Station 2: Deduplication

Elizabeth Moreton

Health Sciences Library
Clinical, Academic, & Research Engagement (CARE)
Why are duplicates missed?
<table>
<thead>
<tr>
<th>Problem</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple parts of last names (especially last names that begin with Al, De, De La, Du, Dos, Von, Van, etc.)</td>
<td>Souza, B. M. L. vs. Lima de Souza, Bruna Marcela</td>
</tr>
<tr>
<td>Last &amp; first name switched</td>
<td>Huang, Zhe vs. Zhe, H.</td>
</tr>
<tr>
<td>Missing authors</td>
<td>Barrett M, Ymer B, Moesbergen AM vs. Barrett M, Ymer B</td>
</tr>
<tr>
<td>Special characters</td>
<td>Välimäki</td>
</tr>
<tr>
<td>Square brackets</td>
<td>[&quot;Scientific evidence not considered&quot;]</td>
</tr>
<tr>
<td>Misplaced fields</td>
<td>Booth, A. %J Research synthesis methods</td>
</tr>
<tr>
<td>Journal column names</td>
<td>Research reflections. Promoting excellence...</td>
</tr>
</tbody>
</table>
Previous approaches:

- Manual
- Excel
- Citation Manager/ SR Screening tool
- Some sort of mix

New approach:

Use machine learning to predict likely duplicates
Super DeDuper

Phase 1: EndNote Enhanced (Maximum Precision)

Phase 2: Machine Learning (Maximum Recall)
Phase 1: Maximum Precision

**Input:**
RIS files – separated in priority order

**Two new fields added:**
Starting page # & First author’s last name

**Dupes removed:**
Match title, year, 1st author last name, starting page

**Output:**
Two RIS files: Keeps & Dupes (no checking needed)
Phase 2: Maximum Recall

Input:
‘Keeps’ from Phase 1

Dupes removed:
Uses machine learning to predict likely duplicates

Cluster ID added:
Matches groups of duplicates to one cluster ID #

Output:
Two RIS files: Unique Records & To Review
Next Steps…

• Consider what you have available at your workplace…
  • Do you have something that works well for you?
  • Can your process be made more efficient or more effective with small changes?

• Visit Station 3