AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts

Róbert Péter[1], Zsolt Szántó[2], József Seres[2],
Vilmos Bilicki[2], Gábor Berend[2,3]

[1] University of Szeged, Institute of English and American Studies
[2] University of Szeged, Institute of Informatics
[3] MTA-SZTE, Research Group on Artificial Intelligence
robert.peter@ieas-szeged.hu
{szantozs,bilickiv,berendg}@inf.u-szeged.hu,
seres.jozsef.1@stud.u-szeged.hu

Abstract: The objective of this paper is to demonstrate the different functions and features of the multilingual AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) data-driven digital tool. This new web application enables digital humanists to critically analyse the bibliographic data and texts of large corpora including entire library repositories and digital collections at scale. The implemented state-of-the-art text analytical and visualization tools such as topic modeling and network analysis provide interactive close and distant reading of texts and bibliographic data. The export functions of AVOBMAT facilitate the reproducibility of the results and transparency of the preprocessing and text analysis.

1 Introduction

Text and data mining methods have become increasingly widespread in the natural, social sciences and the humanities in recent years (Prescott, 2015; Graham et al., 2016; Moreux 2016; Schiuma and Carlucci 2018; Mahmoudi and Abbasalizadeh, 2019).Vast amount of humanities resources has been digitalized and encoded in the last three decades in various quality and made available in numerous open access and subscription-based databases with many restrictions including the lack or limited access to the digital collections for exploratory analysis with data-driven methods and tools. It can be observed that the often rich bibliographic (meta)data of documents are scarcely and partially made use of during the corpus preparation and computational text analysis process. This is not surprising since, as Franco Moretti, Matthew L. Jockers and Mikko Tolonen have argued, bibliographic metadata has been unmapped as a means of investigating (literary) history (Jockers, 2013; Moretti, 2013; Tolonen, 2015). Recent research has demonstrated that, like text mining, the critical examination of bibliographic (big) data can also offer many new insights, unveil hitherto overlooked patterns and trends, provide novel type of evidence and findings as well as question old hypotheses in the humanities (Varlamis and Tsatsaronis, 2011; Péter,

2011, Fenlon et al., 2012; Jockers, 2013; Prescott, 2013; Péter, 2015; Hill et al. 2019; Lahti et al., 2019). That is why the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) research tool combines these two data-driven approaches in one integrated, interactive and user-friendly web application. It hopefully assists non-programmer researchers in their critical text and data mining of open access text corpora, individual digital collections and bibliographic datasets. The implemented state-of-the-art analytical and visualization tools ranging from network analysis to topic modelling can provide close and distant reading of texts and bibliographic data at scale. The export of the preprocessed texts and analysis results facilitates the reproducibility of the findings and foster critical thinking about the text preparation and analytical process.

Since March 2017 researchers, library IT specialists and students in the Institute of Informatics, Klebelsberg Library and Department of English Studies at the University of Szeged have been working on the development of the AVOBMAT research tool. The aim of this paper is to briefly demonstrate the workflow and the different analytical functions and features of the AVOBMAT web-based application designed for digital humanities research.

2 Related work

Text mining and natural language processing techniques have been utilized for the analysis of linguistic corpora and databases for a long time. However, with very few exceptions such as the Media Monitoring of the Past project[1], most publicly available historical, literary and cultural databases do not allow researchers to analyse the included texts with data-driven and natural language processing methods in a critical manner. In commercial products such as the Gale Digital Scholar Lab[2] (inaccessible at most universities and research institutions) individual or library databases, (pretrained) NLP models and dictionaries cannot be uploaded. Furthermore, most currently available web-based text analysis tools such as Voyant Tools[3] cannot cope with large corpora of texts (Sinclair and Rockwell, 2019). The relatively few browser-based digital humanities applications only allow users to set a reduced number of processing parameters for the NLP algorithms they offer. Paper Machine[4] as a plugin for Zotero Standalone bibliographic management software once managed to combine basic bibliographic metadata (date, title and location) and topic modeling analysis but it is not compatible with the current version (5.0) of Zotero and thus it is no longer maintained. The Interactive Text Mining Suite[5] web application pre-processes texts (txt, pdf) and provides cluster, topic and frequency analyses but it can handle few metadata fields such as author, year, title and category (Scrivner and Davis, 2016; Scrivner and Davis, 2017; Scrivner et al., 2017). Among these browser-based applica-

[1] https://impresso-project.ch/app/#/
[2] https://www.gale.com/primary-sources/digital-scholar-lab
[3] https://voyant-tools.org/
[4] http://papermachines.org/
[5] https://languagevariationsuite.wordpress.com/2016/03/18/interactive-text-mining-suite-itms/

tions, Lexos[6] provides the greatest number of tools for pre-processing and segmenting the uploaded texts (TXT, HTML and XML). Lexos tokenizes texts, identifies n-grams, generates statistical summaries, visualize the results (word cloud, multicloud, bubbleviz), analyses the digitized texts (frequency, k-means, clustering, cosine similarity ranking). It also provides a "topic cloud" based on the output format created by MALLET (McCallum, 2002) data but one cannot perform topic modeling within Lexos itself (Kleinman et al, 2019). Unlike these tools, AVOBMAT can enrich, analyse bibliographic data and filter the uploaded digital collections for built-in computational text analysis with the help of metadata or (full-text) keyword searches including fuzzy and proximity queries. The latter filtering can also be used in the AVOBMAT system to explore and visualize the bibliographic data of the filtered digital collections in an interactive and dynamic way.

3 Upload, preprocessing and metadata enrichment

Users can currently upload Zotero collections in CSV and RDF (with full texts) formats as well as EPrints (library) repositories (EP3 XML with urls to the full texts). Zotero can import 20 different types of bibliographic formats (e.g. MARC, BibTex) that users can organize in collections. In Zotero users can modify metadata and texts of these collections, for example, by manually adding new items. These collections can be imported in AVOBMAT in CSV and RDF formats. The Zotero-based csv structure containing 87 bibliographic (meta)data fields (e.g. author, publisher, title) was expanded with 20 other new fields such as bookseller, printers, gender (of the authors) and frequency of publications that researchers can analyse in various ways (see chapter 4) and make use of them when creating their own digital collections to be explored in AVOBMAT.

The differences between the various bibliographic metadata schema standards are reconciled. For example, the EP3 XML 'Publication' field is identical with the Zotero 'Publication Title' field, thus they are both imported into a common Elasticsearch field as publicationTitle. Through the csv upload AVOBMAT can also import full texts of documents via either the added "Full Text" field or the "Url" field pointing to the full texts of the documents. Hence it can import texts in several formats including TXT, PDF, DOC(X) and XML since the Apache Tika Python library converts them to plain text.

In the preprocessing phase the user can set the individual parameters for the different types of content analyses including N-gram viewer, topic modeling, significant text and TagSphere. Users can create different configurations for the different analyses and visualizations. The following preprocessing steps are implemented: (i) lowercase the text; (ii) remove numbers; (iii) remove non-alphabetical tokens; (iv) replace expressions and characters; (v) remove hyphens. The (vi) context-filter can be used to keep the context of a keyword or keywords, and remove all other parts of the document. Users can specify the search terms and the length of the context.

[6] http://lexos.wheatoncollege.edu/upload

The preprocessing interface also provides three optional functions (vii) lemmatization, (viii) punctation and (ix) stopword filtering) that can be configured for each analysis separately. In these cases the outcome of the module depends on the language of the texts to be uploaded. There are two ways to assign a language to a document—researchers can either manually select a language for the full dataset or choose the automatic language detection option. In case of automatic language detection, the system chooses a language independently for each document by using the langdetect language detection library[7]. Based on the selected or detected language one can select stopword, punctuation filtering and lemmatization. For the stopword and the punctuation filtering we use the spaCy library[8]. Extra stopword and punctuation lists can also be added. In case of Bulgarian, Czech, Estonian, Hungarian, Italian, Romanian, Serbian, Slovenian, Polish and Spanish we use the LemmaGen (Juršic et al., 2010) – an open source multilingual tool – for lemmatization, and for other languages with spaCy support – including English, French and German – we use spaCy. All the content analyses are performed online based upon the preprocessed texts. The raw and the preprocessed texts are stored in distinct fields on the server. Hence, one can access and search the raw full texts after the preprocessing phase.

The metadata enrichment includes the automatic identification of the gender of the authors and automatic language detection of the documents. We utilize a further developed version of Python sexmachine package for automatic identification of the gender of the authors primarily based on their forenames.

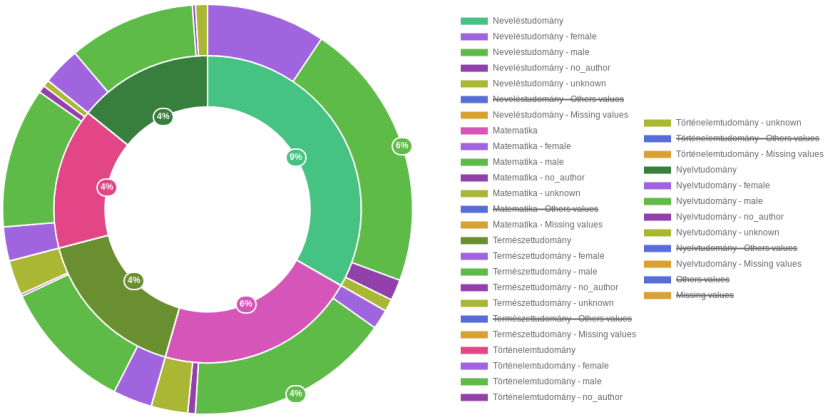

Fig. 1. Distribution of female, male, unknown and no author according to major disciplines as recorded in the Acta repository of the University of Szeged (43300 articles)

[7] https://github.com/shuyo/language-detection
[8] https://spacy.io/

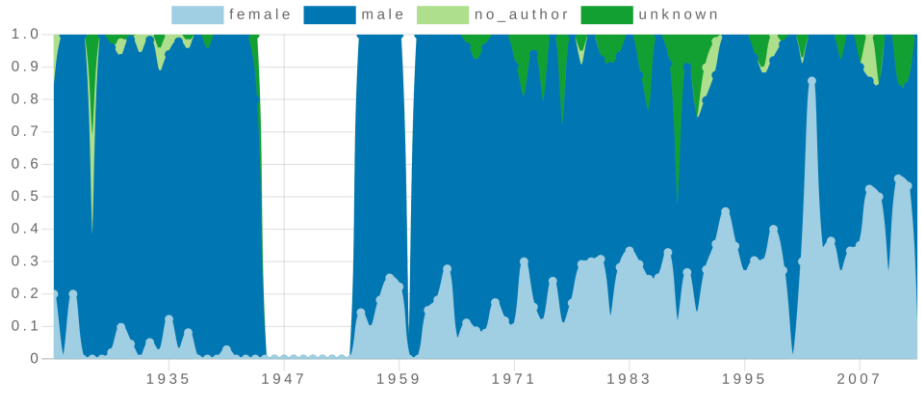

Fig. 2. Distribution of female, male, unknown and no authors in 1653 linguistics journal articles as recorded in the Acta repository of the University of Szeged between 1924 and 2013.

Empowered by the Elasticsearch engine users can search and filter the uploaded and enriched bibliographic data and preprocessed texts in faceted, advanced and command line modes and perform all the subsequent analyses on the filtered dataset. Through the search results interface it is possible to access the uploaded texts and related metadata directly. Our framework also supports fuzzy and proximity search as well as regular expression queries.

The reproducibility and transparency of the experiments and results are enhanced by the export of the parameter settings, preprocessed texts, generated tabular statistical data of the analytical results and visualizations in different formats. The application also enables researchers to use the processed texts and bibliographic data in other software. AVOBMAT does not retain any data after a session has expired.

4 Bibliographic data analysis

Having filtered the uploaded databases and selected the metadata field(s) to be explored, users can, among others, (i) analyse and visualize the bibliographic data chronologically in line and area charts in normalized and aggregated formats; (ii) create an interactive network analysis of maximum three (meta)data fields; (iii) make pie as well as horizontal and vertical bar charts of the bibliographic data of their choice according to the provided parameters: researchers can choose the metadata field(s) and the number of top items for visualization. For example, AVOBMAT can create an interactive network graph of authors, publishers and booksellers. All the analytical results and visualizations such as graphs, time series and charts can be exported in csv and png formats. To foster the critical investigation of the bibliographic records it also presents the number of missing and other values (not included in the dataset limited by the selected number of top items parameter) in the filtered corpus.

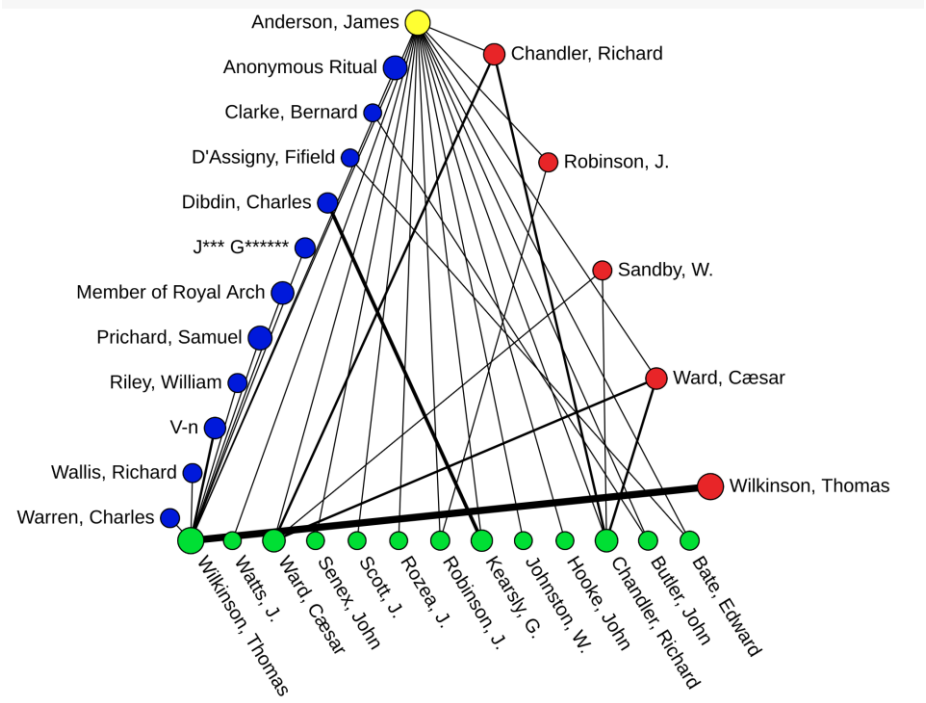

Fig. 3. James Anderson's (author of the *Constitutions* of Freemasonry) eighteenth-century publisher-bookseller network

5 Content analysis

5.1 N-gram viewer

The diachronic analysis of texts is supported by the N-gram viewer in AVOBMAT. It shows the yearly count of the specified n-grams. The inputs are the list of n-grams and a selected metadata field (e.g. the full texts or the titles of the documents). The n-grams with a maximum 5-word length are generated at the preprocessing stage. The chart also provides a normalized view where the number of the n-grams is divided by the total number of words in the given years.

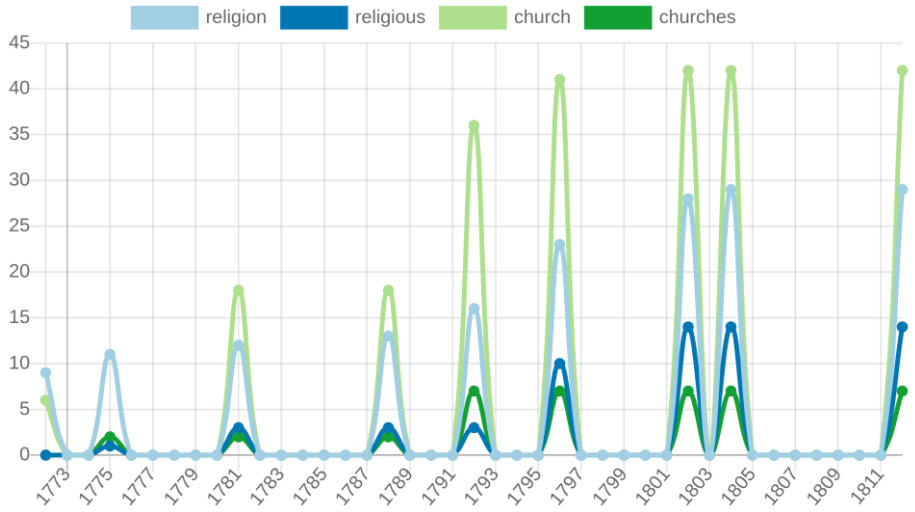

Fig. 4. The terms religion, religious, church and churches in the nine different editions of William Preston's *Illustrations of Masonry*, 1775-1812.

5.2 Word clouds

Word clouds can be efficient tools to highlight prominent terms in a corpus. We implemented two special types of word clouds: the significant text cloud showing what differentiates a subset of the documents from others[9] and the TagSpheres (Jänicke and Scheuermann, 2017) enabling users to investigate the context of a word. There are bar chart versions of the different word clouds that present the applied scores and frequencies.

The significant text visualization highlights the most related terms to a special query. If the user filters a time period or selects an author by using the search functions of the AVOBMAT, this tool highlights the words that are most strongly related to this selected subset of documents.

Traditional word clouds treat the words independently and lose the contextual information between them. For the graphical representation of the context of a word in a corpus the TagSpheres was integrated. It creates tag clouds that show the co-occurring words of a specified search term in a corresponding word distance. Besides the search term, the user can specify the minimum frequency and the maximum distance of the co-occurring words. While the minimum frequency handles the rare, uncommon words and maximum word distance controls mostly the size of the chart.

[9] https://www.elastic.co/guide/en/elasticsearch/reference/master/search-aggregations-bucket-significanttext-aggregation.html

Fig. 5. TagSpheres: the context of the word secret without stopwords in the nine editions of William Preston's *Illustrations of Masonry*, 1775-1812.

5.3 Topic modeling

Topic modeling can help us find hidden semantic information about selected corpora. Statistical methods are used to discover the themes that are embedded in the texts and to reveal the connections of these themes and their changes over time (Blei et al., 2003). The AVOBMAT has an in-browser Latent Dirichlet Allocation function that draws on the jsLDA library[10] to calculate and graphically represent topic models. The LDA gives a predefined number of latent topics where each document may be viewed as a mixture of these topics. Besides the topic analysis AVOBMAT can also visualize the results of the modeling in various ways. As an improvement over the standard jsLDA tool, AVOBMAT allows for the adjustment of the LDA hyperparameters α and β that control the topic diversity of the documents and words. Bibliographic in-

[10] https://mimno.infosci.cornell.edu/jsLDA/

formation (i. e. authors, publication titles, dates) that plays a crucial role in knowledge discovery is hardly reflected in topic modelling. Unlike jsLDA, when listing topic documents, AVOBMAT displays the afore-mentioned basic bibliographic data, along with the probabilistic values for each document. It also presents the evolution of the different topics in an interactive time series. Moreover, before running the topic modelling, the user can filter and categorize the text corpus with the help of the metadata (currently 97 optional fields) associated with each document. The results can be exported in six different ways: document topics, topic words, topic summaries, topic-topic connections, doc-topic graph file (for Gephi) and complete sampling state (docID, word and topic number).

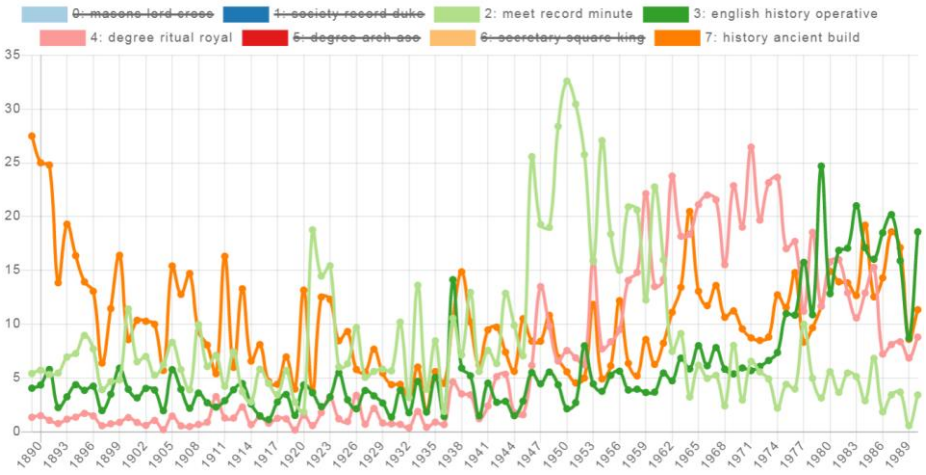

Fig. 6. Time series of topics with lemmatized words in the *Ars Quatuor Coronatorum* journal between 1890 and 1990. For instance, topic 4 (degree, ritual, royal, arch [Royal Arch: name of the 4th degree [rank] in masonic ritual hierarchy], ceremony. lecture, question] is clearly related to ritual studies in the journal. Topic 2 (meet, record, minute, pay, masons, elect, secretary apprentice) is concerned with the lodge meetings of Freemasons as recorded in their minutes. Topic 3 is about the history of operative (stonemason) Freemasonry.

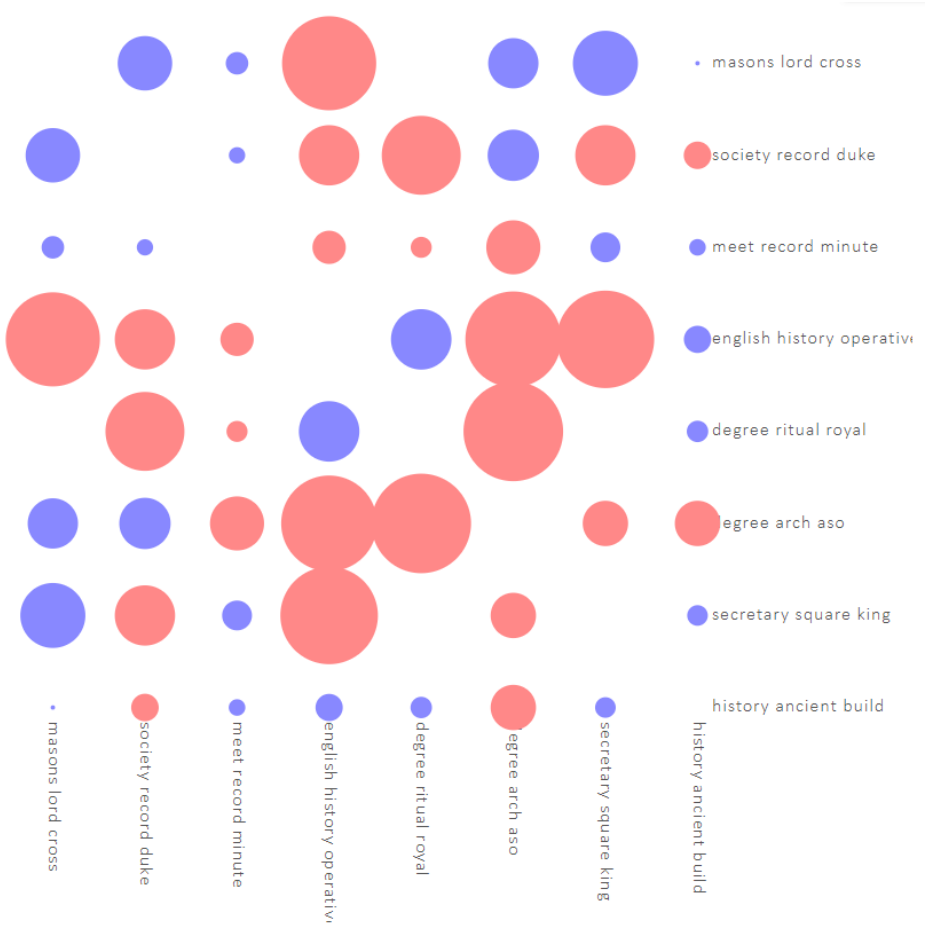

Fig. 7. Topic correlations in the *Ars Quatuor Coronatorum* journal.

5 Conclusion and future work

This paper has introduced the platform-independent and multilingual AVOBMAT system, which was primarily developed for scholars and students who are not familiar with command-line scripting. We have seen that this exploratory web application allows for a range of dynamic text and data mining tasks. The simple user-friendly web-based interface provides interactive parameter tuning and control from the pre-processing to the analytical stages.

The unique feature of the AVOBMAT toolkit is that it combines cutting-edge bibliographic data and computational text analysis research. It allows users to filter the uploaded datasets by metadata and full-text searches of various types and perform all the bibliographic, network and natural language analyses on the filtered datasets. In this way researchers can easily and interactively experiment with the different metada-

ta and content analyses, the parameters of which can be set online. Hence, it helps users realize the epistemological challenges, limitations and strengths of computational text analysis and visual representation of digital texts and datasets.

Besides revealing hitherto unknown connections of the different metadata fields and highlighting overlooked trends over time, the bibliographic data analysis can also highlight selection biases, errors in the bibliographic (meta)data (e.g. incorrect classifications) and reveal missing values and gaps in the data. Most data providers including libraries and profit-oriented companies have either not realized these, or if they did, they are reluctant to make this information public. Identifying these shortcomings helps researchers make informed decisions about their projects and critically analyse their datasets. It can also assist librarians in identifying the historical development regarding the creation of bibliographic records and improving the quality of the metadata of the repositories.

By combining distant and close reading approaches in our analytical framework, researchers can identify new perspectives for bibliographic data and textual analysis, discover novel insights, hidden patterns, themes and trends in digital collections. For instance, the use of bibliographic data allows scholars to perform diachronic topic analysis revealing wider semantic patterns in language use than a close reading of massive digital collections would provide. With the help of the bibliographic metadata we can filter our text corpora for advanced and comprehensive content analyses and carry out network analysis of the different metadata fields. The traditional contextual close reading examination is fostered by the TagSphere visualization tool and the keyword in context (KWIC) representation of the search queries.

The AVOBMAT application will be made available for the public in 2020 with some restrictions concerning the size of the digital collections to be uploaded and explored due to limited server capacities.

We plan several developments in the near future. For instance, we intend to extend the AVOBMAT multilingual system with named entity recognition (with disambiguation), parts of speech tagger, lexical richness and sentiment analysis tools. We would also like to increase the number of supported languages for lemmatization as well as the input file types. If server capacities permit, users will be allowed to upload their own pre-trained models.

Acknowledgements

The research was supported by the EU-funded Hungarian grant EFOP-3.6.1-16-2016-00008. It was also supported by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002. The project has been supported by the European Union and co-funded by the European Social Fund. We are grateful to Károly Kokas, Ákos Sándor, Gyula Nagy and Zoltán Erdődi of the Klebelsberg University Library for providing access to library repositories as well as their professional and technical support. Our thanks are due to Tamás Ficand, Gábor Simon and Gergely Dér who participated in the devel-

opment of the previous versions of the AVOBMAT. Simon and Dér devoted their BSc and MSc theses to this topic.

Bibliography

Blei, D. M., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning. Research. 3, 993–1022 (2003).

Fenlon, K. et al.: Tooling the Aggregator's Workbench: Metadata Visualization Through Statistical Text Analysis. Proceedings of the American Society for Information Science and Technology. 49, 1, 1–10 (2012).

Graham, S. et al.: Exploring Big Historical Data: the Historian's Macroscope. Imperial College Press, London (2016).

Hill, M. J. et al.: Reconstructing Intellectual Networks: From the ESTC's Bibliographic Metadata to Historical Material. In: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference: Copenhagen, Denmark, March 5–8, 2019. pp. 201–219. CEUR-WS.org (2019).

Jänicke, S., Scheuermann, G.: On the Visualization of Hierarchical Relations and Tree Structures with TagSpheres. In: Braz, J. et al. (eds.) Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 199–219. Springer International Publishing, Cham (2017).

Jockers, M. L.: Macroanalysis: Digital Methods and Literary History. University of Illinois Press., Champaign, IL. (2013).

Juršic, M., et al. Lemmagen: Multilingual Lemmatisation with Induced Ripple-down Rules. Journal of Universal Computer Science 16, 9, 1190-1214 (2010).

Kleinman, S., LeBlanc, M.D., Drout, M., and Feng, W. (2019). Lexos. v4.0 https://github.com/WheatonCS/Lexos/.

Lahti, L. et al.: Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly. 57, 1, 5–23 (2019).

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Mahmoudi, M.R., Abbasalizadeh, A.: How Statistics and Text Mining can be Applied to Literary Studies? Digital Scholarship in the Humanities. 34, 3, 536–541 (2019).

Moretti, F.: Distant Reading. Verso, London; New York (2013).

Moreux, J.-P.: Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment. In: IFLA News Media Section, Lexington, United States (2016).

Péter, R.: Researching (British Digital) Press Archives with New Quantitative Methods. Hungarian Journal for English and American Studies. 17, 2, 283-300 (2011).

Péter, R.: Digitális és módszertani fordulat a sajtókutatásban: A 17–18. századi magyar vonatkozású angol újságcikkek 'távolságtartó olvasása' [Digital and Methodological Turn in the Study of the Press: a 'Distant Reading' of 17^{th}-18^{th} c. English Newspapers Concerning Hungary]. Aetas 29, 1, 5-30 (2015).

Prescott, A.: Bibliographic Records as Humanities Big Data. In: 2013 IEEE International Conference on Big Data. pp. 55–58 (2013).

Prescott, A.: Big Data in the Arts and Humanities: Some Arts and Humanities Research Council Projects. University of Glasgow, Glasgow (2015).

Schiuma, G., Carlucci, D.: Big Data in the Arts and Humanities: Theory and Practice. Taylor and Francis, Boca Raton, FL (2018).

Scrivner, O. et al.: Building Customized Text Mining Tools via Shiny Framework: The Future of Data Visualization. In: MAICS. (2017).

Scrivner, O., Davis, J.: Interactive Text Mining Suite: Data Visualization for Literary Studies. In: CDH@TLT. (2017).

Scrivner, O., Davis, J.: Topic Modeling of Scholarly Articles: Interactive Text Mining Suite. Presented at the Computational Linguistics and Intellectual Technologies Conference, Moscow, June 1-4, 2016 (2016).

Sinclair, S., and Rockwell, G. (2019). Voyant Tools. Web. http://voyant-tools.org/.

Tolonen, M. et al.: A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800. Liber quarterly. 25, 2, 87–116 (2015).

Varlamis, I., Tsatsaronis, G.: Visualizing Bibliographic Databases as Graphs and Mining Potential Research Synergies. In: 2011 International Conference on Advances in Social Networks Analysis and Mining. pp. 53–60 (2011).