Digital Approaches Towards Serial Publications

Joke Daems, Gunther Martens, Seth Van Hooland, and Christophe Verbruggen

Journal of European Periodical Studies, 4.1 (Summer 2019)

ISSN 2506-6587

The *Journal of European Periodical Studies* is hosted by Ghent University

Website: ojs.ugent.be/jeps

To cite this article: Joke Daems, Gunther Martens, Seth Van Hooland, and Christophe Verbruggen, 'Digital Approaches Towards Serial Publications', *Journal of European Periodical Studies*, 4.1 (Summer 2019), 1–7

Digital Approaches Towards Serial Publications

Joke Daems*, Gunther Martens*, Seth Van Hooland^, and Christophe Verbruggen*

*Ghent University and ^Université Libre de Bruxelles
joke.daems@ugent.be

> For periodical studies to activate its potential fully, therefore, we will need dedicated institutional sites that furnish the necessary material archives as well as the diverse expertise these rich materials require. The expanding digital repositories, which often patch up gaping holes in the print archive, have begun to provide a broad array of scholars with a dazzling spectrum of periodicals. This dissemination of the archive, in turn, now challenges us to invent the tools and institutional structures necessary to engage the diversity, complexity, and coherence of modern periodical culture.[1]

This volume was inspired by the *Digital Approaches Towards 18th–20th Century Serial Publications* conference, which took place in September 2017 at the Royal Academies for Sciences and Arts of Belgium. The conference brought together humanities scholars, social scientists, computational scientists, and librarians interested in discussing how digital techniques can be used to uncover the different layers of knowledge contained in serial publications such as newspapers, journals, and book series. In this introduction, we discuss some of the key concepts the reader will find throughout this volume, how they fit into the digitization and analysis workflow a digital humanities scholar might employ, and where the different contributions to this volume come into play.

The availability of periodicals in a digital format allows researchers to move from the anecdotal to the statistical through computational analysis. More so than humans, computers can detect patterns in large amounts of data, making it easier to test hypotheses or to observe developments over time in a systematic way. An often-cited concept borrowed from literary studies is 'distant reading'.[2] As opposed to 'close reading', where a literary scholar hypothetically reads every literary piece of work of relevance to their research, 'distant reading' allows scholars to identify recurring patterns and evolutions in large amounts of data without having to read anything. Computational text analysis can even lead to the discovery of new connections. In the medical world, for example, text analysis made it possible to discover relations between findings from

1 Sean Latham and Robert Scholes, 'The Rise of Periodical Studies', *PMLA*, 121.2 (2006), 517–31.
2 Franco Moretti, *Distant Reading* (London: Verso Books, 2013).

seemingly unrelated fields, an association that would have remained unnoticed because field specialists would not otherwise have been exposed to the findings in other fields.[3]

It is the goal of digital humanities to answer questions from a humanities perspective using computational tools as an aid. This does not mean a digital humanist has blind faith in the output of a system; on the contrary, scholarly knowledge is necessary to correctly interpret results and to detect discrepancies or gaps in the data that is used. As Hearst wrote: 'I suggest that to make progress we do not need fully artificial intelligent text analysis; rather, a mixture of computationally-driven and user-guided analysis may open the door to exciting new results.'[4] It is important to stress the need to revise and refine the results of the distant reading processes at the lowest level of granularity, so that the historically relevant concepts that the researcher is looking for throughout the entire corpus can be displayed in context, combining the power of both distant reading and close reading.[5]

From Physical Object to Digital Analysis

A standard digitization and research workflow follows the following pattern: first, a researcher needs access to the periodicals of interest. These can usually be found in libraries or archives. If the researcher is lucky, the same library or archive has access to the entire collection of periodicals the researcher wishes to study. If not, the information must be collected from a variety of locations. Not all resources that are physically available are digitally available. For a researcher interested in digital analysis, the second step is to gain digital access to the resources. Generally, libraries have scanning facilities to make a digital copy of the physical resource. In other cases, the researchers themselves have to take scans or even pictures of the resources they wish to study. This phase leaves the researcher with a collection of images. While already easier to process than physical issues, this is not sufficient for many researchers. The assumption that digitization is about making images has long been refuted. On the one hand, a digital copy needs to be made for preservation purposes, on the other, one has to take into account the complexity of resources to correctly (re)present them to an audience and, possibly, researchers. The following are some of the aspects of digitization that, according to Zdeněk Uhlíř, of necessity move beyond 'making images': The relationship between different images needs to be expressed, metadata has to be added to contain descriptive and structural information, orientation and navigation through the database and collections has to be possible and straightforward, a digital document should consist of images and related text(s), for example, translations, or even other data files such as audio documents, secondary documents, and others. Uhlíř further highlight the following necessary technical conditions: 'dividing data from software', 'standardization of data', and 'interoperability of tools and systems'.[6]

The current markup language of choice to represent data in a standardized way is XML (extensible markup language). For metadata, popular XML standards are

3 Don. R. Swanson and Neil R. Smalheiser, 'Undiscovered Public Knowledge: A Ten-Year Update', *KDD-96 Proceedings* (Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 1996), pp. 295–98.

4 Marti A. Hearst, 'Untangling Text Data Mining', *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (College Park, MD: Association for Computational Linguistics, 1999), pp. 3–10.

5 Stephen McGregor and Barbara McGillivray, 'A Distributional Semantic Methodology for Enhanced Search in Historical Records: A Case Study on Smell', *Proceedings of the 14th Conference on Natural Language Processing — KONVENS* (Vienna: Austrian Academy of Sciences, 2018), pp. 1–11.

6 Zdeněk Uhlíř, 'Digitization Is Not Only Making Images: Manuscript Studies and Digital Processing of Manuscript', *Book Science*, no. 51 (2008), 148–62.

MODS, 'Metadata Object and Description Schema (MODS), [... an] XML schema for encoding descriptive data', and METS, 'for packaging the descriptive metadata and various other important types of metadata needed to assure the use and preservation of digital resources'.[7] For textual markup and the creation of digital critical editions, the TEI (text encoding initiative) standard is commonly used.[8]

For the purpose of text analysis, the most important of these aspects is naturally the access to the text itself, ideally the whole text. One possible way to get access to the full text is through transcription. In this scenario, volunteers or researchers are shown an image and they are given a text box where they can type the text they see. This approach can be relatively accurate though very time-consuming, and — depending on whether or not volunteers get paid — costly.[9] An alternative approach is the use of OCR, or optical character recognition. OCR is a way of automatically detecting text on images.[10] Computer systems are trained on a large collection of images with known text (for example, collected through transcription), after which they are able to identify the text on new images. This technique works fairly well for modern print, because the printed characters look alike and there is a lot of data available that the systems can be trained on. For certain types of historic print or special fonts, however, this technique still leaves much to be desired.[11] The benefits of using OCR is that it is fast and relatively cheap. Researchers can choose between free open source solutions or commercial systems. Still, researchers need to be aware of the limitations of OCR software. Incorrectly recognized characters can greatly impact search results or subsequent analysis as perhaps not all relevant mentions in a corpus are found due to OCR errors.[12] Excellent OCR quality is vital to meet the expectations of both specialist (e.g. automated analysis) and non-specialist users who are often frustrated with recognition results.[13] In addition to certain characters not being recognized correctly, OCR software often struggles with layout. This is crucial in particular for periodicals where multiple-column formats are used,[14] or where images are placed randomly inside a text. Many OCR systems work line by line, meaning that they continue horizontally on a page, even when the column format requires a reader to move to the next line halfway. If it is truly important that the text is entirely correct, a researcher can perform postcorrection on the OCR output.[15] This is done manually — that is, a person corrects the characters that were misinterpreted by the system — or (semi-)automatically — that is, a system learns from the changes

7 Rebecca Guenther and Sally McCallum, 'New Metadata Standards for Digital Resources: MODS and METS', *Bulletin of the American Society for Information Science and Technology*, 29.2 (2005), 12–15.

8 Edward Vanhoutte and Ron Van den Branden, 'Text Encoding Initiative (TEI)', in *Understanding Information Retrieval Systems: Management, Types, and Standards,* ed. by Marcia Bates (Boca Raton, FL: Auerbach Publications, 2011), pp. 671–84.

9 Jacquelyn Slater Reese, 'Transcribing the Past: Crowdsourcing Transcription of Civil War Manuscripts', *Midwest Archives Conference*, 37.2 (2016), 59–74.

10 For a summary, see Noman Islam, Zeeshan Islam, and Nazia Noor, 'A Survey on Optical Character Recognition System', *Journal of Information and Communication Technology*, 10.2 (2016), 1–4.

11 Paul Thompson, John McNaught, and Sophia Ananiadou, 'Customised OCR Correction for Historical Medical Text', *Digital Heritage*, 1.1 (2015), 35–42.

12 Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux, 'Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information', *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (Piscataway, NJ: Institute of Electrical and Electronic Engineers, 2017), pp. 249–52.

13 Andrew Prescott, 'I'd Rather be a Librarian', *Cultural and Social History*, 11.3 (2015), 335–41.

14 Rupinder Pal Kaur and Manish Kumar Jindal, 'Headline and Column Segmentation in Printed Gurumukhi Script Newspapers', in *Smart Innovations in Communication and Computational Sciences: Advances in Intelligent Systems and Computing*, ed. by Bijaya Ketan Panigrahi, Munesh Trivedi, Krishn Mishra, Shailesh Tiwari, and Pradeep Kumar Singh (Singapore: Springer, 2019), pp. 59–67.

15 For a discussion of potential improvements to the OCR output, see Rose Holley, 'How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs', *D-Lib Magazine*, 15.3/4 (2009), no page.

a human makes and corrects the same issues throughout the text systematically. OCR correction remains an important issue today, as evidenced by the annual Post-OCR Text Correction competition hosted at ICDAR, the International Conference on Document Analysis and Recognition.

As soon as a researcher has access to the full text, it is possible to perform digital text analysis. Getting information from text using digital tools is often called text mining or NLP (natural language processing). The method greatly depends on the researcher's goals. Some basic techniques are lemmatization (reducing every word in a sentence to its base form in a dictionary), part-of-speech tagging (determining the part of speech of a word in a sentence), and parsing (finding the components in a sentence and identifying their syntactic roles). A specific NLP use case is named entity recognition or NER. An NER system automatically recognizes named entities (organizations, place names, person names, etc.) in a text. Just like OCR systems, such systems need to be trained before they can be used. The basic principle is the same: a system is given a corpus (a collection of text) with added information (for example, the manually added part-of-speech) and based on the many examples it finds in that corpus, it can apply the same categorization to new text it has never seen before. For most of these basic NLP tasks, researchers can find open-source, pretrained systems for many different languages. For more specific use cases, however, it can be necessary for researchers to train their own systems. In such a case, a corpus first has to be manually annotated or coded. As this is a very difficult process, it is often done by more than one person and the annotations made by the different people are compared to verify whether the so-called inter-coder reliability or inter-annotator agreement is high enough. If sufficient data is manually annotated, a system can be trained to automatically detect the same patterns in unseen text.

This is where the importance of collaboration between domain experts and computational linguists comes in. Text mining tools always need a certain level of adaptation to make them suitable for application to a given text type or subject area: domain specific terminological resources and so called 'gold standard' annotated corpora, that is, 'collections of domain-specific texts in which domain experts have manually marked-up various levels of semantic information that are relevant to the domain in question, such as Named Entities and relationships between them'.[16] For instance, once a corpus of nineteenth-century medical journals has been annotated by experts and/or volunteers (with, for example, relevant names or medical concepts), the gold standard can be used for testing the validity of (semi) automated analysis of other published documents on medically-related matters.

Unlocking European Serial Publications through Digitization and Text Analysis

In the opening contribution to this volume, Michael Piotrowski discusses the importance of the 'human' in 'digital humanities'. Before analysis can take place, it is crucial a researcher thinks about the questions they want to answer, and which underlying models they build their assumptions on. Source criticism becomes an important research skill, in particular on the corpus level: how balanced is the corpus one has collected and wishes to study, how meaningful is the corpus given the assumptions? In the case of periodicals, the question becomes whether periodicals can be used as a model for, for

16 Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou, 'Text Mining the History of Medicine', *PLoS ONE*, 11.1 (2016), e0144717.

example, discourses, ideas, and ideologies. Piotrowski's contribution is intended as a position paper. We challenged the other contributors to take some of the ideas and reflect upon them in their own work.

The important role libraries fulfill in the digital humanities ecosystem is demonstrated in the contribution by Manfred Nölte and Martin Blenkle. They discuss the digitization of historical collections, more specifically, the German journal *Die Grenzboten*, as performed by the State and University Library Bremen. While libraries often consider images and metadata to be sufficiently digitized material, this is often not sufficient for researchers. A collaboration between libraries and research infrastructures as presented in this contribution is a logical next step and echoes the sentiments expressed by Latham and Scholes.[17] It would certainly be interesting to see comparable collaborations develop in other parts of Europe in the future. The main accomplishment was the development of a fully automated OCR postcorrection, to allow researchers to move beyond keyword search and actually perform textual analyses. The authors raise important questions about the desired quality of OCR output for further analysis.

Jani Marjanen's work is situated in a phase that comes before text analysis. They use the metadata information from newspapers in Finland (publication places, language, number of issues, number of words, size of papers, and publishers) to start thinking about a model for the historical development of the public sphere. While their focus is on Finland and the tensions between the Swedish-language and Finnish-language newspapers from 1771 to 1917, their methodology can be applied to other collections as well, to model how public discourse was shaped throughout Europe (and beyond).

Giovanni Colavizza and Matteo Romanello describe the digitization and analysis process for a specific type of text mining: citation mining, or the automatic extraction of citation or reference information from text. They discuss two projects: Cited Loci and Linked Books. Cited Loki attempts to retrieve canonical references from journal articles on JSTOR, Linked Books explores the history of Venice by extracting citations from books as well as journals. This shows that techniques developed for serial publications can in some cases be applied to other resources as well. Citation mining can be used to link the present to the past, and to study the exchange of ideas across publications, regions, and through time. In line with Nölte and Blenkle, Colavizza and Romanello stress the need for infrastructural collaboration, on a national and European level.

Another practical application of text mining is outlined in the contribution by François Dominic Laramée. He studies three prominent French periodicals during 1740 and 1761 to analyze the potential impact of the press on the French colonial ambitions (or lack thereof). In addition to outlining the importance of each of the three publications, relevant issues with OCR quality for the q-subsequent analysis are discussed. The method includes a combination of metadata information and textual information, bringing together keyword searches and collocation frequencies with, for example, the country of origin of a news article.

The contribution by Joke Daems and others contains a section on named-entity recognition and on collocation analysis. The authors examine two socialist newspapers (one French, one Dutch) to study how 'international' the Belgian socialist movement was between 1885–1940. Although conceptually certainly a compelling case study, the more important contribution of this article lies in its methodological evaluation. A large section is dedicated to a discussion of OCR quality and its likely impact on the quality of the subsequent NER analysis. The authors further highlight key issues encountered during the NER analysis. This goes to show once more that digital methods are not

17 Latham and Scholes, pp. 517–31.

perfect just yet, and that it is crucial to apply the knowledge of humanities scholars when interpreting digital text analysis results.

The notion that computational approaches can complement historical and philological research is further expanded upon by Mike Kestemont, Gunther Martens, and Thorsten Ries in the final contribution to this volume. They apply stylometric authorship verification techniques to identify contributions of Goethe to the *Frankfurter gelehrte Anzeigen* of 1772–73. The extent on Goethe's contribution to this journal has been at the center of a long-standing debate, and previous attempts at attribution based on philological and stylistic indicators have proven indecisive. Novel techniques can be used now because of the availability of large corpora of, for example, digitized periodicals. The contribution also outlines the practical challenges of working with old print fonts and 'dirty' OCR, and concludes that stylometry opens up interesting pathways for studying periodicals and the rise of the modern notion of authorship.

With the contributions included in this issue, we hope to showcase some of the potential of digital approaches for periodical studies and the types of research questions these techniques can help answer. The works included inevitably present but a fraction of the possibilities computational techniques can offer. In the past few years, linguists have made a lot of progress in the development of multilingual automatic term extraction from comparable corpora, such as cross-lingual word sense disambiguation and the automatic construction of multilingual semantic networks,[18] but both applied and fundamental (computational) research remains necessary. Moreover, many NLP tools and workflows have been developed for linguists and computer scientists whose research interests differ from those focusing on cultural heritage and long-term dynamics of conceptual change. On the other hand, libraries and museums are increasingly investing in technologies such as IIIF, the International Image Interoperability Framework, that help with digital preservation and presentation of cultural heritage, and that could also be of benefit to researchers, yet many researchers active in text mining and computational linguistics are unaware of their existence. As a consequence, further development of new methods in text mining and computational linguistics that can be applied to historical resources such as periodicals can only be realized when (historical) expert knowledge is combined with expertise in computational linguistics.

Bibliography

Chiron, Guillaume, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux, 'Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information', *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (Piscataway, NJ: Institute of Electrical and Electronic Engineers, 2017), pp. 249–52

Guenther, Rebecca, and Sally McCallum, 'New Metadata Standards for Digital Resources: MODS and METS', *Bulletin of the American Society for Information Science and Technology,* 29.2 (2005), 12–15

Hearst, Marti A., 'Untangling Text Data Mining', *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (College Park, MD: Association for Computational Linguistics, 1999), pp. 3–10

18 Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever, 'A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents', *Proceedings of the 11th International Conference on Language Resources and Evaluation* (Miyazaki: European Language Resources Association, 2018), pp. 1803–08; Roberto Navigli and Simone Paolo Ponzetto, 'BabelNet: The Automatic Construction, Evaluation, and Application of a Wide-Coverage Multilingual Semantic Network', *Artificial Intelligence*, vol. 193 (2012), 217–50.

Holley, Rose, 'How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs', *D-Lib Magazine*, 15.3/4 (2009), no page

Islam, Noman, Zeeshan Islam, and Nazia Noor, 'A Survey on Optical Character Recognition System', *Journal of Information and Communication Technology*, 10.2 (2016), 1–4

Kaur, Rupinder Pal, and Manish Kumar Jindal, 'Headline and Column Segmentation in Printed Gurumukhi Script Newspapers', in *Smart Innovations in Communication and Computational Sciences: Advances in Intelligent Systems and Computing*, ed. by Bijaya Ketan Panigrahi, Munesh Trivedi, Krishn Mishra, Shailesh Tiwari, and Pradeep Kumar Singh (Singapore: Springer, 2019), pp. 59–67

Latham, Sean, and Robert Scholes, 'The Rise of Periodical Studies', *PMLA*, 121.2 (2006), 517–31

McGregor, Stephen, and Barbara McGillivray, 'A Distributional Semantic Methodology for Enhanced Search in Historical Records: A Case Study on Smell', *Proceedings of the 14th Conference on Natural Language Processing — KONVENS* (Vienna: Austrian Academy of Sciences, 2018), pp. 1–11

Moretti, Franco, *Distant Reading* (London: Verso Books, 2013)

Navigli, Roberto, and Simone Paolo Ponzetto, 'BabelNet: The Automatic Construction, Evaluation, and Application of a Wide-Coverage Multilingual Semantic Network', *Artificial Intelligence*, vol. 193 (2012), 217–50

Prescott, Andrew, 'I'd Rather be a Librarian', *Cultural and Social History*, 11.3 (2015), 335–41

Reese, Jacquelyn Slater, 'Transcribing the Past: Crowdsourcing Transcription of Civil War Manuscripts', *Midwest Archives Conference*, 37.2 (2016), 59–74

Swanson, Don. R., and Neil R. Smalheiser, 'Undiscovered Public Knowledge: A Ten-Year Update', *KDD-96 Proceedings* (Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 1996), pp. 295–98

Terryn, Ayla Rigouts, Véronique Hoste, and Els Lefever, 'A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents', *Proceedings of the 11th International Conference on Language Resources and Evaluation* (Miyazaki: European Language Resources Association, 2018), pp. 1803–08

Thompson, Paul, John McNaught, and Sophia Ananiadou, 'Customised OCR Correction for Historical Medical Text', *Digital Heritage*, 1.1 (2015), 35–42

Thompson, Paul, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou, 'Text Mining the History of Medicine', *PLoS ONE*, 11.1 (2016), e0144717

Uhlíř, Zdeněk, 'Digitization Is Not Only Making Images: Manuscript Studies and Digital Processing of Manuscript', *Book Science*, no. 51 (2008), 148–62

Vanhoutte, Edward, and Ron Van den Branden, 'Text Encoding Initiative (TEI)', in *Understanding Information Retrieval Systems: Management, Types, and Standards,* ed. by Marcia Bates (Boca Raton, FL: Auerbach Publications, 2011), pp. 671–84