp.140-154.

Siemens, L., Cunningham, R., Duff, W., & Warwick, C. (2010). "More Minds are Brought to Bear on a Problem": Methods of Interaction and Collaboration within Digital Humanities Research Teams. Digital Studies/le Champ Numérique, Vol. 2 No. 2. DOI: http://doi.org/10.16995/dscn.80

Siemens, L., Cunningham, R., Duff, W., and Warwick, C. (2011). "A tale of two cities: Implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities", *Literary and linguistic computing,* Vol. 26 No. 3, pp. 335-348.

Siemens, R., Dobson, T., Ruecker, S., & Cunningham, R. (2016). Human–computer interface/interaction and the book: a consultation-derived perspective on foundational e-book research. In Collaborative research in the digital humanities (pp. 175-202). Routledge.

Su, F. (2020), "Cross-national digital humanities research collaborations: structure, patterns and themes", *Journal of Documentation,* Vol. 76 No. 6, pp.1295-1312.

Sula, C.A. (2013), "Digital humanities and libraries: A conceptual model", *Journal of Library Administration,* Vol. 53 No. 1, pp.10-26.

Tang, M.C., Cheng, Y.J., and Chen, K.H. (2017). "A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses", *Scientometrics,* Vol. 113 No. 2, pp.985-1008.

Tomasi, F. (2018). "Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain", *Historical Social Research/Historische Sozialforschung.* Vol. 31, pp.170-179.

Van Eck, N.J. and Waltman, L. (2019). "VOSviewer manual. Leiden, Netherlands: Univeristeit Leiden", available at https://wwwvosviewer.com/documentation/Manual_VOSviewer_1.6.10.pdf (accessed 20 March 2021)

Wang, X. and Inaba, M. (2009), "Analyzing structures and evolution of digital humanities based on correspondence analysis and co-word analysis", *Art research,* Vol. 9, pp.123-134.

Wang, Q. (2018), "Distribution features and intellectual structures of digital humanities", *Journal of Documentation.* Vol. 74 No. 1, pp.223-246.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445-1456.

Zeng, L.M., and Chen, S.S.J. (2018), "Knowledge Organization and Cultural Heritage in the Semantic Web-A Review of a Conference and a Special Journal Issue of JLIS", *DHQ: Digital Humanities Quarterly,* Vol. 12 No. 1, pp.1-6.

Table 1. Most frequent stemmed terms in each corpus

Abstract terms				Subject terms			
Rank	Term	Frequency	Percent	Rank	Term	Frequency	Percent
1	digit	5361	2.37%	1	digit	2886	7.02%
2	human	3574	1.58%	2	human	2201	5.35%
3	research	2865	1.27%	3	data	784	1.91%
4	data	2035	0.90%	4	comput	724	1.76%
5	studi	1651	0.73%	5	inform	669	1.63%
6	project	1644	0.73%	6	research	657	1.60%
7	develop	1267	0.56%	7	system	553	1.35%
8	paper	1232	0.54%	8	librari	520	1.26%
9	work	1197	0.53%	9	histori	508	1.24%
10	articl	1179	0.52%	10	analysi	408	0.99%
11	inform	1156	0.51%	11	cultur	373	0.91%
12	text	1151	0.51%	12	visual	369	0.90%
13	present	1055	0.47%	13	languag	362	0.88%
14	cultur	1048	0.46%	14	social	344	0.84%
15	analysi	1007	0.45%	15	semant	328	0.80%
16	tool	1000	0.44%	16	network	308	0.75%
17	method	988	0.44%	17	technolog	306	0.74%
18	collect	984	0.43%	18	scienc	294	0.72%
19	model	944	0.42%	19	archiv	263	0.64%
20	scholar	931	0.41%	20	histor	260	0.63%
21	technolog	923	0.41%	21	process	259	0.63%
22	histori	914	0.40%	22	model	251	0.61%
23	librari	894	0.40%	23	text	248	0.60%
24	histor	883	0.39%	24	learn	246	0.60%
25	approach	875	0.39%	25	knowledg	241	0.59%
26	provid	814	0.36%	26	web	239	0.58%
27	discuss	799	0.35%	27	educ	234	0.57%
28	comput	778	0.34%	28	studi	219	0.53%
29	archiv	769	0.34%	29	heritag	218	0.53%
30	practic	757	0.33%	30	manag	187	0.45%

Table 2. Most frequent bigrams

Abstracts			Subject terms		
Rank	Bigram	Frequency	Rank	Bigram	Frequency
1	digital humanities	2244	1	digital humanities	1690
2	cultural heritage	220	2	humanities digital	222
3	humanities research	193	3	digital libraries	188
4	case study	167	4	information systems	141
5	paper presents	124	5	geographic information	110

6	humanities dh	116	6	semantic web	105
7	digital scholarship	108	6	natural language	105
8	big data	104	8	language processing	104
9	humanities scholars	101	9	cultural heritage	90
10	social sciences	94	10	humanities computing	88
10	social media	94	11	information retrieval	83
12	field digital	91	12	data mining	82
13	digital tools	90	13	humanities research	78
14	article discusses	89	14	cultural heritages	74
15	research data	78	15	big data	73
16	digital technologies	77	16	processing systems	72
17	humanities projects	76	17	linked data	62
18	case studies	75	18	history digital	61
19	digital libraries	74	19	character recognition	59
20	linked data	73	20	text mining	56
20	digital media	73	21	network analysis	56
22	humanities project	72	22	heritage digital	54
23	open data	70	23	linked open	53
24	use digital	69	23	open data	53
25	computer science	68	25	data visualization	51
25	research project	68	26	machine learning	50
27	historical research	67	27	user interfaces	47
28	paper present	65	27	artificial intelligence	47
29	design methodology	64	29	social sciences	46
30	methodology approach	63	29	social networking	46
30	computational methods	63	31	social media	45
30	originality value	63	31	united states	45
30	digital resources	63	33	humanities historical	44

Table 3. Analysis of bigram trends (document frequency)

2010–2014 (665 documents)		2015–2017 (1078 documents)			2018–2020 (974 documents)		
Bigram	Rank	Bigram	Rank	Score	Bigram	Rank	Score
digital humanities	1	humanities dh	1	25.93	machine learning	1	17.96
humanities research	2	digital scholarship	2	23.03	cultural heritage	2	17.44
article discusses	3	case study	3	21.61	natural language	3	16.83
cultural heritage	4	humanities projects	4	19.79	also mentions	4	13.00
case study	5	paper presents	5	18.58	data available	5	12.10

field digital	6	humanities project	6	18.03	one hand	6	11.58
humanities scholars	7	language processing	7	15.76	language processing	7	10.83
paper present	8	text analysis	8	15.14	studies digital	8	10.58
paper presents	9	distant reading	9	14.14	topic modeling	8	10.58
social media	10	textual data	10	14.00	virtual reality	10	10.39
digital technology	11	research data	11	13.90	widely used	10	10.39
use digital	12	research projects	12	13.52	paper argues	12	10.29
big data	13	digital humanities	13	13.24	allow us	13	10.19
paper describes	14	purpose paper	14	13.14	second part	14	10.10
digital libraries	15	recent years	15	12.89	wide range	15	9.87
linked data	16	open data	16	12.65	article focuses	16	9.36
article presents	16	computational methods	17	12.52	data science	17	9.19
arts humanities	16	using digital	18	12.52	american studies	18	9.10
research project	16	natural language	18	12.52	archives museums	18	9.10
digital library	16	digital humanists	20	12.27	humanities dh	20	8.53
new media	16	supporting digital	21	11.00	spatial humanities	21	8.19
digital resources	22	purpose purpose	22	10.38	results show	22	8.16
historical research	22	article describe	23	10.00	topic modelling	23	8.10
computer science	22	domain experts	24	9.38	libraries museums	23	8.10
within digital	22	network analysis	25	9.27	practical implications	25	7.77
digital media	22	project management	26	9.00	social sciences	26	7.41
research questions	27	makes possible	26	9.00	large amounts	27	7.39
higher education	28	two case	26	9.00	neural networks	28	7.29
digital archives	28	results show	29	8.76	new perspectives	29	7.29
humanities social	28	text mining	30	8.65	open access	30	7.25

Table 4. Topic modeling result (ordered by probability)

Topic	Topic label	Most probable stemmed terms
T20	DH research	digit, research, human, librari, articl, includ, discuss, univers, develop, inform
T35	Linked open data	data, research, link, use, open, human, paper, big, applic, service
T29	Text mining	analysi, method, text, use, tool, techniqu, studi, corpus, mine, methodolog
T7	DH librarianship	digit, librari, librarian, human, scholar, scholarship, role, academ, research, support
T48	Collaboration work	project, collabor, digit, work, practic, team, initi, research, studi, manag
T9	Visualization	visual, human, digit, research, present, issu, uncertainti, within, support, design
T63	DH technology	digit, human, technolog, practic, work, within, critic, way, engag, object
T31	Semantic web and ontology	knowledg, semant, ontolog, link, use, web, paper, base, present, entiti
T15	Literature	studi, digit, field, human, literatur, will, approach, disciplin, practic, archaeolog
T36	History	histori, histor, historian, scienc, digit, articl, studi, research, oral, new
T42	Digital archives and preservation	archiv, digit, web, preserv, record, materi, sourc, document, collect, access
T33	Text annotations	annot, tool, corpus, text, digit, human, user, use, process, paper
T18	Text digitization and OCR	text, use, charact, histor, name, digit, document, ocr, ancient, can
T30	Medieval manuscripts	scholar, work, edit, manuscript, digit, mediev, new, author, articl, scholarship
T27	System model	model, use, system, approach, human, differ, propos, implement, inform, present
T52	Learning and education	student, learn, cours, teach, use, educ, digit, experi, human, undergradu
T40	University and libraries	univers, librari, educ, digit, collect, public, institut, school, faculti, partnership
T4	Literary criticism	literari, read, text, literatur, studi, close, digit, critic, method, distant
T62	Social network analysis	network, social, can, differ, relationship, studi, charact, use, graph, structur
T44	Cultural heritage	cultur, heritag, museum, object, digit, use, inform, technolog, collect, system
T23	DH resources	digit, human, resourc, impact, research, studi, use, develop, field, technolog
T25	Linguistics	languag, linguist, translat, semant, one, use, context, corpus, natur, present
T34	Diversity (Black and women)	digit, human, new, black, articl, practic, studi, women, chang, research
T50	Infrastructure	research, infrastructur, use, paper, digit, human, scholar, twitter, practic, scienc
T39	Cultural heritage and community engagement	digit, cultur, collect, user, heritag, engag, communiti, differ, interact, content
T3	Computational science	comput, human, scienc, copyright, develop, work, use, field, new, research
T24	Questions and answers	human, digit, question, author, method, use, articl, answer, shakespear, comput
T10	Arts	art, imag, artist, digit, work, collect, use, histori, experi, artwork

T12	Maps and spatial	spatial, map, place, space, histor, geograph, use, represent, narrat, deep
T60	Document and image collections	document, imag, collect, data, qualiti, histor, newspap, digit, extract, use
T37	Social media	media, social, cultur, digit, platform, technolog, new, emerg, use, product
T32	African American	american, cultur, nation, space, articl, african, world, urban, steampunk, map
T51	Crowdsourcing	task, learn, effect, use, particip, crowdsourc, studi, perform, result, motiv
T55	Information technology	inform, technolog, develop, knowledg, digit, infrastructur, research, polici, organ, human
T17	Collections and databases	collect, databas, paper, digit, set, use, brows, imag, process, can
T14	Digitization projects	project, digit, innov, univers, lab, develop, use, digitis, will, transcrib
T38	Geographic data and GIS	geograph, gis, histor, data, geographi, map, inform, use, spatial, develop
T11	Books and citations	book, use, citat, digit, inform, librari, publish, cite, resourc, imag
T26	DH project	digit, human, project, process, focus, research, also, design, platform, differ
T45	History: war, politics	war, memori, histor, cultur, polit, use, world, period, past, state
T19	Journals	journal, period, websit, digit, index, issu, articl, studi, perform, theatr
T2	Metadata	format, workshop, describ, metadata, process, tool, digit, can, human, tei
T49	Rhetoric and essays	rhetor, digit, human, univers, practic, essay, theori, press, modern, paper
T54	Music and musicology	music, abstract, may, copyright, copi, user, publish, email, articl, musicolog
T46	Content management	content, manag, assign, present, student, event, servic, issu, support, resourc
T41	Language analysis	word, use, chang, studi, result, languag, mean, predict, differ, set
T61	Time analysis (history)	time, analysi, use, histor, method, chang, can, quantit, studi, word
T6	Conference papers	human, confer, research, digit, new, scienc, paper, address, comput, intern
T64	Video games and virtual reality	game, video, cultur, realiti, use, virtual, play, studi, effect, educ

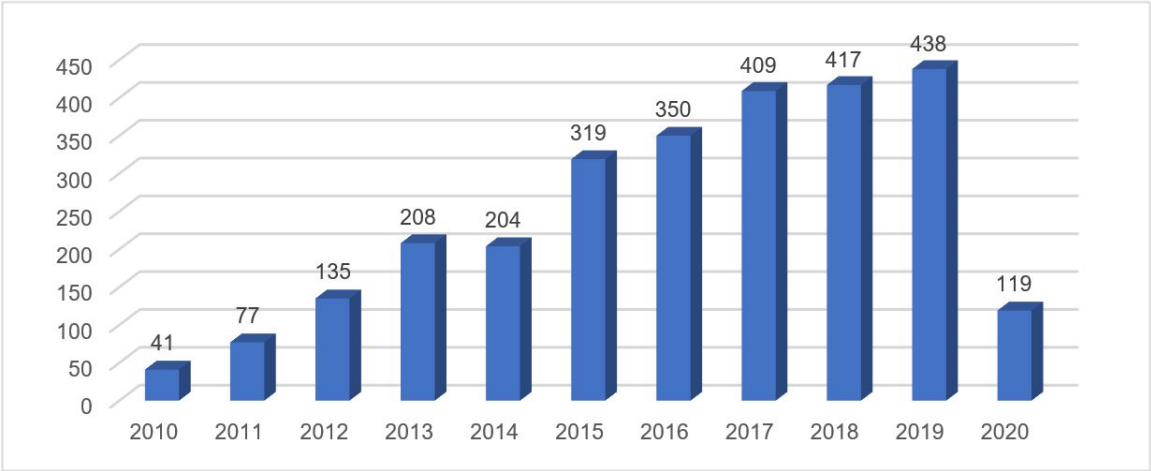

Figure 1. Number of articles by year

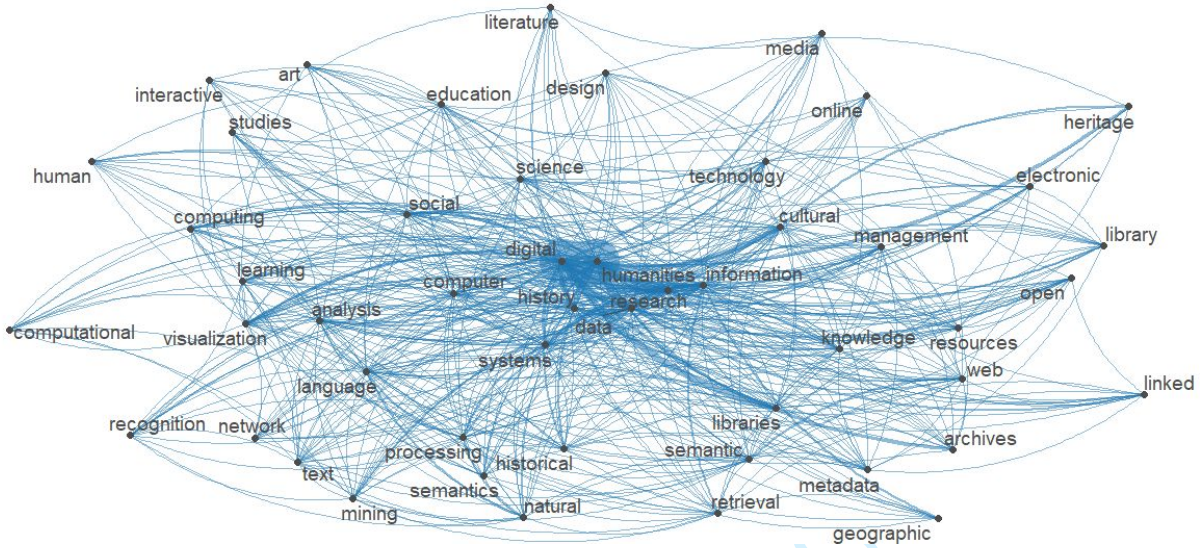

Figure 2. A network visualization of keyword co-occurrences

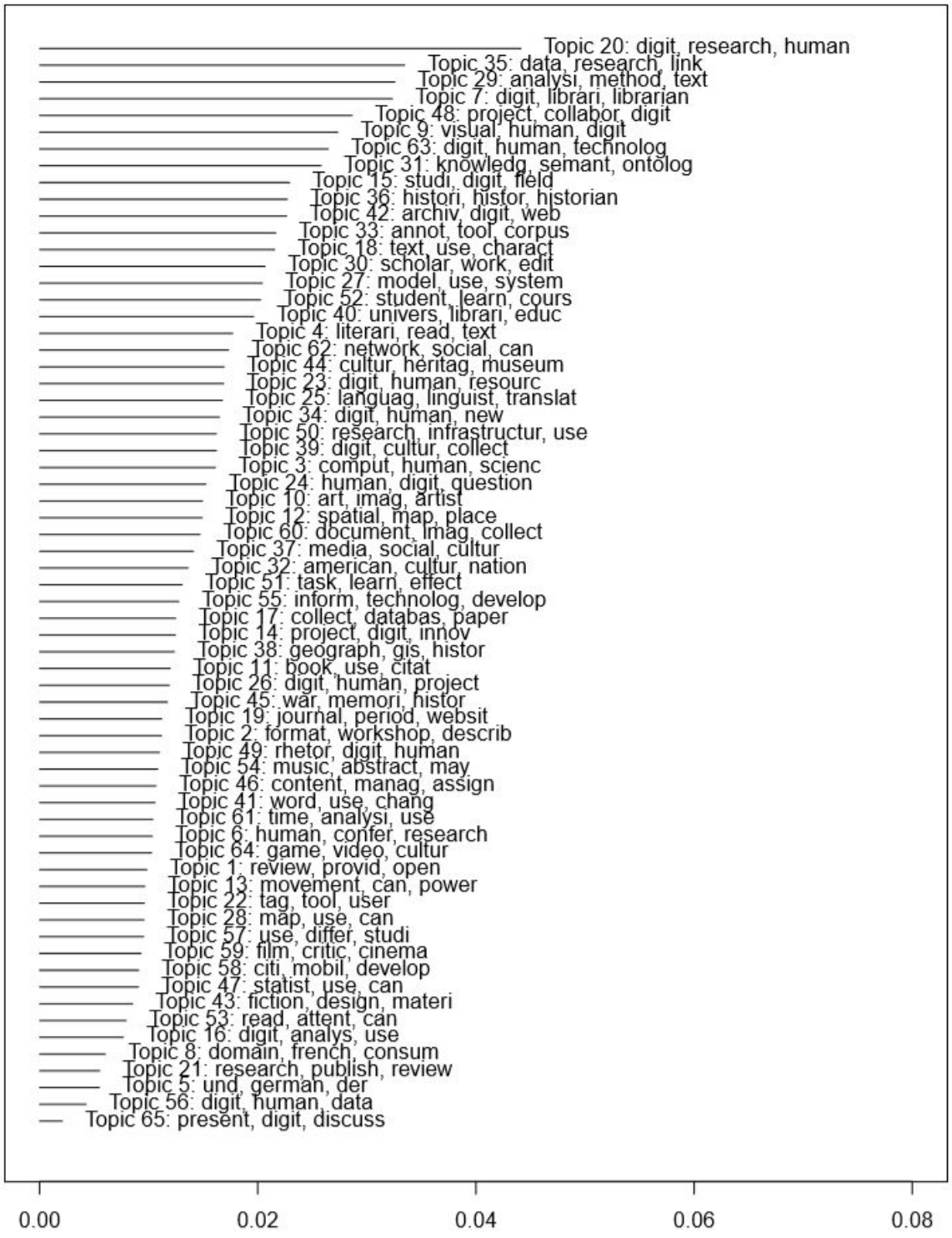

Figure 3. Estimated probabilities of topics

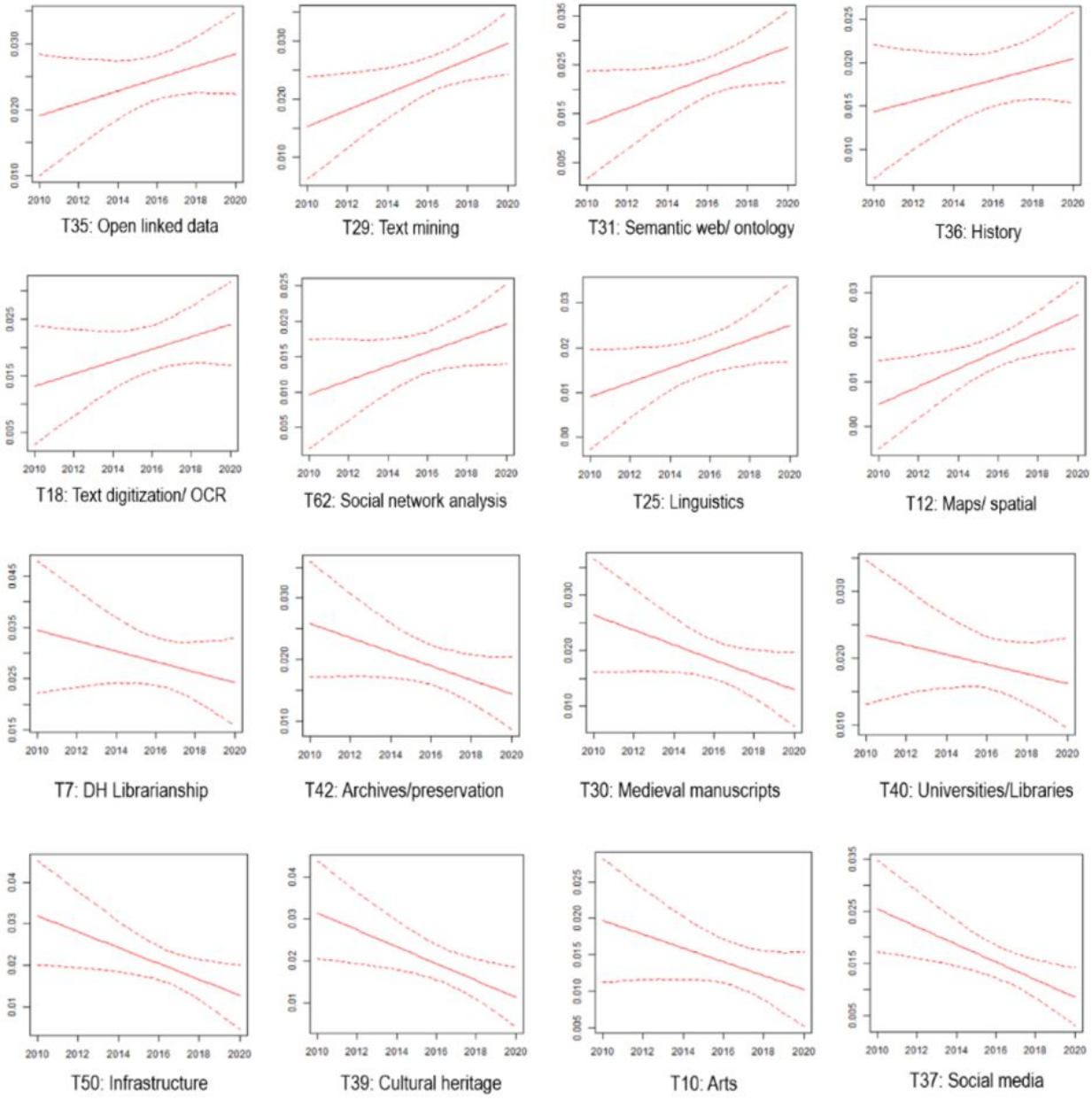

Figure 4. Select hot and cold topics

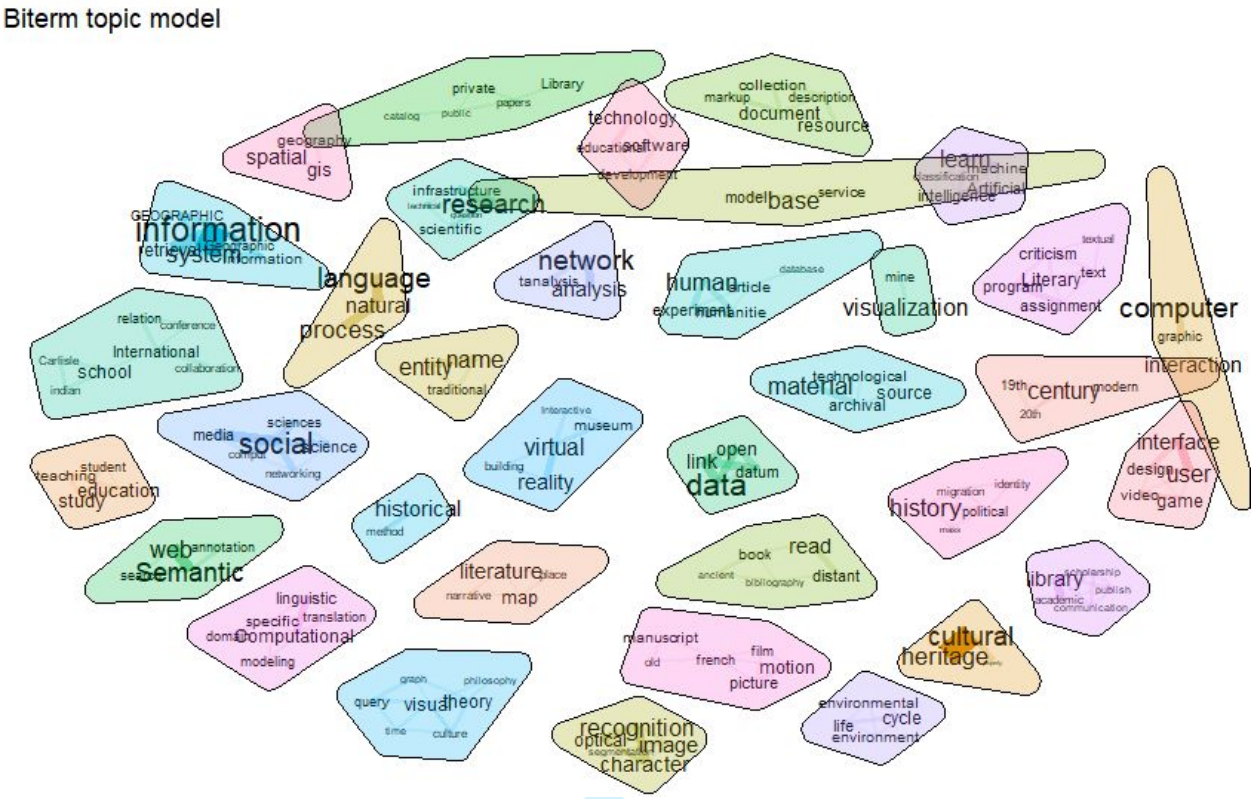

Figure 5. Bi-term topic model based on the subject terms