Mapping Cultures in the Big Tent: Multidisciplinary Networks in the Digital Humanities Quarterly

Dulce Maria de la Cruz, Jake Kaupp, Max Kemman, Kristin Lewis, Teh-Hen Yu

Abstract—Digital Humanities Quarterly (DHQ) is a young journal that covers the intersection of digital media and traditional humanities. In this paper, we explore the publication patterns in DHQ through visualizations of co-authorship and bibliographic coupling networks in order to understand the cultures the journal represents. We find that DHQ consists largely of sole-authored papers (66%) and the authorship is dominated (75%) by authors publishing from North American institutions. Through the backbone of DHQ's bibliographic coupling network, we identify several communities of articles published in DHQ, and we analyze their collective abstracts using term frequency-inverse document frequency (TF-IDF) analysis. The extracted terms show that DHQ has wide coverage across the digital humanities, and that sub areas of DHQ can be identified through their citation behavior.

Index Terms—Digital Humanities, Information Visualization, Co-author network, Bibliographic Coupling, big tent

INTRODUCTION

Digital Humanities (DH) is a field of research difficult to define due to its heterogeneity[1]. With its inclusionary ambitions, DH is regularly referred to as a 'big tent' [1] encompassing scholars from a wide variety of disciplines such as history, literature, linguistics, but also disciplines such as human-computer interaction and computer science. This collaborative, multidisciplinary approach to digital media makes DH an interesting field, but also difficult to grasp. A question is to what extent the big tent of DH represents a single, or actually a variety of cultures [1, 2].

The Digital Humanities Quarterly (DHQ) journal is arguably one of the largest journals aimed specifically at DH research, and covers all aspects of digital media in the humanities, representing a meeting point between digital humanities research and the wider humanities community [3]. Articles published in DHQ involve authors of multiple countries, institutions and disciplines who work on several subjects and areas related to digital media research. Under a recent grant from NEH (National Endowment for Humanities), DHQ has developed a centralized bibliography which supports the bibliographic referencing for the journal. To gain an understanding of the diversity of culture(s) in the DH, we are interested in how unique disciplinary cultures are represented in DHQ. Considering cultures are self-referential systems, we might expect that scholars from a certain culture are more likely to cite scholars from their own culture rather than from others [2]. As such, we expect citation behaviour to reflect disciplinary cultural norms. Therefore, visualizing and analysing the bibliographic data of DHQ not only gives insights into the specific bibliographies from DHQ, it might give insight into the way the different epistemic cultures in the DH big tent interact with one another, and how this interaction and collaboration impacts the networks over time.

This paper reports on a project undertaken in the Information Visualization MOOC from Indiana University[2]. We have analysed the DHQ bibliographic data and created visualizations in order to discuss the following questions provided by the DHQ editors:

1. how citations reflect differences in academic culture at the institutional and geographic level
2. the changes to that culture over time.
3. correlations between article topics (reflected in keywords) and citation patterns

1 METHOD

1.1 Data

Two tables were extracted from the Client dataset:

1. dhq_articles (178 records)
2. works_cited_in_dhq (3823 records)

The attributes for both tables are: article id, authors, year, title, journal/conference/collection, abstract, cited references, and isDHQ.

The raw dataset posed several problems, including:

- missing articles,
- duplicate authors,
- double affiliations and inconsistencies,
- duplicated articles and citation self-loops,
- special characters, and
- incomplete information (lack of information regarding affiliation and country for each DHQ paper, and disciplines for authors).

The DHQ website[3] was therefore scraped using the tool Import.io[4] to find missing articles and to obtain information about affiliations for each author. Once that information was known, it was used to obtain the country associated with each institution by searching in the web. Custom programs in the R language were then used to create paper IDs (cite me as) similar to those used for the references and to

- *Dulce Maria de la Cruz is Freelance Data Analyst. E-mail: Dulce.Maria.delaCruz@gmail.com.*
- *Jake Kaupp is Engineering Education Researcher in Queen's University, Canada. E-mail: jkaupp@gmail.com.*
- *Max Kemman is PhD Candidate in University of Luxembourg, Luxembourg. E-mail: maxkemman@gmail.com.*
- *Kristin Lewis is Science & Technology Policy Fellow at AAAS. E-mail: kristin.l.m.lewis@gmail.com.*
- *Teh-Hen Yu is IT Professional. E-mail:tehhenyu@hotmail.com.*

[1] See e.g. http://whatisdigitalhumanities.com for a wide variety of definitions from different scholars

[2] http://ivmooc.cns.iu.edu/

[3] http://www.digitalhumanities.org/dhq/

[4] https://www.import.io/

calculate the number of times each DHQ paper has been cited (times cited) and the number of references cited by each DHQ paper (count cited references). Furthermore, we assigned a discipline to each paper based on the first author's departmental affiliation as described in [4]. In order to produce a more detailed list of disciplinary culture, departmental affiliation was manually mapped to web of science subject areas. This information was eventually not used for the final visualizations, but left in the dataset for further exploration by others.

After validations, data mining/scraping, data processing with custom programs coding and a lot of manual work, we have come up with a master dataset with additional info added (cite me as, times cited, affiliation, country, count cited references, geocode, discipline, affiliations including departments info, and community, plus the keywords provided by editors of DHQ). To provide sufficient resolution, and categorical variables, for visualizations an author look-up table was created which contained the additional information outlined above but for each separate author for each article ID. Both the master datafile and the author lookup table are our primary sources of data to load for visualization and analysis.

The source code, final datasets, and resulting visualizations are available through github[5].

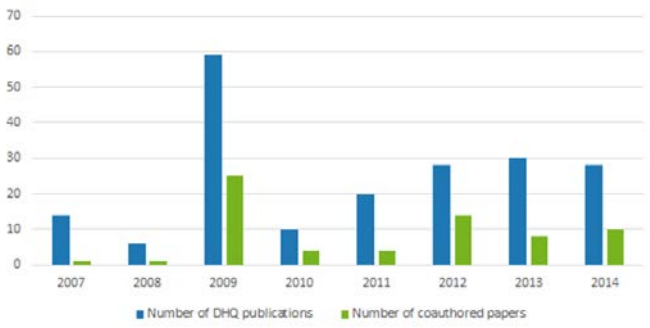
The final dataset provides the following statistics as in Table 1.

Table 1. DHQ dataset statistics

Attribute	Count	Note
DHQ articles	195	
Unique cited articles	4718	
Unique DHQ author	276	
Affiliations	148	Including all institutions + independent scholars
WOS subject areas	29	
Countries	17	
Publication years	8	2007-2014

Figure 1 provides an overview of the number of DHQ publications and number of co-authored papers per year, revealing a surprisingly uneven temporal distribution.

Fig. 1. DHQ (co-authored) publications per year.



1.2 Co-author network

People are the key inputs in determining and understanding cultural differences. Therefore, in order to better understand the cultures within DHQ, we explored the authors who published within DHQ. Using Sci2 [5], we created yearly cumulative time slices of the master dataset and extracted co-author networks for each time slice. Columns for author country were added, and each time slice was imported into Gephi to create a dynamic co-author network [6]. The network was laid out using the Force Atlas 2 algorithm [7], with nodes colorized by country. Each time slice was visualized, and compiled into comprehensive visualizations using Adobe Illustrator and Adobe Photoshop.

In addition to a co-author network, we explored a bibliographic coupling network of authors, in which nodes (authors) would be linked based on the number of cited articles in common. This analysis however introduced a strong bias towards co-authors who cite large numbers of articles. In order to derive userful insights from this type of visualization, a de-biasing operation must be identified and applied. Without an established method for these, we chose to focus on the geographic information in the co-authorship network and analyse bibliographic coupling of articles

1.3 Bibliographic coupling & Backbone identification

In order to investigate the bibliographies of DHQ articles, we analysed the data using Sci2 by extracting the paper-citation network, followed by extracting the reference co-occurrence network, also known as "bibliographic coupling" [8]. By doing so, we create a network of DHQ articles with co-occurring references. To simplify the visualization, we created a minimum spanning tree using the MST Pathfinder algorithm whereby articles are connected to the network only by their strongest relation [9], also called Backbone identification. As such, the network becomes a tree that is easier to read. Finally, all articles with zero references were removed from the network in order to remove non-DHQ articles, as well as DHQ articles that could not be analysed due to a lack of references.
This network was then analyzed using the SLM community detection algorithm with undirected and weighted edges [10]. The network with community attributes was then imported into Gephi and ordered using the Force Atlas 2 algorithm [6], after which we colorized the nodes by their identified community.

1.4 Word clouds

In order to investigate the correlations between article topics (reflected in keywords) and the citation patterns, word clouds of keywords were obtained for each of the communities identified via SLM detection in the bibliographic coupling network. For this purpose, community-based abstracts were obtained by combining the abstracts associated with the DHQ papers belonging to each community. These community-wide abstracts were normalized to lower case, tokenized, and stop words were removed. Words were not stemmed in order to differentiate between words like digital and digitized. Unique keywords were extracted from the community-based abstracts with custom R programs (using the R packages stringr[6] and tm[7]). The most significant keywords for each community were then identified through the Term frequency - Inverse Document Frequency (TF-IDF) method [11]. Terms with high TF-IDF values imply a strong relationship with the document in which they appear.

In this specific case, the terms are the unique keywords and the corpus of documents is the set of community-based abstracts. Therefore, the higher the TF-IDF value of a keyword in a

[5] Available at https://jkaupp.github.io/DHQ. Please cite as Kaupp, J., De la Cruz, D.M,, Kemman, M., Lewis, K., Yu, T.-H. (2015) Mapping Cultures in the Big Tent: Multidisciplinary Networks in the Digital Humanities Quarterly. GitHub, https://jkaupp.github.io/DHQ

[6] http://cran.r-project.org/web/packages/stringr/index.html
[7] http://cran.r-project.org/web/packages/tm/index.html

community, the more representative the keyword is of that community.

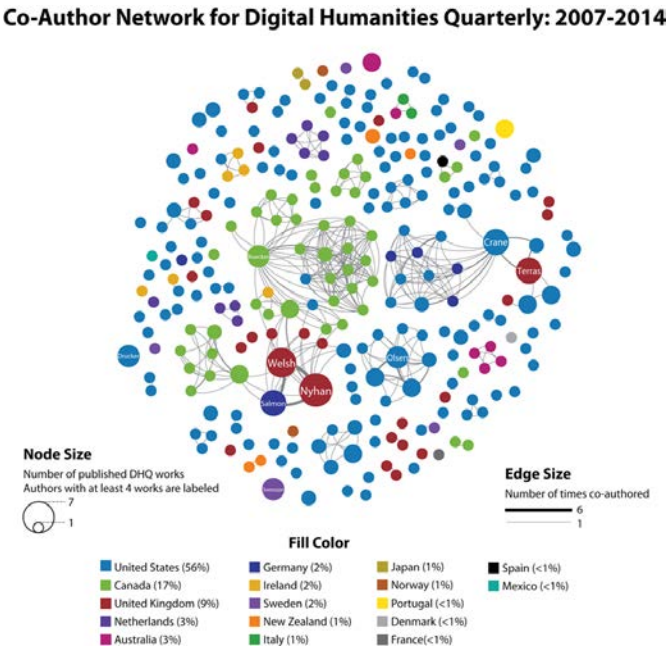
The ten top-scoring words from each community were put into a word cloud and the words were sized by TF-IDF score. The word clouds were manually adjusted to unify the appearance of terms (plural vs. singular, infinitive vs. gerund, etc.) and were added to the bibliographic coupling network visualization.

2 Result[8]

2.1 Co-author Network

Figures 2 and 3 represent the co-author network for DHQ, both comprehensively (Figure 2) and through cumulative time slices (Figure 3). Nodes are sized by the number of works published in DHQ, and in Figure 2, authors with at least 4 DHQ publications are labeled with the author's last name. Nodes are colorized by the country of the author. The edges are weighted by the number of times each pair co-

Fig. 2 Co-Author network, 2007-2014



authored a DHQ publication together. The maximum number of authored works (articles) for a single author is 7: Julianne Nyhan from UK. The maximum co-authored articles for two authors are 6: by Anne Welsh and Julianne Nyhan from UK. The most active year is 2009, as also shown in Figure 1, with several authors publishing multiple papers in this year.

2.2 Bibliographic coupling network with word clouds

Figure 4 shows the backbone bibliographic coupling network for DHQ, representing the strongest connections in the larger bibliographic coupling network (not shown). Nodes are colored by community, as identified through SLM detection, and sized by the number of articles cited in each article. Edges are weighted by the number of cited articles in common. Alongside each community is a word cloud of keywords in the same color extracted from the abstracts of each article in the community.

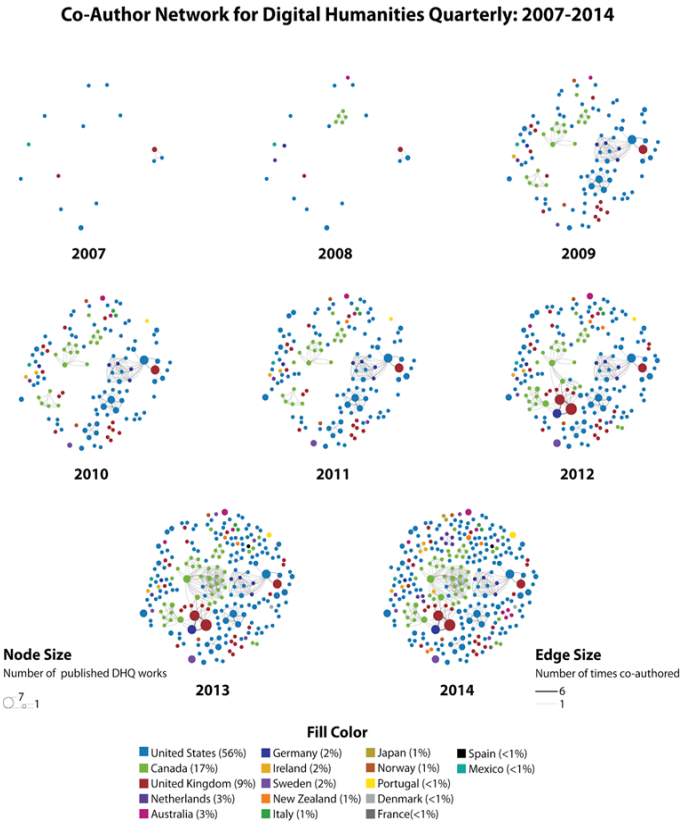
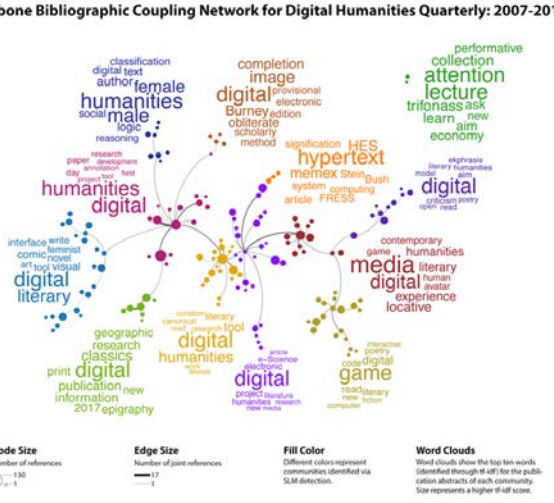


Fig. 3 Co-Author network by year

Figure 5 shows key papers in the backbone bibliographic coupling network, that is, the papers that link each of the communities in the giant component. The labels are shown in the same colors as the communities in Figure 4.

After we removed articles with zero references, the network contained 170 articles (out of 195), of which 23 are without a connection to other (i.e. they remained isolate). These 23 are not shown in the final visualization above, showing 147 articles and 145 connections. The bibliographic coupling network contains twelve communities, of which one consists of two articles not otherwise connected to the

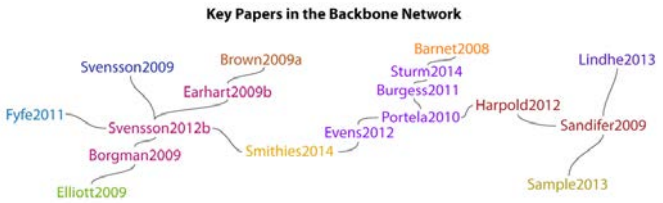
Fig. 4 Backbone bibliographic coupling network for DHQ.



[8] Larger versions of all visualizations are available in the github repository.

major component (see dark green at the upper right). The other eleven communities are all connected in the large component and shown with their respective word clouds.

Fig. 5 Key papers in the backbone bibliographic coupling network.articles



There are a total of 4880 documents, including the 195 articles from DHQ itself. Together all the DHQ articles contain 5330 references. The highest cited document is Matthew Kirschenbaum's *"Mechanisms: New Media and the Forensic Imagination"* (2008) , cited 15 times. The DHQ article with the most references is Christine Borgman's *"The Digital Future is Now: A Call to Action for the Humanities"* (2009), with 130 references.

3 Discussion

3.1 Co-author Network

The co-author network suggest that DHQ publications follow the patterns of the humanities community, with many single-authored papers (128 out of 195, 65.6%). Moreover, its origins are in North America, and three quarters of the authors are from either the US (58%) and Canada (17%). A distant third is the UK (9%), further demonstrating the Anglo-Saxon nature of DHQ.

The largest co-author network component consists of 43 authors; which is about 16% of all authors (276 authors in all) who contributed to DHQ during this period. The second largest co-author network component consist of 18 authors.

Canadian authors show the most collaborative behavior: the article with the most co-authors: *"Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET)"* has 14 co-authors. The most collaborative author in this period from Canada is Stan Ruecker; he co-authored 4 articles with 25 others.

There does not seem to be a growth of co-authorship after 2008. Overall, articles have on average a little under two authors per paper, and in 2012 a bit above two on average (2.18). When we remove all the single-authored papers, the average number of authors per article is above three, but there is no trend that this is growing with the years.

3.2 Bibliographic coupling network with word clouds

From the word clouds we see that several communities explicitly discuss terms such as *digital* and *humanities* as well as *tool*, which is unsurprising. At the centre of the large component, the communities (magenta, yellow, purple) of articles are related to (textual) tools and discussing DH itself, with terms such as *curation, e-Science, project,* and *research.* The communities further to the left (light blue & dark blue) are related to textual analysis and tools, with terms such as *classification, author, write, annotation, interface,* and *literary.* The communities to the right however (dark purple, dark red, moss-green) suggest articles related to artistic subjects, with terms such as *poetry, ekphrasis, games,* and *fiction*.

4 Conclusion

We return to the questions provided by the DHQ editors:

1. how citations reflect differences in academic culture at the institutional and geographic level
2. the changes to that culture over time.
3. correlations between article topics (reflected in keywords) and citation patterns.

With respect to the first question, we focus on the geographic level of academic culture. The co-author network shows that despite DH being a collaborative culture, over half of all publications are single authored, something demonstrated earlier for other journals[9]. Moreover, DH as represented by DHQ is largely an Anglo-Saxon North American undertaking. With respect to the second question; there is no visible trend regarding co-authorship between 2007-2014. However, authors from non-Anglo Saxon countries are emerging, showing DH is slowly becoming a more global phenomenon as also evidenced by the DH conferences[10].

With respect to the third question, we find that the references present in the DHQ articles lead to a large number of communities. The boundaries are however diffuse, making it difficult to describe clear cut communities. However, from the word clouds we do see at least three different patterns emerge: 1) article related to tools and DH itself, 2) articles related to textual analysis with tools, and 3) articles related to artistic subjects.

While we have provided an exploration of the articles and authors within DHQ, additional insights may be learned from further analysis. In particular, interactive visualizations will provide the user with a more comprehensive understanding of the data. These may allow the user to explore communities via institution or discipline as well as country. In addition, we believe a properly de-biased authorial bibliographic coupling network may provide further insight into the academic cultures within DHQ. Lastly, our analysis focused on DHQ articles alone. Further analysis may allow us to explore the non-DHQ articles cited by DHQ papers.

In sum, we see DHQ fairly represents the heterogeneity of DH, critically examining DH itself and discussing computational analyses of research questions from different backgrounds. On the other hand, however, we see DHQ representing a somewhat homogeneous view of DH, with strong representation from Anglo-Saxon scholars and those from North America in particular. Here, DHQ can be challenged to provide a better representation of scholars from other backgrounds, as well as the 'big tent' of DH in general.

Acknowledgments

The authors wish to thank Professor Julia Flander, Professor Katy Börner, Dr. Andrea Scharnhorst, and the participants of Indiana University's Information Visualization MOOC for providing us valuable feedback during the process of the project work.

References

[1] Svensson, Patrik. (2012) Beyond the big tent. Debates in the Digital Humanities, 36-49.

[2] Knorr Cetina, K. (2007). Culture in Global Knowledge Societies: Knowledge Cultures and Epistemic Cultures. The Blackwell

[9] http://blogs.lse.ac.uk/impactofsocialsciences/2014/09/10/joint-authorship-digital-humanities-collaboration

[10] See http://www.scottbot.net/HIAL/?p=41064

Companion to the Sociology of Culture, 32(4), 361–375. doi:10.1002/9780470996744.ch5

[3] Digital Humanities Quarterly (n.d.). About DHQ. Retrieved from http://www.digitalhumanities.org/dhq/about/about.html

[4] Ortega, L., & Antell, K. (2006). Tracking Cross-Disciplinary Information Use by Author Affiliation: Demonstration of a Method. College & Research Libraries, 67(5), 446–462. Retrieved from http://crl.acrl.org/content/67/5/446.

[5] Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, https://sci2.cns.iu.edu.

[6] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." *ICWSM* 8 (2009): 361-362.

[7] Jacomy, Mathieu, et al. "Forceatlas2, a continuous graph layout algorithm for handy network visualization." *Medialab center of research* 560 (2011).

[8] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American documentation, 14(1), 10-25.

[9] Schvaneveldt, R. W., D. W. Dearholt, and F. T. Durso. "Graph theoretic foundations of pathfinder networks." *Computers & mathematics with applications* 15.4 (1988): 337-345.

[10] Waltman, Ludo, and Nees Jan van Eck. "A smart local moving algorithm for large-scale modularity-based community detection." *The European Physical Journal B* 86.11 (2013): 1-14.

[11] Blázquez, M. (n.d). Frecuencias y pesos de los términos en un documento. Retrieved from: http://ccdoc-tecnicasrecuperacioninformacion.blogspot.com.es/2012/11/frecuencias-y-pesos-de-los-terminos-de.html