The Helsinki Digital Humanities Hackathon: Two Perspectives on Multidisciplinary Historical Newspapers Research in a Hackathon Context.

Ruben Ros[1] [0000−0002−5303−2861] and Sarah Oberbichler[2] [0000−0002−1031−2759]

[1] Universiteit Utrecht, Utrecht, The Netherlands
ruben@rubenros.nl
[2] Universität Innsbruck, Innsbruck, Austria
sarah.oberbichler@uibk.ac.at

Abstract. This paper describes the 2019 edition of the Helsinki Digital Humanities Hackathon from the perspective of two of its participants. As (digital) historians they were part of the group that investigated the history of medical advertisements in British nineteenth-century newspapers. The paper describes the research process, as well as the data and methods used during the research. The paper also considers the Hackathon as a laboratory for Digital Humanities research and reflects on the nature of the collaboration as experienced during the Hackathon. As such, the paper describes the challenges of multidisciplinary research and identifies the factors that hinder and foster collaboration in a Digital Humanities context.

Keywords: Hackathon · Multidiscilinary Collaboration · Historical Newspapers · Text Mining.

1 Introduction

In this Twin Talk we discuss concept of the Helsinki Digital Humanities Hackathon (DHH) and the research done during the Hackathon from the perspective of two of its participants (two historians familiar with computational methods). We aim to show how the Hackathon concept brings together researchers and offers an excellent opportunity for exploring the potential of multi- and interdisciplinary Digital Humanities research. We report on the research process, including the methods we used, the questions we answered and challenges we faced. Hereby, this paper also reflects on the nature of collaboration in Digital Humanities research. We will argue that specifically shared vocabularies, "bridge-building" capacities of individual researchers and leadership make a difference in promoting and facilitating multidisciplinary research.[3]

The Helsinki DHH is organized yearly by the Helsinki Centre for Digital Humanities (HELDIG)[1]. Five Hackathons have been organized so far. The

[3] This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

Helsinki DHH offers a chance to experience an interdisciplinary research project from start to finish within the span of 1.5 weeks. For researchers and students with a computer science or data science background, the Hackathon gives the opportunity to test abstract knowledge against complex historical problems. For people from the humanities and social sciences, the DHH shows the potential of computational methods and multidisciplinary research.

Every edition, around forty participants are divided into four groups. The group themes are established in advance and the participants can indicate their prefered group. This year, the group titles were: *Newspapers & Capitalism*, *Genre and Style in Early Modern Publications*, *The Many Voices of the European Parliament* and *Brexit in Transnational Social Media*.

In earlier years the Hackathon was visited predominantly by Finnish and European students and researchers. Recently, the Hackathon has widened its scope with the help of Common Language Resources and Technology Infrastructure (CLARIN) and Digital Research Infrastructure for the Arts and Humanities (DARIAH), and the Hackathon participants now include students and researchers from all over the world.

The team whose work is reported in this paper focused on the history of nineteenth century British newspapers and consisted of historians (7), linguists (2), computer scientists (3), data scientists (1) and literary scholars (1).

In the remainder of this paper we describe the research in the *Newspapers & Capitalism* group. We discuss the research topics and questions, as well as the data and methods used in the research. Lastly, we reflect on the Hackathon as a laboratory for DH-research and we identify factors that challenge and/or promote multidisciplinary collaboration.

2 Research Topics and Questions

During the first days the group decided to develop several lines of research based on the personal interests, individual expertise and academic background of the group members. To develop research questions, the whole team got together and collected ideas. This process was supported by literature research and close reading of the newspapers. The group decided to focus on the topic of nineteenth-century medical advertisements and the language of persuasion employed in those advertisements.

Throughout the century, medical advertisements occupied an important place in periodical culture [2][3]. Pills, lotions and ointments were regularly promoted, not seldom by so-called “quacks”: charlatans who promised to cure every disease imaginable. The link between sellers of patient medicines and the publishing industry was intimate. Without the steady demand for advertisement space many newspapers would have gone bankrupt, and without constant advertising, the patient medicine industry would not have been able to sustain itself [4].

In the early nineteenth century, the well-organized patent medicines industry had replaced the small-scale quacks, who lacked the skills and resources to take part in a market that became increasingly national in scope. The literature identifies four categories of nineteenth-century advertisers: market leaders, tradesmen, medical practitioners, elite and locals [5]. During the Victorian period there were four leading pill-makers who together represented most of the newspaper advertising; James Morison, the creator of Universal Pills, a 'venerable' Salopian called Thomas Parr who sold "Parr's Life Pills" to increase the beauty of women, Thomas Holloway who is also identified as the first world-wide advertiser, and Thomas Beecham who invented "Beecham's Pills" and claimed to cure "bilious and nervous disorders" [6]. During the nineteenth century, quackery slowly disappeared from the newspaper pages as scientific insights reached the broader public and legal action against health fraud was organized [7][8]. This transformation of medical advertising thus ties in with broader questions about the rise of consumer society, the professionalization of medicine and the newspaper industry itself.

For the group, the literature also provoked several questions, such as: what diseases and cures were advertised, whether the advertisers distinguished between different (gendered) publics, and how the cures were rhetorically marketed in the advertisements. These questions formed the basis for the Hackathon research in the *Newspapers & Capitalism* group. They were chosen because they invited for the use of computational methods, but also benefited from 'traditional' close reading.

Later in the research process, several other group members explored another significant group of advertisements: job ads. Using the HISCO classification of jobs, they were able to shed light on the long-term evolution of jobs that were marketed in advertisements [9].

3 Data

The project used the British Library Nineteenth Century Newspapers collection, provided by Gale Cengage. The data consist of 304 unique newspapers, 270.744 issues, 26.295.841 articles and 1.560.916 advertisements. The data was indexed by the HELDIG team and made accessible through an API. In light of the limited amount of time, the group chose to focus on a specific newspaper:"The Morning Post", a British newspaper that was successfully established in 1772 and existed until 1937 [10]. The Morning Post advertisements, that ran throughout the century, were used for the questions on diseases/cures, persuasion and gender. The statistical analysis was applied to all newspapers in the dataset.

The newspapers were accessible in machine-readable form and through an API developed by the HELDIG researchers. Since the quality of the photographed pages varies, the quality of the so-called "Optical Character Recognition" (OCR) as provided by Gale does so as well. Especially in the first two decades of the nineteenth century, the OCR confidence levels (included in the metadata) proved to be relatively low (Figure 1). Another pressing problem was the article segmentation. Because advertisements appear in all forms and sizes, it is hard to draw boundaries between advertisements and articles, and between individual advertisements. In the process of segmenting newspaper pages, advertisements are often grouped together, resulting in low-quality segmentation. In order to reduce the effect of lacking segmentation and low-quality OCR the analysis used issue-based frequency, meaning that for example the frequency of a word was not normalized by the number of advertisements in a newspaper but by the number of newspaper issues. In response to lower OCR confidence levels in the first decade of the century, keyword frequency was broadened by taking into account potential "alternatives" that were automatically generated with word embeddings ("fchool", "schoof, when searching for "school") [11].

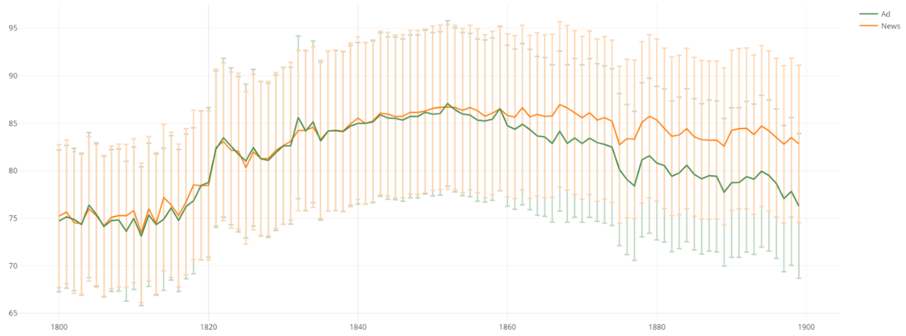

Fig. 1. Mean OCR Confidence of News articles (Orange) and Advertisements (Green) in British newspapers.

4 Methods and Research Process

After defining the research questions and research topics, the group divided itself in subgroups that would work on specific tasks. First, two computer scientists set out to gather metadata statistics. Three aspects were particularly important: the 'quality' of the data as measured with OCR (Optical Character Recognition) confidence levels, the changing composition of newspapers as measured through the share of categories such as "Advertisements and Notices" or "Business News", and, lastly, the original locations of the newspapers. The statistical insights were presented in Tableau, an online interface that allowed the group to explore the results.

Informed by the features of the data as presented in Tableau, a group of historians and data scientists set out to investigate advertisements by using frequency analysis. First, they inquired into the relation between advertisements and gender by creating subsets of male- and female-oriented ads. The historians manually composed a vocabulary of gender-related keywords (e.g. "male", "female", "boy", "girl") and a vocabulary of cures (e.g. "pills", "ointments", “balsam”). These vocabularies were used to extract a subset of medical advertisements from the digitized Morning Post and to extract ads that mentioned gender-related terms. Ads that did not mention any of the vocabulary terms were excluded. The subset that emerged from this pipeline showed how the ratio of male- and female-oriented ads changed significantly during the century (Figure 2).

The group then investigated the relation between gender and medicine by extracting cures and diseases from the ads. To do so, the group employed word embeddings, a method used to embed words in a ’vector space’ that is subsequently used to inquire into semantic relations between words [12]. Embeddings were trained on samples from the Morning Post for every decade by using the popular Python library word2vec [13]. This allowed the group to filter words that were semantically close to words such as “disease”, “cough” and “pain”. In this way, a typology of diseases could be created. A similar method was used to extract the advertised cures.

By combining the extracted information on cures and diseases with the vocabularies of gender-related keywords, the ads could be classified into groups of male- and female-oriented ads. This allowed us to inquire into two different aspects of nineteenth-century medical advertising. First, we could track the changes in the target audience. Especially later in the century, female-oriented ads became more present in the corpus. Also, we could investigate whether different cures and diseases appeared in advertisement aimed at different publics. By looking separately at male- and female-oriented ads, we were able to find differences between the targeted diseases. Mental illnesses, for example, were slightly more associated with women. We related this trend to existing literature on the history of female “hysteria” [14].

These data-driven insights were verified by historians in the group who systematically collected samples of newspapers through the Gale Search Engine and its keyword search interface [15]. For each decade, twenty newspapers were subjected to close reading. This allowed the group not only to complement the frequency analysis, but also pointed at potential problems, such as ads that were not specifically aimed at men or women, or ads that consisted primarily of images.

A similar iterative approach was taken to the strand of research that looked into the language of persuasion. The literature on the language of early-modern advertising provided several insights into the linguistic features of advertising in the eighteenth century [16,17,18]. Several linguists in the group investigated these features in the nineteenth century, hereby focusing on modal verbs (w.g. ”wishes highly to recommend”) and the use of repetition to draw attention and testimony. During the research, it appeared that many of these fine-grained rhetorical tropes were hard to quantify as a result of lacking article segmentation and OCR-errors. For this reasons, the digital methods in this line of research were restricted to the relative frequency of specific parts-of-speech such as modal verbs and adjectives.

The last subquestion, on the changing prominence of occupations in job advertisements was investigated by using a list of English-language job titles and comparing those to the full-text advertisements. Using the associated classification codes as a way to cluster specific occupations together, we were able to shed a light on the changing job market: low-skilled and production-related jobs rose in prominence, a pattern that was also visible in the category of professional and technical workers.

5 Results

The Hackathon research yielded surprising outcomes given the limited amount of time. With regard to the medical advertisements, the changing proportion and character of male- and female-oriented advertisements was a striking finding. In the first decades of the nineteenth century, ads were mainly targeted at a male audience. Gradually, female-oriented ads advanced. Especially in the last two decades of the century, the proportion of those advertisements increased significantly.

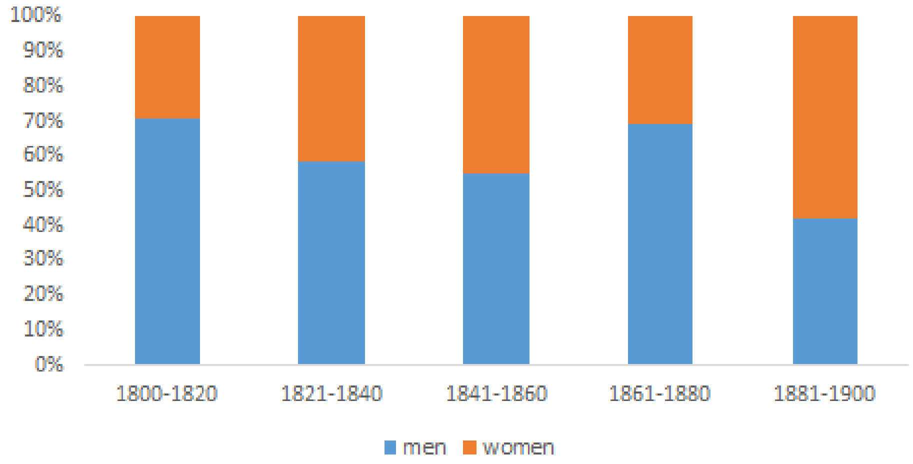

Fig. 2. The share of medical advertisements in the Morning Post containing male- and female-related terms, 1800-1900.

The advertisements also differed in terms of their contents. Although distant reading techniques were not able to fully capture the complexity of this issue, close reading revealed the gendered nature of specific illnesses and cures. The investigation of persuasiveness in newspaper language similarly produced surprising results. Computational analysis was not very important for the outcomes, but the results invite for further analysis into for example modal verbs and the repetition of specific product names. Lastly, the research into job advertisement made good use of the HISCO-classification and was able to gain an insight in nineteenth-century transformations on the job market with relatively simple methods.

6 Experiences

The Hackathon was first and foremost a learning experience. Humanists learnt new skills and methods to do research computationally and computer/data scientists were confronted with the complexities of messy historical data. Throughout the week, the group formed a comprehensive understanding what kind of multidisciplinary research could be done, and how such research should be designed and executed. The group also learned about the challenges of collaboration within an interdisciplinary projects; How can I communicate what I need and what I want and how can I explain digital tools and methods to humanities researchers?

If we look back at the Hackathon experience from the perspective of broader questions about collaboration in Digital Humanities research one aspect stands out. During the Hackathon, crossing disciplinary boundaries remained a constant effort for all the researchers. Despite the favourable "geography of practice" and the unique circumstances offered by the Hackathon format, practically all the group members frequently 'drifted' back to methods and traditions that were familiar to them [19]. Computer scientists and data scientists reverted to data curation and visualization while for the humanists close reading the newspapers and surveying the literature was something they could and would do easily. The short time span pressured the group members to come up with clear results, and resorting to familiar research practices was considered to be an easy way out. For example, if the data harmonization or model training would take longer than planned, going back to close reading was a tempting solution. The design of the research questions to a large extent facilitated this reflex to return to familiar territory. Because the questions posed could be answered by both computational methods and close reading, splitting those tasks was tempting.

This dynamic of "disciplinary isolationism" surfaced frequently. Three factors can be identified as having a positive effect on promoting collaboration and preventing this reflex of "methodological isolationism". First, a shared vocabulary: throughout the Hackathon, researchers visibly (and audibly) integrated their methodological vocabularies. Initially, terms such as "modelling" and "data cleaning" put the humanists at a distance from the computer scientists' practices. Similarly, the latter group had a completely different understanding of the language of hermeneutics employed by historians and linguists. Towards the end of the Hackathon, however, terms such as "harmonization", "close reading" and "language models" were used by historians and computer scientists alike.

Second, a group of what could be called "bridge-builders", "intermediaries" or "translators" proved essential in facilitating and actually doing the research [20][21][22]. A small number historians in the group who had worked in multidisciplinary research groups before or were proficient in coding and data management themselves formed important links between the disciplines. After two to three days, a workflow emerged that evolved around these (coding) bridge-builders who had access to the data and were able to translate historical questions into for example statistical tests. Whereas this inter-disciplinary group members were important, the Hackathon also showed how vital steps in the analysis could be easily outsourced to this limited number of “specialists”, ultimately preventing integration and collaboration. For example the mass-scale extraction of diseases and cures from the digitized ads was an essential task, but was done by programming historians.

Lastly, leadership was an vital factor in promoting collaboration. The group leaders were especially important in the first stage of designing the research plan. Because they had experience with DH-research, they were able to indicate the limits of what was possible. During the research itself, the group leaders provided focus. Because the Hackathon provided the participants with an full arsenal of tools and method, it proved important to decide on what was *not* to be done. Additionally, the group leaders were also the key links between the computer scientists and the humanists. They differed from the earlier mentioned category of ”bridge-builders” in the sense that they remained at a distance, not doing the research themselves.

7 Conclusion

The Helsinki Digital Humanities Hackathon is a fascinating laboratory for Digital Humanities research. Participants from a diverse range of backgrounds engage in cutting-edge research. Professors and bachelor students, humanities scholars and computer scientists worked together on the same research question and learned from each other. The research done in the *Newspapers & Capitalism* group crossed borders but the group also encountered challenges. The mixing of the different disciplines was slow and it would have taken even more time to establish collaboration that truly integrates existing traditions and methods. Given the impressive results and the increase in mutual understanding, however, the Hackathon gives cause for optimism when it comes to the future of multidisciplinary Digital Humanities research.

References

1. Helsinki Centre for Digital Humanities, HELDIG, https://www.helsinki.fi/en/helsinki-centre-for-digital-humanities. Last accessed 20 Feb. 2020.
2. Hindley, D., & Hindley, G., 1972. Advertising in Victorian England, 1837-1901. Wayland, London.
3. Church, R., 2000. Advertising consumer goods in nineteenth-century Britain: reinterpretations. Economic History Review 53.4, 621-645.
4. Carter, K. C., 1993. The concept of quackery in early nineteenth century British medical periodicals. Journal of Medical Humanities 14.2, 89-97.

5. Mackintosh, A., 2017. The Patent Medicines Industry in Late Georgian England: A Respectable Alternative to Both Regular Medicine and Irregular Practice. Social History of Medicine 30.1, 22–47.
6. Porter, R., 1989. Health for sale: quackery in England, 1660-1850. Manchester University Press, Manchester.
7. Barker, H., 2009. Medical advertising and trust in late Georgian England. Urban History, 36.3, 379-398.
8. Sackville Turner, E., 1965. The Shocking History of Advertising, Great Britain, 61-64. Ballantine, New York.
9. Van Leeuwen, M. H., Maas, I., & Miles, A., 2002. HISCO: Historical international standard classification of occupations. Leuven, Leuven University Press.
10. Hindle, W. H., 1937. The Morning post, 1772-1937: portrait of a newspaper. Routlegde, London.
11. Hládek, D., Staš, J., Ondáš, S., Juhár J., & Kovács, L., 2017. Learning string distance with smoothing for OCR spelling correction. Multimedia Tools Application 76, 24549–24567.
12. Mikolov, T., Chen, K., Corrado, G. and Dean, J.,2013. Efficient estimation of word representations in vector space. arXiv:1301.3781v3 [cs.CL].
13. Kutuzov, A., Øvrelid, L., Szymanski, T. and Velldal, E., 2018. Diachronic word embeddings and semantic shifts: a survey. arXiv:1301.3781v3 [cs.CL].
14. Arnaud, S., 2015. On Hysteria: The Invention of a Medical Category between 1670 and 1820. The University of Chicago Press, Chicago.
15. Gale Search Engine, http://gdc.galegroup.com. Last accessed 17 Feb 2020.
16. Leech, G. N., 1966. English in advertising: A linguistic study of advertising in Great Britain. Longmans, London.
17. Gotti, M., 2005. Advertising discourse in eighteenth-century English newspapers. In: Skaffari, J., Peikola, M., Carroll, R., Hiltunen, R. and Wårvik, B., 2005. Opening Windows on Texts and Discourses of the Past, pp. 23-38. John Benjamins, Amsterdam.
18. Percy, C., 2012. Early Advertising and Newspapers as Sources of Sociolinguistic Investigation. In: Hernández-Campoy, J.M. & Conde-Silvestre, J.C., The Handbook on Historical Sociolinguistics, pp. 191-211. Blackwell, Hoboken.
19. Kemman, M., 2019, Boundary Practices of Digital Humanities Collaborations. Digital Humanities Benelux 2019 1, 1-24.
20. Edmond, J., 2005. The role of the professional intermediary in expanding the humanities computing base. Literary and Linguistic Computing 20.3, 367-380.
21. Siemens, L., Cunningham, R., Duff, W., and Warwick, C., 2011. A tale of two cities: implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities. Literary and Linguistic Computing 26.3, 335-348.
22. Edmond, J., Collaboration and Infrastructure. In: Schreibman, S., Siemens, R. and Unsworth, J., 2015. A New Companion to the Digital Humanities. Wiley, Hoboken.