Removing duplicates in retrieval sets from electronic databases: comparing the efficiency and accuracy of the Bramer-method with other methods and software packages

Wichor M. Bramer (1), Leslie Holland (2), Jurgen Mollema (3), Todd Hannon (4), Tanja Bekhuis (5)

(1) Medical Library, Erasmus MC, University Medical Center, Rotterdam, The Netherlands
w.bramer@erasmusmc.nl
(2) Library, Southern College of Optometry, Memphis, TN, USA
(3) Library, Faculty of Health Care, HU University of Applied Sciences Utrecht, The Netherlands
(4) Library, Oregon Health & Science University, Portland, OR, USA
(5) Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, PA, USA

Introduction. Evidence-based medicine rests, in part, on exhaustive reviews of the research literature given topics of interest. This typically requires thoroughly searching several electronic databases. However, database content often overlaps and storage of bibliographic data differs. Information specialists and researchers who execute comprehensive searches face problems when removing duplicate records in their retrieval sets. Depending on the number of records and databases, this process can be quite challenging. Consider, the same research article may be represented in different ways depending on the database (e.g., journal titles and page numbers may be complete or abbreviated). In contrast, different articles may be represented as if equivalent in some fields (e.g., the title and author list may be the same, but the journal may differ). Therefore, removing duplicate records is quite time consuming and prone to errors. Many software packages for managing bibliographic references, such as EndNote, Reference Manager, Refworks, Mendeley, Papers, and Zotero, enable de-duplication, but how well do they perform?

Aim. In this research, we compare the performance of different software packages ‘out-of-the-box,’ i.e., with default settings. We consider how efficient is the process and how accurate the results? Can users easily change the settings of the various de-duplication algorithms? If so, how well can an EndNote algorithm with good performance characteristics be translated in other packages? Which software is the best choice for de-duplication and with which settings, and are improvements in software design needed or desirable?

Methods. We developed an algorithm in EndNote that speeds de-duplication with good precision by normalizing field contents when needed. We consider how efficient is the process and how accurate the results? Can users easily change the settings of the various de-duplication algorithms? If so, how well can an EndNote algorithm with good performance characteristics be translated in other packages? Which software is the best choice for de-duplication and with which settings, and are improvements in software design needed or desirable?

Conclusions. Preliminary results suggest that performance associated with default settings of popular bibliographic software packages varies considerably. Moreover, translation of an efficient and accurate EndNote algorithm is often difficult or even impossible. We offer suggestions regarding de-duplication for informationists and researchers based on our experience.