UK DATA ARCHIVE KEYWORD INDEXING
WITH A SKOS VERSION OF HASSET THESAURUS

PURPOSE AND MOTIVATION
Apply automatic indexing tool, KEA, to some of the UK Data Archive’s document collection using HASSET thesaurus with aims to:
• see whether KEA could potentially be used to aid metadata creation.
• develop recommendation for the future use of automatic indexing with an existing thesaurus.

DATA COLLECTION
<table>
<thead>
<tr>
<th>Corpus Name</th>
<th>Whole Corpus</th>
<th>Training Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reactive bank of variables/ questions</td>
<td>26,634</td>
<td>21,307</td>
</tr>
<tr>
<td>Survey Question Bank (SQB)</td>
<td>1,353</td>
<td>1,082</td>
</tr>
<tr>
<td>EDS partial data catalogue records</td>
<td>5,610</td>
<td>4,488</td>
</tr>
<tr>
<td>Case Studies / Support guides</td>
<td>243</td>
<td>194</td>
</tr>
</tbody>
</table>

KEA (KEYWORD EXTRACTION ALGORITHM)
• an algorithm for extracting keywords from text documents
• calculates feature values for each candidate (TF.IDF, First Occurrence, Length)
• uses a machine-learning algorithm to predict which candidates are good keywords.

INDEXING PROCESS
1. Get PDFs
   - Extract Metadata (Manual-Keywords)
   - Convert PDFs to Text

2. Prepare In/Out (.txt/.key)
   - Apply KEA
   - Extract Auto Keywords

3. Automatic Evaluation
   - Manual Evaluation (Experts)

KEYWORDS EXTRACTION
Training data + Controlled vocabulary + Testing data = Automatic keywords

Manual keywords

CONCLUSIONS AND RECOMMENDATIONS
• KEA is a useful tool for indexers of full text social science materials
• KEA would work best as a suggester of new terms, with moderation from a human indexer
• KEA could also be used as a quality assurance tool, to ensure that terms are not overlooked
• more work is needed to investigate KEA further and to see how it could be incorporated technically, and in terms of process, into ingest systems.

RESULTS
Individually, the best performance overall was seen in the SQB corpus. Catalogue records’ low F1 score was to be expected, given that KEA had relatively little text to index from, compared to the manual indexers who indexed from the full catalogue records.

<table>
<thead>
<tr>
<th>Corpus Name</th>
<th>Auto F1</th>
<th>Strict F1</th>
<th>Broad F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reactive bank of variables/ questions</td>
<td>0.12</td>
<td>0.14</td>
<td>0.34</td>
</tr>
<tr>
<td>Survey Question Bank (SQB)</td>
<td>0.14</td>
<td>0.33</td>
<td>0.43</td>
</tr>
<tr>
<td>EDS partial data catalogue records</td>
<td>0.11</td>
<td>0.19</td>
<td>0.21</td>
</tr>
<tr>
<td>Case Studies / Support guides</td>
<td>0.06</td>
<td>0.27</td>
<td>0.36</td>
</tr>
</tbody>
</table>

As expected there was relatively little overlap between KEA keywords and manual keywords:
• on average KEA extracted 18.60 keywords per document across the four corpora.
• only 2.33 were exact matches with the manual keywords.
• a high percentage of KEA keywords were considered relevant/suitable even if they were not exact matches - 33 per cent for the SQB corpus, with an average of 25 per cent across all four corpora.
This suggests that KEA could be a very useful tool for indexers. The average number of manual keywords varies from 1.63 for Nesstar to 62.86 for catalogue records.