# Exploring data quality and retrieval strategies for Mendeley reader counts

Zohreh Zahedi, Stefanie Haustein and Timothy D. Bowman

z.zahedi.2@cwts.leidenuniv.nl; Stefanie.haustein@umontreal.ca; tim.bowman@gmail.com

**Submission for presentation** at Metrics14 - ASIS&T Workshop on Informetric and Scientometric Research

## Introduction

Mendeley is a popular social reference manager with currently about 2.8 million users and 535 million user documents (Haustein & Larivière, 2014). Mendeley helps users to organize scientific documents in private libraries but at the same time publishes aggregated readership counts per documents to indicate their popularity among Mendeley users. The platform thus represents a rich source for readership counts of scientific documents.

Readership counts are accessible both on the Mendeley website via manual searches in the catalog and via an open API, which allows automatic retrieval based on queries using different metadata fields including unique identifiers (e.g., DOI, PMID, arXiv id) and bibliographic data (title, author, journal, etc.). Quite a number of studies have used both the online catalog (Bar-Ilan, 2012; Bar-Ilan, 2014; Haustein et. al, 2013) as well as the API (Haustein & Larivière, 2014; Haustein, Larivière, Thelwall, Amyot & Peters, in press; Costas, Zahedi & Wouters, 2014; Mohammadi & Thelwall, 2014; Zahedi, Costas,& Wouters, 2013 & 2014) to retrieve Mendeley reader counts as statistics about article usage; however, little is known about different retrieval strategies and how results are affected by data, such as the availability of DOIs and quality of metadata. Therefore, assessing data quality and accuracy in Mendeley is crucial before applying reader counts as impact metrics for research evaluation purposes.

The focus of this paper is to study the quality of the data extracted from the Mendeley catalog as well as the API and to test both the accuracy and completeness of data by using multiple retrieval strategies such as the DOI or paper and journal title and author names publication identifiers to identify differences and determine the best retrieval strategy. The following research questions are addressed:

1. How accurate is the metadata on Mendeley for a random sample of publications?
2. In how far do results differ between online catalog searches and API requests?
3. What are the most frequent error types in the bibliographic data for the sampled publications on Mendeley?
4. What retrieval strategy will provide the most accurate and complete results for the sampled publications?

## Methods

We used a random sample (384) representative of all 2012 publications (1,873,759) from Web of Science (WoS). For each of the sampled papers, title and DOIs were used separately to search both manually in the Mendeley online catalog and automatically via Mendeley API in July 2014. Manual search results were saved in Excel and API results stored in a My SQL database for further analysis. Bibliographic information (i.e., DOI, first author, title, journal, year, volume, issue, page, ISSN) of the publications retrieved from Mendeley was compared with WoS data to identify potential errors that might affect the retrieval of reader counts for papers in Mendeley.

## Preliminary results

Since this is a work-in-progress paper, we focus on the results of the manual search. Out of the 384 sampled publications, 182 (47.4%) were found in the Mendeley catalog and compared to WoS metadata to test the accuracy and completeness of bibliographic information in Mendeley and identify the most common types of errors.

*Data accuracy.* Figure 1 (left) displays the number and percentage of correct and incorrect metadata per field for the 182 publications found in the Mendeley catalog. The field with the highest share of incorrect, i.e. different from WoS, data is source (i.e. journal title) as 28% of publications with data differing from WoS, followed by page number

(20%), paper title (15%), ISSN (13%), journal issue (10%), DOI (8%), name of first author (7%), and journal volume (6%). Except for once, the publication year was always correct in the Mendeley data (1%).

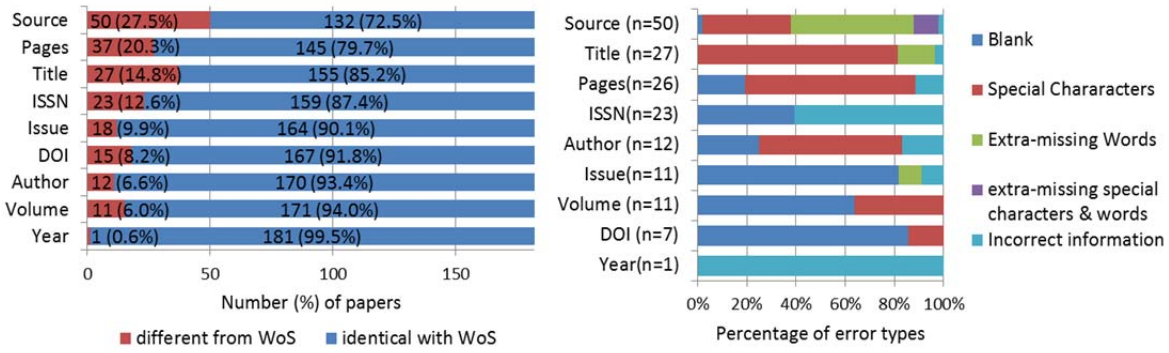


**Figure 1.** Number of papers with different and identical to WoS metadata (*left*) and percentage of types of errors in metadata (*right*)

*Error types.* In order to understand more about the types of errors in order to develop robust retrieval strategies, errors were classified into the following categories: blank (refers to publications without any data in the particular fields); special characters (fields with extra or missing special characters such as '*z*', ',', '.', '[]'); extra or missing words (fields with additional or missing words); extra or missing special characters and extra or missing words (publications with both extra or missing special characters and words); incorrect information (fields with wrong information). According to Figure 1 (*right*), the most common error types in Mendeley are special characters and blank fields, other error types varies across different fields. For example, in case of source, 96% of errors consist of additional or missing words (50%) and special characters (36%) and both (10%), while as few as 2% of journal titles were either incorrect or missing. Special characters are the most prevalent errors in article titles (81%), page numbers (69%), and author names (58%). DOI (86%), issue number (82%), volume (64%) and ISSN (39%) are the fields with the highest shares of missing information.

**Conclusion and outlook:**

We found that a significant number of metadata of papers in Mendeley differs from the bibliographic information provided in WoS. In particular, journal and article titles as well as page numbering are most often erroneous so that reader counts retrieved based on a direct match of these fields could be missed. DOIs, journal issues and volumes were most often missing, so that these fields should thus not be used as the only basis of retrieval strategies. In our future work we will compare manual with API search results and optimize automated data retrieval by limiting false negatives and false positives. For example, we will ignore special characters, as these was the most frequent error source in article titles, pages and author names.

**Acknowledgments**

This research visit has been partially funded by Leiden University Fund/ van Walsem and the Alfred P. Sloan Foundation.

**References**

Bar-Ilan, J. (2014): JASIST@Mendeley Revisited. altmetrics14: expanding impacts and metrics An ACM Web Science Conference 2014 Workshop, 23-26 June, Indiana University, Indiana, USA.
http://dx.doi.org/10.6084/m9.figshare.1031681

Bar-Ilan, J. (2012). JASIST@Mendeley. ACM Web Science Conference 2012 Workshop, *Evanston, IL, 21 June 2012.* *http://altmetrics.org/altmetrics12/bar-ilan/*

Costas, M., Zahedi, Z. & Wouters, P. (2014). Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology, DOI: 10.1002/asi.23309

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2013). Coverage and adoption of altmetrics sources in the bibliometric community. Scientometrics. DOI: 10.1007/s11192-013-1221-3

Haustein, S. & Larivière, S. (2014). Mendeley as a Source of Readership by Students and Postdocs? Evaluating Article Usage by Academic Status. *Proceedings of the IATUL Conferences*, Paper 2. http://docs.lib.purdue.edu/iatul/2014/altmetrics/2

Haustein, S., Larivière, V., Thelwall, M., Amyot, Didier & Peters, I. (in press). Tweets vs. Mendeley readers: How do these two social media metrics differ? *it – Information Technology*.

Mohammadi, M. & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, DOI: 10.1002/asi.23071

Zahedi, Z., Costas, R. & Wouters, P. (2013). How well developed are Altmetrics? Cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics, DOI:10.1007/s11192-014-1264-0

Zahedi, Z., Costas, R. & Wouters, P. (2014). Assessing the impact of the publications read by the different Mendeley users: Is there any different pattern among users? *Proceedings of the IATUL Conferences*, Paper 4. http://docs.lib.purdue.edu/iatul/2014/altmetrics/4