Submitted on: 8/16/2014

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

Abstract:

The presentation describes a cultural preservation project that involved digital conversion of backlogs of three popular music journals issued in Denmark in the second half of the XXth 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*

The project (2007-2008), was funded by the Danish Agency for Culture, Libraries. The following other organizations took part in the project:

- *Gaffa A/S, Gaffa Nordic, CEO Robert Borges*
- *Musikbibliotek.dk, Gentofte Centralbibliotek, former Editor-in-Chief Amalie Ørum Hansen 2008-2010 (now Bibzoom.dk, State and University Library, Editor-in-Chief Niels Mark Pedersen); the idea was initiated by former Editor-in-chief Susanne Kier in 2007.*
- *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, CEO Sergey Borovoy*

The goal of the project was to digitize full collections of all three magazines, believed to be valuable documents of their epoch that convey the cultural "flavor" of the 80-s and 90-s in the Nordic countries; including information about local music acts, the attitude of Nordic music community to worldwide sensations, and snapshots of associated subculture movements. This important knowledge was to become available to people online.

The specifics of the material (from the digitization point of view) included:

- *Wide format and multi-column layout*
- *Colored and textured backgrounds*
- *A great variety of fonts and font colors within one page; designer fonts used*
- *Multilingual content*

The project volume was over 16 000 pages in total. To ensure excellent searchability of the collection, high recognition accuracy of key materials was a requirement.

Keywords: digitisation, music magazines, cultural preservation.

The project goals

The musical periodical digitisation project (2007-2008), was funded by the Danish Agency for Culture, Libraries. The following other organizations took part in the project:

- Gaffa A/S, Gaffa Nordic, CEO Robert Borges
- Musikbibliotek.dk, Gentofte Centralbibliotek, former Editor-in-Chief Amalie Ørum Hansen 2008-2010 (now Bibzoom.dk, State and University Library, Editor-in-Chief Niels Mark Pedersen); the idea was initiated by former Editor-in-chief Susanne Kier in 2007.
- 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, CEO Sergey Borovoy

Within the scope of the project, a series of musical periodicals were to be made available online; two of them – on the site of Det Virtuelle Musikbibliotek (namely, MM and Nordic Sounds) - the national Danish project aimed at creating "a dynamic and permanent central web portal to music resources in public institutions in Denmark" (http://dvm.nu/). The third magazine in focus was GAFFA – a free music magazine, presently all-Nordic, published 1983-present, and also known to the community for its online concert overview (http://gaffa.dk/live) and its user-written music encyclopedia, GAFFApedia (a project closed in 2010). GAFFA was to be available on GAFFA own website.

The criteria of magazine selection were closely aligned with the project goals, which were:

- To preserve and convey 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.
- To make this important knowledge available online to researchers and wide public, free of charge

Apart from "born-digital" issues of the magazines (those of the newer times), there existed the archives of old issues, available only in paper form. To make all three periodical archives available online, these backlog issues of all three magazines were to be digitized as well.

At the time the project started, it was clear for the cultural preservation community that digitization was no longer just about scanning and publishing page images; it has become text oriented, ensuring that the valuable data obtained through a digitization project would be searchable, and not only by metadata, but also by keyword on the text level. And surely, there was the typical challenge of most digitization projects: keeping the optimal balance between cost-efficiency and digitization quality.

In response to this challenge the Library developed a digitization workflow in which the most costly phase of the job (text recognition and proofreading) was outsourced to a third party, the process involving project coordination and quality control on the Library side. Other phases (carried out on the Library side) were: scanning, preparation (dividing into

batches), quality control, and online publishing. This approach allowed to effectively digitize more than 16 000 pages of magazine backlogs and to achieve the project goals.

The material challenges:

First of all, we faced the specifics of periodicals as input material:

- Wide format and multi-column layout. This was truly challenging for automatic segmentation by most OCR packages available at the time
- Colored and textured backgrounds
- A variety of fonts and font colors within one page
- Designer fonts used
- Skewed fragments + fragments with normal text orientation present in one page
- Inverted and «normal» text present on one page

These were challenges for both automatic segmentation and OCR of the text

The input and output formats also presented certain specifics:

- Input material: TIFF, PDF. The PDF files often contained a text layer, which was sometimes partially unrecognizable (appeared in vector format).
- Output requirements: the XML output was aligned with industry standards for digital libraries. ABBYY FineReader did not natively support the required XML output format; therefore additional format conversion was required.

There were also the language specifics:

- Several languages (Danish & English) on one page (additional challenge for OCR even with correct dictionaries enabled)
- Critical information was often present in Danish (Russia-based employees might find this challenging).

The scope of work

On the Library side, the workload was prepared: scanned, organized in batches and sent over to ATAPY by FTP.

The digitization process included the following phases**:**

1. Analysis/segmentation of pages (automatic)

This was done at the side of ATAPY by means of ABBYY FineReader OCR package by ABBYY Software. Due to the challenges presented by the material (discussed above), this phase did not always produce excellent results. In many cases, manual correction of segmentation was required*.

* Having segmented the page incorrectly, the software would produce poor recognition accuracy, as the results depend largely on the segmentation accuracy.

2. Segmentation cross-check & correction (manual)

In the digitization workflow, much effort was invested into optimization of the process. Surely, we did not check the segmentation for each file. Only pages showing recognition accuracy less that the certain agreed threshold were analyzed for segmentation issues. If it was observed that segmentation was the reason of poor recognition, the operator corrected the segmentation manually and then sent the page for re-recognition.

3. OCR (automatic)

At this phase, full-text OCR was performed by the software package.

4. **Verification/correction or KFI if unrecognized (manual)**

Since the material was challenging, all pages were proofread (single verification) for possible recognition mistakes. In cases when poor accuracy occurred due to some specific reasons (other than incorrect segmentation), operators applied certain techniques to increase automatic readability of the text. We will dwell on such techniques in detail below. In other cases, operators performed manual KFI of unrecognized symbols.

The software package used in the project provided for high productivity of this phase, highlighting the incorrectly recognized symbols in color and providing convenient synchronization of the document image and the recognize text in one screen by means of highly ergonomic GUI, which also allowed us to minimize manual effort.

5. **Export to Microsoft Word (automatic)**

As the XML format required by the Digital library standards differed from that the XML produced by the OCR software, it was more efficient to output the information to Microsoft Word and then perform semi-manual markup with the help of MICROSOFT Word Macros, and then convert the results to XML. Export to Microsoft Word was performed automatically by the OCR software.

6. **XML markup (manual, semi-automated)**

Described below

7. **XML file aggregation (merging several files related to one article)* (manual)**

XML validation (automatic)

This phase was carried out with the help of a specialized software package. Every file sent to the Library had passed the validation process.

* Optional phase (is case of issues with automatic article segmentation)

Output

The specifics of the XML format included:

- 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.)

The process organization in detail

The process, from the technology point of view, looked the following way:
The page image was OCRed. As a part of the process, the layout of scanned page images was automatically analyzed. At this phase, the logical structure of the magazine page was determined through an automated structure analysis process. Column-organized text was

united into logical sequences corresponding to the appropriate articles. Also, at this phase the software builds up zone information: the page was divided into sections (elements) identified as text blocks, tables, or illustrations. OCR was then performed on the page zones containing text, as determined by the layout analysis. The OCR engine supports a wide range of languages and recognizes standard and historical fonts. The OCR results are stored together with the coordinates of the text block.

There was a user interface allowing manual reviewing and correction of both page segmentation and recognition results. The system was empowered with a spellchecking mechanism (based on multi-language user dictionaries); the uncertainly recognized symbols (which are mistakes in 90% cases) are highlighted in color for user's attention. Unknown words can be added to user dictionaries.
In case of incorrect segmentation, the operator re-zoned the page and sent it for re-recognition. In case of recognition mistakes, the operator corrected them manually via the ergonomic interface allowing zoom, scrolling, synchronization between the source page image and the recognized text block (the same section of both showed in adjacent screen views), operator hot keys, etc.

The operators paid special attention to the metadata-critical information, such as headings, captions, etc.

After the correction/review finished, the operator exported the results into Microsoft Word.

In MS Word, ATAPY engineers have programmed a series of VB macros that allowed operators to tag the text in a semi-manual mode by selecting the text and marking it with a tag. The items to be tagged were: the article heading, author, type (interview, news piece, musical review, etc.), abstract, subheadings, formatting characteristics of the text (bold, italics, etc.). Such approach was necessary since the OCR package did not allow export into the customer-defined XML format; only into its on XML format. Some re-formatting effort was considered, including the effort of building a custom XML converter, but this option appeared to be more costly than the option of semi-manual mark-up in Microsoft Word that was finally chosen, and involved less quality control options. The process of semi-manual mark-up was so streamlined that such organization of work proved to me the most effective in the project.

The descriptive and structural metadata for each article (including the magazine issue/date, page, etc.) were added to each article automatically (major properties set up in the beginning of issue processing).

According to the project requirements, some information was to be excluded from the results: for example, the advertisements. In some batches, illustrations were to be excluded as well, in others – were to be saved separately in the appropriate directory of the delivery, with a hyperlink placed in the resulting XML file.

Also, the batch was checked for correct article splitting, cropping out the appropriate information, page sequence, article segmentation, metadata integrity and consistency. After that, the resulting XML files from the batch were validated using a specialized XML validation software tool.

Solutions for the challenges presented by the material

1. The specifics of periodicals as input material:
- Wide format and multi-column layout (challenge for automatic segmentation)

Solution: manual after-correction of automatic segmentation, with consequent re-recognition

2. Colored and textured backgrounds (challenge for both segmentation and OCR)

Solution: manual after-correction of segmentation (re-recognition with altered block properties), semi-automated image preprocessing in graphic packages (increasing contrast, etc.). By «semi-automated» we mean batch processing of images featuring the same irregularities; such images were sorted out and accumulated by the operators while processing the batch and then processed together.

3. A variety of fonts and font colors within one page
4. Designer fonts used
5. Skewed fragments + fragments with normal text orientation present in one page

Solution: image preprocessing in graphic packages (increasing contrast, deskew, etc.), in worst cases - KFI of poorly recognized or unrecognized text

6. Inverted and «normal» text present on one page

Solution: multi-attempt OCR with varied text block properties (inverted/normal text)

7. Format specifics:
- Input material: partially PDF, sometimes containing a text layer, which was sometimes partially unrecognizable (appeared in vector format)

Solution: OCR of PDF as image-only (sometimes), merging of the extracted text layer into the «rawly» recognized text

8. Language specifics:
- Several languages (Danish & English) on one page (additional challenge for OCR even with correct dictionaries enabled)
- Critical information in Danish (GAFFA)

Solution: operators with linguistic background, capable of understanding text in Danish

Quality control and management on the Library side

All delivered material was checked against the delivery list to ensure the batch has been processed in full. The inspection also included manual checking of randomly selected pages from the batch.

To ensure appropriate acceptance procedure, certain acceptance criteria were discussed and then fixed in the Agreement. Minor recognition discrepancies that did not affect text searchability seriously were to be ignored, while mistakes in headings or other metadata-critical substances were considered crucial.

In case of test failure, the batch was rejected and returned to ATAPY with appropriate comments. At the end of each contract phase, which included a series of batches, the Acts of Acceptance were signed.

The exchange of materials (unprocessed and processed batches) was organized by FTP.

Publishing

Once documents have fully completed the conversion and quality-assurance processes, the resulting XML files were generated, in full compliance with the industry standards, where international metadata standard METS hosted and supported by the Library of Congress was primarily used. The files were then published online on the website of the Det Virtuelle Musikbibliotek project, among other digitized periodicals (all available at www.dvm.nu/periodical). For GAFFA magazine, the materials were published on GAFFA website (http://gaffa.dk/arkiv), and, according to the project requirements, also the PDF versions of the issues were published.

What would be the next steps?

Currently libraries all over the world are considering new strategies aimed at reducing manpower costs and spending a higher percentage of the budget on cheaper digitization. Some crowdsourcing projects are underway, though the effect of such initiatives is not yet clear.

The shift from a paper-based library to a digital library presents a huge challenge for libraries' technical infrastructure as well. At the same time, the software packages aimed at mass digitization are evolving as well. At the moment, the wide-format and column-organized layouts that used to be a challenge for most OCR packages, can be processed by newer versions of the software much better than before. Software packages have appeared that can provide smart automatic indexing, which, doubled with full-text recognition results, allow achieving results that are searchable enough to serve a valuable source of information for researchers and for the wide public. This certainly implies a compromise between the quality of our services and our digitization ambitions. But the ultimate goal – saving Europe's printed heritage for future generations and transforming it into a digital resource, accessible from anywhere in the world – is probably worth it.

Acknowledgments

The following organizations have contributed to the project:

Danish Agency for Culture, Libraries (the funding organization)

Gaffa A/S, Gaffa Nordic, CEO Robert Borges

Musikbibliotek.dk, Gentofte Centralbibliotek, former Editor-in-Chief Amalie Ørum Hansen 2008-2010 (now Bibzoom.dk, State and University Library, Editor-in-Chief Niels Mark Pedersen); the idea was initiated by former Editor-in-chief Susanne Kier in 2007.

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, CEO Sergey Borovoy

References

The websites of Det Virtuelle Musikbibliotek and GAFFA:

http://gaffa.dk/

www.dvm.nu/periodical

Speakers

Amalie Ørum Hansen
Development Consultant at Gentofte Centralbibliotek (2008-present)

ahan@Gentofte.dk

Area of expertise: Strategic competence development, particularly concerning Libraries of the Capital Region of Denmark

Representative of the libraries in Danish Agency for Culture, Libraries, Udvalget vedr. Fonogrammer (Committee of Public Lending Right) (2011-2015)

Work experience:

Project manager of the fusion between netmusik.dk and musikbibliotek.dk (now: bibzoom.dk) (2010-11)
Musikbibliotek.dk, Editor-in-Chief (2008-2010)
Producer, Danish National Chamber Orchestra (2007-08)
Production Manager, Danish National Opera (2006)
Producer, Mogens Dahls Koncerthus (2005-06)
Assistant Teacher, Royal Danish Academy of Music (Aarhus), (2005-06)

Education:

Aarhus University, MA, Musicology and History of Art, 1996 – 2004
Royal Danish Academy of Music (Aarhus), Teachers Degree, Ear Training and Music Theory, 1998 – 2003

Sergey Borovoy
CEO of ATAPY Software (2001-present)

sergeyb@atapy.com

Areas of expertise: Computer programming, business development, sales, project management, business administration, technical consulting in the fields of image analysis, OCR and digitization/data capture, including software tools and technologies used in cultural preservation projects.

Work experience:

2001-present: Founder and CEO of ATAPY Software – a company specializing in document imaging, digitization of printed documents and associated software development services. Under Mr. Borovoy's leadership, ATAPY completed a large number of projects for libraries and cultural heritage organizations across Europe, and took part in some notable projects in the field (METAe, IMPACT), contributing both software development and human-aided data entry effort.
Project manager, later Deputy VP of Corporate projects and Integration of Technologies, ABBYY Software House (1999-2001)

Education:

1994 - Novosibirsk State University, BSc, Computer Science
1997 - Arizona State University, MSc, Computer Science