BiblioDAP'21: The 1st Workshop on Bibliographic Data Analysis and Processing

Zeyd Boukhers
boukhers@uni-koblenz.de
Institute for Web Science and Technologies
University of Koblenz-Landau
Koblenz, Germany

Philipp Mayr
philipp.mayr@gesis.org
GESIS – Leibniz-Institute for the Social Sciences
Cologne, Germany

Silvio Peroni
silvio.peroni@unibo.it
Research Centre for Open Scholarly Metadata
Department of Classical Philology and Italian Studies
University of Bologna, Italy
Bologna, Italy

ABSTRACT

Automatic processing of bibliographic data becomes very important in digital libraries, data science and machine learning due to its importance in keeping pace with the significant increase of published papers every year from one side and to the inherent challenges from the other side. This processing has several aspects including but not limited to I) Automatic extraction of references from PDF documents, II) Building an accurate citation graph, III) Author name disambiguation, etc. Bibliographic data is heterogeneous by nature and occurs in both structured (e.g. citation graph) and unstructured (e.g. publications) formats. Therefore, it requires data science and machine learning techniques to be processed and analysed. Here we introduce BiblioDAP'21: The 1st Workshop on Bibliographic Data Analysis and Processing.

CCS CONCEPTS

• **Information systems** → *Information retrieval*; • **Computing methodologies** → *Machine learning*; • **Applied computing** → **Digital libraries and archives**.

KEYWORDS

Bibliographic data, Digital libraries, Machine Learning, Data Science

ACM Reference Format:
Zeyd Boukhers, Philipp Mayr, and Silvio Peroni. 2021. BiblioDAP'21: The 1st Workshop on Bibliographic Data Analysis and Processing.

1 INTRODUCTION

The aim of the Workshop on Bibliographic Data Analysis and Processing (BiblioDAP[1]) is to address open challenges in digital libraries and to attract the attention of the research communities to present novel approaches to bibliographic data analysis, processing and understanding. Consequently, we invite the submission of original and high-quality research papers and reports of live demonstrations and prototypes on the following and any related topics:

- Reference Analysis
- Citation Network Analysis
- Author Name Disambiguation
- Scientific data management
- Metadata extraction
- Plagiarism detection
- Bibliographic data quality improvement
- Entity linkage in bibliographic data

Author Name Disambiguation (AND) is an open challenging problem and its effects are growing as the number of authors sharing the same names arises significantly [4]. To the best of our knowledge, the last big AND challenge was held in KDD Cup in 2013, which was very successful as several original implemented techniques were presented and available until nowadays. However, since then, the area of Data Science and Machine Learning has significantly evolved together with the availability of computational resources. To fill the gap between the challenges of AND caused by the continuous increase of authors sharing the same name from one side and the remarkable advancement of artificial intelligence from the other side, we are also organizing a shared task on Author Name Disambiguation [2]. To this end, two annotated datasets are released; one for development and one for testing. The workshop invite researchers to submit their results and request those with high accuracy to submit their original research papers, live demonstrations and source codes that tackle this particular problem.

As SIGKDD is a premier Data Science conference that gathers passionate scholars specialized in data science and its related fields, organizing the workshop in conjunction with KDD 2021 aims to attract the attention of an important part of the data science community that is interested in challenging topics related to bibliographic data.

2 MOTIVATION

The purpose of BiblioDAP'21 is to share knowledge and techniques in data science to be applied to bibliographic data and to overcome its inherent challenges. Data science and machine learning techniques are evolving and progressing significantly being applied to various types of data and for different objectives. BiblioDap' 21 aims to benefit from this progress by bringing together bibliographic data and data science capabilities to overcome the open

[1] https://bibliodap.uni-koblenz.de/

[2] https://github.com/BiblioDap/AND-Task

BiblioDAP, The 1st Workshop on Bibliographic Data Analysis and Processing

challenges. The organizers have been managing several projects [3] and infrastructures [4] related to processing and analysing bibliographic data which allow them to have a broad expertise in this discipline and a wide familiarity with the challenges inherent with this data.

3 WORKSHOP SETUP

BiblioDAP' 21 consists of one invited talk followed by technical presentations and a panel. The technical presentations are given by the authors of the accepted papers which are reviewed by at least two seasoned reviewers based on the originality and clarity of the paper. BiblioDAP' 21 supports open peer review and allows reviewers to disclose their identities and publish their reviews. Using this approach, it is up to the reviewers to decide either to go through an open peer-review process or to keep their names anonymous. The authors of submitted papers were aware that their reviewers may decide to disclose their names and to publish the review openly online in the platform of their choice. In particular, we pointed to specific guidelines (available at https://open-sci.github.io/review/) that have been devised to foster the adoption of Open Science practices, to publicly acknowledge the effort researchers spend in reviewing, and to enable the reviewers to take public responsibility for the content of the reviews they write to help the authors to improve their work. Since BiblioDAP' 21 is non-archival, it accepts submissions of papers that are under review to other venues, including KDD' 21.

3.1 Keynote

Our keynote was given by Alberto Laender (Universidade Federal de Minas Gerais, Brazil)[5] in collaboration with Anderson A. Ferreira (Universidade Federal de Ouro Preto, Brazil) and Marcos André Gonçalves (Universidade Federal de Minas Gerais, Brazil). Alberto and his team talked about "**Automatic Disambiguation of Author Names: Foundations, Methods and Open Issues**" (see their recent book on AND [4]).

Abstract of their talk: *Author name disambiguation is a well-known hard problem, that has profound impacts on services provided by bibliographic repositories and similar platforms. Despite almost 20 years of research on the topic and efforts such as ORCID, there are still several open issues to be solved. In this talk, we will revisit this problem, presenting an overview and related taxonomy, and elaborate on some methods developed by our own research group that follow distinct approaches and tackle fundamental aspects of the problem such as self training and incremental disambiguation. Finally we will briefly discuss recent approaches and issues still open.*

3.2 Program Committee

We appreciate the reviewers' efforts and would like to thank the members of the program committee for their valuable support.

[3] EXCITE: https://excite.informatik.uni-stuttgart.de/ [5]
[4] OpenCitations: http://opencitations.net/
[5] https://homepages.dcc.ufmg.br/~laender/

4 RELATED WORKSHOPS

The chairs of BiblioDAP'21 have been involved, in the past, in the organisation of several events and workshops dedicated to scholarly data, including bibliographic and citation data. Among these events there are:

- the SAVE-SD workshop series (https://save-sd.github.io/) is a series of workshops dedicated to technologies aiming at enhancing scholarly dissemination;
- the Workshop on Open Citations and Open Scholarly Metadata (https://workshop-oc.github.io/) has reached its second edition in 2020 and involved researchers, scholarly publishers, founders, policy makers, and opening citations advocates, interested in the creation, reuse, and improvement, of open citation data and open scholarly metadata;
- the Bibliometric-enhanced IR (BIR) workshop series (https://sites.google.co tackles issues related to academic search, at the crossroads between Information Retrieval and Bibliometrics [2].
- the Scholarly Document Processing workshop (https://sdproc.org/2021/) covers research and shared task tracks to work on enhancing search, summarization, and analysis of scholarly documents [3].

REFERENCES

[1] Zeyd Boukhers, Nada Beili, Timo Hartmann, Prantik Goswami, and Muhammad Arslan Zafar. 2021. MexPub: Deep Transfer Learning for Metadata Extraction from German Publications. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 1–4.

[2] Guillaume Cabanac, Ingo Frommholz, and Philipp Mayr. 2020. Bibliometric-Enhanced Information Retrieval 10th Anniversary Workshop Edition. In *Advances in Information Retrieval*. Vol. 12036. Springer International Publishing, Cham, 641–647. http://link.springer.com/10.1007/978-3-030-45442-5_85

[3] Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the First Workshop on Scholarly Document Processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, 1–6. https://www.aclweb.org/anthology/2020.sdp-1.1.pdf

[4] Anderson A. Ferreira, Marcos André Gonçalves, and Alberto H. F. Laender. 2020. *Automatic Disambiguation of Author Names in Bibliographic Repositories*. Morgan & Claypool Publishers. https://doi.org/10.2200/S01011ED1V01Y202005ICR070

[5] Azam Hosseini, Behnam Ghavimi, Zeyd Boukhers, and Philipp Mayr. 2019. EXCITE - A toolchain to extract, match and publish open literature references. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2019*. ACM, 432–433. https://doi.org/10.1109/JCDL.2019.00105

[6] Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics* 124, 3 (2020), 1907–1922.

[7] Alexander Tekles and Lutz Bornmann. 2020. Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies* 1, 4 (2020), 1510–1528.

[8] Martijn Visser, Nees Jan van Eck, and Ludo Waltman. 2021. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies* 2, 1 (2021), 20–41.