

Preliminary Results on Mapping Digital Humanities Research

Gianmarco Spinaci
DHDK - FICLIT University
of Bologna
gianmarco.spinaci
@studio.unibo.it

Giovanni Colavizza
University of Amsterdam
g.colavizza@uva.nl

Silvio Peroni
DHARC - FICLIT University
of Bologna
silvio.peroni@unibo.it

Abstract

English. Maps of research could provide key insights in the current and future development of the digital humanities. Two obstacles to this end are the definition of what can be considered digital humanities research in the first place, and the mostly unknown coverage of digital humanities publications in citation indexes. In view of addressing these challenges, we release a list of digital humanities journals developed via an iterative approach combining manual curation with citation data. Based on this list, we further assess the journal article coverage of Web of Science, Scopus, Crossref and Dimensions. We find that Crossref has the best coverage of exclusively digital humanities journals, while Dimensions has the best coverage of related journals. Furthermore, we use Dimensions data to map digital humanities research via a directed citation network, finding connections with fields such as computational linguistics and digital libraries, and a strong correlation between journals and citation clusters.

Italiano. Le mappe della ricerca scientifica possono aiutare a comprendere gli sviluppi attuali e futuri delle *digital humanities*. I maggiori ostacoli alla loro realizzazione sono la definizione di cosa includere nelle pubblicazioni *digital humanities*, e la copertura degli indici citazionali al riguardo. In questo contributo, pubblichiamo una prima lista di riviste digital humanities, curata tramite un metodo iterativo che include revisione manuale e uso di dati citazionali. Sulla base della lista, verifichiamo la copertura di pubblicazioni *digital humanities* in Web of Science, Scopus, Crossref e Dimensions. Crossref risulta avere la migliore copertura rispetto alle riviste esclusivamente focalizzate alle *digital humanities*, Dimensions rispetto alle riviste adiacenti. Utilizzando i dati citazionali di Dimensions, costruiamo infine una mappa/rete delle *digital humanities* mostrando una relazione tra queste ultime e aree limitrofe quali la linguistica computazionale e le biblioteche digitali, e una forte correlazione tra gruppi di articoli individuati tramite *community detection* e le riviste in cui sono pubblicati.

1 Introduction

It is a scientometrics trope to consider humanities research as poorly indexed in citation databases, and thus poorly understood in terms of research outputs (Hammarfelt, 2016). Several studies have pointed out to the limitations of indexes such as Web of Science and Scopus with respect to the humanities, both in terms of quantity and quality (e.g., lack of books) (Nederhof, 2006). Nevertheless, in recent years more indexes have become available, such as Dimensions and Microsoft Academic, while coverage has been improving (Harzing and Alakangas, 2016). In view of these developments, a comprehensive and cross-index map of research in the (digital) humanities is still pending.

Some previous work has considered the intellectual and social organization of the digital humanities using bibliometrics. Nyhan and Duke-Williams (2014) focused on collaboration patterns in the journals *Computers and the Humanities* and *Literary and Linguistic Computing* (up to 2011), finding a propensity to collaborate within small, tight groups and a persisting tendency for single-author publishing. It is worth noting that more recent work on the humanities as a whole found an increasing propensity for

collaboration, albeit with high variation among different disciplines/departments (Burroughs, 2017). Citation analyses based on *Computers and the Humanities*, *Digital Scholarship in the Humanities* and *Digital Humanities Quarterly* highlighted instead a sparser organization, around thematic areas such as information studies, historical literature, linguistics, natural language processing and statistical text analysis (Gao et al., 2017, 2018). Further work based on the *Journal of Digital Humanities*, *Digital Humanities Quarterly*, *International Journal of Humanities and Arts Computing*, *Digital Medievalist*, *Digital Studies*, *Literary and Linguistics Computing* assessed co-authorship, co-citation and bibliographic coupling networks. The authors found a sustained growth in digital humanities publications, coupled with increasing integration with respect to citation networks, and persisting fragmentation with respect to collaborations as mapped by co-authorship relations (Tang et al., 2017).

A recent bibliometric comparison considered the annual conference of the Italian Association of Digital Humanities and Digital Culture (AIUCD) and the annual Italian conference on computational linguistics (CLiC-it). Results show how collaborations are sparser in digital humanities, how research methodologies usually are introduced in the computational linguistics conference and then readopted in the digital humanities one, and how the citation behaviour in the latter one closely resembles that of humanities scholarship (e.g., higher ratio of references to books) (Sprugnoli et al., 2019). Altmetrics data has also been used to map the digital humanities community worldwide. In particular, Twitter follower and co-retweet networks were used to show how the community is organized around few “influencers” and according to language and geographical region (Grandjean, 2016; Gao et al., 2018). Finally, some scholars attempted to position the digital humanities within the broader context of humanities scholarship (Leydesdorff and Akdag Salah, 2010; Salah et al., 2015).¹

In this paper, we address the following research questions: *a) what qualifies as digital humanities research, from a bibliometric point of view? b) What is the coverage of citation indexes with respect to digital humanities research? c) What is the organization of the resulting map of research?* We propose an iterative method to individuate digital humanities publications by combining manual journal classification and automatic citation clustering. One of the outcomes of our work is the first version of a list which includes digital humanities journals. We use this list to assess the number of digital humanities journal publications indexed by Web of Science (WoS), Scopus, Crossref and Dimensions. Finally, we use the citation data included in the index with most digital humanities publications, i.e., Dimensions (Hook et al., 2018), to present a map of digital humanities research based on journal articles. It is worth noticing that our results are still preliminary and stem from ongoing work to create a comprehensive map of humanities research.

2 Data and methods

Database coverage limitations notwithstanding, individuating digital humanities (DH) publications is problematic in itself. First of all, there is little agreement on what constitutes DH research among practitioners. Secondly, DH research tends to be highly interdisciplinary, so much so that clear-cut classifications would be intrinsically arbitrary. We adopt here a combination of top-down journal level classification, in view of expanding the ERIH-Plus journal list ², and a bottom-up clustering approach, where we use citation clusters to find candidate journals to be added to the list.

More in detail, we perform the following steps:

1. create a seed list of known journals in DH, by disseminating a survey to the participants to DH 2019 and in the Humanist mailing list, which resulted in obtaining 14 replies;
2. consider a fine-grained clustering of all publications within each citation index, obtained by using the Leiden algorithm (Traag et al., 2019) and following the heuristics proposed in Waltman and van Eck (2012);

¹A reasoned review of quantitative analyses of the digital humanities is maintained by Scott Weingart at <http://scottbot.net/dh-quantified>

²See <https://dbh.nsd.uib.no/publiseringskanaler/erihplus>

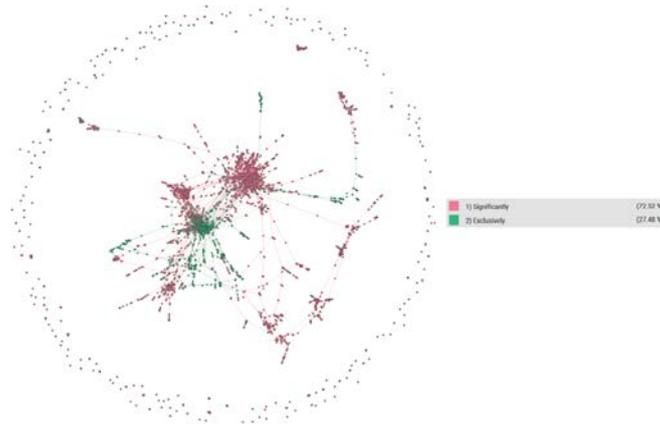


Figure 1: DH articles in Dimensions according to their journal category: “exclusively” (green) and “significantly” (pink). Only articles with with one citation or more (34.4% of the total) were considered. The network was visualized using Gephi 0.9.2 and the Force Atlas 2 layout. The size of the nodes (i.e., the articles) is proportional to number of citations they receive.

3. detect which clusters contain a relatively high proportion of publications from the journals in the list, in order to identify those which are also highly represented in the detected clusters but that are not part of the list obtained in point 1. We do this by considering the top 5 journals (by number of articles) per cluster with more than 5% of articles already from DH journals within the list;
4. manually assess each journal in the set identified in the previous point so as to add it in the original list; and
5. iterate again from point 2 until a convergence criterion is met.³

As convergence criterion, we iterate the proposed method twice (i.e., seed plus first iteration) and focus exclusively on research articles or review articles as publication typologies, published from the year 2000. This approach follows previous work in citation indexing for the humanities (Colavizza et al., 2018).

2.1 Journal classification

Given the highly interdisciplinary character of most publications in DH, we classify journals in the list using three categories: *exclusively*, if we deem a journal to be solely devoted to DH; *significantly*, if we deem that at least 50% of publications in the journal can be considered DH; *marginally*, if the journal contains an estimated 5% to 50% publications in DH. Categories were assigned by survey participants (iteration 1) or the authors (iteration 2) independently, and disagreements solved by majority. We acknowledge as a limitation the subjective perspective and biases this approach might have introduced in the resulting list of journals.

3 Results

The first outcome of our study is a list of DH journals (Spinaci et al., 2019), arranged according to the proposed categories, and containing 19 “exclusively”, 17 “significantly” and 64 “marginally” classified journals.⁴

³Due to data access constraints, we worked with the following versions of the citation indexes under consideration: Web of Science: December 2018, Scopus: May 2019, Dimensions: December 2018, Crossref: August 2018. Coverage results might be affected accordingly.

⁴The list has been also made available in a Google sheet (<https://tinyurl.com/y6rfrsuw>) which can be commented for further feedback.

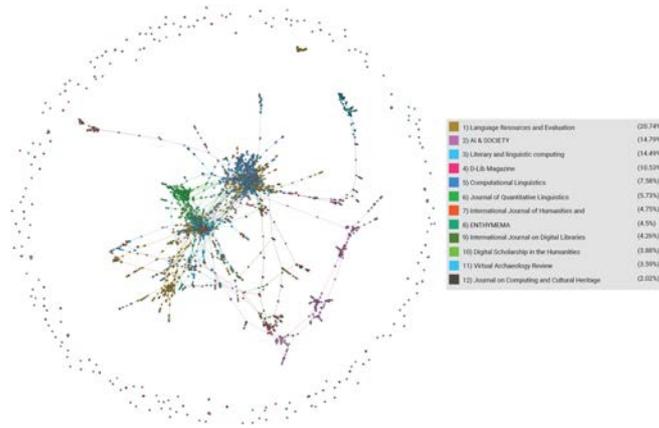


Figure 2: DH articles in Dimensions according to their related journal, if it contained more than 2% of the visible articles. The underlying network and layout is created as in Figure 1. The journal names are: 1) *Language Resources and Evaluation*; 2) *AI & Society*; 3) *Literary and Linguistic Computing*; 4) *D-Lib Magazine*; 5) *Computational Linguistics*; 6) *Journal of Quantitative Linguistics*; 7) *International Journal of Humanities and Arts Computing*; 8) *Enthymema*; 9) *International Journal on Digital Libraries*; 10) *Digital Scholarship in the Humanities*; 11) *Virtual Archaeology Review*; 12) *Journal on Computing and Cultural Heritage*.

Table 1: Database coverage for journals in the three categories.

Journal	WoS	Scopus	Crossref	Dimensions
Exclusively	1243	1858	3989	2751
Significantly	1096	4421	5395	7259
Marginally	39,655	40,439	55,126	117,782
Total	41,994	46,718	64,510	127,792

3.1 Coverage

The overall database coverage, expressed as the number of indexed articles per category, is shown in Table 1. Crossref has the best coverage with respect to “exclusively” journals, while Dimensions has better coverage in the “significantly” and “marginally” categories.

The publication coverage of journals in the “exclusively” category is shown in Table 2. Only few journals show a good coverage, including some non-active ones: *Computers and the Humanities*, *Digital Scholarship in the Humanities*, *International Journal of Humanities and Arts Computing*, *Journal on Computing and Cultural Heritage*, and *Literary and Linguistic Computing*. Many other DH journals we collected in (Spinaci et al., 2019) were either poorly represented or even not present in the indexes we used for the analysis. Crossref appears to be the most comprehensive database in this respect.

3.2 Map of research

We further present a preliminary map of DH research, focusing on journal articles from the “exclusively” and “significantly” categories. We chose Dimensions for this analysis in order to better explore its apparently complementary coverage with respect to both categories and with respect to previous work (Gao et al., 2017; Tang et al., 2017; Gao et al., 2018). Coverage in the “significantly” category is mostly due to work in computational linguistics and digital libraries: *Journal of Quantitative Linguistics* (574 articles), *Computational Linguistics* (759), *D-Lib Magazine* (1054), *Language Resources and Evaluation* (2076). We also highlight the presence of almost 1500 articles from the journal *AI & Society*, a topic of increasing interest in DH. The map shown in Figures 1, 2, 3 considers all articles with at least one (given or received) citation, that is to say articles with a degree of one or more. The network initially contains 10,010 articles and 5,283 citation edges, while the number of articles with citations is 3,446 (34.4% of the

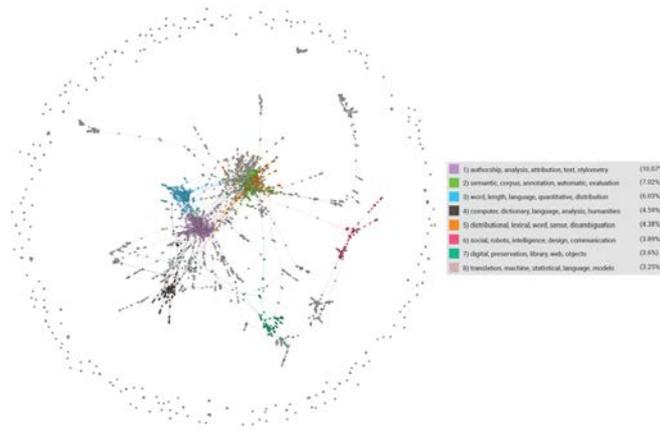


Figure 3: DH citation clusters calculated using the citations available in Dimensions, considering only the clusters containing more than 2% of the visible articles in the dataset. The legend contains the five most frequently occurring words in the titles of the articles within each cluster, after filtering out uninteresting ones. The underlying network and layout is created as in Figure 1. The journal coverage for each cluster is as follows (we only include journals accounting for more than 10% of the articles within a cluster): 1) *Literary and Linguistics Computing* (39.7%), *Language Resources and Evaluation* (19.3%), *Digital Scholarship in the Humanities* (17.6%), *Journal of Quantitative Linguistics* (17%), 2) *Computational Linguistics* (55.4%), *Language Resources and Evaluation* (34.3%), 3) *Journal of Quantitative Linguistics* (91.2%), 4) *Language Resources and Evaluation* (91.1%), 5) *Language Resources and Evaluation* (55%), *Computational Linguistics* (32.5%), 6) *AI Society* (99%), 7) *International Journal on Digital Libraries* (50.8%), *D-Lib Magazine* (37.1%), 8) *Computational Linguistics* (53.6%), *Language Resources and Evaluation* (34.8%)

total). Two thirds of the articles are not connected to any other article through citations. The network’s layout was created using Force Atlas 2 (Jacomy et al., 2014).

Figure 1 shows the articles assigned to the categories “exclusively” and “significantly”. The bulk of the articles in “exclusively” journals include, somewhat predictably, the articles in *Literary and Linguistic Computing* or its successor, *Digital Scholarship in the Humanities*. However, we noticed that, when considering the most represented journals in our dataset (i.e., those with most articles) in Figure 2, the articles are visually arranged by journal and tend to follow field-specific patterns: digital libraries (*IJDL*, *D-Lib*), computational linguistics (*Computational Linguistics*, the *Journal of Quantitative Linguistics*), artificial intelligence and society (*AI & Society*), and DH (*Literary and Linguistic Computing*, *Digital Scholarship in the Humanities*). Instead, the articles in *Language Resources and Evaluation* are more evenly spread across different field clusters.

When we consider citation clusters detected using a modularity-maximizing method (Blondel et al., 2008), in Figure 3, we observe a modular structure with a high correlation with respect to the publication venue. The main focus of the DH cluster (number 1 in Figure 3) are quantitative literary studies, e.g., stylometry and authorship attribution. Other clusters cover the DH-related areas of computational linguistics and natural language processing (2,3,4,5,8), digital libraries (7) and AI and society (6). As far as we could notice from this graph, DH publications tend to connect to related disciplinary areas, even if each area maintained its distinctiveness. The publication venue remains a key trait of the intellectual structure of the DH.

4 Conclusion

In this article, we proposed an approach to find digital humanities publications by iterating between a list of journals (top-down) and its expansion using citation clustering (bottom-up). In this way, we were able to propose a first version of a list of digital humanities journals split in three categories: those that are “exclusively”, “significantly” and “marginally” related to the digital humanities. We assessed the

Table 2: Database coverage for journals exclusively devoted to digital humanities scholarship.

Journal	WoS	Scopus	Crossref	Dimensions
Computers and the Humanities	663	806	1465	
Digital Humanities Quarterly (DHQ)		38		
Digital Medievalist			67	66
Digital Scholarship in the Humanities (DSH)	254	237	327	388
Digital Studies / Le champ numérique			230	
Digitális Bölcsészet / Digital Humanities				19
Frontiers in Digital Humanities			58	66
International Journal of Digital Humanities				
International Journal of Humanities and Arts Computing		15	253	475
Journal of Cultural Analytics			13	27
Journal of Data Mining and Digital Humanities				
Journal of Digital Archives and Digital Humanities				
Journal of Digital Humanities				
Journal of the Japanese Association for Digital Humanities			15	21
Journal of the Text Encoding Initiative			73	
Journal on Computing and Cultural Heritage (JOCCH)	75	178		202
Literary and Linguistics Computing	251	584	1465	1450
Revista de humanidades digitales			23	37
Umanistica Digitale				
Total	1243	1858	3989	2751

coverage of Web of Science, Scopus, Crossref and Dimensions in this respect, finding that Crossref has the best coverage of “exclusively” digital humanities journals, while Dimensions has the best coverage of the number digital humanities-related articles overall. We discussed a first map of research using citation data from Dimensions, highlighting how just one third of the articles in Dimensions are connected with each other via citations. We further found that digital humanities articles are connected via citations to computational linguistics and natural language processing, digital libraries and other developing areas such as AI and society. Nevertheless, we also found that the venues (i.e., the journals) strongly overlap with citation clusters, and are a key trait of the intellectual organization of digital humanities research.

We acknowledge that our work is still in progress, and thus it has a set of limitations which we plan to address in the future. In particular, we plan to include additional bibliographic entity types in addition to journal articles (e.g., books), and to also include the COCI (Heibi et al., 2019) and Microsoft Academic citation indexes to the comparison. Coverage will also be assessed at the article level (i.e., which citation index contains which articles) and chronologically. Lastly, we will elaborate on the map of research by including a comparison across all indexes.

Acknowledgements

The authors would like to thank the Centre for Science and Technology Studies (CWTS), Leiden University, for providing access to their databases and computing facilities. This work was in part conducted when Spinaci was visiting CWTS.

References

- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. *Fast unfolding of communities in large networks*. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Jennie M. Burroughs. 2017. *No Uniform Culture: Patterns of Collaborative Research in the Humanities*. *portal: Libraries and the Academy* 17(3):507–527. <https://doi.org/10.1353/pla.2017.0032>

- Giovanni Colavizza, Matteo Romanello, and Frédéric Kaplan. 2018. The references of references: a method to enrich humanities library catalogs with citation data. *International Journal on Digital Libraries* 19(2-3):151–161. <https://doi.org/10.1007/s00799-017-0210-1>
- Jin Gao, Oliver Duke-Williams, and Simon Mahony. 2017. The Intellectual Structure of Digital Humanities: An Author Co-Citation Analysis. In *Digital Humanities Conference Proceedings*.
- Jin Gao, Julianne Nyhan, Oliver Duke-Williams, and Simon Mahony. 2018. Visualising the Digital Humanities Community: A Comparison Study Between Citation Network and Social Network. In *Digital Humanities Conference Proceedings*.
- Martin Grandjean. 2016. A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities* 3(1). <https://doi.org/10.1080/23311983.2016.1171458>
- Björn Hammarfelt. 2016. Beyond Coverage: Toward a Bibliometrics for the Humanities. In Michael Ochsner, Sven E. Hug, and Hans-Dieter Daniel, editors, *Research Assessment in the Humanities*, Springer International Publishing, Cham, pages 115–131.
- Anne-Wil Harzing and Satu Alakangas. 2016. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics* 106(2):787–804. <https://doi.org/10.1007/s11192-015-1798-9>
- Ivan Heibi, Silvio Peroni, and David Shotton. 2019. COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics* <https://doi.org/10.1007/s11192-019-03217-6>
- Daniel W. Hook, Simon J. Porter, and Christian Herzog. 2018. Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics* 3. <https://doi.org/10.3389/frma.2018.00023>
- Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 9(6):e98679. <https://doi.org/10.1371/journal.pone.0098679>
- Loet Leydesdorff and Alkim Almila Akdag Salah. 2010. Maps on the basis of the Arts & Humanities Citation Index: The journals Leonardo and Art Journal versus “digital humanities” as a topic. *Journal of the American Society for Information Science and Technology* 61(4):787–801. <https://doi.org/10.1002/asi.21303>
- Anton J. Nederhof. 2006. Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics* 66(1):81–100. <https://doi.org/10.1007/s11192-006-0007-2>
- Julianne Nyhan and Oliver Duke-Williams. 2014. Joint and multi-authored publication patterns in the Digital Humanities. *Literary and Linguistic Computing* 29(3):387–399. <https://doi.org/10.1093/llc/fqu018>
- Alkim Almila Akdag Salah, Andrea Scharnhorst, and Sally Wyatt. 2015. Analysing an Academic Field through the Lenses of Internet Science: Digital Humanities as a Virtual Community. In Thanassis Tiropanis, Athena Vakali, Laura Sartori, and Pete Burnap, editors, *Internet Science*, Springer International Publishing, Cham, volume 9089, pages 78–89. https://doi.org/10.1007/978-3-319-18609-2_6
- Gianmarco Spinaci, Giovanni Colavizza, and Silvio Peroni. 2019. List of digital humanities journals. <https://doi.org/10.5281/zenodo.3406564>
- Rachele Sprugnoli, Gabriella Pardelli, Federico Boschetti, and Riccardo Del Gratta. 2019. Un’Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale. *Umanistica Digitale* 5. <https://doi.org/10.6092/issn.2532-8816/8581>
- Muh-Chyun Tang, Yun Jen Cheng, and Kuang Hua Chen. 2017. A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses. *Scientometrics* 113(2):985–1008. <https://doi.org/10.1007/s11192-017-2496-6>
- Vincent A. Traag, Ludo Waltman, and Nees J. van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* 9(1). <https://doi.org/10.1038/s41598-019-41695-z>
- Ludo Waltman and Nees Jan van Eck. 2012. A new methodology for constructing a publication-level classification system of science: A New Methodology for Constructing a Publication-Level Classification System of Science. *Journal of the American Society for Information Science and Technology* 63(12):2378–2392. <https://doi.org/10.1002/asi.22748>