

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



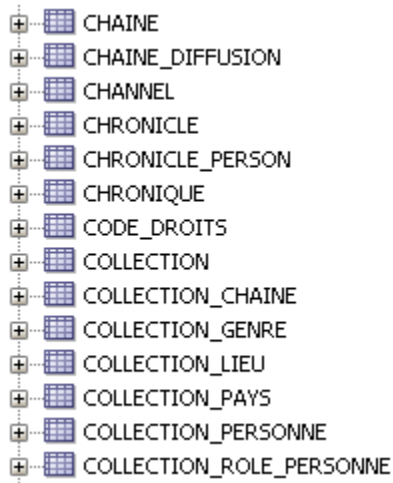
- **Effective and global approach**
 - Knowledge of the collection (inventories, dictionaries...)
 - Needs analysis and demands
 - Public tendering procedure
- **Selection of the skilled partners**
 - Restoration & digitization
 - Automatic indexation, metadata management
- **15 FTE divided in 4 departments**
 - 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



DIMANCHE 11 MARS 73

(COULEURS) JEU

15H15

VISA POUR LE MONDE

Georges Désir interrogera le premier candidat sur Mysore.

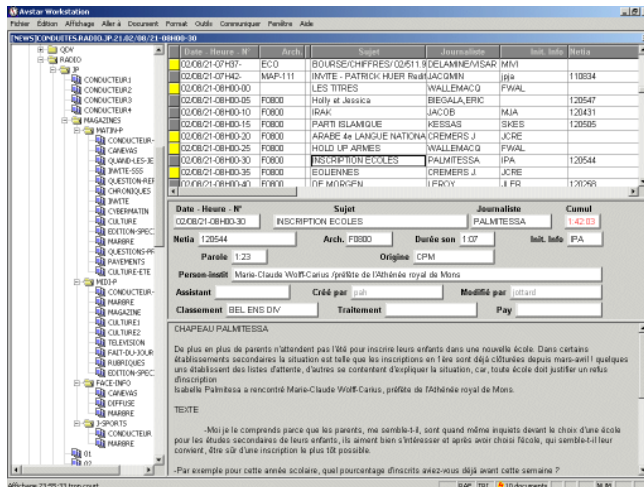
L'état de Mysore (192.203 km², 23.586.000 habitants) est situé au sud de la péninsule de l'Inde.

La sidérurgie, les huileries, l'or et l'industrie du tapis constituent ses principales activités économiques.

Incorporé à l'Empire Mongol en 1326, cet état fut occupé par les Anglais au 19^e siècle.

En 1956, l'état de Coorg et quelques fractions des états voisins lui ont été annexés, ce qui a plus que doublé sa superficie et sa population.

- Metadata harvesting
 - Databases from archives, newsroom, layout 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**



- 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

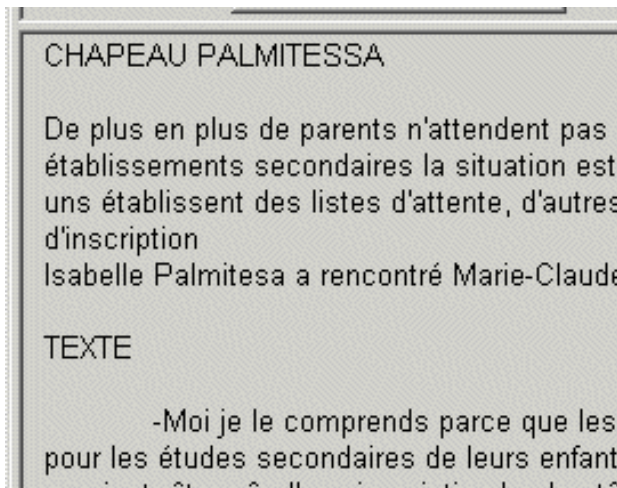
Sonuma's demand :

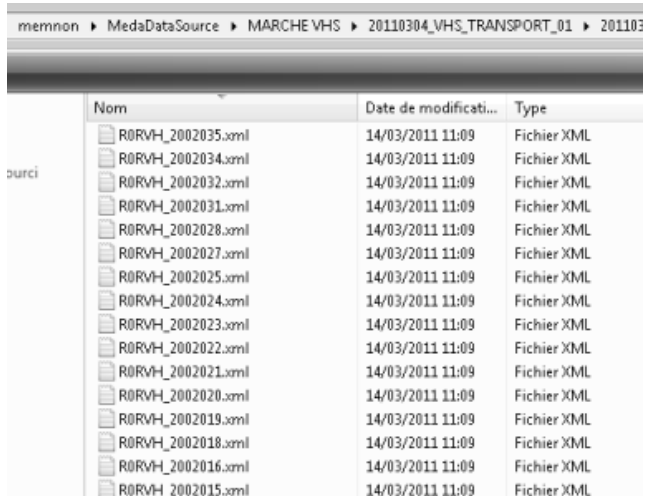
- Segmentation based on the video track informations : one file by recording,
- Text alignment on the audio track



- 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





memnon ▶ MedaDataSource ▶ MARCHE VHS ▶ 20110304_VHS_TRANSPORT_01 ▶ 20110304

Nom	Date de modificati...	Type
R0RVH_2002035.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002034.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002032.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002031.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002028.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002027.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002025.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002024.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002023.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002022.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002021.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002020.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002019.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002018.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002016.xml	14/03/2011 11:09	Fichier XML
R0RVH_2002015.xml	14/03/2011 11:09	Fichier XML

- 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, ...)



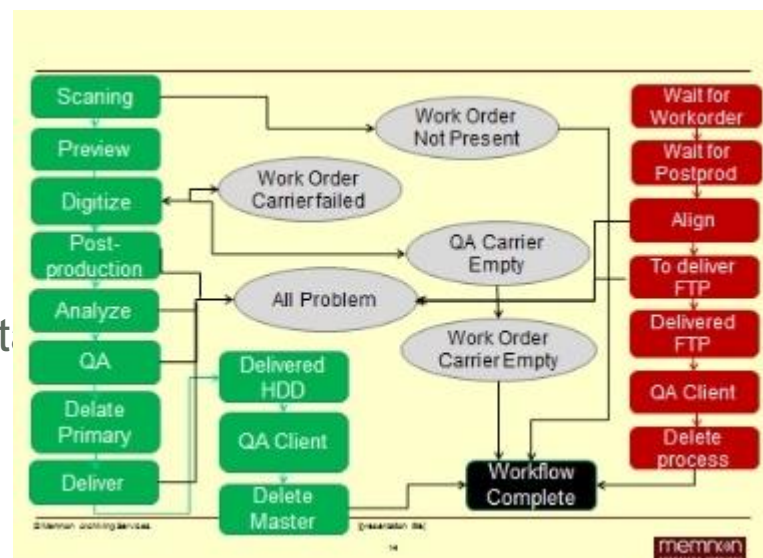
Video :

All types of carriers (1inch, U-matic, Digit Beta, Beta SP, DVcam ...)



How did Memnon approach the issue

- A Global Workflow
 - Automate 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 Aligement 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)

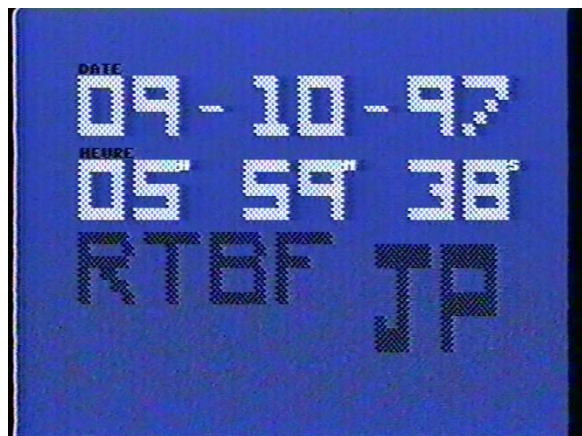
- OCR
 - Clean picture
 - Locate the digits
 - Match digit

Dynamic OCR (1)

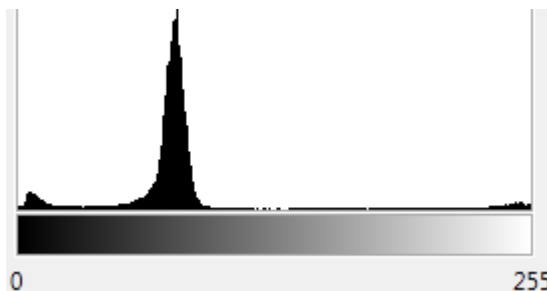
- OCR
 - Clean the picture
 - Build a black and white version of the picture by using simple assumptions and features to isolate the digits.
 - Locate the digits
 - Scan through the cleaned image to find areas which could contain digits. First start by scanning the rows and then isolate each area within each row.
 - Match digit
 - Compare the detected areas with some prebuilt models to recover the digit. A high score of confidence is required to avoid false positives.

Dynamic OCR (2) : Clean & Locate

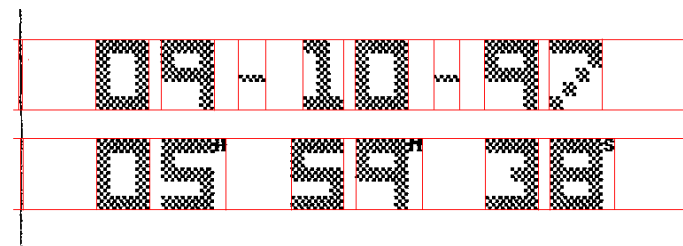
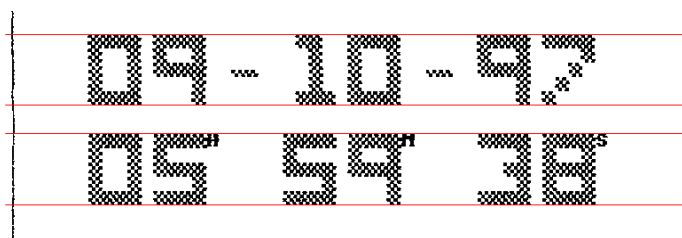
- Source image :



- Result :



- Scan :

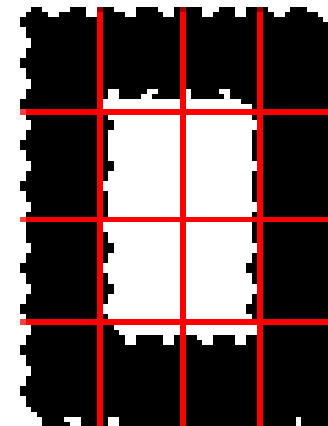
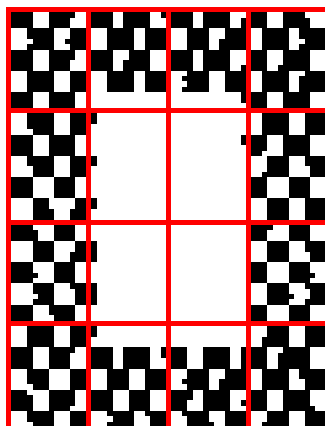


Dynamic OCR (3): Match the digit

- Models :



- Matching Sample :



Text Alignment Algorithm

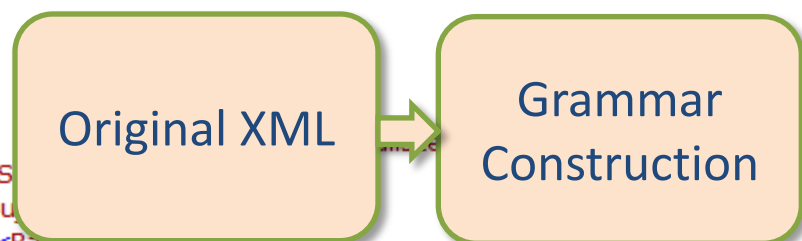
- Original XML associated to the tape:
 - Matching to the actual to be confirmed by the process
 - Associate file reference to be filled
 - Segmentation of programs according to the OCR match to be processed within the workflow
 - Timecode of each Headline to be filled

Original XML

```
desannonce</TexteJournaliste>
</Sujet>
- <Sujet>
  <Rang>4</Rang>
  <NumeroFichier>*** A DETERMINER A LA SYNCHRONISATION
  ***</NumeroFichier>
  <RelTcIn>*** A DETERMINER A LA SYNCHRONISATION ***</RelTcIn>
  <Validation />
  <Titre>BELGIQUE/JUDICIAIRE/AFFAIRE JONATHAN</Titre>
  <TexteJournaliste>chapeau procureur _____ Voila
  deux ans que le petit Jonathan mourait sous les coups portés par
  l'ami de sa mère. Ce petit garçon avait trois ans. Ce drame
  s'était déroulé à Obourg, dans la région montoise. Il avait déjà
  été jugé en correctionnelle. Il a été rejugé en appel. Attention,
  pour la justice, il peut y avoir de sérieuses différences d'une
  affaire d'enfant battu à l'autre. Jean Paul procureur. bande 1'30"
  fin: mais aussi d'une certaine indifférence</TexteJournaliste>
</Sujet>
- <Sujet>
  <Rang>5</Rang>
  <NumeroFichier>*** A DETERMINER A LA SYNCHRONISATION
```

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;

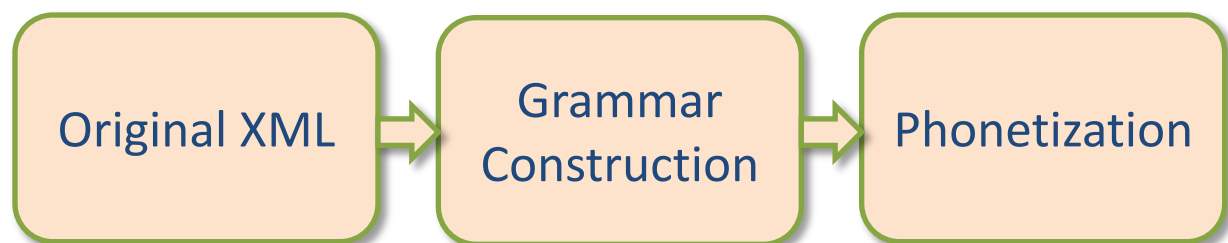


```
</S>
- <Su
  <Rang>1</Rang>
  <NumeroFichier>*** A DETERMINER A LA SYNCHRONISATION
    ***</NumeroFichier>
  <RelTcIn>*** A DETERMINER A LA SYNCHRONISATION ***</RelTcIn>
  <Validation />
  <Titre>BELGIQUE/JUDICIAIRE/AFFAIRE JONATHAN</Titre>
  <TexteJournaliste>chapeau procureur _____ Voila
    deux ans que le petit Jonathan mourait sous les coups portés par
    l'ami de sa mère. Ce petit garçon avait trois ans. Ce drame
    s'était déroulé à Obourg, dans la région montoise. Il avait déjà
    été jugé en correctionnelle. Il a été rejugé en appel. Attention,
    pour la justice, il peut y avoir de sérieuses différences d'une
    affaire d'enfant battu à l'autre. Jean Paul procureur. bande 1'30"
    fin: mais aussi d'une certaine indifférence</TexteJournaliste>
  </Sujet>
- <Sujet>
  <Rang>5</Rang>
  <NumeroFichier>*** A DETERMINER A LA SYNCHRONISATION
```

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;

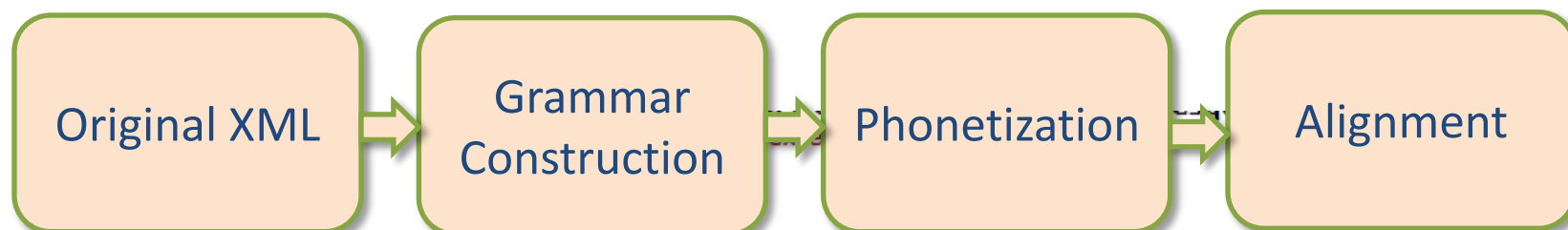


Headline1_1 – Voila deux ans
Headline1_2 – petit Jonathan
Headline1_3 – mourrait sous les cou
Headline1_4 – porté par l’ami
Headline1_5– petit garçon
Headline1_6 – avait trois ans
Headline1_7– drame s’était déroulé
Headline1_8 – Obourg dans la région

Headline1_1 – vwallaaddeuzzan
Headline1_2 – ppeettiijoonnaattan
Headline1_3 – mmuurraissuulleekkuu
Headline1_4 – ppOOrtteppaarrllaammii
Headline1_5– ppeuttiiggaarrson
Headline1_6 – aavvEettrroizzan
Headline1_7– ddrraammsseettEEddeerruulle
Headline1_8 – oobbuurrddanllaarreeGGjjon

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 →

```
<NumeroFichier>01</NumeroFichier>
<RelTcIn>00:04:44.420</RelTcIn>
<Validation xsi:type="xsd:boolean">>true</Validation>
<Titre>BELGIQUE/JUDICIAIRE/AFFAIRE JONATHAN</Titre>
<TexteJournaliste>chapeau procureur _____ Voila
deux ans que le petit Jonathan mourait sous les coups portés par
l'ami de sa mère. Ce petit garçon avait trois ans. Ce drame
s'était déroulé à Obourg, dans la région montoise. Il avait déjà
été jugé en correctionnelle. Il a été rejugé en appel. Attention,
pour la justice, il peut y avoir de sérieuses différences d'une
affaire d'enfant battu à l'autre. Jean Paul procureur. bande 1'30"
fin: mais aussi d'une certaine indifférence</TexteJournaliste>
</Sujet>
- <Sujet>
<Rang>5</Rang>
<NumeroFichier>01</NumeroFichier>
```

Original XML →

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

```
<?xml version="1.0" encoding="utf-8" ?>
- <SonumaMasterFile xmlns:i="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="http://sonuma.be/MasterFileTechnicalData">
  <FileName>RORVH_1990038_21.wav</FileName>
  <MD5>EB73C3A2CC98D1CB2E501C3A72D1B3DD</MD5>
  <OCRDate>1990-03-29 07:59:29Z</OCRDate>
  <Duree>00:35:11.6800000</Duree>
  <DateTransfert>2/03/2011 16:53:55</DateTransfert>
  <Operateur>110301_0122_hbliciek</Operateur>
  <VCR_Id>VHS-008</VCR_Id>
  <ParametresNumerisation>DeEssing;</ParametresNumerisation>
  <AudioInfo>Sifflantes;Débalance;</AudioInfo>
  <PositionDansLaCassette>06:15:29.4400000</PositionDansLaCassette>
  <OCRNextSegmentDate>1990-03-29 08:59:29Z</OCRNextSegmentDate>
- <OCRStatus>
  <Status>OK</Status>
  <HasOCRException>>false</HasOCRException>
  <Is4GoCut>>false</Is4GoCut>
</OCRStatus>
</SonumaMasterFile>
```

```
<?xml version="1.0" ?>
- <Support xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  <IdSonuma>RORVH_1990038</IdSonuma>
+ <Plages>
  <Commentaire>Date fausse sur l'écran avant 13h06</Commentaire>
  <Description>Emissions d'information du 26/03/1990 13:00 au 29/03/1990 13:00</Description>
+ <Programme>
+ <Programme>
- <Programme>
  <IdOrigine>1990/03/29-0800F08000008</IdOrigine>
  <Rang>*** A DETERMINER A LA SYNCHRONISATION ***</Rang>
  <Titre>JP FEDERAL 08H00</Titre>
  <Chaine>LA PREMIERE</Chaine>
  <DateDiff>29/03/1990 08:00</DateDiff>
+ <Sujet>
+ <Sujet>
+ <Sujet>
+ <Sujet>
+ <Sujet>
- <Sujet>
  <Rang>6</Rang>
  <NumeroFichier>21</NumeroFichier>
  <RelTcIn>00:06:44.700</RelTcIn>
  <Validation xsi:type="xsd:boolean">true</Validation>
  <Titre>BREVES BELGES</Titre>
  <TexteJournaliste>Quelques télégrammes de Belgique avec Alain Gerlache GERL Les
  manifestation fin avril et peut-être une grève en mai. Le problème, c'est que la C
  des infirmières indépendantes des maisons de repos. Elles s'inquiètent d'un projet
  avril. Seuls les soins indispensables seraient assurés. Les TCT wallons sont inquiets
  à statut précaire. Résultat : on parle du licenciement d'un tiers d'entre eux. Alertes
  hollandaises. Seule consolation : le Zwin serait épargné. Immeuble à vendre. C'est
  Tout un symbole.</TexteJournaliste>
</Sujet>
- <Sujet>
  <Rang>7</Rang>
  <NumeroFichier>21</NumeroFichier>
  <RelTcIn>00:07:38.650</RelTcIn>
  <Validation xsi:type="xsd:boolean">true</Validation>
  <Titre>GRANDE BRETAGNE - ETATS-UNIS - IRAK - NUCLEAIRE</Titre>
  <TexteJournaliste>Les douanes britanniques ont arrêté trois personnes hier à l'aéroport
  nucléaire au Proche Orient rebondisse. Les britanniques et les américains accusent
  douanes américaines. D'ailleurs, la piste était suivie depuis Los Angeles où les dénoncent
  prolifération nucléaire du Proche Orient. Depuis quelques temps, plusieurs experts
  qu'ils respectent le traité sur la non prolifération des armes nucléaires. L'Irak a d'ailleurs
  question demeure pourquoi importer des détonateurs nucléaires. Ce ne sont pas
  </Sujet>
```

Import into the Sonuma's DAM

Home > R0RVH_1990038_21_1.vux Imprimer


\\archmedia\archmedi...e\R0RVH_1990038_21.wav

Fiche Médias

Nom de la fiche : R0RVH_1990038_21_1
Dossier : Home
Modèle associé :

Début : 07:59:29:00 Fin : 08:13:49:00
Durée : 00:14:20:00

Statut : **verrouillée** par sba
Date de création : mardi 30 août 2011 14:13:54
Date de modification : jeudi 1 septembre 2011 13:52:23



Fixer l'image de référence

Métadonnées Segments Palette d'annotations Rafraîchir

METADONNEES PROGRAMME

Titre du programme : JP FEDERAL 08H00
Type de média :
Date de diffusion : 29/03/1990
time code corrigés :
Métadonnées corrigées :
Droit(s) d'utilisation :

Description Informations de diffusion Informations de production Informations techniques Informations pour le web

Heure de début : 07:59:29
Heure de fin : 08:13:49
Durée du programme : 00:14:20

Chaîne : LA PREMIERE;
Date de diffusion imprécise :
Date(s) de rediffusion :

```
<Titre>JP FEDERAL 08H00</Titre>  
<Chaîne>LA PREMIERE</Chaîne>  
<DateDiff>29/03/1990 08:00</DateDiff>  
+ <Sujet>
```

```
<URL>http://sonuma.be/masterfile/recordings/  
<FileName>R0RVH_1990038_21.wav</FileName>  
<MD5>EB73C3A2CC98D1CB2E501C3A72D1B3DD</MD5>  
<OCRDate>1990-03-29 07:59:29Z</OCRDate>
```


Import into the Sonuma's DAM

<OCRDate>1990-03-29 07:59:29</OCRDate>



```
- <Sujet>
  <Rang>6</Rang>
  <NumeroFichier>21</NumeroFichier>
  <RelTcIn>00:06:44.700</RelTcIn>
  <Validation xsi:type="xsd:boolean">true</Validation>
  <Titre>BREVES BELGES</Titre>
  <TexteJournaliste>Quelques télégrammes de Belgique
  manifestation fin avril et peut-être une grève en m
```

```
- <Sujet>
  <Rang>7</Rang>
  <NumeroFichier>21</NumeroFichier>
  <RelTcIn>00:07:38.650</RelTcIn>
  <Validation xsi:type="xsd:boolean">true</Validation>
  <Titre>GRANDE BRETAGNE - ETATS-UNIS - IRAK - NUCLEAIRE</Titre>
  <TexteJournaliste>Les douanes britanniques ont arrêté trois personnes
  nucléaire au Proche Orient rebondisse. Les britanniques et les amér
```

Calque : Métadonnées séquence

Rafraîchir | Zoom: x2

:39:14	08:05:27:08	08:06:15:03	08:07:02:22	08:07:50:17	08:08:38:11	08:09:26:06	08:10:14:00	08:11:01:19	08:11:49:00
urn:guid:{7d4ce36-11e0-994}	urn:guid:{7d9e30-ce36-11e0-...}	urn:guid:{7dae09b-ce36-11e0-...}	urn:guid:{7d6dfbb4-ce36-11e0-b794-...}	urn:guid:{7da62d18-ce36-11e0-a97e-00199990b994}	urn:guid:{7db5c840-ce36-11e0-ba28-...}	urn:guid:{7db5c840-ce36-11e0-ba28-...}	urn:guid:{7db5c840-ce36-11e0-ba28-...}	urn:guid:{7db5c840-ce36-11e0-ba28-...}	urn:guid:{7db5c840-ce36-11e0-ba28-...}
CHAMBRE				RFA					SPORT


Sélection

TCin: 00:00:00:00

TCout: 00:00:00:00

Durée: 00:00:00:00

Informations sur le segment sélectionné



Calque : 1
 Id dans le calque : 8
 Niveau dans le calque : 1
Début : 08:08:08:00
 Fin : 08:09:48:00
 Durée : 00:01:40:00

ID séquence : urn:guid:{7da62d18-ce36-11e0-a97e-00199990b994}
 Titre de la séquence : RFA

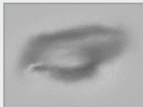
- Palette d'annotations
- Ajouter au panier
- Ajouter les segments au panier
- Supprimer ce segment
- Supprimer les segments
- Copier ce segment
- Fixer le photogramme

Import into the Sonuma's DAM

Métadonnées Segments Palette d'annotations Rafraîchir

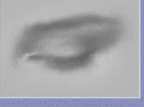
Calque : Métadonnées séquence

Segment précédent TC In: 08:07:07:00 Segment suivant



Id dans le calque 6
Niveau dans le calque 1

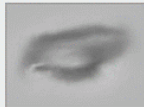
Début 08:06:13:00
Fin 08:07:07:00
Durée 00:00:54:00



Id dans le calque 7
Niveau dans le calque 1

Début 08:07:07:00
Fin 08:08:08:00
Durée 00:01:01:00

Modèle de données : METADONNEES SEQUENCE



Id dans le calque 8
Niveau dans le calque 1

Début 08:08:08:00
Fin 08:09:48:00
Durée 00:01:40:00

METADONNEES SEQUENCE

📄 Titre de la séquence : GRANDE BRETAGNE - ETATS-UNIS - IRAK - NUCLEAIRE 📅 Date de diffusion : ▼ Droit d'utilisation :

Description Informations de diffusion Informations de production Informations techniques Information pour le web

📄 Sous-titre de la séquence :

📄 Titre chronique :

📄 Rang de la séquence : 07

▼ Casting :

▼ Matière :

▼ Lieu :

▼ Thématique :

▼ Fonds : RTBF;

▼ Public :

📄 Récompense :

▼ Sous-titrage :

📄 Fichier de sous-titre :

📄 Résumé documentaliste :

📄 Analyse des vues :

📄 Interview :

📄 Texte du journaliste :
MENU. Les douanes britanniques ont arrêté trois personnes hier à l'aéroport de Londres. Elles s'apprêtaient à exporter vers l'Irak des détonateurs d'armes nucléaires. Il n'en fallait pas plus pour que la polémique sur la prolifération nucléaire au Proche Orient rebondisse. Les britanniques et les américains accusent l'Irak de vouloir se doter de l'arme atomique. LANGE. C'est une enquête longue d'un an et demi qui a abouti hier. Elle était menée en collaboration avec les douanes américaines. D'ailleurs, la piste était suivie depuis Los Angeles où les détonateurs avaient été embarqués vers Londres. La fin du voyage se situait en Irak. Ces arrestations remettent sur l'avant de la scène le rôle de l'Irak dans la prolifération nucléaire du Proche Orient. Depuis quelques temps, plusieurs experts occidentaux accusent l'Irak de vouloir se doter de l'arme atomique. Le président américain Georges Bush a lancé un appel aux pays du Proche Orient pour qu'ils respectent le traité sur la non prolifération des armes nucléaires. L'Irak a d'ailleurs signé ce traité. Les irakiens se défendent. L'ambassadeur d'Irak à Washington dément toute volonté de se doter de l'arme atomique. Cependant une question demeure pourquoi importer des détonateurs nucléaires. Ce ne sont pas spécialement des objets décoratifs.

📄 Résumé commercial :

▼ Langue originale :

memnon
ARCHIVING SERVICES

BAAC conference, 4 October 2012, Helsinki

24

sonuma
LES ARCHIVES AUDIOVISUELLES

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