Benito-Santos, A., & Sánchez, R. T. (2020). A data-driven introduction to authors, readings, and techniques in visualization for the digital humanities. IEEE Computer Graphics and Applications, 40(3), 45-57. https://doi.org/10.1109/MCG.2020.2973945

A Data-Driven Introduction to Authors, Readings, and Techniques in Visualization for the Digital Humanities

Alejandro Benito-Santos and Roberto Therón Sánchez
University of Salamanca

***Abstract*—The newly rediscovered frontier between data visualization and the digital humanities has proven to be an exciting field of experimentation for scholars from both disciplines. This fruitful collaboration is attracting researchers from other areas of science who may be willing to create visual analysis tools that promote humanities research in its many forms. However, as the collaboration grows in complexity, it may become intimidating for these scholars to get engaged in the discipline. To facilitate this task, we have built an introduction to visualization for the digital humanities that sits on a data-driven stance adopted by the authors. In order to construct a dataset representative of the discipline, we analyze citations from a core corpus on 300 publications in visualization for the humanities obtained from recent editions of the InfoVis Vis4DH workshop, the ADHO Digital Humanities Conference, and the specialized digital humanities journal Digital Humanities Quarterly. From here, we extract referenced works**

and analyze more than 1900 publications in search of citation patterns, prominent authors in the field, and other interesting insights. Finally, following the path set by other researchers in the visualization and Human–Computer Interaction (HCI) communities, we analyze paper keywords to identify significant themes and research opportunities in the field.

The collaboration between the digital humanities (DH) and the data visualization communities has grown larger in recent years. This fact is attracting scholars from both areas of knowledge who are keen on designing tools that can reveal insight on humanistic data in an increasingly broader range of disciplines.

However, precisely due to its novelty and inherent interdisciplinary character, this collaboration often is hard to articulate, as it poses very particular challenges in the visualization design process and the construction of shared design spaces.[1,2] For these reasons, the scholars' initial excitement may soon become disenchantment if these challenges are not addressed from the very initial stages of the collaboration. In this work, we attempt to provide new researchers with an interest in the field with a series of recommended readings, authors, and terminology derived from a meta-analysis of the discipline's current state in a very concise yet effective manner.

In this regard, we hope our work succeeds at the task of orienting interdisciplinary visualization researchers, and that the contents of this article can lead them to examples, best practices, and resources to ease the production of future quality research on visualization for the humanities.

In order to produce a critical summarization of a scholarly field, it is a recurring first step in mappings studies, surveys, and literature reviews to invest time in clearly defining what exactly is to be considered in the study. Once a definition of the subject of the study has been reached, the researcher employs it to systematically retrieve publications from a selection of sources (e.g., online scientific databases or search engines) that are further analyzed at later stages. However, defining the DH is a challenging task that inevitably builds on rather shaky epistemological grounds, and therefore it has been (and still is) the subject of important discussions in the community. For example, the 2012 edition of *Debates in the Digital Humanities* accounted for 21 definitions of the DH alone.[3] Indeed, some authors argue that this continuous process of questioning the self-identify is one of the core values of the DH, and therefore this question may never be resolved. For these reasons, producing a definition of "visualization in the DH," that satisfied both humanities and visualization scholars at the same time seemed overwhelming to us. Not only this, but capturing this definition into a textual query string that could be used to query an online scientific database to retrieve relevant publications was something that we wanted to avoid.

Instead, we decided to adopt a more practical stance to address this issue, which brought us to rely on data-driven, quantitative techniques that supported the foundations of this work. To this end, we analyzed visualization contributions to two core venues intimately related to visualization for the humanities: the Vis4DH InfoVis workshop[*] and the ADHO Digital Humanities Conference[†]. From these works, we extracted referenced publications to construct a dataset of more than 1900 journal articles, conference submissions, books, and web pages (see Figure 1) to which we applied several bibliometric techniques to answer a set of research questions that we outline as follows:

- *RQ.1.* What are the most influential...
  - *RQ.1.1.* publications?
  - *RQ.1.2.* authors?
- *RQ.2.* How long is the community's collective memory and how is it distributed in time?
- *RQ.3.* What are the concepts generating a more significant number of publications?
- *RQ.4.* What are the main themes in the DH Visualization practice?
- *RQ.5.* How do these themes relate to each other?

RELATED WORK

To answer the research questions that were proposed at the beginning of the study, and to gain insight into the discipline of visualization in the DH,

[*] http://www.vis4dh.org/
[†] http://adho.org

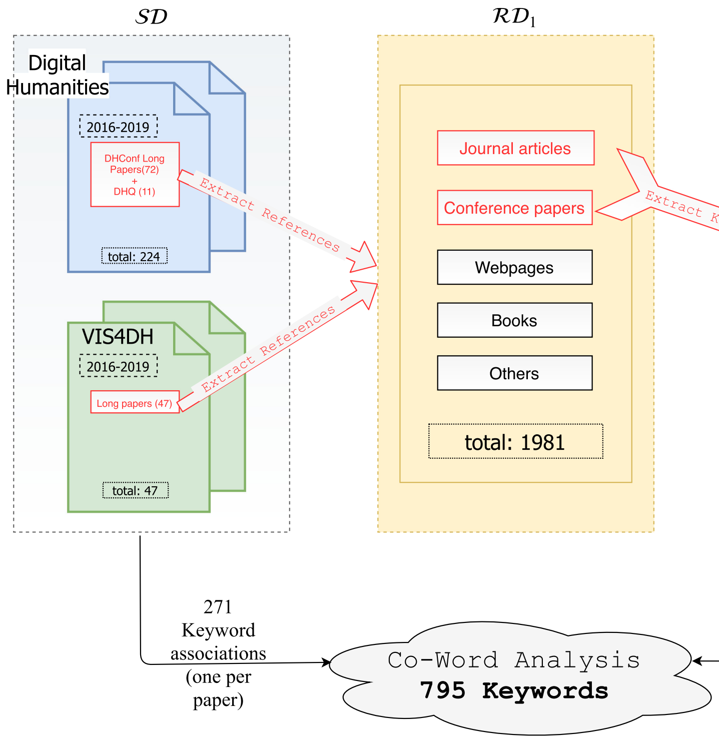


Figure 1. Construction process of the keywords dataset $\mathcal{K}$ and the intermediate publication datasets originating at $\mathcal{SD}$. The final result is a dataset of 1942 unique keywords related to the DH visualization practice.

our study relies on previous works in the visualization, human–computer interaction (HCI), and bibliometric domains that we introduce in this section.

Mapping Visualization

Many scholars have attempted to map the scientific landscape of visualization in different knowledge domains employing keywords.[4] Moreover, the task of understanding vast amounts of research papers is a longstanding HCI problem that has produced significant contributions in the past.[5] Regarding the mapping of the discipline, an important recent advance is the work by Isenberg *et al.*,[6] who compiled a dataset of visualization research papers presented at IEEE VIS (VisWeek) in the period 1990–2018. Since its publication, different visual solutions to explore the dataset have been proposed, ranging from the visualization of co-citation and co-authorship patterns[7] to the visualization of topic models,[8] or a combination of approaches based on network analysis and natural language processing (NLP) techniques.[9] More related to our study, the authors of the dataset performed co-word analysis on the research paper keywords [10] that has greatly inspired our work.

Surveying Visualization in the DH

There exist notable previous attempts to produce reviews on visualization for the DH. For example, Jänicke *et al.*[11] evaluate past visualization approaches to support distant and close reading tasks on a variety of textual data. This review was later extended by Jänicke *et al.*[12] to include other kinds of text visualization. More recently, Windhager *et al.* review visual solutions to explore cultural heritage collections.[13] As opposed to our study, these works focus on specific subdomains of the DH practice, and therefore are not able to offer a complete view of the discipline.

Co-Word Analysis

Co-word analysis is a bibliometric quantitative technique that is rooted in the idea that a paper's keywords are able to describe its contents correctly. Therefore, it can be assumed that the co-occurrence of keywords in a publication denotes a kind of implicit conceptual link between the ideas represented by such terms. The study of the frequency patterns emerging from these links has been long employed to measure the development of science in a wide variety of knowledge domains such as chemistry, software engineering, consumer behavior, patent analysis, ubiquitous computing, library and information science, to name a few. In particular, the works by Liu *et al.*,[14] who successfully analyzed publication keywords in 20 editions of the CHI conference, and Isenberg *et al.*, with *vispubdata*[6] and their study on visualization keywords,[10] are good examples of the validity of co-word analysis as a tool to produce comprehensive studies on these areas.

Strategic Diagrams

Strategic diagrams combine co-word with network and clustering techniques and have been typically employed to produce maps of the intellectual structure of a discipline in a variety of topics.[10,14] The process to generate these diagrams is straightforward: First, a network of keywords is generated employing different methods, which can be simple co-occurrence (two keywords are connected if they appear on the same paper) or correlation (two keywords are connected if they are positively correlated).[10] The network is then partitioned into clusters (or subnetworks), usually making use of unsupervised hierarchical clustering algorithms. For each of the resulting clusters, two key measures are calculated: *density* and *centrality*. The first measure, density, "characterizes the strength of the links that tie the words making up the cluster together"[15] and depicts the ability of a cluster to constitute a coherent and integrated whole, which can be understood as a measure of the theme's development. Therefore, the higher the density of the links of the cluster, the more likely it is to contain inseparable expressions. The second measure, centrality, measures the strength and number of interactions of the cluster with other parts of the network and it is employed to quantify the importance of a theme in the research field under study. The more and stronger connections a cluster has, the more central the theme is in respect to the whole network.

The combination of these two concepts is then plotted in the *strategic diagram*, a two-dimensional representation of density (y-axis) and centrality (x-axis). The space is usually divided into four quadrants with separation lines corresponding to the median density and centrality values of all the previously calculated clusters. This disposition is presented in Figure 2.

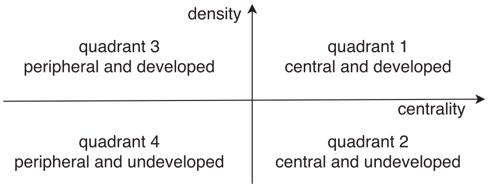


Figure 2. Strategic diagram with its four main quadrants explained. The location of a cluster in the diagram characterizes the theme it represents in the context of the discipline.

Below we provide details on how these areas usually are interpreted in the study of a given research field.

- *Quadrant 1* (see top-right of Figure 2): Internally coherent (high density) and central (strongly connected to other subnetworks) themes to the research network. These clusters are considered to be the "motor themes" of the discipline. They are dealt with systematically and over a long period, probably by a well-defined group of researchers.
- *Quadrant 2* (see bottom-right of Figure 2): Clusters in quadrant two are strongly connected to other clusters, but the density of their internal links is low. They are interpreted as connectors of other clusters or emerging themes that are starting to become central but have not yet been the object of a significant number of contributions.
- *Quadrant 3* (see top-left of Figure 2): These clusters are not well communicated with other parts of the network, but the strength of their internal links denotes research problems whose study is already well-developed. It is often the case that these clusters were central in the past, but their relative importance has decayed in recent times.
- *Quadrant 4* (see bottom-left of Figure 2): Within this category fall the clusters that are peripheral and underdeveloped. They are considered marginal in the global research network.

DATASETS

A critical step of bibliometric studies is the selection of publications to consider. To this end, researchers usually rely on online scientific databases from which this information is extracted via query strings. Query strings result from the application of a search strategy that is aligned with the aim of the study. There exist different methods to construct a query string in a systematic manner, although some authors have noted that they might be difficult to apply when the subject of the study is hard to define.[16] DH-specific publications are hard to find in scientific databases since many of them are not indexed (e.g., DH conference papers). All datasets collected in this study can be consulted in the supplementary materials, which are available in the IEEE Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/MCG.2020.2973945.

Our study is based on a core set of publications in visualization for the humanities ($\mathcal{SD}$ for "Seed Dataset") which, in turn, is twofold. First, it contains publications from the VisWeek Vis4DH workshop that represent the engineering/visualization community. Second, it also contains publications from the ADHO DH, which is completed with the addition of selected works from Digital Humanities Quarterly (DHQ). These two sources are meant to represent the humanities side of the community. The collection process is outlined as follows:

1) *Engineering Dataset:*
   a) *Vis4DH Workshop*: This workshop is a colocated event with IEEE VIS conferences that kicked off in 2016. The workshop, initially supported by visualization researchers with common experience in the DH, attracted stakeholders from different academic backgrounds in the humanities and science, promoting a series of publications, debates, and panel discussions framed under the particular interdisciplinary collaboration setting that is characteristic of the discipline.[17] Initially, a total of 38 publications published in the 2016 (17), 2017 (10) and 2018 (11) and 2019 (9) editions of the workshop were included in $\mathcal{SD}$.
2) *Humanities Dataset:*
   a) *ADHO DH Conference*: The DH Conference is an annual event organized by the umbrella organization known as the Association of Digital Humanities Organizations (ADHO). Due to the popularization and increasing availability of visualization techniques in recent years, there has been a great surge in the number of papers of this kind submitted to the conference.[1] In order to select a sufficient number of papers, we employed the following strategy: first, we downloaded the conference abstracts in the period 2016–2019 (4 editions, to overlap in time with the years the Vis4DH workshop has taken place). Then, we included all papers matching the regular expression *"[Vv]isua*"* in their title, list of keywords, or list of topics. This resulted in 214 candidate contributions (see Figure in the supplementary materials, available online).
   b) Digital Humanities Quarterly: Following the same rationale as we did with publications on the DH conference, we included works from the DH-centric journal DHQ in our seed dataset. We included 15 extra works with this procedure for a total of 229 papers representing the humanities side in our seed dataset.

Before moving on to other sections of the paper, here we acknowledge some limitations related to the research methodology that was adopted in this work. For example, it is worth noting that, due to limiting publications in $\mathcal{SD}$ to only those appearing in the Vis4DH workshop, DH Conference and DHQ journal, we might have left out certain works that should have been initially included. Although we believe the citations dataset can (and should be enhanced by the community in the future: see for example the work by Isenberg *et al.*[6]), we believe the citation analysis captures a majority of relevant works for the DH practice that are good enough to propose an initial analysis.

By composing $\mathcal{SD}$ of a mix of humanities-related publications presented in a visualization conference and visualization-related papers presented in a DH venue, we ensured enough representative works of scholars pertaining to both areas of knowledge were included in the study while avoiding to employ a query search string, which would have been very difficult to construct, given the problematic previously presented in this article.

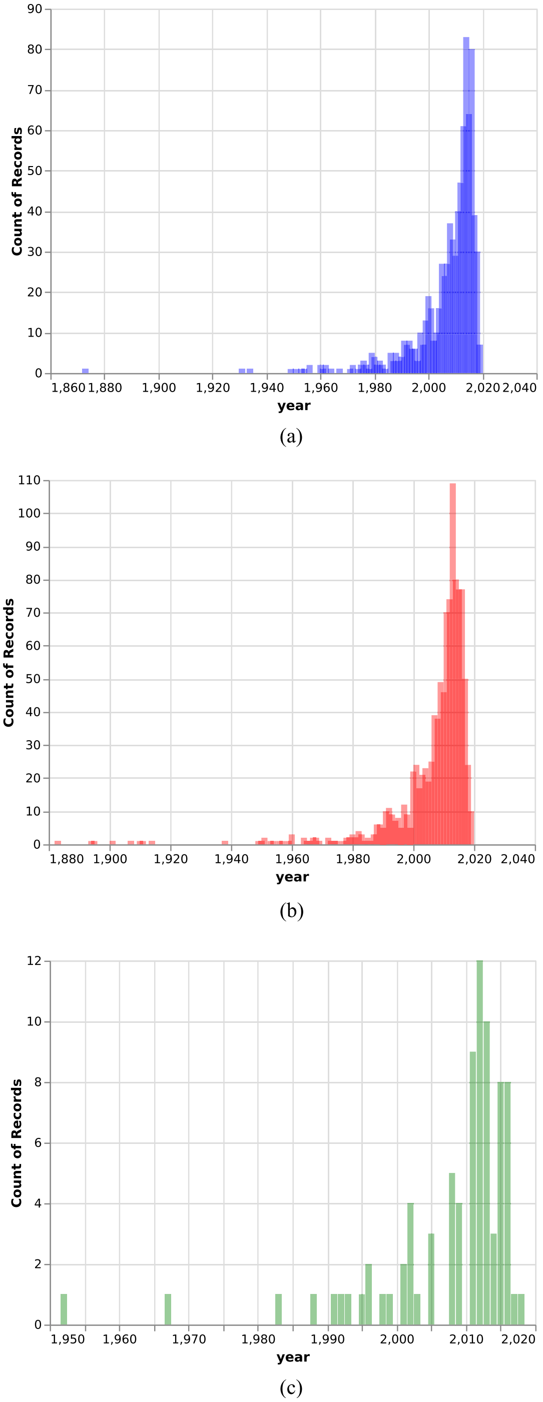
Additionally, we extracted all references found in long papers in $\mathcal{SD}$ to construct a new dataset,

$\mathcal{RD}_1$, which originally contained 1981 referenced works (excluding self-references) including journal publications, conference papers, books/book chapters, webpages, blogposts, and others information. In total, we obtained 830 citations from works in the humanities subset of the seed dataset, whereas 1068 could be traced to works in contributions to any of the Vis4DH workshops. Eighty-three publications (4% of the total) were referenced from both subsets of publications. From this list of publications, we extracted author keywords (when applicable, note that some works in $\mathcal{RD}_1$ are books or blog posts that do not contain author-assigned keywords), forming $\mathcal{RD}_2$, obtaining 571 papers or keyword associations (mainly from journal and conference papers). These 571 keywords were merged with those from the seed dataset (224), to base our co-word analysis in a total of 795 keywords.

Insights on Cited Publications

An analysis of the temporal distribution of the publications cited by works published on VIS or DH venues reveals very similar citation patterns, including works that go as far back as the last decades of the nineteenth century (see Figure 3). In Table 1 the top cited papers, up to rank 5, are displayed. The main themes found in these key works are: text visualization/distant reading, poetry visualization, graphs/network visualization, and visualization design theory and best practices.

Following a similar process, from $\mathcal{RD}_1$ we extracted the most cited books (listed, up to rank 4, in Table 2). Unsurprisingly, in this publications set we can find Franco Moretti's pioneer works on distant reading *Graphs, Maps, Trees: Abstract Models for a Literary History* and *Distant Reading,* which supposed a turning point in the modern development of the DH. Also worth noting is Johanna Drucker's *Graphesis,* in which the author critically comments on different aspects of the DH from visualization and design theory perspective. Also, two classic data/information visualization books are shown at the bottom of the table, Card's *Readings in Information Visualization: Using Vision to Think* and Tufte's *The Visual Display of Quantitative Information.* Interestingly, these two volumes, which may sound more familiar to data visualization practitioners,



(a)

(b)

(c)

Figure 3. Temporal distribution of works cited from (a) vis/engineering publications in the seed dataset (blue), (b) humanities publications in the seed dataset (red), and (c) both (in green). The collective memory of both communities seems to follow a similar pattern in both cases. Interestingly, the oldest work cited by publications in the vis and humanities seed datasets is "The origins of intelligence in children" (1952) by J. Piaget.

Table 1. Top cited papers in dataset $\mathcal{RD}_1$.

#	Author	Title	Venue	Year
15	S. Jänicke *et al.*	On close and distant reading in digital humanities: A survey and future challenges	EuroVis	2015
10	B. Shneiderman	The eyes have it: A task by data type taxonomy for information visualizations	InfoVIS	1996
	J. Drucker	Humanities approaches to graphical display	DHQ	2011
8	A. Thudt *et al.*	The Bohemian bookshelf: Supporting serendipitous book discoveries through information visualization	CHI	2012
	N. McCurdy *et al.*	Poemage: Visualizing the sonic topology of a poem	TVCG	2016
6	A. Gibbs and T. Owens	Building better digital humanities tools: Toward broader audiences and user-centered designs	DHQ	2012
5	M. Whitelaw	Generous interfaces for digital cultural collections	DHQ	2015
	M. Dörk *et al.*	Critical InfoVis: Exploring the politics of visualization	CHI	2013
	U. Hinrichs	In defense of sandcastles: Research thinking through visualization	DH	2015
	S. Jänicke	Visual text analysis in digital humanities	EuroVis	2016

are the oldest in the listing. We hypothesize this fact may be due to a certain degree of stagnation in the DH visualization community and may be indicative of the need for novel techniques resulting from renovated visualization design processes conceived for the DH practice.

Keywords Dataset

As it has been explained before, the dataset of keywords that was used to perform the co-word analyses contains author-assigned keywords from publications in the seed and $\mathcal{RD}_2$ datasets. As it is usual in this kind of approaches, we removed domain stopwords using the following regular expression: *“(data—information).?visuali[sz]ation [s]?,” “visual analytics”* and *“digital humanities.”* After removal, the 795 papers containing author-assigned keywords yielded a total of 2511 unique keywords, occurring 4015 times (5.05 author keywords per paper).

ANALYSIS PROCESS

In this section, we provide details on the calculations and algorithms that were applied to the keywords dataset obtained in the previous step in order to create the strategic diagram and the keywords network. All code was implemented in a *Jupyter Python* environment employing the libraries *nltk, pandas, bokeh*, and *networkx*.

Preprocessing of Keywords

To group keywords of similar themes, some authors have relied in the past on an expert coding of the keywords.[10] To accelerate the analytic process, we, instead, designed an automatic method that yielded similar quality results and also worked well in a smaller corpus such as ours. The procedure, which is well known in the NLP literature, involved the tokenization and stemming of keywords, in which the multiterm words are split into their constituent parts and reduced to their root form. We employed Porter’s stemming algorithm as it yielded the most satisfactory results. In a similar manner as we did with the original

Table 2. Top cited books in dataset $\mathcal{RD}_1$.

#	Author	Title	Year
13	F. Moretti	Graphs, maps, trees: Abstract models for a literary history	2005
7	F. Moretti	Distant reading	2013
6	M.L. Jockers	Macroanalysis: Digital methods and literary history	2014
5	J. Drucker	Graphesis: Visual forms of knowledge production	2014
4	E. R. Tufte and P. Graves-Morris	Graphesis: Visual forms of knowledge production	2012
	S. Rücker *et al.*	Visual interface design for digital cultural heritage	2011

Table 3. Three examples of hierarchy groups resulting from the stemming of keywords. In bold, tokens that were matched to an upper element of the hierarchy.

stem	keywords
american	**american** culture, **american** history, **american** television, c19 **american** literature, nineteenth century **american**, wright **american** fiction corpus
corpu	cbeta **corpus**, **corpus** analysis, **corpus** analysis tool, **corpus** examples, **corpus** linguistics, **corpus** studies, **corpus** visualization, **corpus** workbench, diachronic **corpus**, n-gram **corpus**, wright american fiction **corpus**
cultur	american **culture**, **cultural** artefacts, **cultural** collections, **cultural** differences, **cultural** heritage/ history, **cultural** probe, **cultural** studies, digital **cultural heritage**, online **cultural** heritage, personalized access to **cultural** heritage, popular **culture**, virtual **cultural** heritage, visual **culture**

keywords, the following stems were also removed from the analysis: "analysi," "digit," "human," "visual," "analyt," "dh," "data," "algorithm," "comput" as they referred to generic elements of computer science and the domain under study. Furthermore, uninformative keyword stems, appearing less than six times were also removed. At the end of this process, all individual keyword tokens had been translated into their correspondent root form (see Table 3), yielding a total of 106 tokens (dataset $\mathcal{K}$) that were employed to construct a correlation matrix of keywords.

Ultimately, the tokenization and stemming of the keywords modified the distributional model of the keywords (see Figure 4) in the corpus by organizing them in a *hierarchy*. This change is key to reveal insights that could not be reached from studying the distribution function from the previous situation (see Figure 5). For example, in the new situation it can be seen that the particle "network" has been promoted to the first position in the new distribution. However, in the previous case the occurrences of the term appeared in the 5th ("network analysis"), 8th ("networks"), and 13th ("network visualization"). This grouping promoted the term to the #1 frequency rank in the new distribution, highlighting the key role of "networks" as a transversal theme in the discipline. In a similar effect, the term "gis" is removed from the list of top words after preprocessing, giving way to the more general concept "map" that is now placed at rank #6.

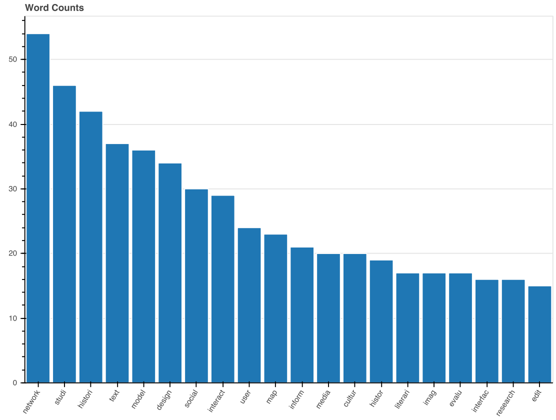
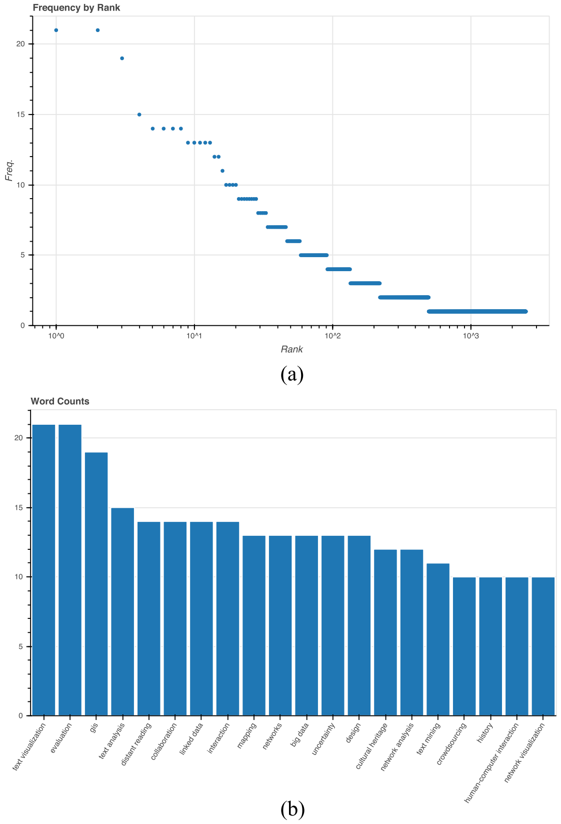


Figure 4. 20 most common roots after stemming of the keywords. The stemming effectively changed the distributional model of the keywords, revealing different patterns to what could be observed in the prestemming situation (shown in Figure 5).



(a)

(b)

Figure 5. (a) Keywords frequency by rank (log-scaled), (b) 20 most common keywords. The observed distributional model seems to be in line with findings from similar studies.[10,14]

Table 4. Hierarchical cluster results for the $\mathcal{K}$ dataset. Members are sorted by frequency, with the two most popular terms in bold.

ID	Members	N	#	cw-#	centr.	dens.
1	**user, inform**, interfac, retriev, search	9	10.222	1.639	0.544	0.163
2	**languag, process**, natur	3	7	3.333	0.483	0.485
3	**imag, annot**, graphic, tool	4	10.5	2.5	0.682	0.239
4	**semant, link**, web	3	11.333	3.333	0.723	0.278
5	**studi, literari**, literatur, linguist, corpu	5	17.6	2.8	0.764	0.132
6	**recognit, relat**, extract, featur, name	5	4.2	1.4	0.225	0.325
7	**evalu, graph**, chart, multipl	4	7.5	1.333	0.413	0.322
8	**cultur, collect**, heritag, explor	4	12.75	3.167	0.749	0.222
9	**histor, ontolog**, place, servic, event	5	6.2	1	0.306	0.199
10	**text, mine**, vector, word	4	12.5	2.167	0.530	0.199
11	**model, edit**, topic, scholarli	4	15	2.833	0.698	0.181
12	**manag, databas**, plan, architectur, project	5	4.6	0.8	0.299	0.177
13	**design, research**, scienc, knowledg, technolog	7	11	1.14286	0.533	0.116
14	**mediev, align**, dynam, program	4	4.75	1.167	0.330	0.241
15	**histori, collabor**, art, archiv, learn	8	12.25	1.149	0.625	0.0796
16	**map, media**, spatial, 3 d, archeolog	9	10.111	0.889	0.585	0.069
17	**represent, classif**, narr, detect	4	3.75	0.667	0.181	0.182
18	**network, social**, commun, critic, cartographi	7	16	1.429	0.642	0.069
19	**interact, video**, uncertainti, document, method	16	6.25	0.142	0.356	0.011

Correlation Matrix and Clustering

After we applied the preprocessing step outlined in the last section, we constructed a boolean document-term matrix in which we annotated when a certain token was contained in a document. After, we used this matrix to calculate a correlation matrix on the keywords. Finally, the keywords were hierarchically clustered using Ward's method and a squared Euclidean metric. Instead of relying in a predefined number *k*, we employed a maximum distance criterion to form clusters. Under this assumption, any two observations in a cluster shall not have cophenetic distance greater than 95% of the maximum total distance between two any two pairs in the dataset.

RESULTS

In this section, we discuss, in light of research questions *RQ.4* and *RQ.5*, the keyword clusters, network, and strategic diagram that were built following the procedure introduced in previous sections. In Table 4, we display the results of the semisupervised hierarchical clustering process that was applied to the keywords. For each cluster, we show the following.

- *Members:* The set of keyword stems that form the cluster. The two top keywords of each clusters are written in bold.
- *Size (N):* The number of keywords that are in the cluster.
- *Frequency (F):* Average frequency for all terms in the cluster.
- *Co-Word Frequency (CW-F):* Average number of times any two given keywords of the cluster can be seen together in the documents collection.
- *Centrality:* Degree of the interaction of the cluster with any other parts of the network.

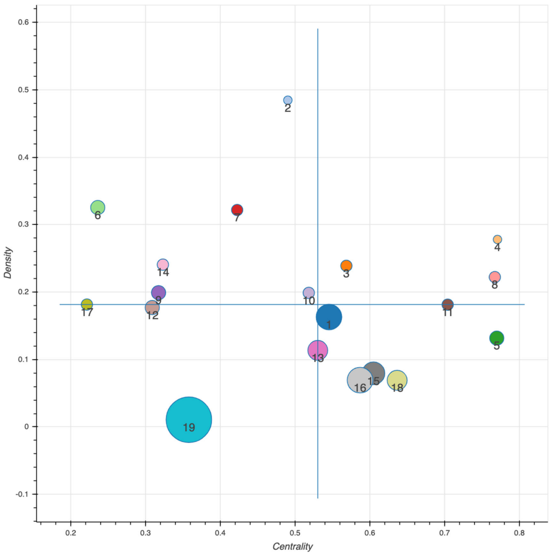


Figure 6. Strategic diagram for the 19 hierarchical clusters that were created. Blue lines indicate the medians.

It measures how well communicated the topic is with other themes. We calculated it as the average betweenness centrality using the standard value two for the *K*-step reach.

- *Density:* Topic's degree of internal cohesion. It measures how strong the connections between members of the same cluster are.[15] This is calculated as the average correlation between all member pairs in the cluster.

The 19 clusters are plotted in the strategic diagram of Figure 6, according to their centrality and density measures. Additionally, we complement this information with the keywords network (see Figure 7), which aims to highlight structural patterns and other interesting information not easily identifiable in the strategic diagram. The network links depict positively correlated pairs of keywords, which were obtained from the correlation matrix. Correlations ≤ 0.20 were omitted. The visualizations were created using the Python library Bokeh[‡]. The network layout employs *networkx's* implementation of the Kamada–Kawai graph layout algorithm. In the network visualization, the circle sizes and edge thickness follow a logarithmic scale that is dependant on the term's frequency and correlation strength, respectively.

DISCUSSION

The algorithm successfully organizes keywords in 19 main themes that were found in the corpus. Remarkably, the smaller clusters showing higher densities (and therefore appearing on the upper side of the strategic diagram) are easily interpretable. This can be observed for example in cluster 2 ("natural language processing"), a cluster that from its position in the graph seems to be of major importance in the discipline. In a similar manner, we can find cluster 4 ("semantic web" and "open linked data") in the first quadrant. Cluster 10, that is placed right at the crossing of the medians can be easily interpreted as text analysis based on word embeddings, a discipline that has attracted much interest from the community due to its recent popularization. Perhaps in the short future, we will see novel techniques in the DH practice that employ more modern and powerful linguistic models beyond word2vec in a variety of DH research contexts beyond text summarization, such as translation of ancient languages or others. Cluster 19, the largest of it all appearing in quadrant 4, contains terms that are more difficult to relate. Interestingly, it catches our attention the word "uncertainty," which is becoming a hot topic among data visualization practitioners in recent years. As it happens, two out of nine papers submitted to Vis4DH 2019 contained themes related to the management and display of uncertainty in visualization for the humanities, a trend that we are expecting to continue in forthcoming years.

Looking at the right of the chart, cluster 11 (topic models and scholarly editing software) appears to be a central and well-established theme in the discipline by looking at their position in the diagram. In Figure 7, it can be seen how topic models do not seem to be particularly attached to any other themes of the discipline, which means they maintain a relatively constant high correlation with other terms shown (at least 0.2). Therefore, it is reasonable to think that topic models are employed in a broad range of DH applications due to their summarization capabilities and close relationship to distant reading. We invite the reader to explore the dataset[§] using the visualization notebook[¶] set up for the purpose.

[‡] https://bokeh.pydata.org

[§] https://docs.google.com/spreadsheets/d/1TCnEIfbyow7s7_qnl_KZs4cUZjrt4bpz5C8VJLe-XIA/

[¶] https://github.com/ale0xb/vis4dh-analysis

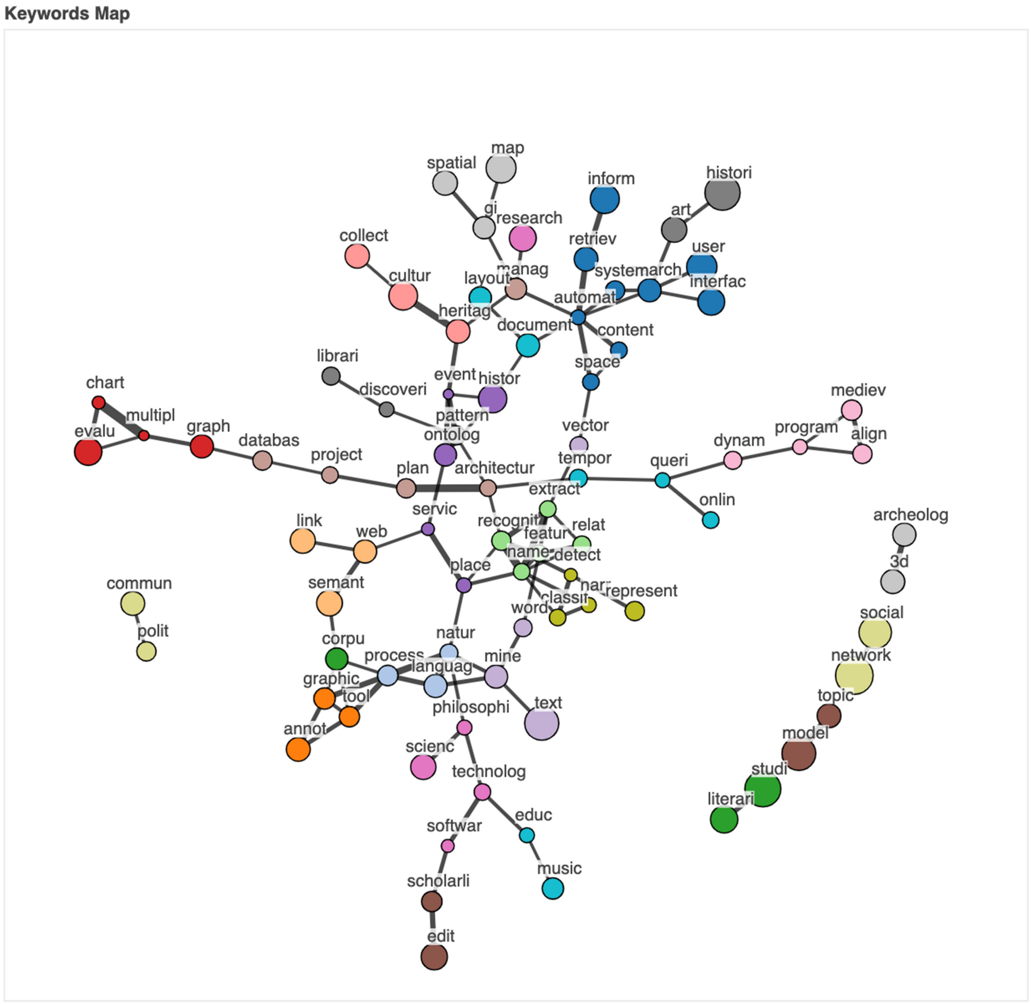


Figure 7. Keyword stems map. Edges represent a correlation strength of ≥ 0.20 between two nodes. Thicker links depict higher correlation values. Absolute keyword frequency was encoded in circle size.

LIMITATIONS

Regarding the preprocessing step that was applied to the keywords, we acknowledge its capacity to explain the themes may be suboptimal when compared to other approaches employing an expert coding of the keywords. Therefore, although its analytical capacity may be inferior, it presents other advantages, such as a minimum time investment or unsupervised character, that may make it a good fit in a broad range of research contexts. Also, in relation to this technique, we are aware that the stemming algorithm that was applied to the data may introduce noise as nonexistent connections between concepts. This issue is hard to resolve since it is related to the disambiguation of terms that share the same root but they are semantically and etymologically different. While some authors in the NLP literature address this by including part-of-speech tagging in the analysis, this can be very hard to achieve in case of stand alone expressions such as keywords. Therefore, it might be worth looking into alternative language/co-word models or even move to full-text kinds of analysis.[18]

Finally, some publications had to be excluded from the co-word analysis because the authors did not add keywords to their work, rendering this kind of analysis inadequate for these publications. Although automatic keyphrase/keyword extraction techniques exist in the literature, substantial more work is required to understand the implications of substituting human-generated keywords with their machine-generated counterparts, especially in the task of extracting knowledge from vast amounts of scientific literature.

CONCLUSION

In this article, we have presented a systematic, data-driven approach to provide an introduction to an uncharted interdisciplinary research field as visualization for the DH. By combining numerical overviews with unsupervised data science and bibliometric techniques, we were able to capture the discipline's current state while avoiding common pitfalls of more traditional analysis workflows, which would have been hard to apply in this context. Furthermore, we share our dataset (accessible at https://docs.google.com/spreadsheets/d/1TCnEIfbyow7s7_qnl_KZs4cUZjrt4bpz5C8VJLeXIA/) with researchers who might be willing to use it and expand it in future research. Ultimately, we hope the findings of this research may be of help to humanities and visualization scholars stepping on such a vibrant and interesting discipline in the future.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful suggestions in improving the paper and Ines Holiday for the proofreading. This work was supported by the CHIST-ERA programme under National (MINECO-Spain) Grant PCIN-2017-064. This article has supplementary downloadable material at http://ieeexplore.ieee.org, provided by the authors.

REFERENCES

1. S. Jänicke, "Valuable research for visualization and digital humanities: A balancing act," in *Proc. 1st Workshop Vis. Digit. Humanities*, 2016, pp. 1–5.
2. K. Coles, "Show ambiguity: Collaboration, anxiety, and the pleasures of unknowing," in *Proc. 1st Workshop Vis. Digit. Humanities*, 2016, pp. 38–42.
3. M. K. Gold, *Debates in the Digital Humanities*. Minneapolis, MN, USA: Univ. Minnesota Press, 2012.
4. M. Meyer, T. Munzner, and M. Sedlmair, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012.
5. C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr, "Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 12, pp. 2351–2369, 2012.
6. P. Isenberg *et al.*, "Vispubdata.org: A metadata collection about IEEE visualization (VIS) publications," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 9, pp. 2199–2206, Sep. 2017.
7. R. Vuillemot and C. Perin, "Investigating the direct manipulation of ranking tables for time navigation," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput., Syst.*, Seoul, South Korea, 2015, pp. 2703–2706.
8. M. Abdelaal, F. Heimerl, and S. Koch, "ColTop: Visual topic-based analysis of scientific community structure," in *Proc. Int. Symp. Big Data Vis. Analytics*, Nov. 2017, pp. 1–8.
9. Z. Zhou, C. Shi, M. Hu, and Y. Liu, "Visual ranking of academic influence via paper citation," *J. Vis. Lang. Comput.*, vol. 48, pp. 134–143, Oct. 2018.
10. P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Visualization as seen through its research paper keywords," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 771–780, Jan. 2017.
11. S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann, "On Close and distant reading in digital humanities: A survey and future challenges," in *Eurographics Conference on Visualization (EuroVis)—STARs*, R. Borgo, F. Ganovelli, and I. Viola, Eds. Norrköping, Sweden: The Eurographics Association, 2015.
12. S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann, "Visual text analysis in digital humanities," *Comput. Graph. Forum*, vol. 36, no. 6, pp. 226–250, Sep. 2017.
13. F. Windhager *et al.*, "Visualization of cultural heritage collection data: State of the art and future challenges," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 6, pp. 2311–2330, Jun. 2019.
14. Y. Liu *et al.*, "CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2014, pp. 3553–3562.
15. M. Callon, J. P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry," *Scientometrics*, vol. 22, no. 1, pp. 155–205, Sep. 1991.
16. K. El-Arini and C. Guestrin, "Beyond keyword search: Discovering relevant scientific literature," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2011, pp. 439–447.
17. A. J. Bradley *et al.*, "Visualization and the digital humanities:," *IEEE Comput. Graph. Appl.*, vol. 38, no. 6, pp. 26–38, Nov. 2018.

18. A. Benito-Santos and R. Therón Sánchez, "Cross-domain visual exploration of academic corpora via the latent meaning of user-authored keywords," *IEEE Access*, vol. 7, pp. 98144–98160, 2019.

Alejandro Benito-Santos is currently a Research Assistant and Lecturer with the Department of Computer Science and Automation, University of Salamanca, Salamanca, Spain, which he joined in 2016. He received the B.Sc. degree in computer engineering and the M. Sc. degree in intelligent systems in 2016 from the University of Salamanca. He is a member of the Visual Analytics Group VisUSAL (within the Recognized Research Group GRIAL), where he is currently working toward the Ph.D. degree under the supervision of Dr. Roberto Therón Sánchez. In his thesis, he applies visual analytics in a broad range of interdisciplinary research contexts such as the digital humanities, sports science, or linguistics. His interests lie in the areas of human–computer interaction, design, statistics, and education. He has taught HCI and Introduction to Python Programming for Statisticians with the Faculty of Sciences of Salamanca in the past. He is a Student Member of IEEE since 2018. Contact him at abenito@usal.es.

Roberto Therón Sánchez is currently the Manager of the VisUSAL Group (within the Recognized Research Group GRIAL), University of Salamanca, Salamanca, Spain, which focusses on the combination of approaches from computer science, statistics, graphic design, and information visualization to obtain an adequate understanding of complex datasets. He received the Diploma degree in computer science from the University of Salamanca, the B.S. degree from the University of A Coruña, the B.S. degree in communication studies and the B.A. degree in humanities from the University of Salamanca, and the Ph.D. degree from the Research Group Robotics, University of Salamanca. His Ph.D. thesis was on parallel calculation of the configuration space for redundant robots. He has authored more than 100 articles in international journals and conferences. In recent years, he has been involved in developing advanced visualization tools for multidimensional data, such as genetics or paleoclimate data. In the field of visual analytics, he develops productive collaborations with groups and institutions internationally recognized as the Laboratory of Climate Sciences and the Environment, France, or the Austrian Academy of Sciences, Austria. He was the recipient of the Extraordinary Doctoral Award for his Ph.D. thesis. Contact him at theron@usal.es.