De-duplication of database search results for systematic reviews in EndNote

Wichor M. Bramer, Dean Giustini, Gerdien B. de Jonge, Leslie Holland, Tanja Bekhuis

See end of article for authors' affiliations.

DOI: http://dx.doi.org/10.3163/1536-5050.104.3.014

When conducting exhaustive searches for systematic reviews, information professionals search multiple databases with overlapping content [1–4]. They typically remove duplicate records to reduce the reviewers' workload associated with screening titles and abstracts; sometimes the reviewers remove the duplicates. Several articles have been published recently on de-duplication methods. In the authors' opinion, these methods are either very time consuming [5] or impractical, as they require uploading large files to an online platform [6, 7]. A recent overview article compared existing software programs but found that none was truly satisfactory [8].

Unique identifiers for journal articles are digital object identifiers (DOIs) and PubMed IDs (PMIDs). However, these identifiers are not present in every database. When they are present, they often cannot be exported easily. Thus, they cannot be relied upon to identify duplicates. An alternative involves using pagination, because the often large page numbers in scientific journals, in combination with other fields, can serve as a type of unique identifier. However, this is complicated by variations in the way page numbers are stored. Most biomedical databases use a long format (e.g., 1008–1012), but two important databases (MEDLINE and the Cochrane Library) use an abbreviated one (e.g., 1008–12).

BETTER WAY

The de-duplication method presented in this article was previously described in a brief conference paper [9]. The method consists of three stages.

1. Settings are changed for the displayed fields, and custom filters and export formats are installed.
2. Several databases are imported into a temporary library and exported in an adapted format before being imported for de-duplication.
3. Several subsequent changes in the settings for fields are used to detect duplicates, followed by removal of probable duplicates.

In this article, we describe this method in detail for EndNote [10], a popular reference manager.

Field settings and filters

Settings must be changed at the outset to optimize the EndNote configuration for de-duplication. Because page numbers play an important role, it is vital to show page numbers in the *library window*.

1. Go to Edit > Preferences > Display Fields.
2. Under Field, select Pages for one of the larger-numbered columns.
3. Click OK.

We customized an EndNote style to create export files in which abbreviated page numbers are expanded. We also made an import filter to import the modified files. These filters should be installed prior to de-duplication.

1. Go to http://bit.ly/emcendnote.
2. Open the zip file.
3. Double-click _Correct Pages.ens (the file will open in EndNote).
4. In EndNote, click File > Save As.
5. Remove the text "copy" from the file name and click save.
6. Close the style.
7. Repeat steps 3 through 6 for the file _Import Corrected Pages.enf.

Importing references

The exported references from PubMed [11], MEDLINE via Ovid [12], and the Cochrane Library [13] are modified to adapt the page number format of references to the format used in most other databases.

1. Import all references from PubMed, Ovid MEDLINE, and the Cochrane Library into an empty EndNote library.
2. Select all references in this library (Ctrl-A).

3. Go to File > Export.
4. Select the output style: _Correct Pages and save the resulting file with the extension ".txt."
5. Close the temporary library, and create a final EndNote library where records from databases are imported as usual.
6. When importing MEDLINE and Cochrane reference sets, choose the newly created file and use the import filter _Import Corrected Pages.

De-duplication

1. Go to Edit > Preferences > Duplicates, and select the fields to match the ones mentioned in row A under "Set field preferences" of Table 1 and click [OK].
2. Click on All References, and select one reference at random.
3. Go to References > Find Duplicates.
4. Click on [Cancel].
5. Follow the steps as described in row A under "Steps to remove duplicate" of Table 1.
6. Repeat the process again from step 1 onward for each row in Table 1.

DISCUSSION

Although the de-duplication method that we designed for EndNote resembles procedures regularly carried out by other information professionals, it is more systematic, rigorous, and reproducible. The steps may be somewhat challenging to master at first, but they become easy to carry out over time. The time spent de-duplicating references and the error rate are significantly reduced because just a small subset of the search results has to be assessed manually.

To enhance efficiency and accuracy, the steps described here should be followed closely in the order presented and without omission. The method's strength is based on the specificity of the first two steps, which require no manual assessment. The next three steps require checking a small subset, in other words, the references that lack page numbers. The last two steps require some additional manual assessment, but screening by page numbers expedites the work, and the number of references to assess is lower than in other methods.

The limitation of this method is that it is tuned to EndNote; however, EndNote is commonly used to manage bibliographic records. The only alternative software to EndNote in which the fields that are to be compared in the de-duplication process can be changed is Reference Manager, also provided by Thomson Reuters [14]. Reference Manager allows comparisons by start pages. However, when we tried our method in Reference Manager, it failed: too many false duplicates were removed. The comparison of start page numbers in Reference Manager appears to be flawed; therefore, tailoring by our method for Reference Manager is not a good alternative. Additionally, Reference Manager is no longer for sale, and support for Reference Manager will likely be discontinued. We hope that EndNote will adopt some of Reference Manager's useful features, such as the option to regulate the amount of overlap in the title and other fields, and a comparison on start pages, albeit more robust than in Reference Manager.

The method described in this paper is for the most recent version of EndNote for Windows, version X7. It will also work in earlier versions (versions X3 and higher); however, in older versions, step 2 in the third column of rows C and D in Table 1 will not work as desired because duplicate references are not highlighted. Before executing the steps in row C for older versions, go to All References and sort this group by Page Numbers. Next, instead of clicking on the column heading "Pages" as is described in step 2, go to All References and then go back to Duplicate References. Now, the Duplicate References group will be sorted by Page Numbers. Then, click on one of the scroll bars to reactivate the highlighting of duplicates and follow the other steps as described.

A requisite of this method is that for efficient de-duplication, page numbers should include both a start and an end page. This is the reason that we advise exporting the data from several databases into temporary files, which are then exported and reimported into the final EndNote library for de-duplication. Databases also differ in the format of the exported journal titles. Some databases use abbreviations, while others provide full journal titles. We use customized import filters for several databases and interfaces—including Embase.com, Web of Science, CINAHL, and Scopus—to import the abbreviated journal titles into EndNote. Although this is not strictly necessary, it improves the sensitivity of the first step and, thus, reduces the number of references that have to be checked manually. If databases would standardize their page numbers and journal titles, it would be possible to compare these data without the extra steps. To complicate matters, Cochrane recently switched from exporting full page numbers to

	Set field preferences	Steps to remove duplicates
A.	Author | Year | Title | Secondary Title (Journal)	Press <Delete> to remove all selected duplicates without manual assessment.
B.	Author | Year | Title | Pages	Press <Delete> to remove all selected duplicates without manual assessment.
C.	Title | Volume | Pages	1. Manually assess the top references with blank title or author fields, using Ctrl-Click to deselect false duplicates. 2. Click on the column heading "Pages" to sort all duplicate references by descending order of page numbers. 3. Review the top references without page numbers and those with page numbers, starting with number 1 for equivalent author names. If author names of subsequent references differ, deselect the marked false duplicates with <Ctrl-Click>. 4. Remove the selected duplicates with <Delete>.
D.	Author | Volume | Pages	1. Repeat steps 1–2 as described in row C. 2. Deselect the top references without page numbers by pressing <Ctrl-Click> on the first highlighted reference and <Ctrl-Shift-Click> on the first highlighted reference with a starting page number greater than 1. Remove the remaining selected duplicates with <Delete>.
E.	Year | Volume | Issue | Pages	1. Right click on My Groups > Create Group and press <Enter>. 2. In the group Duplicate References, click on the column heading "Pages" to sort all duplicate references by descending order of page numbers. 3. Select all references with page numbers by clicking on the top reference, holding <Shift>, and then clicking on the last reference with page numbers. 4. Drag the selected references to the just created temporary "New Group." 5. Click on "New Group." Check the group for references with just one page and page numbers starting with 1 or with a letter. Select false duplicates from those references, and press <Delete> to remove them from the group. (They remain in All References but are not de-duplicated in this step.) 6. Select one of the references in "New Group," click References > Find Duplicates, click Cancel, and press <Delete> to remove all selected duplicates.
F.	Title	1. Compare page numbers of consecutive references. If page numbers are present and different, examine journal titles and authors. Deselect false duplicates with <Ctrl-Click>. References with blank pages or pages starting with the number 1 are usually true duplicates, but check journal titles and author names when in doubt, especially when multiple consecutive blank pages are selected. 2. After checking the entire list, remove the remaining selected duplicate references with <Delete>.
G.	Author | Year	If a true duplicate is found, deselect all references by clicking the first true duplicate reference without holding <Ctrl>. Compare subsequent references on page numbers: if two adjacent references have the same page numbers, select the one with the largest record number with <Ctrl-Click>. After checking the complete list, remove the remaining selected references with <Delete>.

Table 1

De-duplication field settings and removal of duplicates

abbreviated page numbers, and CINAHL recently appended the length of the article in the fields for page numbers (e.g., in the format 1008–1012 5p).

This de-duplication method is rather complicated, and the learning curve is steep. However, simplification of the method (e.g., by reducing the number of different field combinations or by not normalizing page numbers) increases the workload by increasing the number of references to manually assess. If performed frequently, librarian-mediated de-duplication services can be faster than current methods and less error prone.

REFERENCES

1. Levay P, Raynor M, Tuvey D. The contributions of MEDLINE, other bibliographic databases and various search techniques to NICE public health guidance. Evid Based Libr Inf Pract. 2015;10(1):50–68.

2. Ahmadi M, Sarabi RE, Orak RJ, Bahaadinbeigy K. Information retrieval in telemedicine: a comparative study on bibliographic databases. Acta Inform Med. 2015; 23(3):172–6.

3. Lorenzetti DL, Topfer LA, Dennett L, Clement F. Value of databases other than MEDLINE for rapid health technology assessments. Int J Technol Assessment Health Care. 2014;30(02):173–8.

4. Beyer FR, Wright K. Can we prioritise which databases to search? a case study using a systematic review of frozen shoulder management. Health Inf Libr J. 2013; 30(1):49–58.

5. Qi X, Yang M, Ren W, Jia J, Wang J, Han G, Fan D. Find duplicates among the PubMed, EMBASE, and Cochrane Library Databases in systematic review. PLOS One. 2013; 8(8):e71838.

6. Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of systematic review assistant-deduplication module. Systematic Rev. 2015;4(1):6.
7. Jiang Y, Lin C, Meng W, Yu C, Cohen AM, Smalheiser NR. Rule-based deduplication of article records from bibliographic databases. Database: J Biological Databases Curation. 2014;2014:bat086.
8. Kwon Y, Lemieux M, McTavish J, Wathen N. Identifying and removing duplicate records from systematic review searches. J Med Libr Assoc. 2015 Oct; 103(4):184–8. DOI: http://dx.doi.org/10.3163/1536-5050.103.4.004.
9. Bramer WM, Holland L, Mollema J, Hannon T, Bekhuis T. Removing duplicates in retrieval sets from electronic databases: comparing the efficiency and accuracy of the Bramer-method with other methods and software packages. Proceedings of the 14th European Association for Health Information and Libraries (EAHIL) Conference, 57-9; Rome, Italy; 11–13 Jun 2014.
10. EndNote. Version X7. Thomson Reuters; 2015.
11. National Library of Medicine. PubMed [Internet]. The Library; 2016 [cited 4 Feb 2016]. <http://www.ncbi.nlm.nih.gov/pubmed/>.
12. Ovid MEDLINE [Internet]. Wolters Kluwer; 2016 [cited 4 Feb 2016]. <http://site.ovid.com/pdf/OvidMedline_fs.pdf>.
13. Cochrane. Cochrane Library [Internet]. John Wiley & Sons; 2016 [cited 4 Feb 2016]. <http://www.cochranelibrary.com>.
14. Reference Manager. Version 12. Thomson Reuters; 2008.

AUTHORS' AFFILIATIONS

Wichor M. Bramer, w.bramer@erasmusmc.nl, Biomedical information Specialist, Medical Library, Erasmus MC–Erasmus University Medical Centre, Rotterdam, The Netherlands; **Dean Giustini,** dean.giustini@ubc.ca, UBC Biomedical Branch Library, University of British Columbia, Vancouver, Canada; **Gerdien B. de Jonge,** g.dejonge@erasmusmc.nl, Biomedical information Specialist, Medical Library, Erasmus MC–Erasmus University Medical Centre, Rotterdam, The Netherlands; **Leslie Holland,** lholland@sco.edu, Manager, Library, Southern College of Optometry, Memphis, TN, USA; **Tanja Bekhuis,** tcb24@pitt.edu, Assistant Professor and Director, Department of Biomedical Informatics and Department of Dental Public Health, School of Medicine and School of Dental Medicine, University of Pittsburgh, PA, USA

Received February 2016; accepted February 2016