OPTICAL CHARACTER RECOGNITION (OCR) IN LINKING ENTOMOLOGICAL LABELS WITH FIELD NOTEBOOK DATA

Tero Mononen, Riitta Tegelberg, Janne Karppinen, Mira Sääskilahti, Hannu Saarenmaa – Digitarium, University Of Eastern Finland

Tommi Koskinen, Jyrki Muona – Finnish Museum Of Natural History
FIELD NOTEBOOKS

• Field notebooks have been used for recording specimen data: taxonomic name, date, locality, host plant, method of collection...

• Labels of insect specimens are small and contain very basic information – especially during the times of ink pens
NOTEBOOK LABEL

• In pin, specific label with referring notebook number
• Number: variance in font style and size, colours, lines, colours of lines, sub- and superscripts…
• Label: variance in colour
• Differences signify specific year, area…
DIGITISATION OF ENTOMOLOGICAL NOTEBOOKS

• Around 400 entomological notebooks are archived at the Finnish Museum of Natural History (Luomus)

• Notebooks were digitised by Luomus during project “Digitisation of entomological notebooks”

• http://digit.luomus.fi/

• Workflow:
  • Imaging using cameras
  • Cataloguing of notebook information
  • Entering the text content into a text field in Drupal
  • Proofreading
  • Structured data entry (Excel, ABCD schema)
  • XML conversion, SQL database
# Digital Notebooks

<table>
<thead>
<tr>
<th>Technical Contact</th>
<th>Content Contact</th>
<th>Metadata Language</th>
<th>Title</th>
<th>Taxonomic Term</th>
<th>Informal Name String</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koskinen, Tommi</td>
<td>Broman, Ann-Charlotte</td>
<td>fi</td>
<td>Broman, Gunner (1898-1970)</td>
<td>Coleoptera</td>
<td>Carabus</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Laho, Juha</td>
<td>fi</td>
<td>Broman, Gunner (1898-1970)</td>
<td>Coleoptera</td>
<td>Conus</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Seppäläinen, Tia</td>
<td>fi</td>
<td>Broman, Gunner (1898-1970)</td>
<td>Coleoptera</td>
<td>Hybocatus</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Laho, Juha</td>
<td>fi</td>
<td>Broman, Gunner (1898-1970)</td>
<td>Coleoptera</td>
<td>Corticaria</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Laho, Juha</td>
<td>fi</td>
<td>Broman, Gunner</td>
<td>Coleoptera</td>
<td>Cercyon</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Broman, Ann-Charlotte</td>
<td>fi</td>
<td>Broman, Gunner</td>
<td>Coleoptera</td>
<td>Acollis</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Seppäläinen, Tia</td>
<td>fi</td>
<td>Broman, Gunner</td>
<td>Coleoptera</td>
<td>Grynochera</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Seppäläinen, Tia</td>
<td>fi</td>
<td>Broman, Gunner (1898-1970)</td>
<td>Coleoptera</td>
<td>Dendrophagus</td>
</tr>
<tr>
<td>Koskinen, Tommi</td>
<td>Seppäläinen, Tia</td>
<td>fi</td>
<td>Broman, Gunner</td>
<td>Coleoptera</td>
<td>Zyas</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ISODateTimeBegin</th>
<th>ISODateTimeEnd</th>
<th>Method</th>
<th>DatasetLocalityText</th>
<th>LocalityText</th>
<th>DatasetCountry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1936-08-21</td>
<td></td>
<td></td>
<td></td>
<td>Jno, Haukkial.</td>
<td>Suomi</td>
</tr>
<tr>
<td>1937-10-10</td>
<td></td>
<td></td>
<td></td>
<td>Jno, v-strand</td>
<td>Suomi</td>
</tr>
<tr>
<td>1943-06-24</td>
<td></td>
<td></td>
<td></td>
<td>Pub, Jno</td>
<td>Suomi</td>
</tr>
<tr>
<td>1944-06-02</td>
<td></td>
<td></td>
<td></td>
<td>Suorkansa, Jno</td>
<td>Suomi</td>
</tr>
<tr>
<td>1946-03-17</td>
<td></td>
<td></td>
<td></td>
<td>Pub, Jno</td>
<td>Suomi</td>
</tr>
<tr>
<td>1947-06-19</td>
<td></td>
<td></td>
<td></td>
<td>runsala</td>
<td>Suomi</td>
</tr>
<tr>
<td>1938-11-06</td>
<td></td>
<td></td>
<td></td>
<td>Tronnable, Vimenoble, Tenaia</td>
<td>Finland</td>
</tr>
<tr>
<td>1939-09-04</td>
<td></td>
<td></td>
<td></td>
<td>Vano, Goteborg, Hette</td>
<td>Suomi</td>
</tr>
<tr>
<td>1951-04-09</td>
<td></td>
<td></td>
<td></td>
<td>Pub, På fabrica karteret (Kvarter?) Jno</td>
<td>Finland</td>
</tr>
<tr>
<td>1952-04-12</td>
<td></td>
<td></td>
<td></td>
<td>Jno</td>
<td>Suomi</td>
</tr>
<tr>
<td>1953-04-04</td>
<td></td>
<td></td>
<td></td>
<td>Tronnable, Vimenoble, Tenaia</td>
<td>Finland</td>
</tr>
<tr>
<td>1954-04-04</td>
<td></td>
<td></td>
<td></td>
<td>Malaykonna, Jno</td>
<td>Suomi</td>
</tr>
</tbody>
</table>
DIGITISATION OF COLLECTION BLOMQVIST

• Amateur entomologist Gunnar Blomqvist collected during 1930-1960s around 14,000 Coleoptera specimens, mostly from Finland and representing over 2200 species

• Collection was digitised by Digitarium, using automated imaging line designed for insects

• http://digitarium.fi/en/content/mass-digitisation-pinned-insects
DIGITISATION OF COLLECTION BLOMQVIST

- Individual insects and labels were imaged
- XML Metadata: collector’s name, taxon, (date)
CAN DATA FROM NOTEBOOKS BE COMBINED WITH LABEL DATA?

• Blomqvist notebooks have been digitised by Luomus.

• Blomqvist was a tempting case to testing optical character recognition (OCR): you’ll need from label images.
  • Notebook number
  • Year (because collector started from number 1 every year...and didn’t use any colours or other markings to signify the year)
OCR?

• It became clear very soon that year, handwritten by Blomqvist, was difficult to read with OCR

• Sometimes numbers were hard to read by us (5 and 7 most difficult)
OCR OF NOTEBOOK NUMBER, METHOD 1

• n = 100 images
• From image, the area of notebook number was defined
• Area was cropped and Tesseract-program was used for character recognition
• Threshold value 40 % was used
• Results: correctly read 14, wrong 29, no recognised number 57
OCR OF NOTEBOOK NUMBER, METHOD 2

• $n = 100$ images

• Generation of several (40, 20+20) images from a cropped image. Turning of 1° step-by-step to both directions

• Tesseract was used for 41 images -> one result

• Results: correctly read 68, wrong 27, not recognised 5
OCR OF NOTEBOOK NUMBER, METHOD 3

- n = 100 images
- From images used by methods 1 & 2, pin was cropped out
- Contrast was increased
- Image was blurred and then the borders of characters were sharpened
- Generation of several (30, 15+15) images from a cropped image. Turning of 2° step-by-step to both directions
- Threshold value of 40, 45 and 50 % were used
- 93 images / image
- Results: correctly read 66, wrong 17, not recognised 17
OCR OF NOTEBOOK NUMBER, METHOD 4

• n = 100

• From the 1 = 93 images (method 3)

• Filtering away character strings that did not represent at least 40% of the character strings recognised

• Some exceptions to the rule, sensors to false false

• Results: correctly read 88, wrong 3, not recognised 9
  • Wrongs: 09 (was 109); label missing; 3 (was 28)
  • Not recognised: no obvious winner among character strings
LINKING OCR-NUMBERS WITH NOTEBOOK DATA

• n = 88 (result from method 4)

• Transcription of year from images – by hand

• Search (notebook number, year) from database of the digitised books

• Results: 87 could be combined with notebook data. 1 was missing (notebook not digitised)

• Conclusion: OCR can be used in linking typed label number with notebook data