Community Structure and Coherence in Digital Humanities Works

Shera Potka
University of Victoria, Canada
spotka@uvic.ca

Alex Thomo
University of Victoria, Canada
thomo@uvic.ca

***Abstract*—Digital Humanities is an interdisciplinary field that connects technical disciplines such as Computer Science and Digital Technologies with a vast array of disciplines in Humanities research. This paper presents a Bibliometrics approach to analysing a decade-long corpus of Digital Humanities works collected from Google Scholar. The goal of the paper is to determine the community structure and the cohesion of the publication network in Digital Humanities as measured by the degree of the interconnectedness between the publications. Previous works focus mainly on co-authorship, co-citation, and co-bibliographic research to study the structure of Digital Humanities research. Such approaches are brittle and do not offer a robust way to achieve the stated goal because they only serve as approximations of the degree of interconnectedness between documents. In contrast to previous work, this paper embarks in a direct study of publication interconnectedness by focusing on the document text similarity networks. The main research questions addressed are as follows. Which are the main research interests manifested in the Digital Humanities literature over the last decade? Are the main research interests found "in the wild" (via Google Scholar) well-recognized by authorities in the field? How have those research interests evolved over time? Are research interests growing in number thus becoming more diverse or are they shrinking and becoming more cohesive? This paper strives to provide answers to these questions based on community discovery algorithms and network cohesion measures.**

***Index Terms*—Digital Humanities, Social Network Analysis, Text Analysis, Community Discovery**

I. Introduction

Over the past decade, there has been a growing interest in using computational methods to analyze and interpret large amounts of data in the humanities, such as text, images, and audio. Digital Humanities brings together a range of disciplines from the Humanities, Computer Science, and Digital Technology, thus forming a multi-disciplinary field of study. In his book, "Humanities Computing," Willard McCarty charts the development of the field from "computing and the humanities" to "computing in the humanities" and finally to "humanities computing." He views these three phases as a relationship that was once aspirational but limited, that then became established, and finally became self-aware but enigmatic [1].

Interdisciplinary research in Digital Humanities emphasizes the cognitive integration of concepts, theories, methods, and results from various fields. According to Rafols and Meyer [2], "knowledge integration" involves high cognitive heterogeneity and an increase in relational structure, referred to as coherence, which denotes the interrelationships among particular topics, concepts, and tools. Bibliometrically, coherence is measured by the tightness or looseness of fundamental bibliographic components, such as authors, articles, keywords, or publication sources, in a literature set. Additionally, the concept of "cohesion" has been used to describe knowledge integration among various subspecialties within a discipline, research field, or scientific community [2].

Despite a shared methodological approach among research initiatives in Digital Humanities, it remains unclear whether the field has become more consolidated or has remained fragmented over time. Prior studies have primarily focused on analyzing the structure of Digital Humanities research through co-authorship, co-citation, and co-bibliographic methods. However, these approaches are limited and do not provide a reliable means of achieving the intended objective, as they only offer an estimate of the level of similarity-based interconnectedness between documents. In contrast, this paper takes a novel approach by directly studying publication interconnectedness through building and analyzing text similarity networks. This approach has been largely avoided in previous works due to the difficulties in compiling a substantial corpus of documents and the computational challenges involved in conducting text similarity analysis and constructing networks.

This study examines a manually collected set of over 2000 Digital Humanities documents from Google Scholar covering the last decade. A number of document similarity techniques were thoroughly analyzed and evaluated, and the most suitable method was selected for further analysis. Using pairwise similarities of approximately two million document pairs from the collected documents, this research creates similarity networks for each of the ten years in focus, offering a comprehensive longitudinal examination of the community structure and cohesiveness of Digital Humanities works over the past decade. The aim of the study is to uncover the dominant research interests in the field of Digital Humanities, determine if these interests are recognized by authoritative figures in the field, and analyze how they have evolved over time. It also seeks to determine whether the research interests are becoming more diverse or more cohesive, either increasing in number or decreasing. To address these objectives, the paper employs community discovery algorithms and network cohesion measures.

II. Research Questions and Related Works

This section presents the research questions addressed in the paper and discusses relevant literature, highlighting the distinctiveness of this work compared to previous studies.

A. Research Questions

Specifically, we address the following research questions:

RQ1 Which of the major research areas of Digital Humanities dominate the literature of the last decade?

RQ2 Do the areas recognized by experts in the field appear in the literature indexed by Google Scholar?

RQ3 How have research interests in specific areas changed over time?

RQ4 Are the research interests expanding, leading to increased diversity, or are they contracting and becoming more unified?

We seek to answer these questions by using text analysis methods, community discovery algorithms, and measures of network cohesion.

B. Related Works

Digital Humanities, being a field in constant flux, has drawn attention from a variety of knowledge domains and expertise, resulting in its diverse disciplinary and institutional makeup [3]. Since large-scale observation of research integration during the research process is challenging, researchers commonly deduce knowledge integration in a field through its resulting literature, specifically, its published works [4].

Prior studies have used network analysis to measure the interconnection and integration within research communities. In particular, [5] analyzed the co-authorship network of Sociology and identified different types of network structures. They believed that a structurally cohesive network indicates permeable theoretical boundaries and cross-fertilization among scholars. [6] used server log data from a prominent Education journal to examine the integration of ideas and practices within the discipline and found a network that shows both small-world and structural cohesive characteristics. [7] studied co-authorship networks to discover patterns of cooperation in the field and found the networks to exhibit a small-world structure.

In the comprehensive work of [8], co-authorship, co-citation, and co-bibliographic networks were generated from literature published in prominent Digital Humanities journals. Social network analysis was then applied to measure their interconnectedness and degree of integration. The network topology was examined to provide a deeper understanding into scholarly practices, collaborative patterns, interdisciplinarity, and the state of ”cognitive consensus” in the Digital Humanities field.

1) How Is the Present Paper Different?: The present paper sets itself apart by utilizing the actual textual content of research articles to measure similarity and generate networks that reveal the diversity and integration of the Digital Humanities field. This is a direct, and more challenging, way to achieve the goal and it stands in clear contrast to the indirect ways that all the aforementioned related works follow.

Areas	Abbrev	Ref
Network Analysis	NetA	[10], [11]
Data Visualization	DataViz	[12], [13]
Machine Learning	ML	[14]
Semantic Web	SemWeb	[15], [16]
Virtual and Augmented Reality	VR-AR	[17], [18]
Digital Education and Online Pedagogy	DEdu	[19], [20]
Digital Mapping and Spatial Analysis	DMap	[21], [22]
Digital Storytelling	DStory	[23], [24]
Digital Art and Design	DArt	[24], [25]
Digital Musicology	DMusic	[26], [27]
Digital Archaeology	DArcha	[28], [29]
Digital History	DHist	[30], [31]
Digital Libraries and Archives	DLib	[32], [33]
Digital Literature, Rhetoric and Writing	DLit	[34], [35]
Digital Sociology and Social Networks	DSoc	[36], [37]
Digital Culture and Philosophy	DCulture	[38], [39]
Digital Gender and Sexuality	DGender	[40], [41]
Digital Economy and Business	DEcon	[42], [43]
Digital Law and Policy	DLaw	[44], [45]
Digital Health and Medicine	DHealth	[46], [47]
Digital Organisation and Access	DOrg	[48], [49]
Digital Urbanism	DUrb	[50], [51]
Digital Design and Architecture	DArchitect	[52], [53]
Digital Media and Communication	DMedia	[54], [55]
Digital Language and Linguistics	DLang	[56], [57]

TABLE I
Research Areas in Digital Humanities

For example, co-authorship, co-citation, and co-bibliographic networks are only a proxy for the real similarity between documents. Documents can be vastly different and still share authors, citations, or be cited by common papers. As such, our approach that considers directly the textual content of the articles is more robust and has the potential to provide better and more well-grounded insights.

C. Research Areas in Digital Humanities

The areas of Digital Humanities can be difficult to define and categorize because the field is interdisciplinary and encompasses a wide range of topics and technologies. It draws from the humanities, social sciences, computer science, information science, and other disciplines, and its scope can include areas such as digital literary studies, digital history, digital art history, digital musicology, digital cultural heritage, and digital media studies, among others. Additionally, the field is rapidly evolving and new areas are constantly emerging, making it difficult to keep track of all the different areas and developments. Furthermore, many digital humanities projects and initiatives blur the boundaries between traditional disciplines, making it challenging to categorize them neatly into distinct areas. A list with collected areas from many sources in Digital Humanities and representative publications is given in Table I. Detailed descriptions of each area and their interconnections are given in [9]. One of the main goals of our research is to investigate which of these major areas are well represented in the literature of the last decade.

III. Methodology

Our methodology involves various algorithms and techniques, including text similarity and social network analysis.

To determine the best text similarity method for Digital Humanities documents, we evaluate Term Frequency and Inverse Document Frequency (TFIDF), Universal Sentence Encoder (USE), and Bidirectional Encoder Representations from Transformers (BERT). In terms of network analysis, the Louvain modularity algorithm is applied to identify tightly clustered groups of research documents, revealing the research areas within the top communities each year. Lastly, this section explains the methods used for automatic topic extraction in order to extract topics from the text content of each community.

A. Document Similarity Methods

TF-IDF stands for term frequency – inverse document frequency. It is a weighting scheme that is used to capture how important a word (term) is to a document in a collection of documents (c.f. [58]). The basic idea behind TF-IDF is to weight words based on how often they appear in a document, and how rare they are across all documents in a collection.

One way to use TF-IDF, or the other document encoding schemes in order to determine document similarity is to compute the cosine similarity between the document vectors. Cosine similarity can range from -1 to 1, with a value of 1 indicating that the vectors are identical, a value of 0 indicating that the vectors are orthogonal (unrelated), and a value of -1 indicating that the vectors are completely dissimilar.

Doc2Vec is a method for representing documents as vectors in a high-dimensional space (c.f. [59]). The main idea behind Doc2Vec is to learn a fixed-length vector representation for each document, such that the vectors for semantically similar documents are close to each other in the vector space, while the vectors for dissimilar documents are far apart. This is achieved by training a neural network to predict the context of a word given the word itself.

Universal Sentence Encoder (USE) is a pre-trained neural network model that generates embeddings for text sentences (c.f. [60]). The model is trained on a wide variety of text data and is able to generate embeddings for a wide range of natural language understanding tasks such as semantic similarity, text classification, and question answering. USE is designed to generate embeddings for individual sentences, as the name implies. However, it is not restricted to only inputting single sentences. The official documentation does not specify a limit on the input size, therefore it can be utilized for tasks such as comparing entire documents. The entire document can be input into the USE model as-is, without the need for language processing.

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained transformer-based neural network model developed for natural language processing tasks (c.f. [61]). It is designed to generate embeddings (vectors) for text input such as sentences, paragraphs or entire documents. BERT uses a transformer architecture which allows it to learn the word context from both the previous and the later part of the input, hence the name "bidirectional". Unfortunately, BERT turned out to be too slow for the relatively long documents in this study, thus not being able to complete encoding of documents, even when restricted to very small subsets. We experimented with one of the state-of-the-art implementations from Sentence-Transformers (https://www.sbert.net), which uses the well-known collection of HuggingFace Model Hub.

B. Similarity Graph of Documents

A similarity graph of documents is a network-based representation of the similarities between documents in a given collection. Each document is represented as a node in the graph, and edges between nodes indicate the similarity between the documents. The strength of the edge between two nodes can be determined using a similarity measure as those described above. The resulting graph can be used to explore the relationships between the documents, identify groups of similar documents, or as input for other natural language processing tasks such as document clustering, classification or retrieval.

Formally, the *similarity graph of documents* is a weighted undirected graph $G = (V, E, S)$, where V is the set of nodes representing documents, E is the set of edges or connections representing similarities between documents, and $S : E \to [-1, 1]$ is a weighting function, where for each edge $e \in E$, $S(e)$ is implemented using the cosine similarity for the pair documents at the endpoints of edge e.

1) Community Detection Via Modularity Maximization: Discovering communities of closely connected people in a social network is one of the most important problems in network analysis (c.f. [11]). The notion of communities as dense connected clusters of nodes in a network can be naturally extended to the analysis of the similarity graph of documents which is the focus of this study. Despite the potential peculiarity of referring to document clusters as "communities," this term is utilized in this study to maintain consistency with the nomenclature used in network analysis literature.

Community detection via modularity maximization is a method for identifying groups of nodes (i.e. communities) in a network that are more densely connected to each other than to the rest of the network. The basic idea behind this method is to divide a network into groups of nodes such that the edges within groups are more numerous than the edges between groups (c.f. [62]). The most common algorithm for finding modularity maximising communities is called the Louvain algorithm [62].

2) Community Topic Extraction: After extracting communities of documents based on their similarities, the text of documents in each community is further processed in order to automatically extract topics. Topic modeling is a method used to discover general topics in a large collection of documents without reading them all. It uses algebraic and statistical techniques to identify topics based on the frequency of words used together. The goal of topic modeling is to uncover the overall structure of a collection of documents, not just to assign topics to individual documents. If the collection has a defined structure, such as categories or keywords, topic modeling can reveal its hidden structure.

Topic 00	Topic 01	Topic 02	Topic 03	Topic 04
digital (1.48)	students (1.08)	dati (0.95)	citizen (0.73)	electronic (0.93)
humanities (1.11)	education (0.67)	verbaalpina (0.67)	museum (0.59)	literature (0.91)
research (0.53)	learning (0.66)	progetto (0.38)	digital (0.49)	digital (0.64)
preservation (0.36)	teaching (0.59)	ricerca (0.36)	heritage (0.35)	humanities (0.45)
data (0.36)	teachers (0.32)	accesso (0.34)	public (0.34)	gutenberg (0.39)
DOrg	**DEdu**	**DOrg**	**DHist**	**DLib**
Topic 05	Topic 06	Topic 07	Topic 08	Topic 09
open (0.71)	language (1.69)	visualization (1.50)	school (2.46)	digitalization (2.58)
publishing (0.68)	reading (1.47)	students (1.20)	administration (2.31)	remembrance (1.92)
access (0.57)	english (1.46)	data (0.85)	personnel (1.78)	codex (1.25)
research (0.49)	learners (0.95)	jänicke (0.80)	staff (1.37)	digital (0.80)
scholarly (0.43)	digital (0.84)	course (0.62)	competence (1.28)	biblical (0.71)
DOrg	**DLang**	**DataViz, DEdu**	**DEdu**	**DHist, DRelig**

Fig. 1. Example of topics extracted using NMF

Year	Number of files	Size in MB
2013	207	331
2014	205	401
2015	201	354
2016	210	312
2017	261	314
2018	190	384
2019	213	487
2020	234	488
2021	222	501
2022	258	590
Total	**2201**	**4162**

Fig. 2. Dataset Statistics

In order to automatically extract topics from documents belonging to a community, Non-negative Matrix Factorization (NMF) was used in this study [63]–[65]. NMF involves the factorization of a non-negative matrix into two lower-dimensional non-negative matrices. The lower-dimensional matrices, also known as factors, can be interpreted as the underlying topics present in the original matrix. The technique has been found to be effective in extracting topics from large collections of text data and is commonly used in natural language processing applications. An example of 10 topics extracted from the top document community of year 2020 is given in Figure 1. Along with each topic, an approximate matching with one of the areas in Section II-C is also given.

We can examine the contribution of words to each topic in terms of percentages. With a large number of words, the individual contributions are relatively small, with the exception of "school", "administration", and "digitalization", in Topic 07 and Topic 09. Still, the percentage of words within a topic provides a valuable insight into the quality of the topic model. If the percentages rapidly decrease within a topic, the topic is well-defined, while a gradual decrease in word probabilities suggests a less distinct topic [66]. In this example we see that the topics make sense in the framework of the Digital Humanities areas identified in Section II-C.

Finally, another popular topic modelling method is LDA (Latent Dirichlet Allocation) [67]. It is a generative probabilistic model that assumes that each document is generated by a mixture of topics, where each topic is defined as a distribution over words. However, LDA has some limitations such as the difficulty in setting the number of topics, the sensitivity to hyperparameters, and the scalability issue even with moderate text collections [67]. The latter was a challenging problem for our collection of relative large documents. The lack of scalability did not allow experiementing with LDA in this study.

IV. Results

A. Dataset Collection

At the conclusion of December 2022, a careful collection process was undertaken to gather a complete and thorough dataset consisting of 2201 Digital Humanities documents from Google Scholar. The method of collection involved the use of a precise query on Google Scholar, which was defined as "Digital Humanities filetype:pdf". To further refine the search results, the inquiry was specifically targeted towards each year within the 2013-2022 decade. Information about the dataset is given in Figure 2.

In addition, to the above comprehensive dataset, another, smaller dataset consisting of 130 Digital Humanities documents was procured from the Canadian HSS Commons - Community for Humanities and Social Sciences website hsscommons.ca/publications. The Canadian HSS Commons offers a collection of carefully curated Digital Humanities documents. This study incorporates these documents with the aim of conducting a detailed analysis and comparison of various document similarity methods as they pertain to Digital Humanities collections. The results of this comparison will then be utilized to determine the most effective method for analyzing the larger collection obtained from Google Scholar.

B. Evaluating Similarity

An evaluation was conducted to determine the effectiveness of similarity measures described in Section III-A for academic works in the field of Digital Humanities. A set of documents from the HSS Commons Collection, containing author-specified keywords, was selected. The similarity measures were then applied to every possible pair of these documents, and the top-10 pairs of most similar documents for each measure were extracted. The results of the evaluation are depicted in Figures 3, 4, and 5.

The quality of the similarity measures was quantitatively evaluated using the sets of keywords provided by the authors of the documents. The overlap between the sets of keywords for each pair of documents was computed by determining the number of common keywords for the pair, with each keyword being treated individually rather than as a phrase. The size of overlap (intersection) between the two columns, Keywords 1 and Keywords 2, corresponding to Document 1 and Document 2, is shown in the last column called Overlap. The average overlap for each similarity measure was then calculated, with a higher average overlap indicating a better quality of the similarity measure. Conclusions were drawn about the effectiveness of each similarity measure and which method was best suited for the study through this process.

After conducting a thorough analysis, it was determined that the TF-IDF measure of similarity outperforms other methods in determining the similarity between documents. In particular, the average keyword overlap for the top-10

Score	Document 1	Keywords 1	Document 2	Keywords 2	Overlap
0.78	Siemens 2012: Embedding Small Business and Entrepreneurship Training within the Rural Context	training; small business; rural entrepreneurship; economic	Siemens 2014: We moved here for the lifestyle	small business; entrepreneurship; rural areas; economic development;	4
0.76	Siemens and The INKE Research Group 2019: Developing an Open Social Scholarship Collaboration Lessons from INKE	collaboration; project management; INKE	Siemens 2016: Faster Alone Further Together	Collaboration; Networked scholarship; Research teams; Digital humanities;	1
0.59	Siemens and The INKE Research Group 2019: Developing an Open Social Scholarship Collaboration Lessons from INKE	collaboration; project management; INKE	Siemens and The INKE Research Group 2019: Joining Voices University Industry Partnerships in the Humanities	collaboration; university industry partnerships; INKE; INKE:NOSS	2
0.65	Arbuckle 2019: Open+: Versioning Open Social Scholarship	open scholarship; open access; community engagement; public humanities; digital	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	2
0.59	Arbuckle et al 2019: Introduction Beyond Open Implementing Social Scholarship	open social scholarship; scholarly communication; open access; open scholarship;	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	3
0.57	Arbuckle and Maxwell 2019: Modelling Open Social Scholarship Within the INKE Community	open access; open scholarship; scholarly communication; publishing	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	2
0.63	Arbuckle et al 2019: Introduction Beyond Open Implementing Social Scholarship	open social scholarship; scholarly communication; open access; open scholarship;	Arbuckle 2019: Open+: Versioning Open Social Scholarship	open scholarship; open access; community engagement; public humanities; digital	3
0.62	Arbuckle and Maxwell 2019: Modelling Open Social Scholarship Within the INKE Community	open access; open scholarship; scholarly communication; publishing	Arbuckle 2019: Open+: Versioning Open Social Scholarship	open scholarship; open access; community engagement; public humanities; digital	3
0.58	Arbuckle et al 2019: Introduction Beyond Open Implementing Social Scholarship	open social scholarship; scholarly communication; open access; open scholarship;	El Khatib et al 2019 Foundations for On Campus Open Social Scholarship Activities	social knowledge creation; open social scholarship; citizen scholar; scholarly	4
0.61	Arbuckle and Maxwell 2019: Modelling Open Social Scholarship Within the INKE Community	open access; open scholarship; scholarly communication; publishing	Arbuckle et al 2019 Introduction Beyond Open Implementing Social Scholarship	open social scholarship; scholarly communication; open access; open scholarship;	4
				Average	2.8

Fig. 3. Top 10 similar pairs of documents using TFIDF and Cosine Similarity.

Score	Document 1	Keywords 1	Document 2	Keywords 2	Overlap
0.70	Siemens 2012: Embedding Small Business and Entrepreneurship Training within the Rural Context	training; small business; rural entrepreneurship; economic	Siemens 2014: We moved here for the lifestyle	small business; entrepreneurship; rural areas; economic development;	4
0.69	Arbuckle and Maxwell 2019: Modelling Open Social Scholarship Within the INKE Community	open access; open scholarship; scholarly communication; publishing	Arbuckle 2019: Open+: Versioning Open Social Scholarship	open scholarship; open access; community engagement; public humanities; digital	2
0.63	Arbuckle and Maxwell 2019: Modelling Open Social Scholarship Within the INKE Community	open access; open scholarship; scholarly communication; publishing	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	2
0.64	Arbuckle 2019: Open+: Versioning Open Social Scholarship	open scholarship; open access; community engagement; public humanities; digital	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	2
0.63	El Khatib et al 2019: Foundations for On Campus Open Social Scholarship Activities	social knowledge creation; open social scholarship; citizen scholar; scholarly	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	4
0.61	Arbuckle et al 2019: Introduction Beyond Open Implementing Social Scholarship	open social scholarship; scholarly communication; open access; open scholarship;	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	3
0.60	Milligan et al 2019: The Initial Impact of the Open Scholarship Policy Observatory	scholarship; collaboration; open science	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	2
0.68	Siemens and The INKE Research Group 2019: Developing an Open Social Scholarship Collaboration Lessons from INKE	collaboration; project management; INKE	Siemens 2016: Faster Alone Further Together	Collaboration; Networked scholarship; Research teams; Digital humanities;	1
0.61	Siemens and The INKE Research Group 2019: Developing an Open Social Scholarship Collaboration Lessons from INKE	collaboration; project management; INKE	Siemens and The INKE Research Group 2019: Joining Voices UniversityIndustry Partnerships in the Humanities	collaboration; university industry partnerships; INKE; INKE:NOSS	1
0.60	Robinson and Saklofske 2017: Connecting the dots Integrating modular networks and narrativity in digital scholarship	Narrative, Networks, Modularity, Digital Scholarship, NewRadial	Saklofske 2015: New Radial Challenging scales and standards of humanities scholarship through new knowledge environment prototypes	Scales, standards, prototype, INKE, NewRadial, ontology, curation	0
				Average	2.1

Fig. 4. Top 10 similar pairs of documents using Doc2Vec and Cosine Similarity.

pairs of similar documents was found to be 2.8 when using the TF-IDF method. However, when utilizing Doc2Vec and USE, the average keyword overlap was only 2.1 and 2.2, respectively. Furthermore, the computation of document encodings using Doc2Vec and USE was significantly slower than that of the TF-IDF method, taking approximately an order of magnitude longer. Based on these findings, it can be concluded that the TF-IDF similarity is the most optimal and preferred method for this study, taking into consideration both accuracy and efficiency.

C. Community Areas

A similarity network of documents was constructed for each year, utilizing the TF-IDF document encoding method and the Cosine Similarity calculation. This network was created in such a way that each document was represented as a node, and the connections or edges between these nodes represented the similarity between the linked documents. In simpler terms, these networks were comprised of undirected but weighted graphs, which were then subjected to further analysis. In total, the construction process resulted in the creation of 10 separate networks.

Subsequently, the Louvain Modularity Algorithm was applied to each of the aforementioned networks in order to identify the partitioning of each network into communities of documents that were densely connected within each year. From each partitioning, the top three communities were carefully selected based on their size, resulting in a total of 30 communities being extracted, three communities for each year.

Score	Document 1	Keywords 1	Document 2	Keywords 2	Overlap
0.89	Siemens and The INKE Research Group 2019: Developing an Open Social Scholarship Collaboration Lessons from INKE	collaboration; project management; INKE	Siemens 2016: Faster Alone Further Together	Collaboration; Networked scholarship; Research teams; Digital humanities;	1
0.89	Siemens and The INKE Research Group 2019: Developing an Open Social Scholarship Collaboration Lessons from INKE	collaboration; project management; INKE	Siemens and The INKE Research Group 2019: Joining Voices University Industry Partnerships in the Humanities	collaboration; university industry partnerships; INKE; INKE:NOSS	1
0.80	Siemens 2016: Faster Alone Further Together	Collaboration; Networked scholarship; Research teams; Digital humanities;	Siemens and The INKE Research Group 2019: Joining Voices UniversityIndustry Partnerships in the Humanities	collaboration; university industry partnerships; INKE; INKE:NOSS	1
0.83	Siemens and The INKE Research Group 2019: Developing an Open Social Scholarship Collaboration Lessons from INKE	collaboration; project management; INKE	Siemens 2010: The Potential of Grant Applications as Team Building Exercises A Case Study	Collaboration, Research Teams, Grant Development, Research Offices, Case Study	1
0.80	Siemens 2016: Faster Alone Further Together	Collaboration; Networked scholarship; Research teams; Digital humanities;	Siemens 2010: The Potential of Grant Applications as Team Building Exercises A Case Study	Collaboration, Research Teams, Grant Development, Research Offices, Case Study	1
0.88	Arbuckle et al 2019: Introduction Beyond Open Implementing Social Scholarship	open social scholarship; scholarly communication; open access; open scholarship;	El Khatib et al 2019: Foundations for On Campus Open Social Scholarship Activities	social knowledge creation; open social scholarship; citizen scholar; scholarly	4
0.83	Arbuckle et al 2019 Introduction Beyond Open Implementing Social Scholarship	open social scholarship; scholarly communication; open access; open scholarship;	El Khatib et al 2019: Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	3
0.81	El Khatib et al 2019: Foundations for On Campus Open Social Scholarship Activities	social knowledge creation; open social scholarship; citizen scholar; scholarly	El Khatib et al 2019 Open Social Scholarship Annotated Bibliography	community; open; scholarship; social; technology	2
0.83	Arbuckle and Maxwell 2019: Modelling Open Social Scholarship Within the INKE Community	open access; open scholarship; scholarly communication; publishing	Arbuckle 2019: Open+: Versioning Open Social Scholarship	open scholarship; open access; community engagement; public humanities; digital	3
0.82	Siemens 2012: Embedding Small Business and Entrepreneurship Training within the Rural Context	training; small business; rural entrepreneurship; economic	Siemens 2014: We moved here for the lifestyle	small business; entrepreneurship; rural areas; economic development;	5
				Average	**2.2**

Fig. 5. Top 10 similar pairs of documents using USE and Cosine Similarity.

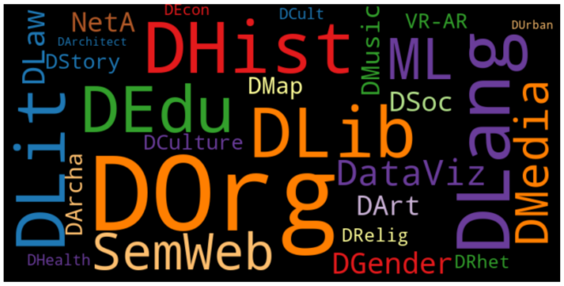

Fig. 6. Wordmap of areas for all the years, 2013-2022

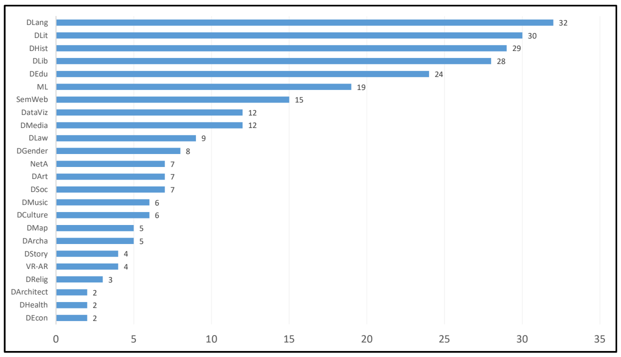

Fig. 7. Barchart of areas for all the years, 2013-2022

Afterwards, the text of the documents in each community was extracted, merged, and processed through Negative Matrix Factorization in order to automatically extract topics based on TF-IDF vectors. Finally, the automatically extracted topics were manually mapped to the areas identified in Section II-C. Aggregated results are shown here in terms of wordmaps, charts, and tables in Figures 6, 7, 8, 9, and 10.

The wordmap in Figure 6 lists different areas and determines their size using the corresponding frequency of each area over all the communities of all the years, 2013-2022. The areas with the highest frequency are "DOrg" with 76, "DLang" with 32, and "DLit" with 27, while the areas with the lowest frequency are "DEcon" with 2 and "DHealth" with 2. In Figure 7, the areas and their frequencies are shown as a horizontal barchart with precise numbers shown on each bar. DOrg has been omitted due to its high overall frequency of 76, which would visually shrink significantly all the bars for the other areas.

Figure 8 gives drill-down numbers for each area mapped to topics extracted from the top three communities for each year. We still see DOrg being the most popular area in several years, but not all. For instance, DOrg is not the most popular area in 2018, 2019, and 2022. In 2021, DOrg is tied with DHist, and ML.

The most popular areas besides DOrg, are ML, DEdu, DHist, DLit, DLang, DLib, SemWeb, DataViz, and DMedia. A more detailed analysis of these areas is provided in Figures 9, and 10. They show the best-year-for-each-area and best-area-for-each-year, respectively. For instance, it can observed that ML was most popular in 2022, whereas DEdu was most popular in 2020. These results make sense given the prominence of ML methods in recent times and the situation with the online education during the Covid 19 pandemic.

Area	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total
DOrg	16	17	8	7	7	4	4	5	4	4	76
DLang	1	3	3	1	3	7	4	2	3	5	32
DHist	2	0	4	2	3	1	5	4	4	4	29
DLib	1	2	2	4	4	6	1	3	3	2	28
DLit	2	1	4	0	6	3	6	5	2	1	27
DEdu	2	1	3	3	3	1	3	4	2	2	24
ML	2	1	1	2	1	0	2	2	4	4	19
SemWeb	0	1	2	1	1	3	1	2	2	2	15
DataViz	0	1	0	3	2	1	2	2	1	0	12
DMedia	0	3	2	2	1	2	0	0	1	1	12

Fig. 8. Drill-down numbers for each of top-10 areas each year

ML	DEdu	DHist	DLit	DLang	DLib	SemWeb	DataViz	DOrg	DMedia
2022	2020	2019	2019	2018	2018	2018	2016	2014	2014

Fig. 9. Best year for each of top-10 areas

On the other hand, it can be observed from Figure 10 that DLang has been quite popular in several years, such as 2014, 2015, 2018, and 2022. This is of course also related to the rise of ML, which has important intersection in terms of methodology with DLang given that Computational Linguistics, a subarea of DLang heavily relies on ML methods. DHist and DLit are also best areas for several areas, such 2013, 2015, 2017, 2019, 2020, and 2021.

D. Community Areas Entropy

Entropy is used to quantify the impurity or heterogeneity of a set of data. In this context, entropy is calculated as the sum of the negative probabilities of each unique class (area) in the set, multiplied by the logarithm of the probability.

A set with high entropy has a large number of different classes, which means that it is more heterogeneous (more diverse) and less pure. On the other hand, a set with low entropy has a small number of different classes, meaning that it is more homogeneous (more coherent) and pure.

The line chart in Figure 11 displays the entropy values for the top three communities each year. It is evident that the entropy values have been in an increasing trend over

Area	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
ML	2	1	1	2	1	0	2	2	4	4
DEdu	2	1	3	3	3	1	3	4	2	2
DHist	2	0	4	2	3	1	5	4	4	4
DLit	2	1	4	0	6	3	6	5	2	1
DLang	1	3	3	1	3	7	4	2	3	5
DLib	1	2	2	4	4	6	1	3	3	2
SemWeb	0	1	2	1	1	3	1	2	2	2
DataViz	0	1	0	3	2	1	2	2	1	0
DMedia	0	3	2	2	1	2	0	0	1	1
	ML DEdu DHist DLit	DLang	DLang DHist	DLib	DLit	DLang	DLit	DLit	ML DHist	DLang

Fig. 10. Best area for each year (DOrg excluded)

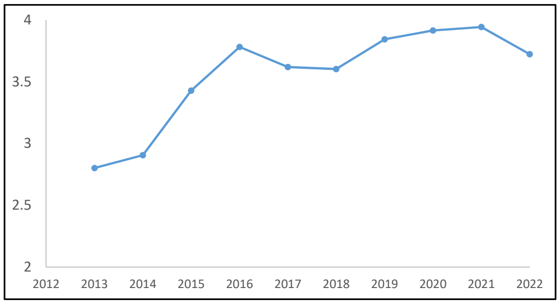

Fig. 11. Entropy of areas over the years, 2013-2022.

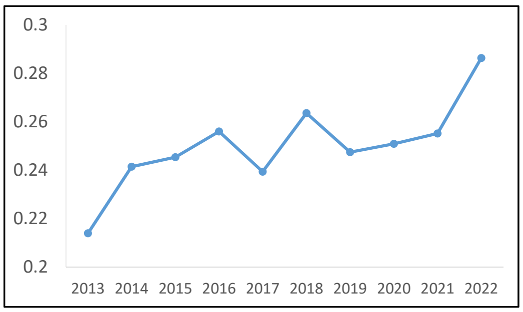

Fig. 12. Modularity of Louvain community partitioning over the years, 2013-2022.

time, suggesting that the diversity of topics covered in the community documents has become more varied as the years have passed. This is a positive trend for Digital Humanities, as it emphasizes the interdisciplinary nature of the field, encompassing a broad range of subjects and disciplines.

E. Community Modularity

Modularity is a measure of the structure of a network or a graph, used to quantify the degree of similarity between the network's structure and the community divisions within it. In the context of community detection in complex networks, modularity is used to evaluate the quality of a partition of the network into communities or clusters. The result of modularity calculation is a value between -1 and 1, where a score of 1 indicates a perfect community structure, and a score close to 0 indicates a random or poorly structured network.

The graph in Figure 12 displays the modularity of the community partitioning produced by the Louvain method. It is evident that the overall trend of modularity is on the rise, indicating that the document network structure is becoming more compact from the perspective of community structure.

It is thus interesting to note that, although the communities are becoming increasingly diverse over the years in terms of the treated areas, they are simultaneously becoming more compactly structured from a network perspective over time.

V. Conclusions

The purpose of Research Question 1 was to investigate the dominant research themes in the field of Digital Humanities over the last decade as extracted from Google Scholar indexed literature. The findings presented in Section IV indicate that not all of the research areas listed in Section II-C experienced equal levels of popularity in the last decade. The top ten most frequently discussed areas were Digital Organization (DOrg), Digital Language (DLang), Digital History (DHist), Digital Libraries (DLib), Digital Literature (DLit), Digital Education (DEdu), Machine Learning (ML), Semantic Web (SemWeb), Data Visualization (DataViz), and Digital Media (DMedia).

With regards to Research Question 2, which explored the alignment between the areas recognized by experts in the field and the areas represented in the literature indexed by Google Scholar, two key observations were made. It was found that the areas extracted from Google Scholar indexed literature correspond well with the areas recognized by experts, however, the reverse is not the case. Several areas recognized by experts, such as Digital Religion (DRelig), Digital Architecture (DArchitect), Digital Economics (DEcon), and Digital Health (DHealth), did not have as much representation in the Google Scholar indexed literature.

With respect to Research Question 3, the data showed clear patterns of growth and increasing prominence for certain areas, such as ML (Machine Learning), DLang (Digital Language), and Digital History (DHist), while Digital Organization (DOrg) remains the leading area of research, but was excluded from the analysis to avoid skewing the results.

Finally, the findings of Research Question 4 revealed two noteworthy insights. Firstly, the results showed that the entropy of research areas in the field of Digital Humanities has been increasing over the years. This suggests that the diversity of research topics within Digital Humanities is growing over time. However, there could be a concern that this growth in diversity may come at the cost of a decreased sense of connection or cohesion among the research publication communities. Secondly, the results showed that this is not the case. In fact, the cohesion among the publication communities within Digital Humanities has been gradually increasing over time.

It is hoped that this study offered a thorough analysis of the recent evolution of the field of Digital Humanities and shed some light on its potential future development. Moving forward, future work could involve expanding the document collection and exploring different time frames through the use of sliding time windows.

References

[1] W. McCarty, *Humanities computing*. Springer, 2005.
[2] I. Rafols and M. Meyer, "Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience," *Scientometrics*, vol. 82, no. 2, pp. 263–287, 2010.
[3] P. Svensson, "The landscape of digital humanities," *Digital humanities quarterly*, vol. 4, no. 1, 2010.
[4] C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols, and K. Börner, "Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature," *Journal of informetrics*, vol. 5, no. 1, pp. 14–26, 2011.
[5] J. Moody, "The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999," *American sociological review*, vol. 69, no. 2, pp. 213–238, 2004.
[6] B. V. Carolan, "The structure of educational research: The role of multivocality in promoting cohesion in an article interlock network," *Social Networks*, vol. 30, no. 1, pp. 69–82, 2008.

[7] P. Liu and H. Xia, "Structure and evolution of co-authorship network in an interdisciplinary research field," *Scientometrics*, vol. 103, pp. 101–134, 2015.
[8] M.-C. Tang, Y. J. Cheng, and K. H. Chen, "A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses," *Scientometrics*, vol. 113, pp. 985–1008, 2017.
[9] S. Potka, "Community structure and coherence of digital humanities academic works in a decade-long corpus: A text and network analysis approach," Master's thesis, Department of Media and Culture, Faculty of Philosophy, University of Cologne, 2023.
[10] M. Cohen, *The Networked Wilderness: Communicating in Early New England*. U of Minnesota Press, 2010.
[11] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press, 2010.
[12] S. Jänicke, "Valuable research for visualization and digital humanities: A balancing act," in *Workshop on Visualization for the Digital Humanities, IEEE VIS*, vol. 7, 2016.
[13] N. Iliinsky and J. Steele, *Designing data visualizations: Representing informational Relationships*. " O'Reilly Media, Inc.", 2011.
[14] C. Bassett, D. M. Berry, B. Fazi, J. Pay, and B. Roberts, "Critical digital humanities and machine-learning," in *ADHO 2017-Montréal*, 2017.
[15] E. Hyvönen, "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, 2020.
[16] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.
[17] J. Hutson and T. Olsen, "Digital humanities and virtual reality: A review of theories and best practices for art history," *International Journal of Technology in Education (IJTE)*, vol. 4, no. 3, pp. 491–500, 2021.
[18] J. Zec, "Augmented reality: A practical guide," *Software Quality Professional*, vol. 12, no. 1, p. 40, 2009.
[19] R. F. Davis, "Digital pedagogy in the humanities: Concepts, models, and experiments," 2015.
[20] M. Condruz-Bacescu, "The impact of digital technologies on learning," in *Conference proceedings of eLearning and Software for Education (eLSE)*, vol. 15, pp. 57–63, Carol I National Defence University Publishing House, 2019.
[21] T. Presner, D. Shepard, and Y. Kawano, *Hypercities thick mapping in the digital humanities*. 2014.
[22] A. Hamraie, "Mapping access: Digital humanities, disability justice, and sociospatial practice," *American Quarterly*, vol. 70, no. 3, pp. 455–482, 2018.
[23] J. Lambert, *Digital storytelling: Capturing lives, creating community*. Routledge, 2013.
[24] R. Martinec and T. Van Leeuwen, *The language of new media design: Theory and practice*. Routledge, 2020.
[25] C. Paul, "Digital art," 2008.
[26] L. Pugin, "The challenge of data in digital musicology," 2015.
[27] N. Collins, *Introduction to computer music*. John Wiley & Sons, 2010.
[28] P. Daly and T. L. Evans, *Digital archaeology: bridging method and theory*. Routledge, 2004.
[29] E. B. Zubrow, T. L. Evans, and P. Daly, "Digital archaeology," *Digital Archaeology. Bridging method and theory, London*, pp. 10–31, 2006.
[30] I. J. Young and T. McCormick, "Digital history: A guide to gathering, preserving, and presenting the past on the web," 2006.
[31] C. Brennan, "Digital humanities, digital methods, digital history, and digital outputs: History writing and the digital revolution," *History Compass*, vol. 16, no. 10, p. e12492, 2018.
[32] G. B. M. Masoodian and S. J. Cunningham, "Digital libraries: Universal and ubiquitous access to information,"
[33] W. Y. Arms, *Digital libraries*. MIT press, 2001.
[34] R. A. Lanham, "Digital rhetoric and the digital arts," in *The Electronic Word*, pp. 29–52, University of Chicago Press, 2010.
[35] M. E. Hocks, "Understanding visual rhetoric in digital writing environments," *College composition and communication*, pp. 629–656, 2003.
[36] N. Marres, *Digital sociology: The reinvention of social research*. John Wiley & Sons, 2017.
[37] D. Miller, "The anthropology of social media," in *Digital anthropology*, pp. 85–100, Routledge, 2021.
[38] A. Feenberg, "What is philosophy of technology?," in *International handbook of research and development in technology education*, pp. 159–166, Brill, 2009.
[39] M. Dascal, "Digital culture: Pragmatic and philosophical challenges," *Diogenes*, vol. 53, no. 3, pp. 23–39, 2006.
[40] J. Hargreaves and E. Anderson, *Routledge handbook of sport, gender and sexuality*. Routledge, 2014.
[41] J. Rak, "The digital queer: Weblogs and internet identity," *Biography*, pp. 166–182, 2005.
[42] A. Bharadwaj, O. A. El Sawy, P. A. Pavlou, and N. v. Venkatraman, "Digital business strategy: toward a next generation of insights," *MIS quarterly*, pp. 471–482, 2013.
[43] M. M. Al-Debi, R. El-Haddadeh, and D. Avison, "Defining the business model in the new world of digital business," *AMCIS 2008 proceedings*, p. 300, 2008.
[44] M. E. Katsh, *Law in a digital world*. Oxford University Press, 1995.
[45] R. K. Sherwin, N. Feigenson, and C. Spiesel, "Law in the digital age: How visual communication technologies are transforming the practice, theory, and teaching of law," *BuJ sCi. & TeCh. L.*, vol. 12, p. 227, 2006.
[46] D. Lupton, *Digital health: critical and cross-disciplinary perspectives*. Routledge, 2017.
[47] P. C. Tang and C. J. McDonald, "Electronic health record systems," *Biomedical informatics: computer applications in health care and biomedicine*, pp. 447–475, 2006.
[48] L. Borek, J. Perkins, C. Schöch, and Q. Dombrowski, "Tadirah: a case study in pragmatic classification," *Digital Humanities Quarterly*, vol. 10, no. 1, 2017.
[49] A. E. Earhart, "Can information be unfettered? race and the new digital humanities canon," *Debates in the Digital Humanities*, pp. 309–18, 2012.
[50] C. Coletta, L. Heaphy, S.-Y. Perng, and L. Waller, "Data-driven cities? digital urbanism and its proxies: Introduction," *TECNOSCIENZA: Italian Journal of Science & Technology Studies*, vol. 8, no. 2, pp. 5–18, 2017.
[51] W. Tao, "Interdisciplinary urban gis for smart cities: advancements and opportunities," *Geo-spatial Information Science*, vol. 16, no. 1, pp. 25–34, 2013.
[52] N. Leach, D. Turnbull, and C. J. Williams, "Digital tectonics," 2004.
[53] L. Krier, *The architecture of community*. Island Press, 2009.
[54] J. Trappel, W. A. Meier, J. Steemers, and B. Thomass, *Media in Europe today*. Intellect Books, 2011.
[55] E. Ellison, "The international encyclopedia of media studies," ***M/C Reviews***, 2013.
[56] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, *et al.*, "Recent advances in deep learning for speech research at microsoft," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8604–8608, IEEE, 2013.
[57] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammus: A survey of transformer-based pretrained models in natural language processing," *arXiv preprint arXiv:2108.05542*, 2021.
[58] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM Press, 2011.
[59] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.
[60] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3802–3811, 2019.
[61] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[62] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.
[63] S. Yuan, J. Du, W. Liu, W. Fan, and J. Ren, "Non-negative matrix factorization for text data: A comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, p. 26, 2019.
[64] X. Guo, X. Liu, and J. Yin, "Topic modeling with non-negative matrix factorization," *arXiv preprint arXiv:2002.06470*, 2020.
[65] W. Liu, W. Fan, J. Liu, J. Ren, and J. Du, "Non-negative matrix factorization for text classification," in *2018 International Conference on Artificial Intelligence (ICAI)*, pp. 359–364, IEEE, 2018.
[66] J. Albrecht, S. Ramachandran, and C. Winkler, *Blueprints for Text Analytics Using Python*. " O'Reilly Media, Inc.", 2020.
[67] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, National Acad Sciences, 2004.