Preserving the Spirit of the Epoch: Digital Conversion of Nordic Music Magazines

Amalie Ørum Hansen
Development Consultant
Gentofte Centralbibliotek

Sergey Borovoy
CEO
ATAPY Software
Quick facts about Gentofte Central Library

Gentofte Central Library (Gentofte Centralbibliotek)

- One of six Central libraries in Denmark
- Responsibility (services) area: the Capital Region of Denmark (28 municipalities and their public libraries)
- Mission and task:
  - providing support to public libraries on buying and obtaining materials
  - providing consultancy services (technical & developmental)
  - servicing all the employees of the region’s public libraries with relevant education (i.e. methodological and skills)
Quick facts about ATAPY Software

- Offices: Russia (Novosibirsk), Germany (Munich)
- Main focus areas:
  - Digitisation bureau
  - On-demand software development services in the fields of OCR & document imaging
- Completed projects in the field: 100+, track record including work for the Royal Danish Library, the National Library of Sweden, Springer, contributions to the METAe and IMPACT projects, etc.
- Strategic partnership with ABBYY and Microsoft
- Ability to handle challenging sources that yield poorly to OCR technology (due to the technology edge)
<table>
<thead>
<tr>
<th>Reaction</th>
<th>Radical generation</th>
<th>Solvent</th>
<th>T[K]</th>
<th>Rate data</th>
<th>Ref/ add. ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N\textsuperscript{5}+</td>
<td>+</td>
<td>NiC\textsubscript{6}H\textsubscript{5}CH\textsubscript{2}OH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NiC\textsubscript{6}H\textsubscript{5}CH\textsubscript{2}OH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pulse rad. reduct. with c\textsuperscript{5} or (CH\textsubscript{3})\textsubscript{3}COH</td>
<td>KAS</td>
<td>H\textsubscript{2}O, pH=7</td>
<td>295</td>
<td>k\textsubscript{1}=6.4 \times 10\textsuperscript{-3} s\textsuperscript{-3}</td>
<td>81Silt</td>
</tr>
<tr>
<td>5(MM phosphat)</td>
<td>+0.1 M 2-propanol</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10.1.7 Bipyrindinium dication + 2pipyridinium dication

Bipy\textsuperscript{2+} + Bipy\textsuperscript{2+} \rightleftharpoons Bipy\textsuperscript{2+} + Bipy\textsuperscript{2+}

<table>
<thead>
<tr>
<th>R-C</th>
<th>R-E</th>
<th>R-O \textsuperscript{2+}</th>
</tr>
</thead>
<tbody>
<tr>
<td>H\textsubscript{2}C</td>
<td>CH\textsubscript{3}</td>
<td>R=H</td>
</tr>
<tr>
<td>\textsuperscript{2}N\textsuperscript{5}-</td>
<td>\textsuperscript{2}N\textsuperscript{5}-</td>
<td></td>
</tr>
<tr>
<td>Pulse rad. reduct. with c\textsuperscript{5} or (CH\textsubscript{3})\textsubscript{3}COH</td>
<td>KAS</td>
<td>H\textsubscript{2}O, pH=7</td>
</tr>
<tr>
<td>5(MM phosphat)</td>
<td>+0.2 M 2-propanol</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{1} For M= Me, p. 165.
\textsuperscript{2} Estimated from a Marcus treatment of rate data for electron exchange with other 2,2'-bipyridinium compounds.
Newspapers and magazines (historic & contemporary)
Fiction of various nature: novel, poetry, theatrical plays and travelogues
Digital Conversion of Nordic Music Magazines

Project duration: 2007-2008

Three popular music journals issued in Denmark in the second half of the XX century:

- *Nordic Sounds* - the magazine of NOMUS, the NORDIC MUSIC COMMITTEE published 1982-2006
- *GAFFA* - a free Danish magazine, published 1983-present
- *MM* - a Danish magazine devoted to Jazz and Rock issued 1968-1989
Digital Conversion of Nordic Music Magazines: Contributors

- **Funded by** the Danish Agency for Culture, Libraries
- **Gaffa A/S, Gaffa Nordic**
  CEO Robert Borges
- **Musikbibliotek.dk, Gentofte Centralbibliotek**
  Former Editor-in-Chief: Amalie Ørum Hansen (now Bibzoom.dk, State and University Library)
  Editor-in-Chief: Niels Mark Pedersen
Digital Conversion of Nordic Music Magazines: Contributors

- Det Virtuelle Musikbibliotek, State and University Library
  Former Editor: Søren Svane Hansen

- The Royal Library, National Library of Denmark and Copenhagen University Library, Dept. of Documentation & Digitisation

- ATAPY Software (digitisation contractor)
  CEO Sergey Borovoy
Digital Conversion of Nordic Music Magazines

Project goals:

- Preserving full collections of the three magazines in digital form
- Preserving and conveying the cultural “flavor” of the late XXth century in the Nordic countries, including
  - information about local music acts
  - the attitude of Nordic music community to worldwide sensations
  - snapshots of associated subculture movements, etc.
- Making this important knowledge available online
Digital Conversion of Nordic Music Magazines

Overall project volume: over 16,450 pages
Digitization workforce: 4 to 8 operators
Timeframe: about 9 months
Digital Conversion of Nordic Music Magazines

Project requirements:

- Full-text recognition
- High recognition accuracy (to ensure excellent searchability of the collection)
- Excluding part of material from recognition (commercials, etc.)

Results: industry-standard XML format

- In some parts of scope – illustrations were extracted & saved in a separate location, hyperlink placed in the XML file
Digital Conversion of Nordic Music Magazines

Complications:

- The specifics of periodicals as input material:
  - Wide format and multi-column layout (challenge for automatic segmentation)
    **Solution:** manual after-correction of automatic segmentation
  - Colored and textured backgrounds (challenge for OCR)
    **Solution:** semi-automated image preprocessing in graphic packages (increasing contrast, etc.)
  - A variety of fonts and font colors within one page
  - Designer fonts used
  - Skewed fragments + fragments with normal text orientation present in one page
    **Solution:** KFI of poorly recognized or unrecognized occurrences
  - Inverted and «normal» text present on one page
    **Solution:** multi-attempt OCR with varied settings (inverted/normal text)
Digital Conversion of Nordic Music Magazines
Digital Conversion of Nordic Music Magazines

Complications:

- **Format specifics:**
  - Input material: TIFF, PDF - the latter sometimes containing a text layer, which was sometimes partially unrecognizable (in vector format)
  
  **Solution:** OCR of PDF as image-only (sometimes), or KFI of unrecognized information

  - Output requirements: XML format, different from that produced automatically by ABBYY FineReader
  
  **Solution:** export to Microsoft Word, semi-automatic XML markup in required format

- **Language specifics:**
  - Several languages (Danish & English) on one page (additional challenge for OCR even with correct dictionaries enabled)
  
  - Critical information in Danish
  
  **Solution:** operators with linguistic background, manual verification
Digital Conversion of Nordic Music Magazines

XML format requirements:
- One XML file per article (including multi-page articles)
- Metadata block: magazine name, issue (e.g. 2007-12), article number on page
- Article level tags: type of article (interview, musical review, news piece, etc.), author, abstract
- Formatting tags: title, subtitle, text type (bold, italic, bold italic), etc.
- Special tags (occurrences of proper and geographical names, etc.)
Digital Conversion of Nordic Music Magazines

Process phases (scope of work):
1. Analysis/segmentation of pages (automatic)
2. Segmentation cross-check & correction (manual)
3. OCR (automatic)
4. Verification/correction or KFI if unrecognized (manual)
5. Export to Microsoft Word (automatic)
6. XML markup (manual, semi-automated)
7. XML file aggregation (merging several files related to one article)* (manual)
8. XML validation (automatic)

* Optional phase (is case of issues with automatic article segmentation)
Digital Conversion of Nordic Music Magazines

Tools and technologies used:

**ABBYY FineReader 8.0** (the latest ABBYY FineReader version at the time):

- Segmentation
- OCR
- Verification (manual, in ABBYY FineReader interface)
- Export to Microsoft Word

**Third-party XML validation software:**

- XML file validation

**In-house-developed macros (VB):**

- XML markup of Microsoft Word files
Digital Conversion of Nordic Music Magazines

Results:

All project requirements fulfilled; results accepted and published

Nordic Sounds and MM magazines: available at the website of Online Music Research Library (www.dvm.nu/periodical)
GAFFA magazine: available online at http://gaffa.dk/arkiv
Thank you!

Amalie Ørum Hansen
Development Consultant
Gentofte Centralbibliotek
ahan@Gentofte.dk
Tel.: +45 39 98 58 47

Sergey Borovoy
CEO
ATAPY Software
sergeyb@atapy.com
Tel.: +7 383 363 96 99