Same same, but different? On the Relation of Information Science and the Digital Humanities

A Scientometric Comparison of Academic Journals Using LDA and Hierarchical Clustering

Manuel Burghardt

Computational Humanities Group
Leipzig University, Germany
burghardt@informatik.uni.leipzig

Jan Luhmann

Computational Humanities Group
Leipzig University, Germany
jan.luhmann@gmx.net

Abstract

In this paper we investigate the relationship of Information Science (IS) and the Digital Humanities (DH) by means of a scientometric comparison of academic journals from the respective disciplines. In order to identify scholarly practices for both disciplines, we apply a recent variant of LDA topic modeling that makes use of additional hierarchical clustering. The results reveal the existence of characteristic topic areas for both IS (*information retrieval*, *information seeking behavior*, *scientometrics*) and DH (*computational linguistics*, *distant reading and digital editions*) that can be used to distinguish them as disciplines in their own right. However, there is also a larger shared area of practices related to *information management* and also a few shared topic clusters that indicate a common ground for – mostly methodological – exchange between the two disciplines.

Keywords: scientometrics; Information Science; Digital Humanities; LDA topic modeling; hierarchical clustering

1 Introduction

In their ISI 2015 paper, Robinson, Priego, and Bawden (2015) discussed if *library and information science* (IS) and the *digital humanities* (DH) might

have a joint future, as there are obviously many connecting factors between both fields. A few years later, the relationship of IS and the DH is explicitly addressed by the conference theme of ISI 2021, as "information science and its neighbors, from data science to digital humanities". Obvious connections between IS and the DH have been widely described in various related works in the past years. One of the earliest examples can be found in an article by Busa (2004), who summarizes the evolution of the DH and notes that there has been a "documentaristic current" and information-infrastructure focus in the DH from its very beginnings, e.g., with the *American Documentation Society* and the *Deutsche Gesellschaft für Dokumentation* in the 1950s. This documentaristic tradition of the DH is also reflected by a number of publications that highlight the relation of DH and libraries or library studies (Koltay, 2016; Millson-Martula & Gunn, 2017; Sula, 2013). In addition, Terras (2013a) stresses the important role of IS scholars as part of the early DH community, as about 15% of the authors of the 2005 DH conference at the University of Victoria (Canada) are actually coming from the field of library and information science. Along the same lines, Sula and Hill (2019) found that 5.4% of the authors from the journal "Computers and the Humanities" (1966–2004) and 7.4% of the authors from "Literary and Linguistic Computing" (1986–2004) are coming from information and library science. Burghardt et al. (2015) point out structural similarities of IS and the DH and make clear where traditional core topics of Information Science research – including information retrieval, information systems, tool science, user interface design and information behavior – can be transferred to novel research questions and applications in the Digital Humanities. Balck et al. (2015) also identify common, central themes of Digital Humanities and Information Science and argue that *digital* core competencies of Digital Humanities (e.g., indexing, databases, information retrieval) are largely covered in existing Information Science curricula. However, there are also more skeptical voices on the relation of IS and the DH. One of those voices is Gladney (2012, p. 203), who compares various definitions of IS and DH and comes to the conclusion that the DH actually are an "unneeded invention".

Despite all the overlaps and synergies between IS and the DH, we believe it is crucial to distinguish the two disciplines from each other, as an overly inclusive approach to academic disciplines without explicit demarcations can have its downsides, as is noted by Terras (2013b) in her critique of the "big tent" metaphor in the DH. Terras describes a "crisis of inclusion" in the DH

that leads to a blurring of the boundaries of the discipline and therefore weakens its position as an emerging field in its own right.

> "[…] if everyone is a Digital Humanist, then no-one is really a Digital Humanist. The field does not exist if it is all pervasive, too widely spread, or ill defined."

This criticism can be easily expanded to the relation of IS and DH: If we just see IS as yet another visitor in the big tent DH, the legitimacy of both disciplines is strongly devalued, which in turn has practical, negative effects on the distribution of third-party funds and the development of study programs and professorships for both IS and the DH. The latter is reflected in a current study of German small subjects ("Kleine Fächer"). In this study, the number of IS professorships has more or less stagnated with 10.5 in 2009 and eleven in 2020, whereas the number in the DH has increased strongly from two in 2009 to 24 in 2020. While this increase is positive from the point of view of the DH, the question arises, whether this development might be at the expense of IS in some cases.

In this paper, we try to shed some more light on the scholarly practices of and the relation between IS and the DH. To this end, we present a scientometric study in which we compare a sample of scholarly articles from typical DH journals to articles that were taken from typical IS journals and compare differences and similarities by means of a topic modeling approach using LDA and hierarchical clustering. Using this approach, we hope to answer the following research questions:

- What are typical topics in a corpus of IS and DH research articles?
- How are these topics distributed over the two disciplines in question?

2 Related work

This study is mainly influenced by related work that uses scholarly publications for scientometric analyses of the DH. For example, Gao et al. (2017) visualize co-citation networks and identify citation patterns based on articles published in the journals "Computers and the Humanities / Language Resources and Evaluation", "Literary and Linguistic Computing / Digital Scholarship in the Humanities" and "Digital Humanities Quarterly" from 1966 to 2016. Tang et al. (2017) present an extensive bibliometric study of a more diverse corpus of DH journal articles, which addresses citation patterns and

their correspondence with article keywords. Weingart (2015, 2016) periodically analyzes trends in the metadata of the DH conferences, focusing on co-authorship and author affiliation as well as critical reflection on diversity in DH. He also identifies trends in research topics, which in this case are not based on topic modeling, but on topic tags that the DH conference authors select for their contribution. A similar quantitative analysis is conducted for example by Sprugnoli et al. (2019) for the Italian conferences AIUCD and CLiC-it from 2014 to 2017. Puschmann and Bastos (2015) present a study of two academic networking platforms, which are related to DH, and analyze user posts based on co-word analysis and topic modeling. They reveal how the discourse on each platform is oriented towards certain sub-areas of DH and towards certain humanities disciplines. A study of Library and Information Science using topic modeling is offered by Figuerola et al. (2017). They collect research papers from the LISA database, show research trends among them and are able to measure the impact of the "movement of digital humanities" on information science, with a rapid increase of humanities-related research starting in 2008.

3 Methodology: LDA and hierarchical clustering

The method used in this study is inspired by related works that have successfully applied topic models to quantify academic journals, for instance Blei and Lafferty (2007), Goldstone and Underwood (2012), Griffiths and Steyvers (2004), Mimno (2012) and Wehrheim (2019). In this section we describe a recent method that combines LDA and hierarchical clustering (Vega-Carrasco et al., 2020), which we adapted from the area of marketing and transaction analysis to the analysis of academic journals. As the method can be considered rather novel for these purposes, we also provide insights from evaluations that accompanied the use of this approach.

3.1 Building the corpus

In order to investigate differences and similarities between IS and the DH, we created a corpus of peer-reviewed publications from both fields. Our corpus (see Tab. 1) comprises the full texts of 6,498 research articles published

between 1990 and 2019 in two journals associated with IS (28 M tokens total) and three journals associated with the DH (15 M tokens total). The selected journals are well-established in the respective disciplines and can be considered top tier.

Table 1: Corpus overview

Discipline	Journal	Time	Articles		
			Decade	Number of articles	Number of tokens
Digital Humanities	*Language Resources and Evaluation* (*LRE*; formerly known as *Computers and the Humanities*)	1990–2019	1990–1999	329	1,746,825
			2000–2009	236	1,283,827
			2010–2019	339	2,647,701
	Digital Scholarship in the Humanities (*DSH*; formerly known as *Literary and Linguistic Computing*)	1990–2019	1990–1999	283	1,334,671
			2000–2009	381	1,988,258
			2010–2019	558	3,267,314
	Digital Humanities Quarterly (*DHQ*)	2007–2019	1990–1999	–	–
			2000–2009	49	363,871
			2010–2019	338	2,534,583
Information Science	*Journal of the Association for Information Science and Technology* (*JASIST*)	1990–2019	1990–1999	252	1,600,561
			2000–2009	1588	10,537,002
			2010–2019	1222	9,298,150
	Journal of Documentation (*J. Doc*)	1990–2019	1990–1999	174	1,141,250
			2000–2009	327	2,260,000
			2010–2019	422	3,357,853

Since we will treat these articles as a sample of the research literature of both disciplines, we must bear in mind that these journals can represent only a fraction of the international discourse which is predominantly oriented towards North American and Western European research, and largely excludes scholarship across the Global South. Also, regarding the many flavors and sub-areas of Digital Humanities research (Huggett, 2012; Sahle, 2015; Burghardt, 2020), our sample of articles is probably biased toward literary computing and computational and corpus linguistics, given the traditions of LRE and DSH.

We obtain most of the above research articles in PDF format via URLs provided by the text mining services of CrossRef[1], except for articles of DSH, which we obtain from *Oxford University Press*[2], and articles of DHQ, which are available as XML straight away[3]. All PDFs are converted to XML using *Grobid*[4]. In addition to the fulltexts we also collect metadata (e.g., author, title, year) for all of the articles. All texts are tokenized, POS-tagged, lowercased and lemmatized using *spaCy*[5]. To improve lemmatization, especially of conjugated verbs, we use the *LemmInflect* dictionary[6], but keep original spaCy lemmata for out-of-vocabulary terms. We applied some manual corrections for a couple of nouns (*humanities*, *linguistics*, *data*, *media*, etc.) that can be considered plural-only nouns or uncountable in the domain context and which spaCy would otherwise reduce to a singular form. To remove texts from the corpus which are not original research articles, but contain, e.g., reviews or organizational matters, we only keep texts which contain at least 1,000 tokens and whose authors are included in the metadata. We also removed any documents with generic titles such as "Editorial", "News and notes", "Book review", etc.

3.2 Building a stabilized, aggregated topic model

Topic modeling is an unsupervised machine learning technique for inferring a set of latent semantic topics from a large collection of documents. For our study we use the popular Latent Dirichlet Allocation (LDA; Blei et al., 2003) with collapsed Gibbs sampling, implemented in the MALLET toolkit (McCallum, 2002). Since LDA is a probabilistic, non-deterministic algorithm that involves random initializations and random sampling, multiple LDA runs on the same data and the same parameters may lead to different results. Furthermore, results are heavily depending on parameter settings (number of topics, alpha and beta priors). To tackle this issue and obtain a stable and robust model of topics, we decided to aggregate topics from several LDA

1 https://www.crossref.org/

2 https://academic.oup.com/

3 http://www.digitalhumanities.org/

4 https://github.com/kermitt2/grobid

5 https://spacy.io/

6 https://github.com/bjascob/LemmInflect/

runs (Blair et al., 2016, 2020) by means of hierarchical clustering (with an agglomerative approach), as proposed by Vega-Carrasco et al. (2020). We justify this approach by evaluating the quality of the clustered topic model in comparison to individual LDA models. The metrics used for evaluation are:

1. **perplexity** on held-out documents to assess the generalizability of a model, estimated by the "iterated pseudo-count" method described by Wallach et al. (2009);
2. **topic coherence** (Aletras & Stevenson, 2013; Mimno et al., 2011), defined as the average value of pairwise collocation probabilities of the top 15 terms of a topic, based on Normalized Pointwise Mutual Information calculated on collocations in our corpus with a window size of 5L and 5R;
3. **topic distinctiveness** (Vega-Carrasco et al., 2020), defined as the minimum of the cosine distances a topic shows to all other topics within a model;
4. **topic stability** (Greene et al., 2014) between two models; utilizing the Hungarian algorithm to find an optimal alignment between the sets of topics of two models based on pairwise cosine distances, topic stability is then defined as the average value of cosine distances of all aligned topic pairs[7].

The approach proposed by Vega-Carrasco et al. (2020), in a first step, involves the identification of a plausible number of topics for LDA models based on the afore-mentioned metrics. In a second step, a series of LDA runs with this number of topics are performed. We leave the optimization of alpha and beta priors to MALLET. The resulting topics of the LDA model series are then merged by hierarchical clustering, based on cosine distance and average linkage. To obtain a final model of clustered topics, two parameters have to be determined: a cosine distance threshold below which topics are merged during the clustering process and a minimum cluster size at which we consider a topic cluster to be included in our aggregated model. Cluster

7 Please note that the cosine distance between topics refers to the cosine distance between the term probability distributions of topics. Our calculation of the topic stability differs from Greene et al. (2014) in that they calculate Average Jaccard distance of the sets of top terms of topics instead of cosine distance. However, the cosine distance outperformed other measures of topic distance when compared to human judgment (Aletras & Stevenson, 2014) and when used for topic matching (Niekler & Jähnichen, 2012).

size is also referred to as *recurrence* by Vega-Carrasco et al. (2020), reflecting the reappearance and therefore stability of a topic among several topic models.

3.3 Preprocessing and sampling

For reasons of efficiency (Martin & Johnson, 2015) we consider nouns and proper nouns as features, on which we perform LDA. We additionally concatenate multi-word units (2- to 5-grams) which are extracted using the NPMI association measure (Bouma, 2009), with ≥ 0.4 being reported to be a suitable threshold. Multi-word units are further filtered to contain at least one noun or proper noun, considering it a shallow approach to extracting noun phrases. These are also included as features. Finally, all terms occurring in less than 1% of texts are excluded, resulting in 6,721 terms.

Since the number of articles varies greatly between journals and decades, each of the following LDA runs is performed with a random subsample of 2,400 texts, stratified by discipline, decade and journal. For each decade, this subsample contains 800 texts, 400 for IS and 400 for DH. These 400 texts per discipline texts are being distributed as evenly as possible among the different journals of the discipline. We consider the resampling of texts on which individual topic models are fitted as an additional way to ensure the generalizability and robustness of our aggregated model. We additionally define a set of 240 test documents, which are stratified in the same way. These documents are never used for LDA training, but only for the calculation of perplexity.

3.4 Estimating the number of topics

In a first step we aim to determine a plausible number of topics. We run LDA using MALLET for a number of 50, 100, 150 and 200 topics, ten runs each. For all the resulting topic models we calculate the above-mentioned metrics (see Fig. 1). Although the perplexity values may suggest the use of a topic model with 200 or more topics, we chose to perform LDA using 100 topics. The 100-topic models show a significant decrease in perplexity, while they perform much better on topic coherence and stability than LDA models with a larger number of topics.

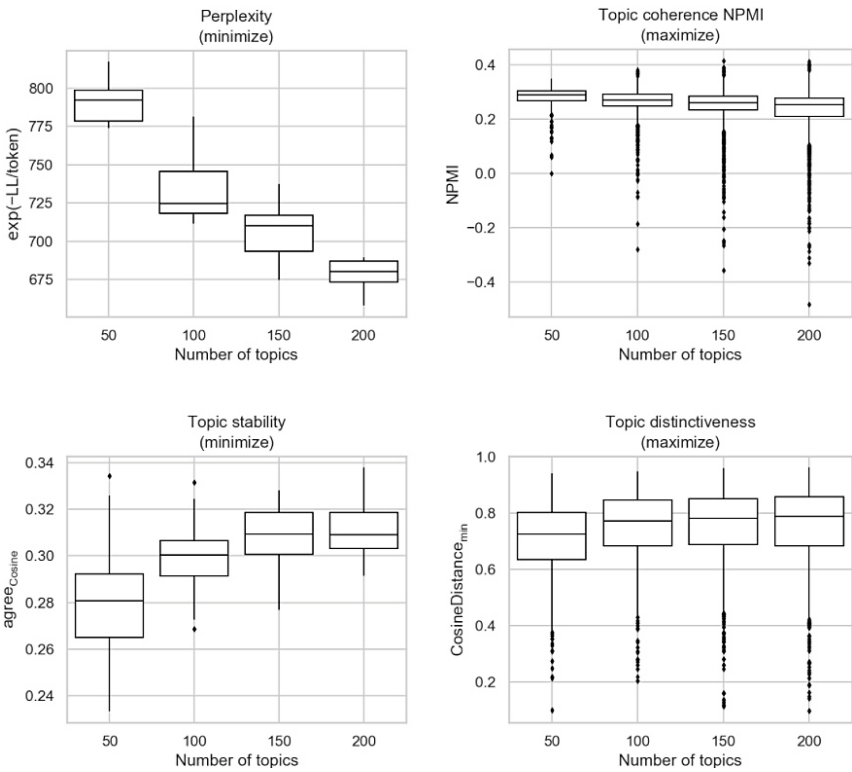

Fig. 1 Evaluation metrics of LDA models with 50, 100, 150 and 200 topics

3.5 Hierarchical clustering of topics

For topic clustering we apply a hierarchical agglomerative clustering algorithm with average linkage and cosine distance metric, providing a distance threshold for the merging of topics. To estimate in which range this threshold should be, we manually annotate 250 randomly selected topic pairs according to their estimated similarity, judging from the top 20 terms of each topic. Figure 2 shows the correlation of annotated topic similarity to cosine distances.

A cosine distance ≤ 0.2 is mostly associated with highly similar topic pairs, while a cosine distance ≥ 0.5 corresponds to dissimilar topic pairs. From this we conclude that a reasonable cosine distance threshold is in the range 0.2 to 0.5. An aggregated topic model is based on topics derived from multiple LDA topic models. In this case, we use 20 individual LDA topic models with 100 topics, meaning we perform a clustering of 2,000 topics. Each

of the 20 models is generated using MALLET, with a hyperparameter optimization of alpha and beta priors every 20th iteration, and a burn-in phase of 50 iterations. The resulting 2,000 topics are then clustered, based on the cosine distance of their term probabilities, once for each distance threshold from 0.25, 0.35 to 0.45. The term probabilities of a topic cluster (or clustered topic) are calculated as the average of the term probabilities of the topics which are part of the cluster. As mentioned above, the size of a topic cluster can be seen as the recurrence of highly similar topics among LDA models. A second parameter is the minimum cluster size at which we include a cluster in our model. To select these two parameters and to evaluate aggregated models, we repeat the entire procedure five times.

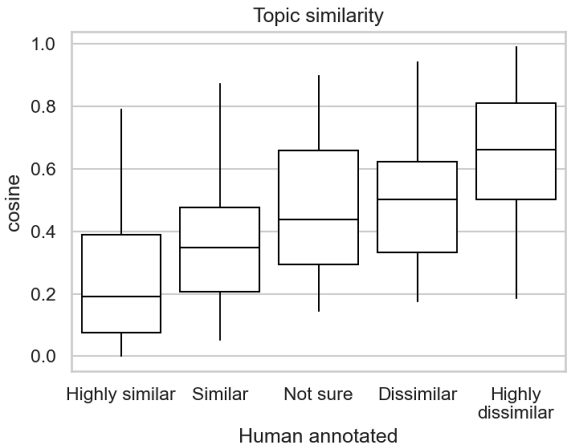

Fig. 2 Human-annotated topic similarity compared to cosine distance

Figure 3 shows the evaluation results of perplexity, topic coherence, topic stability and topic distinctiveness values of the aggregated models, for varying values of cosine distance threshold and minimum cluster size. The average values achieved by the original individual LDA models are displayed as black horizontal lines, one standard deviation as dashed lines. In terms of topic coherence and topic stability, the clustered models clearly outperform the basic topic models, especially if setting a minimum cluster size between 9 and 18.[8] A distance threshold of 0.25 performs slightly better on topic

8 It should be noted at this point that resampling the articles for each LDA run most likely has a negative impact on topic stability among the LDA models, while the aggregated models may have an advantage. This should be evaluated in the future.

coherence, a threshold of 0.45 does slightly better on topic stability and a threshold of 0.35 seems to be a good middle ground for both measures. However, in terms of perplexity, the aggregated models using a distance threshold of 0.25 and a minimum cluster size > 8 perform far worse than original LDA models. This also applies to the models using a distance threshold of 0.35 and minimum cluster size > 10, and those using a distance threshold of 0.45 and minimum cluster size > 12. Regarding topic distinctiveness, the models with 0.25 distance threshold perform best and also better than original LDA models, but only if setting a minimum cluster size > 11. For these parameter settings however, as we have already noted, the perplexity values are far from acceptable. We finally decide on using a distance threshold of 0.45 and a minimum cluster size of 12. The aggregated models using these parameter settings show a comparable perplexity and only marginally lower topic distinctiveness as the original LDA models, and outperform them by far in terms of topic coherence and topic stability. Among the five aggregated topic models we obtained, the one showing the best evaluation results is selected.

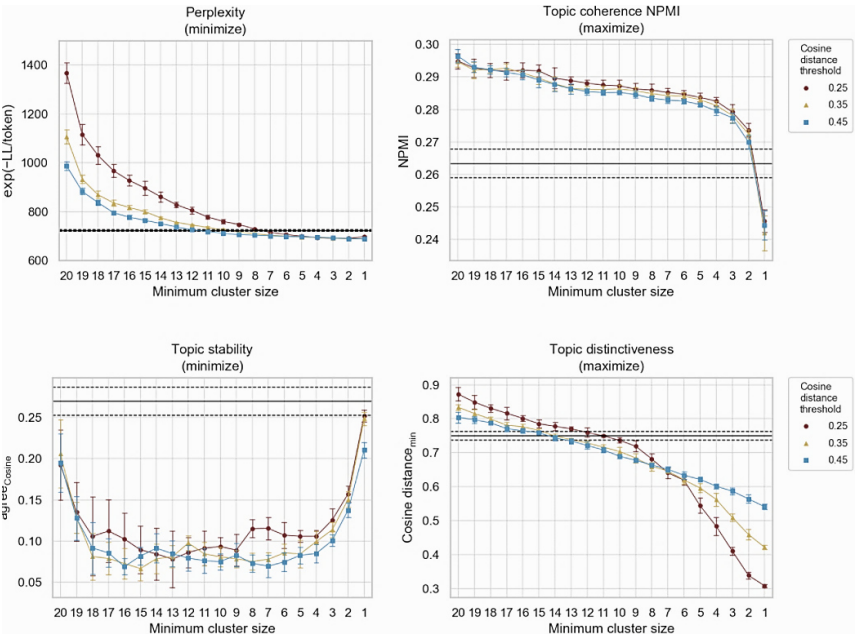

Fig. 3 Evaluation results of aggregated topic models. Horizontal lines show the mean value and standard deviations achieved by standard LDA models.

Figure 4 shows a significant difference in topic coherence values of this model compared to the LDA models it is aggregated from. In order to be able to post-hoc calculate average topic probabilities with respect to individual publication years, journals or disciplines, we require the topic distributions of all documents in our corpus, based on our newly aggregated model. We infer these using the collapsed Gibbs sampler, where we fix the already known topic term probabilities matrix and priors.

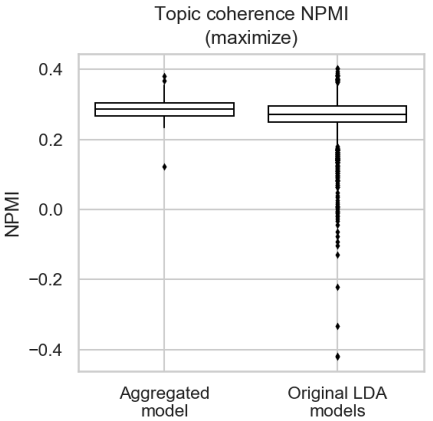

Fig. 4 Distribution of topic coherence values of the final aggregated model vs. the original LDA models

4 Results and discussion

As a result of our clustered topic model, we obtain a total number of 87 topic clusters.[9] These are numbered by their initial cluster index. All clusters with a size < 12 were filtered out. To display the latent semantic content of a topic, we determine the most relevant terms of each topic cluster by applying the term relevance formula proposed by Sievert and Shirley (2014) with weight parameter $\lambda = 0.8$. They show that the top terms of the resulting ranking achieve a better interpretability to humans than simply displaying the most

9 We provide a list of these topic clusters together with their most relevant terms, their actual cluster size (i.e., the number of merged topics), the coherence score as well as average probability for both DH and IS in the Appendix of this article.

probable terms of a topic. The document topic distributions allow us to observe topic probabilities of higher-level groupings of documents. We define the topic probabilities for a given discipline as the weighted average of the mean values corresponding to decade and journal strata of the discipline (for details on stratification and sampling, see 3.3). The mean topic probabilities of a stratum are calculated as the mean values of the topic probabilities of the documents within this stratum.

Figure 5 shows an overview of the distribution of the mean probabilities among IS and DH for all topics. The topics are sorted by relative difference between their discipline-specific probability.

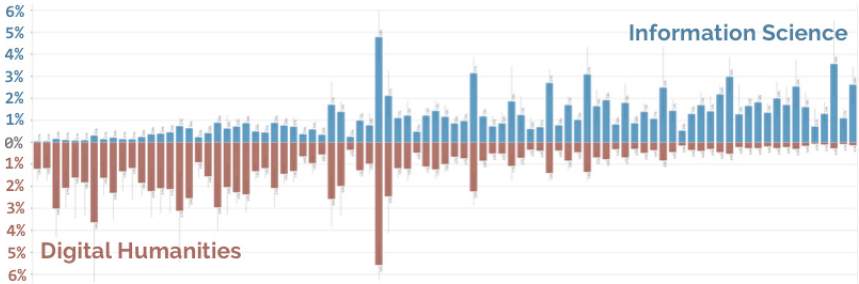

Fig. 5 Distribution of topic clusters among IS and the DH

On the left and right side, the plot clearly shows that there are a number of topic clusters that are rather characteristic for either IS (blue) or the DH (red). In the middle, there are some shared topic clusters that occur in both disciplines.

Figure 6 provides a spatial overview of all topic clusters. It shows a 2D projection of the topic distances based on the transposed document topic probability matrix. This matrix was reduced using singular value decomposition to a dimension of 100, then projected to two dimensions using UMAP (McInnes, 2018). The size of the topic clusters depends on the overall average topic probability. The color of the topic clusters indicates whether they show a statistically significant higher probability for a certain discipline. This was tested via a two-sided Mann-Whitney rank sum test on stratified document topic probability values. If a topic shows no significance for either discipline, it is displayed grey. If it is significant with $p < 0.01$, the intensity of the color depends on the value of the effect size R (rank biserial correlation). The position of topic clusters indicates co-occurrence with other topic clusters in the corpus, and thus allows us to identify superordinate cluster structures, which

we call *topic areas*. In the following we discuss how the topic clusters of Figure 6 can be interpreted to shed some light onto distinctive and overlapping scholarly practices in IS and the DH.

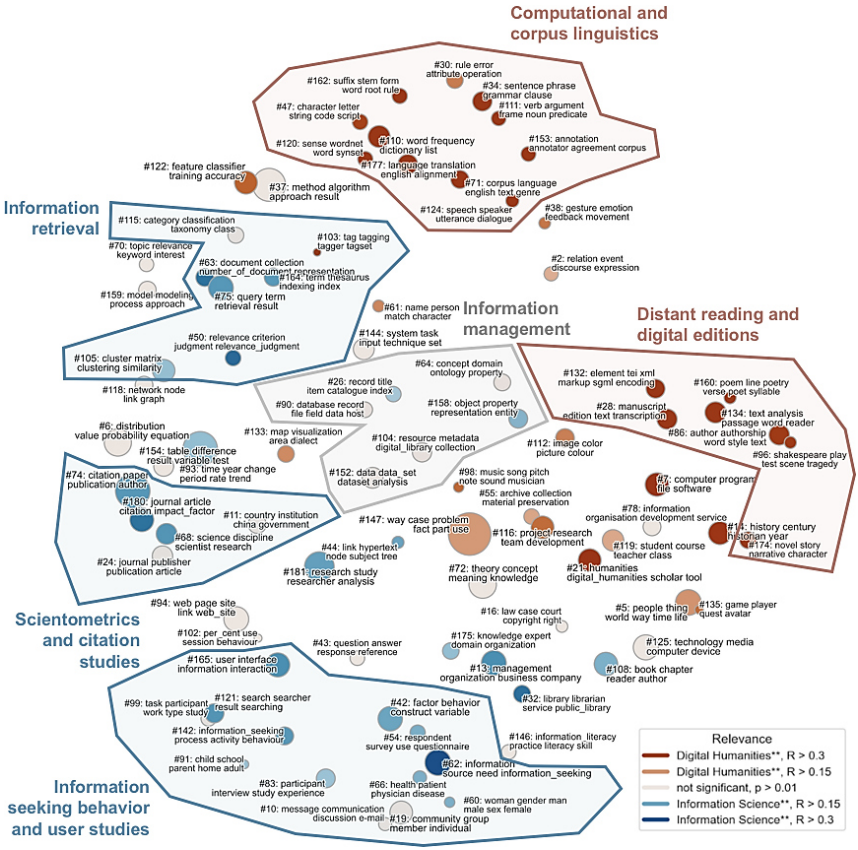

Fig. 6 Spatial overview of all topic clusters (colored bubbles). The size of the topic clusters depends on the overall average topic probability; color indicates whether a topic shows a statistically significant higher probability (Mann-Whitney rank sum test) for IS (blue) or DH (red). Topic clusters that form larger thematical areas are displayed as regions (e.g., "information retrieval", in the top-left).

4.1 General observations on topic clusters

The previous topic clustering approach has resulted in mostly meaningful topics. Only two topic clusters (#5, #147) exclusively contain generic terms

that cannot be used to gain insight about disciplinary characteristics of IS and the DH. A few other topic clusters use more specific terms but eventually describes rather generic concepts, for example: *law* (#16), *time* (#93) or *publishing* (#24). Most of the topic clusters however are very revealing with regard to scholarly practices of IS and the DH, allowing us to show scholarly differences and similarities between the two disciplines.

The following bar charts (Figs. 7–12) show discipline-specific topic probabilities and the confidence interval based on decade-journal strata mean values as error bars. These charts also contain basic diachronic information in the form of a miniature line diagram to the left of the bar chart. These show the trend in topic probability for each discipline (red for DH, blue for IS) for the decades 1990s, 2000s and 2010s from left to right.[10]

4.2 IS topic clusters

Figure 6 shows three large topic areas that can be centrally assigned to IS, i.e., these topics play no significant role in DH publications. The largest area here is *information seeking behavior* and the closely related method *of user studies*. Figure 7 shows that this topic area clearly belongs to IS. We find a similar picture for the topic area of *scientometrics* and *citation studies*, which traditionally has been part of the IS research agenda and does not really seem to be contested by the DH at all (see Fig. 8).

Interestingly, things look similar for the topic area of *information retrieval* (IR; see Fig. 9), which – with its many techniques for text and data analysis – one might have expected to play an important role in the DH too. However, it seems that IR remains a central topic area for the IS and the DH rather seem to have developed their own IR-oriented practices for text analysis. We will see below that the DH indeed seem to borrow many of their text analytics methods – which often involve some form of machine learning – from corpus and computer linguistics. It is also interesting to see that the IR-related topic cluster on *question answering* (#43) is classified as a rather neutral topic that is neither explicitly addressed by IS nor by the DH (see Fig. 6).

10 We do provide these trends as an extra source of information but do not actually discuss diachronic developments in this paper, as we were not able to detect any meaningful trends in the data. For a more comprehensive diachronic perspective, the corpus probably should cover more than just three decades.

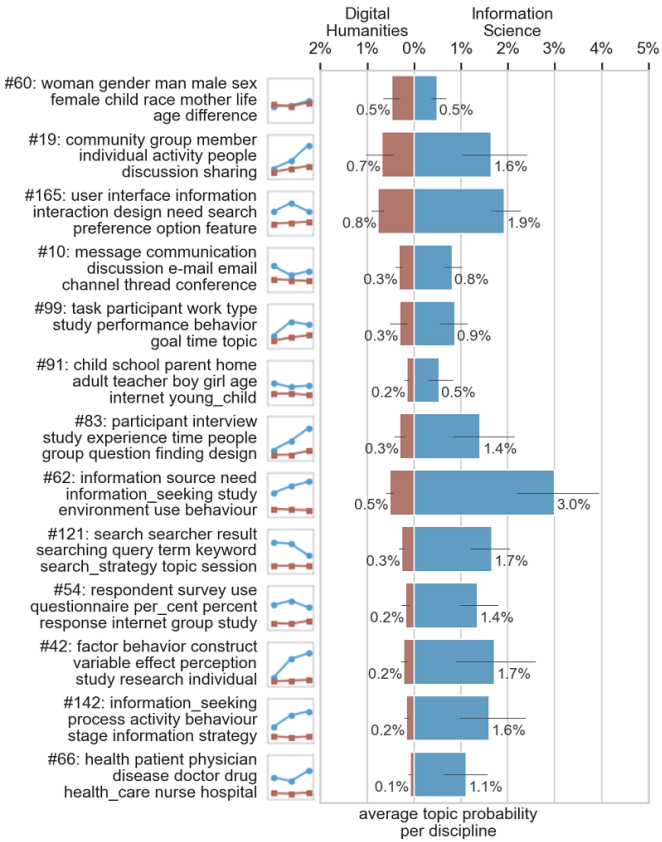

Fig. 7 Topic area: Information seeking behavior and user studies

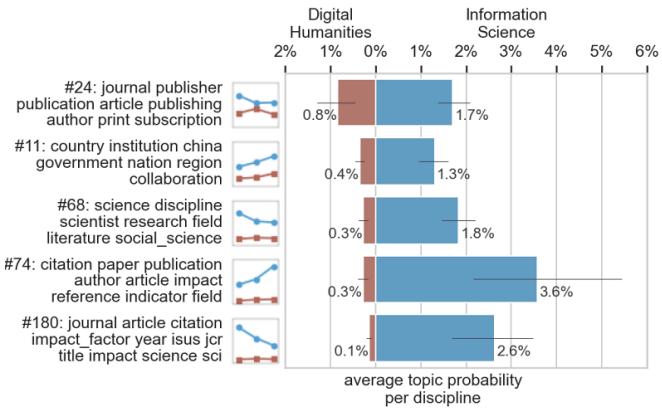

Fig. 8 Topic area: Scientometrics and citation studies

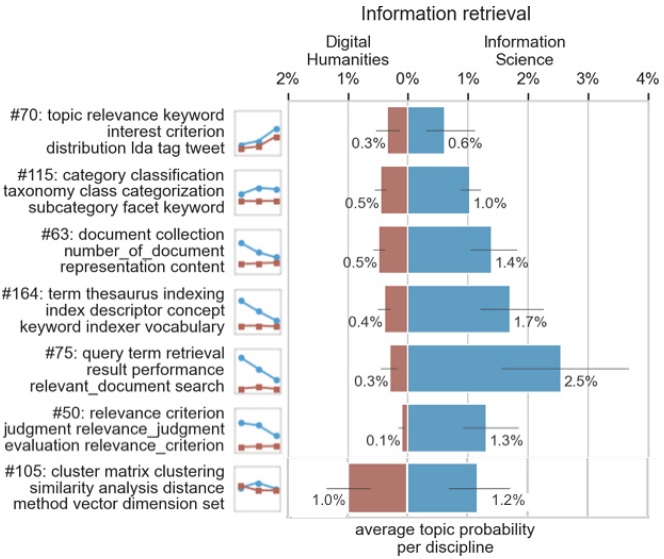

Fig. 9 Topic area: Information retrieval

Apart from these major areas of IS research practices, only a few topic clusters remain that can be clearly assigned to IS. These include a generic *hypertext* topic (#44) as well as *knowledge management* (#175), but also two topic clusters *library* (#32) and *book* (#108) that reflect the *library science* aspect of IS (#32, #108). The two topic clusters *research* (#181) and *business* (#13) illustrate that IS has both an academic, research-oriented perspective, but also a more applied, business-related perspective with industry co-operations etc.

4.3 DH topic clusters

As is the case for IS, the DH too have larger topic areas that are characteristic for the discipline. The DH seem to be strongly influenced by methods and practices from *computational* and *corpus linguistics* (see Fig. 10), which – then again – is not too surprising, as the DH traditionally have had a strong focus on *text*. This focus is also reflected in respective DH journals, which – apart from DHQ – often have a heavy bias toward *linguistics*.[11] The actually

11 We plan to include proceedings of DH-related conferences and workshops for future analyses, as they represent a much broader range of topics and practices in the DH.

interesting thing about this, is that IS has only a very small share of the linguistic clusters, although it has been claimed that IS has many connecting factors to linguistics (Montgomery, 1972, p. 195; Engerer, 2012).

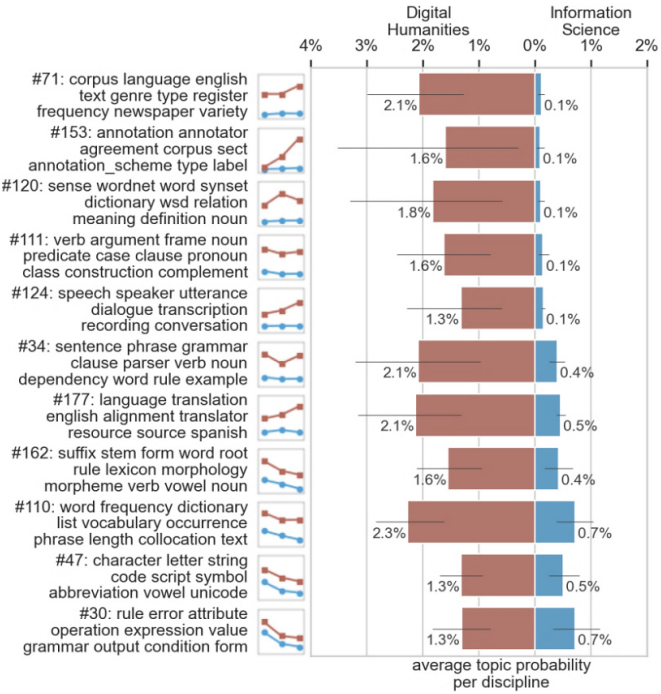

Fig. 10 Topic area: Computational and corpus linguistics

Another large topic area in the DH can be paraphrased as *distant reading* (Moretti, 2000) and *digital editions* (see Fig. 11), entailing the encoding and computational analysis of text documents with a focus on literary rather than on linguistic studies. Digital editions are represented by a topic cluster on manuscript / edition / transcript (#28) as well as by a cluster on TEI / XML / markup (#132). It is also worth mentioning that Shakespeare's plays seem to constitute a topic cluster of its own (#96), underpinning the status of the "Bard" as the *drosophila melanogaster* of the DH. Another topic cluster within this larger area of distant reading that should be highlighted is the method of *stylometry* and *authorship attribution* (#96), which is a frequently used method throughout the DH. The topic cluster on machine learning and

classification (#122) is an important methodological foundation for almost any computational analysis in the humanities.

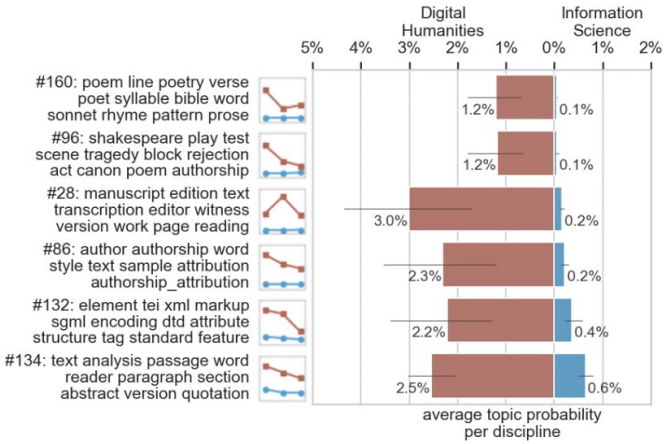

Fig. 11 Topic area: Distant reading and digital editions

Although having a strong focus on text, the DH are not restricted to text-based media, but also take into account *images* (#112) and *music* (#98) as well as *multimodal* genres (#38) such as *games* (#135). While being centered around linguistics and literary studies, the DH also address a wider spectrum of humanities disciplines, including *history* (#14), *geography* (#133) and *translation studies* (#177).

A number of clusters revolve around methodological topics such as named *entity recognition* (#61), *gesture recognition* (#38), and *event and discourse detection* (#2). A highly relevant topic cluster is about rather generic terms, such as *computer*, *program* and *software* (#7). While this may seem odd at first sight, it is actually very revealing for the DH, as they will often use these generic IT terms in order to highlight a fundamental epistemological shift in the humanities. Another fundamental characteristic of the DH is revealed by the *collaboration* (#116) topic cluster: Collaborative research will be the norm in many scientific disciplines, including IS, but it is not for many humanities disciplines. However, with the rise of digital tools and the general interdisciplinary nature of the DH, Digital Humanities have to be much more open to collaborative research practices than traditional humanists, implying another epistemological shift. Another topic cluster in the DH is dedicated to *teaching* (#119). Similar to the two previous topic clusters, this is very telling about the nature of the DH, who not only reflect upon the chances of digital

tools and methods for their research practices, but also for their teaching. One of the most prominent topic clusters for DH actually is dedicated to the reflection and discussion of the *digital humanities as a discipline* (#21), which illustrates the coming-of-age of a young discipline that is still trying to position itself in an interdisciplinary "big tent" setting (see also Terras et al., 2013).

4.4 Shared topic clusters

The previous sections highlighted characteristic topic clusters of IS and the DH. Here, we will highlight some of the topic clusters that are shared by both disciplines. There seems to be a larger topic area that we labeled *information management and library studies* (see Fig. 12). Although some of the topic clusters (#26, #158) here seem to be more contested by IS, more general activities such as *management of resources, data and meta data* as well as *object representation* also seem to be important for the DH.

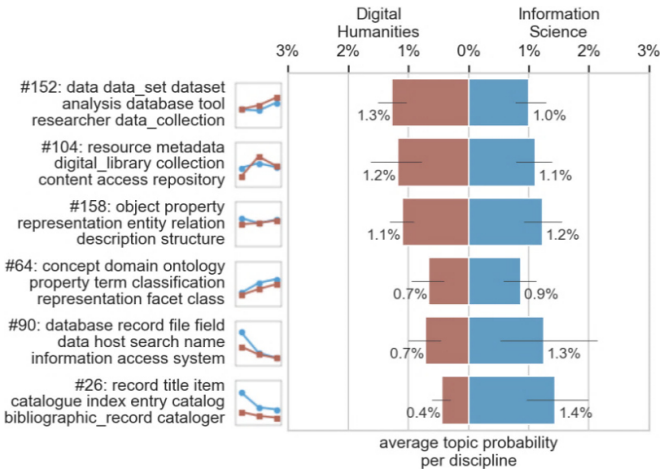

Fig. 12 Topic area: Information management and library studies

While we have seen that both IS and the DH have their specific methods, there are also some shared clusters on the methodological level. These include established *machine learning algorithms* for pattern recognition and information extraction (37) and *statistical tests* (#6, #154), but also techniques for *network analysis* (#118) and *topic modeling* (#70). Besides such methodological topic clusters, we also find two clusters that are rather on the concep-

tual level, as they represent *modeling practices* (#159) as well as *epistemological theory* (#72). Further shared topic clusters are rather general and focus for instance on the *web* as a resource (#94), *digitization and new media* (#125), *software systems and frameworks* (#144), *information institutions* (#78) and *information literacy* (#146) in general.

5 Conclusion

We performed a scientometric analysis of typical IS and DH journals from three decades, using a recent method that combines LDA with hierarchical clustering. The data on which Vega-Carrasco et al. (2020) originally developed and evaluated this novel method are transactions in the grocery retail market, which are treated as bags-of-products. In this paper, we have adapted the method back into the context of text analysis (bags-of-words) and were able to prove its success by means of various evaluation metrics.

The goal of this study was to shed some more light on the relation of IS and DH, which have been said to have a number of overlaps and similarities (see introduction). With the above method we were able to identify topic clusters that are characteristic for both disciplines: IS has three major research areas (*information behavior*, *information retrieval* and *scientometrics*) that are uncontested by the DH. While IS is dedicated to various aspects of libraries, the DH seem to address other cultural heritage institutions, such as archives and collections. The DH is still very heavily focused on text analysis and digital editions, borrowing many concepts and methods from linguistics and literary studies. However, there are also characteristic topic clusters in the DH that indicate that they also deal with other humanities disciplines and also with research objects that go beyond text.

Our results suggest that the generally expected overlap between IS and DH (see, e.g., Sula, 2013; Robinson et al., 2015; Burghardt et al., 2015) seems to be mostly in the broad area of *information management* and on the methodological level, whereby it must be stressed that both IS and DH also have clearly distinguishable research methods and practices. All in all, our results indicate that despite rather occasional overlaps, there is enough uncontested space for both IS and the DH to thrive as individual disciplines and to further develop unique research agendas and study programs.

References

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)* (pp. 13–22). Red Hook, NY: Curran.

Aletras, N., & Stevenson, M. (2014). Measuring the similarity between automatically generated topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (Vol. 2: Short Papers, pp. 22–27). Red Hook, NY: Curran.

Balck, S., Büttner, S., Ducks, D., Lehfeld, A.-S., Schneider, E., & Vietze, E. (2015). Mit den Informationswissenschaften von Daten zu Erkenntnissen. In *DHd 2015*, Graz.

Blair, S. J., Bi, Y., & Mulvenna, M. D. (2016). Increasing topic coherence by aggregating topic models. In *International Conference on Knowledge Science, Engineering and Management* (pp. 69–81). Cham: Springer Nature.

Blair, S. J., Bi, Y., & Mulvenna, M. D. (2020). Aggregated topic models for increasing social media topic coherence. *Applied Intelligence, 50*(1), 138–156.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, R. E. de Castilho, & M. Stede (Eds.): *Von der Form zur Bedeutung: Texte automatisch verarbeiten. Proceedings of the Biennial GSCL Conference 2009* (pp. 31–40). Tübingen: Narr.

Burghardt, M. (2020). Theorie und Digital Humanities – Eine Bestandsaufnahme. AG Digital Humanities Theorie (Blog). https://dhtheorien.hypotheses.org/680

Burghardt, M., Wolff, C. & Womser-Hacker, C. (2015). Informationswissenschaft und Digital Humanities. In: *Information – Wissenschaft & Praxis*, 66(5–6): 287–294.

Busa, R. A. (2004). Foreword: Perspectives on the Digital Humanities. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.): *A Companion to Digital Humanities*. Oxford: Blackwell.

Engerer, V. (2012). Informationswissenschaft und Linguistik: Kurze Geschichte eines fruchtbaren interdisziplinären Verhältnisses in drei Akten. *International Journal for Language Data Processing, 36*(2), 71–91.

Figuerola, C. G., Marco, F. J. G., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics, 112*(3), 1507–1535.

Gao, J., Duke-Williams, O., Mahony, S., Ramdarshan Bold, M., & Nyhan, J. (2017). The Intellectual Structure of Digital Humanities: An Author Co-Citation Analysis. In *Digital Humanities 2017. Alliance of Digital Humanities Organizations (ADHO)*.

Gladney, H. M. (2012). Long-term digital preservation: A digital humanities topic? *Historical Social Research, 37*(3), 201–217.

Goldstone, A., & Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship. *Journal of Digital Humanities, 2*(1), 39–48.

Greene, D., O'Callaghan, D., & Cunningham, P. (2014, September). How many topics? Stability analysis for topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 498–513). Berlin, Heidelberg: Springer.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Huggett, J. (2012). Core or Periphery? Digital Humanities from an Archaeological Perspective. *Historical Social Research / Historische Sozialforschung, 37*(3), 86–105.

Koltay, T. (2016). Library and information science and the digital humanities. *Journal of Documentation, 72*(4), 781–792.

Martin, F., & Johnson, M. (2015). More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop 2015* (pp. 111–115). Red Hook, NY: Curran.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://www.cs.umass.edu/~mccallum/mallet

McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software, 3*(29), 861.

Millson-Martula, C., & Gunn, K. (Eds.) (2017). *The Digital Humanities: Implications for Librarians, Libraries, and Librarianship*. London, New York: Routledge.

Mimno, D. (2012). Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage (JOCCH), 5*(1), 1–19.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Red Hook, NY: Curran.

Montgomery, C. A. (1972). Linguistics and Information Science. *Journal of the American Society for Information Science, 23*(3), 195–219.

Moretti, F. (2000). Conjectures on world literature. *New Left Review, 1*(4), 54–68.

Niekler, A., & Jähnichen, P. (2012). Matching results of latent dirichlet allocation for text. In *Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling* (pp. 317–322). Berlin: Universitätsverlag der TU Berlin.

Puschmann, C., & Bastos, M. (2015). How digital are the digital humanities? An analysis of two scholarly blogging platforms. *PLOS ONE*, 10(2), e0115035.

Robinson, L., Priego, E., & Bawden, D. (2015). Library and Information Science and Digital Humanities: Two Disciplines, Joint Future? *Proceedings of the 14th International Symposium on Information Science, ISI 2015* (pp. 44–54). Glückstadt: Verlag Werner Hülsbusch.

Sahle, P. (2015). Digital Humanities? Gibt's doch gar nicht! In Baum, C. & Stäcker, T. (Eds.), *Grenzen und Möglichkeiten der Digital Humanities* (Sonderband der ZfdG, 1).

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). Association for Computational Linguistics.

Sprugnoli, R., Pardelli, G., Boschetti, F., & Del Gratta, R. (2019). Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale. *Umanistica Digitale, 3*(5).

Sula, C. A. (2013). Digital humanities and libraries: A conceptual model. *Journal of Library Administration, 53*(1), 10–26.

Sula, C. A., & Hill, H. V. (2019). The early history of digital humanities: An analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004). *Digital Scholarship in the Humanities, 34*(1), i190–i206.

Tang, M. C., Cheng, Y. J., & Chen, K. H. (2017). A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses. *Scientometrics, 113*(2), 985–1008.

Terras, M. (2013a). Disciplined: Using Educational Studies to Analyse "Humanities Computing." In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining the Digital Humanities – A Reader* (pp. 67–96). Farnham, UK: Ashgate Publishing.

Terras, M. (2013b). Peering Inside the Big Tent. In M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining the Digital Humanities – A Reader* (pp. 263–270). Farnham, UK: Ashgate Publishing.

Terras, M., Nyhan J., & Vanhoutte, E. (2013). *Defining Digital Humanities: A Reader*. Farnham, UK: Ashgate Publishing.

Vega-Carrasco, M., O'Sullivan, J., Prior, R., Manolopoulou, I., & Musolesi, M. (2020). Modelling Grocery Retail Topic Distributions: Evaluation, Interpretability and Stability. *arXiv:2005.10125*.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1105–1112). New York, NY: ACM Press.

Wehrheim, L. (2019). Economic history goes digital: topic modeling the Journal of Economic History. *Cliometrica, 13*(1), 83–125.

Weingart, S. B. (2015). Submissions to DH2016 (Pt. 1). http://scottbot.net/submissions-to-dh2016pt-1/

Weingart, S. B. (2016). Submissions to DH2017 (Pt. 1). http://scottbot.net/submissions-to-dh2017pt-1/

Appendix

Table A1: All topic clusters in our model, their most relevant terms, cluster size, topic coherence (NPMI), and mean topic probability by discipline

No.	Most relevant terms by log-lift formula (Sievert & Shirley, 2014), λ = 0.8	Cluster size	Topic coherence	Mean prob. DH	Mean prob. IS
2	relation event discourse expression ontology entity	13	0.272	0.96%	0.59%
5	people thing world way time life idea sense day experience	12	0.266	2.58%	1.72%
6	distribution value probability equation function parameter	23	0.286	0.83%	2.50%
7	computer program file software machine computing version	24	0.26	3.11%	0.74%
10	message communication discussion e-mail channel	13	0.281	0.32%	0.81%
11	country institution china government nation region collaboration	17	0.31	0.35%	1.30%
13	management organization business company manager technology	16	0.295	0.46%	2.18%
14	history century historian year period war work time past	17	0.261	2.96%	0.90%
16	law case court copyright right report act authority	16	0.275	0.39%	0.70%
19	community group member individual activity people discussion	14	0.29	0.69%	1.65%
21	humanities digital_humanities scholar tool humanities_computing	21	0.329	3.63%	0.32%
24	journal publisher publication article publishing author	20	0.323	0.84%	1.71%
26	record title item catalog index entry catalog	16	0.275	0.45%	1.44%
28	manuscript edition text transcription editor witness	24	0.313	3.01%	0.16%
30	rule error attribute operation expression value grammar	14	0.269	1.31%	0.72%
32	library librarian service public_library book collection	21	0.316	0.27%	2.01%

No.	Most relevant terms by log-lift formula (Sievert & Shirley, 2014), λ = 0.8	Cluster size	Topic coherence	Mean prob. DH	Mean prob. IS
34	sentence phrase grammar clause parser verb noun dependency	25	0.308	2.09%	0.39%
37	method algorithm approach result experiment performance	16	0.288	2.46%	2.12%
38	gesture emotion feedback movement video hand body multimodal	15	0.25	0.91%	0.24%
42	factor behavior construct variable effect perception study	17	0.261	0.22%	1.71%
43	question answer response reference percent type wikipedia	17	0.261	0.38%	0.78%
44	link hypertext node subject tree hypermedia path marker	14	0.264	0.51%	0.73%
47	character letter string code script symbol abbreviation	14	0.287	1.32%	0.50%
50	relevance criterion judgment relevance_judgment evaluation	16	0.307	0.11%	1.30%
54	respondent survey use questionnaire per_cent percent	19	0.291	0.19%	1.36%
55	archive collection material preservation museum archivist	14	0.302	0.97%	0.77%
60	woman gender man male sex female child race mother life age	14	0.304	0.48%	0.49%
61	name person match character entity surname location place	18	0.241	0.64%	0.38%
62	information source need information_seeking study environment	21	0.299	0.52%	2.99%
63	document collection number_of_document representation	20	0.269	0.49%	1.40%
64	concept domain ontology property term classification	14	0.3	0.66%	0.87%
66	health patient physician disease doctor drug health_care	20	0.382	0.08%	1.11%
68	science discipline scientist research field literature	17	0.335	0.27%	1.83%
70	topic relevance keyword interest criterion distribution lda	14	0.239	0.34%	0.61%
71	corpus language english text genre type register frequency	20	0.284	2.07%	0.11%
72	theory concept meaning knowledge view information_science	20	0.295	1.35%	3.09%
74	citation paper publication author article impact reference	24	0.318	0.28%	3.57%
75	query term retrieval result performance relevant_document	25	0.323	0.31%	2.55%
78	information organization development service center	12	0.265	1.08%	1.87%
83	participant interview study experience time people group	14	0.279	0.30%	1.41%
86	author authorship word style text sample attribution	20	0.302	2.30%	0.22%
90	database record file field data host search name information	18	0.289	0.71%	1.25%
91	child school parent home adult teacher boy girl age	14	0.288	0.16%	0.53%
93	time year change period rate trend over_time figure	14	0.272	0.84%	1.19%
94	web page site link web_site web_page search_engine url	20	0.349	0.69%	1.80%
96	shakespeare play test scene tragedy block rejection act	19	0.26	1.18%	0.06%
98	music song pitch note sound musician composer mode listener	19	0.242	0.56%	0.35%
99	task participant work type study performance behavior goal	18	0.268	0.31%	0.87%
102	per_cent use session behavior log service usage number	14	0.263	0.07%	0.72%
103	tag tagging tagger tagset metadata flickr social_tagging	13	0.123	0.34%	0.25%
104	resource metadata digital_library collection content access	20	0.316	1.18%	1.11%
105	cluster matrix clustering similarity analysis distance	20	0.291	0.99%	1.17%
108	book chapter reader author reading section volume discussion	26	0.272	1.21%	1.22%
110	word frequency dictionary list vocabulary occurrence phrase	21	0.29	2.28%	0.72%

No.	Most relevant terms by log-lift formula (Sievert & Shirley, 2014), λ = 0.8	Cluster size	Topic coherence	Mean prob. DH	Mean prob. IS
111	verb argument frame noun predicate case clause pronoun	17	0.322	1.62%	0.15%
112	image color picture color painting photograph art pixel	22	0.266	1.44%	0.77%
115	category classification taxonomy class categorization	20	0.237	0.45%	1.03%
116	project research team development work resource collaboration	17	0.291	2.37%	0.88%
118	network node link graph connection edge relationship tree	20	0.268	0.52%	0.86%
119	student course teacher class teaching learning classroom	20	0.339	2.08%	0.89%
120	sense wordnet word synset dictionary wsd relation meaning	20	0.323	1.83%	0.10%
121	search searcher result searching query term keyword	19	0.295	0.27%	1.66%
122	feature classifier training accuracy classification set	20	0.321	2.04%	0.63%
124	speech speaker utterance dialogue transcription recording	19	0.295	1.32%	0.15%
125	technology media computer device internet environment	12	0.284	1.99%	1.40%
132	element tei xml markup sgml encoding dtd attribute structure	20	0.355	2.22%	0.37%
133	map visualization area dialect distance location city	16	0.26	1.18%	0.45%
134	text analysis passage word reader paragraph section	20	0.27	2.54%	0.65%
135	game player quest avatar adventure video_game narrative	20	0.274	1.17%	0.14%
142	information_seeking process activity behavior stage	13	0.307	0.17%	1.61%
144	system task input technique set processing evaluation	20	0.273	1.23%	1.44%
146	information_literacy practice literacy skill landscape	13	0.31	0.22%	1.29%
147	way case problem fact part use question example kind form	19	0.251	5.58%	4.79%
152	data data_set dataset analysis database tool researcher	20	0.276	1.28%	1.00%
153	annotation annotator agreement corpus sect annotation_scheme	20	0.295	1.60%	0.09%
154	table difference result variable test value number analysis	23	0.294	2.23%	3.15%
158	object property representation entity relation description	19	0.287	1.10%	1.22%
159	model modeling process approach figure simulation relationship	20	0.243	0.73%	0.98%
160	poem line poetry verse poet syllable bible word sonnet	15	0.277	1.21%	0.06%
162	suffix stem form word root rule lexicon morphology morpheme	19	0.367	1.55%	0.42%
164	term thesaurus indexing index descriptor concept keyword	18	0.302	0.40%	1.70%
165	user interface information interaction design need search	20	0.289	0.77%	1.93%
174	novel story narrative character fiction genre literature	12	0.295	1.85%	0.25%
175	knowledge expert domain organization knowledge_management	20	0.235	0.36%	1.08%
177	language translation english alignment translator resource	19	0.284	2.13%	0.46%
180	journal article citation impact_factor year isus jcr title	20	0.341	0.14%	2.62%
181	research study researcher analysis field method area	17	0.301	1.40%	2.71%

In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th–10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 173–199. DOI: doi.org/10.5283/epub.44944.