To harmonize Quality and Quantity
The collection

130 million pages

70 million pages on microfilm

Growth: 3 million pages/year
The collection

• Publishers deliver three legal deposit copies of all Swedish newspapers printed on monthly basis.

• Duplicate collection
Project Digidaily 1 + 2

- Collaborative project between the Swedish National Library (KB) and the Swedish National Archives in Fränsta (MKC).

- Develop rational and effective methods for mass digitization of historical newspapers.

- Digidaily 1 – Historical material. Approx. 3 million digitised pages

- Digidaily 2 – Modern material and redirection from microfilming to digitising. Approx. 2,4 million pages
Preparation at KB

- 298 newspapers
  Incl. editions, supplements, newsbills
- Register in data base
- REST API = Pre-made posts at MKC
Preperation at MKC

• 650-6000 pages/shift
• Take apart, cut
• Compare editions
• Registration
• Workflow system – registration and metadata holder
Scanning at MKC
Scanning at MKC
Scanning at MKC
Quality

• Technical image quality - resolution, colour depth, levels of tolerance, file format, compression level

• Metadata quality

• Ocular image quality - Artefacts – unwanted effects

• OCR-quality
Quality Work

- **Logg** all quality parameters in a xml-file – the Performance file
  Noise, Illumination Uniformity, colour etc

- **Automated** methods to find quality defects
Quality problems - stripes
Quality problems - bad stitching
Quality problems - wavy text
OCR

• Automated process. No manual correction!

• Semi automated article segmentation

• Alert if confidence figure is less than 90
OCR - headlines
Quality control

- Image quality: manual quality control based on statistical sampling model
- Validated METS- and ALTO files
- Validated Performance file in xml
The elements of image costs
## Costs

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>PRICE</th>
<th>COMMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bound, torn, fragile paper, the biggest format size</td>
<td>6.49 SEK 0.78 € 0.67 £ 1.01 USD</td>
<td>About 4% of the collection</td>
</tr>
<tr>
<td>2. Bound, most of it can be taken apart, only a few are saved still bounded, fair paper quality.</td>
<td>3.23 SEK 0.39 € 0.33 £ 0.51 USD</td>
<td>Main part of the collection, 49%</td>
</tr>
<tr>
<td>3. Tabloids stapled but not bound.</td>
<td>2.17 SEK 0.26 € 0.22 £ 0.34 USD</td>
<td>About 47% of the collection</td>
</tr>
<tr>
<td>4. National copies – not to be taken apart. Or very poor paper quality.</td>
<td>Up to 20,00 SEK 2.39 € 2 £ 3.1 USD</td>
<td>Less than 1% of the collection</td>
</tr>
</tbody>
</table>
Quality
Quantity
= true
Contact

For more information please contact:

Heidi Rosen

heidi.rosen@kb.se

Photos:
Heidi Rosen
Jens Gustafsson