From Mass Digitization to Mass Content Enrichment

Frédéric Beaugendre, Michel Merten, Jean De Gyns: MEMNON
Stéphane Bayot: SONUMA

BAAC conference, 4 October 2012, Helsinki
Sonuma

- Created in 2009 by Wallonia government, RTBF & Wallonia/Brussels Community government
- Private company with capital of €40Mio (24Mio cash)
- SONUMA owns RTBF patrimonial archives from 1930 to end 2007.
- 120,000 hours of TV and Radio archives
- SONUMA preserves, digitizes & discloses archives
- Started from scratch
Sonuma

• Effective and global approach
  – Knowledge of the collection (inventories, dictionnaries…)
  – Needs analysis and demands
  – Public tendering procedure

• Selection of the skilled partners
  – Restauration & digitization
  – Automatic indexation, metadata management

• 15 FTE divided in 4 departments
  – Operational, Editorial, IT, Sales
Digitization milestones

- **Nov. 2010 – April 2012**: 23,000 hours radio news (VHS)
- **Febr. 2011 – March 2013**: 30,000 hours (BETA)
- **July 2011 – October 2011**: 10,000 hours (DAT)
- **Sept. 2011 – July 2012**: 5,000 hours (1’)
- **Jan. 2012 – July 2012**: tenders for 6,000 hours films & 15,000 hours audio magnetic tapes
Metadata approach

- Metadata harvesting
  - Databases from archives, newsroom, playout room & web
  - Paper archives from archives, communication department, newsroom

- First level of indexation asap
  - Integration of electronic metadata in one DB
  - Scanning and OCR of thousands pages
  - Automatic indexation solutions
    - Option in the first tenders: VHS & BETA
VHS

- 23,000 hours of radio news between 1992 to 2002
- Time slots automatically recorded (VHS scheduler)
- 8 hours /VHS (LP)
- Date and hour of broadcasting recorded on the video track
- Texts of the anchorman and the journalists exported from the newsroom system

Sonuma’s demand:
- Segmentation based on the video track informations: one file by recording,
- Text alignment on the audio track
VHS

- **Supports**
  - Inventory of almost 3000 VHS
  - Idsonuma => bar code
  - Date and hour of broadcasting of the first recording

- **Metadata**
  - Import of newsroom data (journalists texts) in the Sonuma database
  - Scripts (filters) to detect and delete « useless informations » in the journalists texts who might reduce the alignment ‘s quality
VHS

- Export in a xml file of all metadata needed for the alignment
- Xml by FTP

- Transport to Memnon
Overview of Memnon Archiving Services

• A European leader in the digitization, migration, and semi-automated indexation and documentation of audiovisual archives, with facilities in:
  – Belgium
  – Also providing “in house” services (set-up, media architecture, training, project-management, metadata issues ... on site)
• Involved in the digitization of more than 700,000 hours of audiovisual archives of most analog and digital formats.
• Large scale audiovisual digitization and migration at a very competitive price through strong project management and optimized workflows
• Expertise in database and metadata management
• Memnon has participated into many projects around audiovisual archives and education
• Memnon has developed tools for automated content enrichment - IPI manager
Large scale audiovisual digitization

**Sound:**
All types of carriers (tape, records, cassette, DAT, …)

**Video:**
All types of carriers (1inch, U-matic, Digit Beta, Beta SP, DVcam …)
How did Memnon approach the issue

• A Global Workflow
  – Automatize as most as possible the post-processing steps after digitization of medias
    ▪ Make the best use of existing metadata
    ▪ Implement automatic extraction algorithms
• House made OCR
  – Automatic extraction of the timecode embedded on the image
• Automatic text Alignment on audio
  – Filtering of metadata
  – Use of Speech Recognition technology for automatic alignment
  – 100% of alignment, if the text corresponds to the actual audio and length of the sentence exceeds 3 sec of audio
Dynamic OCR (1)

- OCR
  - Clean picture
  - Locate the digits
  - Match digit
Dynamic OCR (1)

• OCR
  – Clean the picture
    • Build a black and white version of the picture by using simple assumptions and features to isolate the digits.
  – Locate the digits
    • Scan trough the cleaned image to find areas which could contains digits. First start by scanning the rows and then isolate each area within each row.
  – Match digit
    • Compare the detected areas with some prebuild models to recover the digit. A high score of confidence is required to avoid false positives.
Dynamic OCR (2) : Clean & Locate

- Source image :

- Scan :

Result :
Dynamic OCR (3): Match the digit

- Models:

\[0123456789\]

- Matching Sample:

\[\text{Sample Image} \quad \text{Sample Image}\]
Text Alignment Algorithm

- Original XML associated to the tape:
  - Matching to the actual to be confirmed by the process
  - Associate file reference to be filled
  - Segmentation of programs according to the OCR match to be processed within the workflow
  - Timecode of each Headline to be filled

```
<TexteJournaliste>
  <desannonce>
    <Sujet>
      <Rang>4</Rang>
      <NumeroFichier>*** A DETERMINER A LA SYNCHRONISATION
                    ***</NumeroFichier>
      <RelTcIn>*** A DETERMINER A LA SYNCHRONISATION ***</RelTcIn>
    </Sujet>
  </TexteJournaliste>

<Titre>BELGIQUE/JUDICIAIRE/FAFA JONATHAN</Titre>

<chapeau>
  procureur ______________ Voila
deux ans que le petit Jonathan mourait sous les coups portes par
l'ami de sa mere. Ce petit garcon avait trois ans. Ce drame
s'etait deroule a Obourg, dans la region montoise. Il avait deja
ete jugé en correctionnelle. Il a ete rejugé en appel. Attention,
pour la justice, il peut y avoir de serieuses differences d'une
affaire d'enfant battu a l'autre. Jean Paul procureur. bande 1'30''
fin: mais aussi d'une certaine indifferance</chapeau>
```

Original XML
Text Alignment Algorithm

- Construction of a grammar consisting of “parts-of-speech” from the transcription:
  - part-of-speech starts from a non-grammatical word to the next non-grammatical word;

Original XML

Grammar Construction

Headline1_1 – Voila deux ans
Headline1_2 – petit Jonathan
Headline1_3 – mourrait sous les coups
Headline1_4 – porté par l’ami
Headline1_5 – petit garçon
Headline1_6 – avait trois ans
Headline1_7 – drame s’était déroulé
Headline1_8 – Obourg dans la région
Headline1_9 – avait déjà été jugé
Headline1_10 – rejugé en appel
Text Alignment Algorithm

- Each item of the grammar is processed through a phonetizer;
- These parts-of-speech chunks is used for recognition in the audio files;
- A post-processing algorithm detects the occurrence of 3 consecutive recognized items belonging to the same headline within a window of 20 sec;

Original XML  Grammar Construction  Phonetization

Headline1_1 – Voila deux ans  Headline1_1 – vwallaaddeuzzan
Headline1_2 – petit Jonathan  Headline1_2 – ppeettiijjoonnaattan
Headline1_3 – mourrait sous les cou  Headline1_3 – mmuurraaissuulleekkuu
Headline1_4 – porté par l’ami  Headline1_4 – ppOOrrtteepaarrllaammii
Headline1_5 – petit garçon  Headline1_5 – ppeuttiiggaarrrson
Headline1_6 – avait trois ans  Headline1_6 – aavvEEttrtroizzan
Headline1_7 – drame s’étais déroulé  Headline1_7 – ddrraammsseettEEdddeerrruulllee
Headline1_8 – Obourg dans la région  Headline1_8 – oobbuuurrddanllaarreeGGjjon
Text Alignment Algorithm

- Alignment to the audio file is computed using the corresponding time codes of the recognized parts-of-speech
- The correct file reference is filled in

Original XML → Grammar Construction → Phonetization → Alignment
Other tools for automated enrichment by Memnon

- **Implemented in Bold**, *on the roadmap in italics*

- « Low level » indexation: Distinct Audio & Video algorithms
  - Audio:
    - Acoustic segmentation & classification
    - Speaker segmentation, clustering, tracking & recognition
    - Large vocabulary speech recognition
    - Jingle recognition
    - Music fingerprinting
    - Silence detection
    - …
  - Video:
    - Video shot detection
    - Face recognition
    - Background recognition
    - Caption recognition
    - Logo recognition
    - …

- « High level » structuration
  - Combination of several low level results to:
    - Increase the confidence: e.g. face recognition + speaker recognition results
    - Allow a high level structuration of the content: i.e. use of speech recognition results to extract topics: « in this segment, we are talking about the Olympics, and in this following one, we are talking about the war in Irak…"
Import into the Sonuma’s DAM
Import into the Sonuma’s DAM
Import into the Sonuma’s DAM
Import into the Sonuma’s DAM
Conclusion: Expected time savings versus manual work

- 23000 hours of radio news programs
- Several hundred thousand of anchorman texts available in a DB
- Manually segmentation of 23000 hours of audio files => hard job !!!
- Manually texts alignment => unrealistic !!!
- Unjustifiable costs (« ...and only for radio programs? »)
- Made possible using speech recognition tools
- Excellent indexation level given the metadata provided. Enough criterias to find assets in the DAM (Program title, date and hour of broadcasting, texts, journalists names...)
- Acceptable costs ...even for radio programs!
End of the presentation

Thank you for your attention