

Audio/Video Digital Libraries: the ECHO Experience

ICDAT2006

19-20 October 2006

Pasquale Savino
ISTI-CNR

Consiglio Nazionale delle Ricerche

Talk overview

Provide a review on the techniques and the methodologies for the **organization**, **creation**, and **management** of an Audio/Video Digital Library through the illustration of the results of the ECHO project.

- Characteristics of A/V Digital Libraries
- The ECHO Project and Prototype System
 - Automatic Audio/Video Indexing
 - Content-Based Retrieval of Audio/Video documents
 - The ECHO Metadata Model
 - Visualization and Browsing of search results
- Evolution of the ECHO solution

Consiglio Nazionale delle Ricerche

What is a Digital Library?

A Digital Library is an **organized collection** of digital objects, including **text, images, audio, video** and **services** for its access and retrieval, as well as for **selection, organization and maintenance** of the collection.

The importance of video

- Video can be considered today the primary information and communication channel, due to
 - Richness in information contained
 - Appeal
 - Approximately 50% of what is seen and heard simultaneously is retained
- Video libraries are becoming essential in many application fields
 - Media and entertainment
 - Personal information
 - Distance learning
 - Education
 - Telemedicine
 - Surveillance

Services of A/V Digital Libraries

- Digital Video Libraries are more complex than traditional DLs; they require the integration of several specialized technologies
- They offer the same services of text digital libraries
- Specific characteristics of Indexing and retrieval services
 - Indexing based on the integration of different technologies for the automatic feature extraction
 - Integration of manual and automatic indexing
 - Retrieval based on different video features

Advantages of A/V DLs

- Most of the video material produced is used only once, due to the **difficulty to archive** it, to **protect** and to **retrieve**.
- A large video library of distributed and network searchable videos would enable
 - **Preservation** of precious and expensive material
 - **Reduction of production costs** for new videos, through the reuse of existing material
 - **Diffusion of knowledge**

A/V vs traditional DLs [1/2]

- Library creation
 - Traditional DLs, contain text documents
 - Library creation requires automatic acquisition of text, extraction of document content, and indexing
 - This process is well known and many different techniques have been developed
 - Video is extremely rich in “content” but
 - the indexing of video content is difficult, expensive, and extremely dependent from the user and the application
 - A possible approach consists in an appropriate integration of automatic content extraction (e.g. speech recognition, image analysis, etc.) and manual indexing

A/V vs traditional DLs [2/2]

- Library exploration
 - Traditional DLs, contain text documents
 - Library exploration requires simple interfaces to formulate queries on free text and document metadata.
 - Video libraries should permit
 - To formulate queries on many different “dimensions”
 - Text, as extracted from speech and captions
 - Images extracted as key frames
 - Motion information
 - Other features automatically extracted
 - Metadata provided manually

- Characteristics of A/V Digital Libraries
- **The ECHO Project and Prototype System**
 - Automatic Audio/Video Indexing
 - Content-Based Retrieval of Audio/Video documents
 - The ECHO Metadata Model
 - Visualization and Browsing of search results
- Evolution of the ECHO solution

The ECHO Project

- ECHO: European CHronicles On line
- International research cooperation **funded by the European Union** under the IST fifth Framework Program.
- The project started on **February 2000** and ended on **May 2003**.
- Cooperation among **research institutions, audio/video archives, and industry**.

The Partners

- **Project coordinated** by ISTI-CNR
- **Content providers**
 - Istituto Luce (I), Institut National de l'Audiovisuel (F), Stichting Nederlands Audiovisueel Archief (NL), Memoriav (CH)
- **Industrial partners**
 - Tecmath (D), Eurospider Information Technology (CH)
- **Research Institutions**
 - ISTI-CNR (I), Istituto Trentino di Cultura (I), Centre National de la Recherche Scientifique (F), University of Mannheim (D), Carnegie Mellon University (USA)

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Main project motivations

Large collections of **historical documentary films** owned by private and public institutions

Four main national video archives participated to the ECHO project



Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Main project results

Development of an Audio/Video Digital Library system which makes possible to **access** and **search** by content over the net, **large** and **inter-operating** audio visual archives.

- Management of **historical documentaries**
- Definition and adoption of an appropriate **metadata format**
- Integration of **manual** and **automatic indexing**
- Support of **multiple languages**
- Creation of **automatic video summaries**
- Validation through **test beds** of documentary film archives

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Operations of the ECHO A/V Digital Library

- Video archiving and indexing
- Video storage
- Content-based search
- Visualization of search results
- Video access (visualization and copy)

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

ECHO System Functionality

- **Indexing**
 - Automatic indexing: key frames, object recognition, text extraction, face detection
 - Manual semantic indexing through a specialized video metadata editor
 - Automatic speech recognition for multiple languages (italian, dutch, german, english)
 - Automatic generation of video abstracts
- **Retrieval**
 - Content-based retrieval based on speech transcript, image similarity search, metadata
 - Multilingual retrieval support for free text, and selected metadata fields
- **Metadata model**
 - The model supports the video description at different levels and includes the management of metadata automatically extracted or associated manually
 - Support of multilingual indexing
- **Result presentation**
 - Support of multiple visualization modes of retrieved videos
 - Metadata only
 - Key frames
 - Speech transcripts

Consiglio Nazionale delle Ricerche

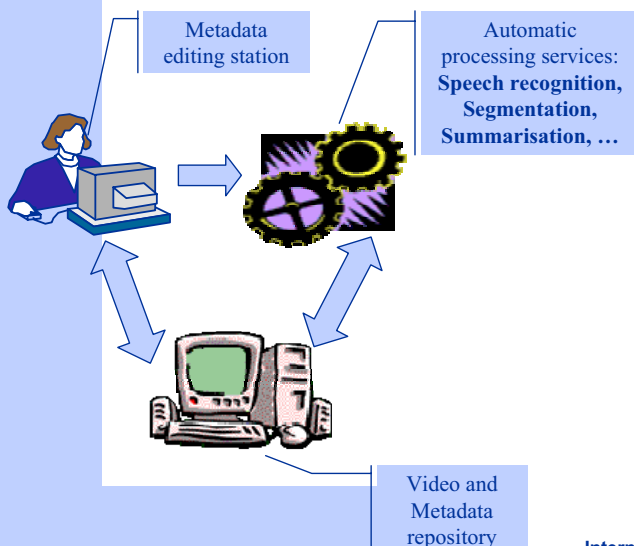
International Conference on Digital Archive Technologies (ICDAT 2006)

Taipei, Taiwan, 19-20 October 2006

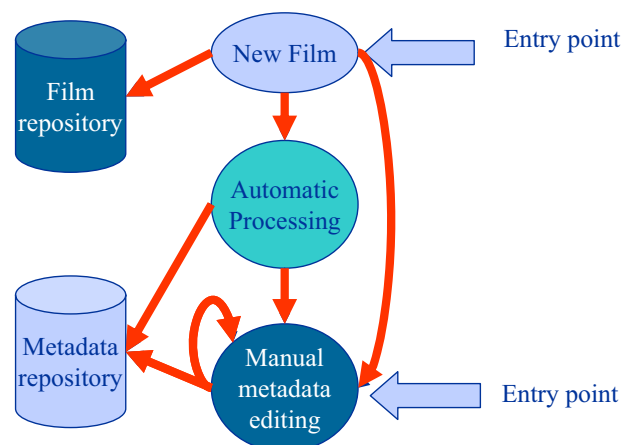
Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Video Indexing

Main components:



Indexing Workflow:



Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Video Indexing

- **Automatic** vs **manual** indexing
- **Manual Indexing** of all bibliographic metadata
 - Background information, e.g.
 - Creation date
 - Author
 - Names of the actors
 - Semantic information
 - Relations among different shots
 - Interpretation of the meaning of a shot
 - Interpretation of the meaning of a frame
 - All this type of information is provided manually, by using a specific tool, the **ECHO metadata editor**

Video Indexing

- **Automatic indexing**
 - Automatic indexing is expected to be
 - Fast
 - Reliable (user independent, error reduction)
 - Features that can be extracted from the **entire video**,
 - e.g. frame rate, resolution, b&w or color video, etc.
 - Features that are associated to the **audio part**
 - e.g. the transcript of the speech
 - Features that can be extracted from each **shot**
 - e.g. object track, camera movement, recognition of specific objects, recognition of faces, text captions, key frames
 - Features that can be extracted from each **frame**
 - typical image features, e.g. color distribution, texture, object's shapes
 - **Object recognition, text extraction, face detection**

Text recognition inside key frames



OCR

IN DIRETTA

L'ADDIO DELL'ITALIA AI
CADUTI DI NASSIRIYA

SKY TG24

11.36

Identification

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Speech recognition

- The purpose of speech recognition is the **generation of a transcript** to be used as a **support for retrieval**
- Main functionality required
 - Speaker independent
 - Multiple languages
 - Operating also with low quality audio
- Does not require perfect recognition
 - Retrieval quality is **acceptable for W.E.R. up to 30-40%**

Consiglio Nazionale delle Ricerche

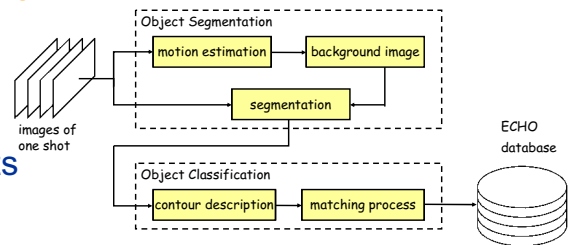
International Conference on Digital Archive Technologies (ICDAT 2006)

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Object detection and recognition

- The system for moving-object recognition consists of two components, a **segmentation module** and a **classification module**.
- For each shot in the video, a background panorama image is constructed. The foreground objects in this background image are removed by means of temporal filtering (median).
- The object is segmented by comparing each frame of the video to the background image.



Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Example: Cars



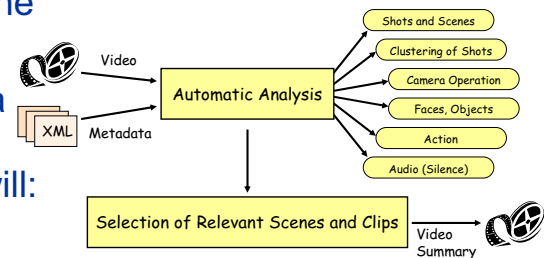
Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Video abstract generation

- A video abstract is a part of a much longer video, which preserves the essential message of the original video.
- A video abstract does not change the presentation medium.
- The users get a quick overview of a much longer video.
- The video abstracting application will:
 - Select relevant clips
 - Order these clips
 - Define a transition between two clips
 - Modify the audio track

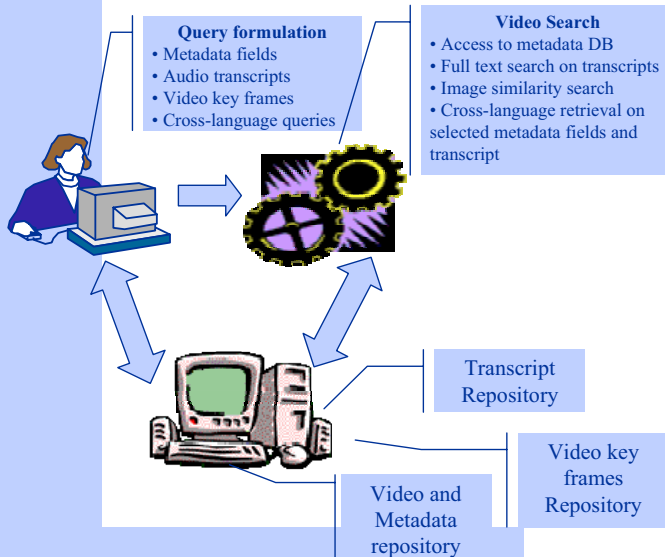


ECHO System Functionality

- **Indexing**
 - Automatic indexing: key frames, object recognition, text extraction, face detection
 - Manual semantic indexing through a specialized video metadata editor
 - Automatic speech recognition for multiple languages (italian, dutch, german, english)
 - Automatic generation of video abstracts
- **Retrieval**
 - Content-based retrieval based on speech transcript, image similarity search, metadata
 - Multilingual retrieval support for free text, and selected metadata fields
- **Metadata model**
 - The model supports the video description at different levels and includes the management of metadata automatically extracted or associated manually
 - Support of multilingual indexing
- **Result presentation**
 - Support of multiple visualization modes of retrieved videos
 - Metadata only
 - Key frames
 - Speech transcripts

Video Retrieval

Main components:



Consiglio Nazionale delle Ricerche

Query examples:

- Metadata associated to the entire video
 - E.g. *find b&w videos produced before II world war by Istituto Luce*
- Metadata associated to video shots
 - E.g. *find a shot where the audio transcript contains the words "Attentato Banca Nazionale dell'Agricoltura"*
- Metadata associated to single frames
 - E.g. *find a video that contains a frame similar to this image [the image is provided as an example]*
- Any combination of the previous cases

International Conference on Digital Archive Technologies (ICDAT 2006)

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

Retrieval characteristics

- Retrieval is based on **queries expressed on metadata values**. Automatically extracted metadata, as well as metadata associated manually to the video can be used.
- Retrieval is based on an **approximate match** between the query and the retrieved videos. This is mainly the case when imprecise query elements are used (e.g. free text, images)
- Retrieved videos are returned to the user in decreasing relevance order, possibly indicating the **degree of relevance** of the retrieved items.

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

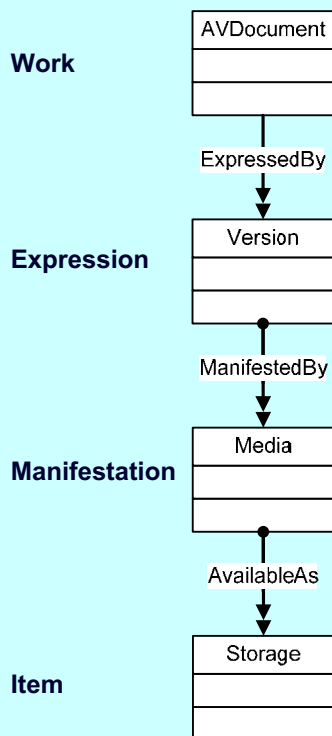
ECHO System Functionality

- **Indexing**
 - Automatic indexing: key frames, object recognition, text extraction, face detection
 - Manual semantic indexing through a specialized video metadata editor
 - Automatic speech recognition for multiple languages (italian, dutch, german, english)
 - Automatic generation of video abstracts
- **Retrieval**
 - Content-based retrieval based on speech transcript, image similarity search, metadata
 - Multilingual retrieval support for free text, and selected metadata fields
- **Metadata model**
 - The model supports the video description at different levels and includes the management of metadata automatically extracted or associated manually
 - Support of multilingual indexing
- **Result presentation**
 - Support of multiple visualization modes of retrieved videos
 - Metadata only
 - Key frames
 - Speech transcripts

The ECHO Metadata Model

- The ECHO metadata model is an **extension of the IFLA/FRBR model**. IFLA/FRBR is structured into four levels which describe all aspects of an intellectual or artistic object: **Work, Expression, Manifestation** and **Item**.
- The entities of ECHO metadata model are organized according to a **hierarchy** which is strictly related to these four levels: there are four classes – **AVDocument, Version, Media** and **Storage** – which correspond to the four levels of IFLA/FRBR.

The structure of the ECHO metadata model



The AVDocument describes a distinct intellectual or artistic creation. It is the abstract idea of a creation. We do not specify if we realize a book, a film, or a cartoon, which is described by the Expression Entity.

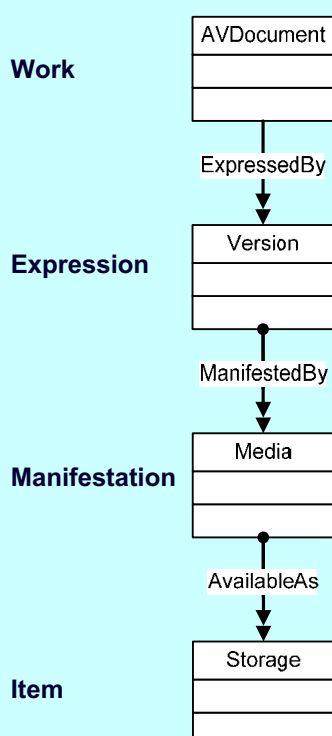
Examples of WORK are a documentary that describes the "Olympic Games in Berlin on 1936", or "2001: A space Odyssey".

Consiglio Nazionale delle Ricerche

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

The structure of the ECHO metadata model



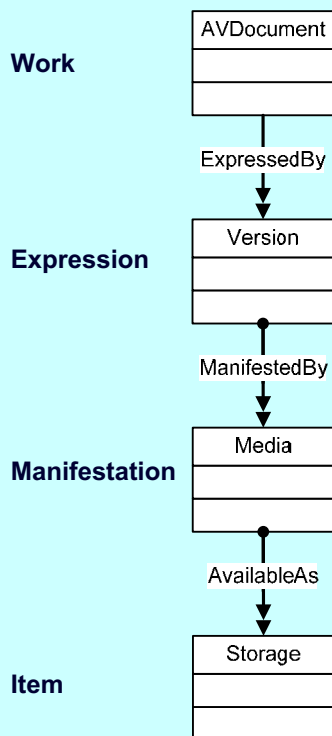
This is the intellectual or artistic realization of a work. It can be used to represent the different versions of a video, e.g. the Italian and English versions, or a reduced and a complete version, etc.

Consiglio Nazionale delle Ricerche

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

The structure of the ECHO metadata model



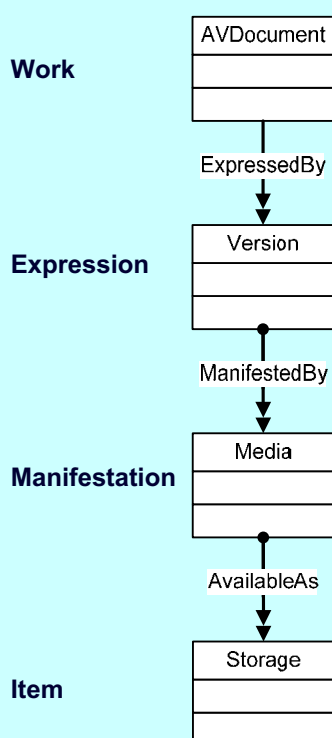
The Version entity does not represent a specific realization of the video. This is expressed by the Media entity. For example, the Italian version of the documentary on Olympic Games, may have two media entities, one coded in MPEG and the second in DivX.

Consiglio Nazionale delle Ricerche

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

The structure of the ECHO metadata model



Media entity does not correspond to the physical implementation, which is described by the Storage entity type. This describes the single copies of the videos.

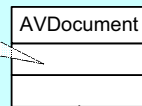
Consiglio Nazionale delle Ricerche

Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

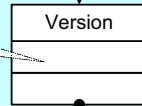
The structure of the ECHO metadata model

Title
EventDate
Director
Censorship
Description
....



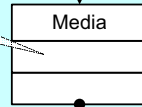
ExpressedBy

VersionTitle
Duration
...



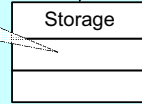
ManifestedBy

Format
Size
...



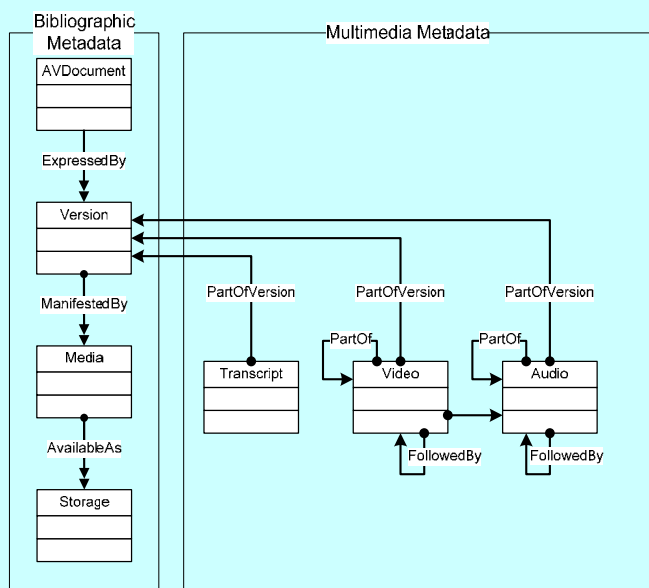
AvailableAs

Collocation
Provider
StorageID
PublicAccess



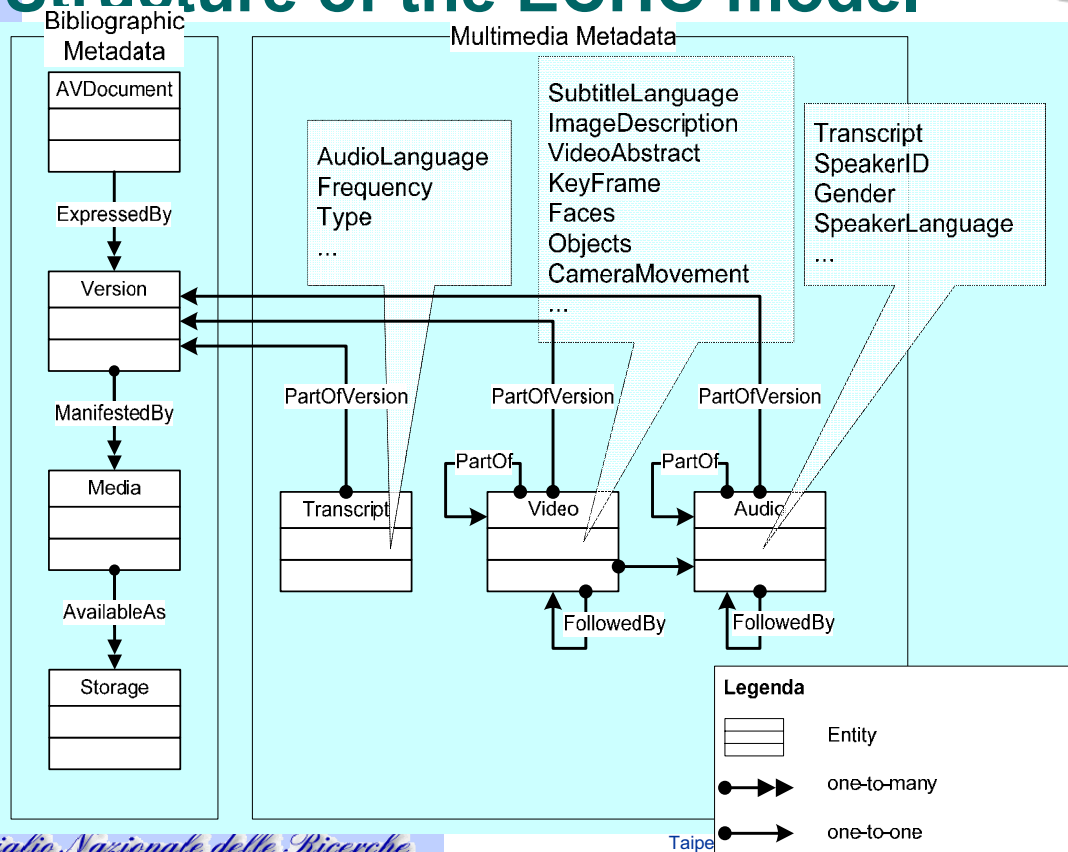
Each entity has a number of metadata attribute fields used to describe the video document.

The (complete) structure of the ECHO metadata model



- Entities are divided in “Bibliographic Metadata” and “Multimedia Metadata”, composed of *Video*, *Audio* and *Transcript*. They belong to the Expression level of the IFLA/FRBR model.
- Each version object is composed of a *Video*, *Audio* and a *Transcript*.
- These objects represent the entire audio/video object.
- Smaller segments (e.g. a shot or a scene) can be represented by a new instance of Video, Audio, and Transcript.
- Objects at the same level are linked (FollowedBy) in order to represent the temporal sequence of different objects.

Structure of the ECHO model

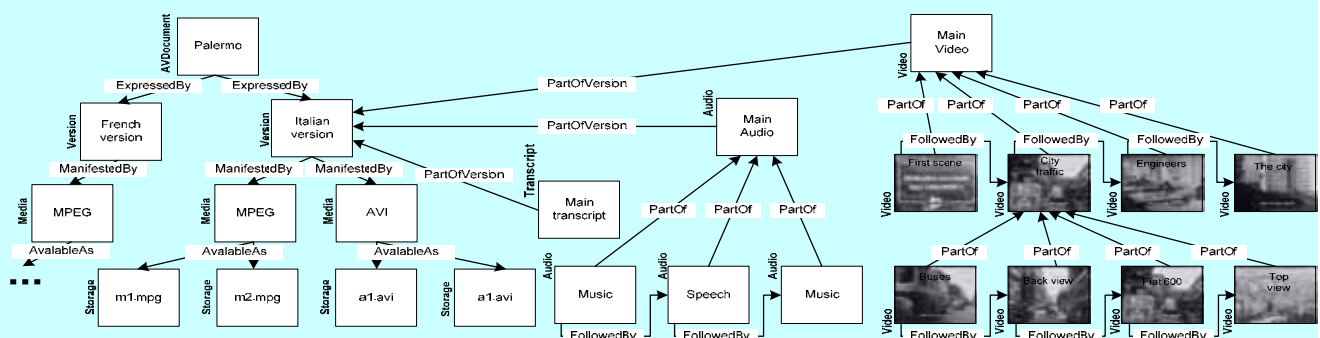


Consiglio Nazionale delle Ricerche

Taipei

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

An example

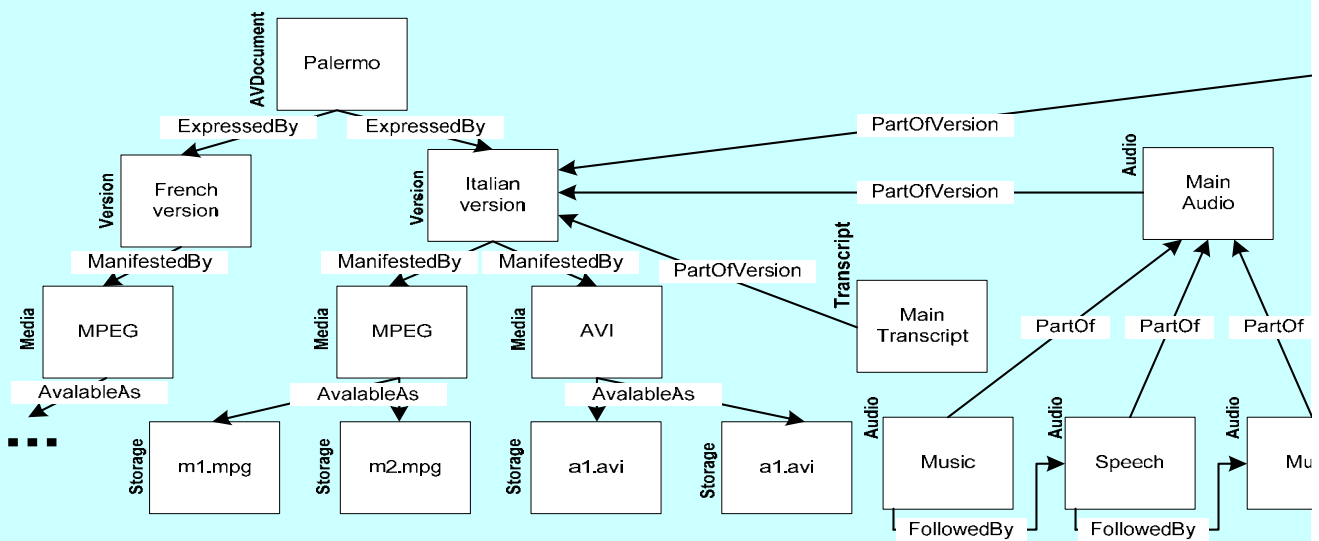


Consiglio Nazionale delle Ricerche

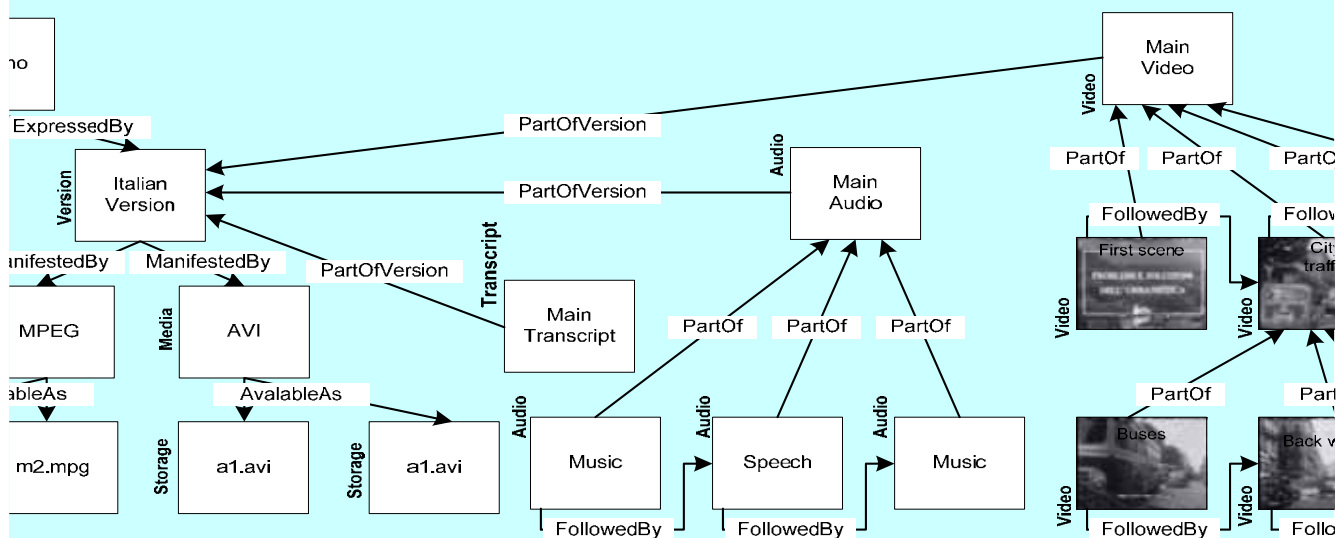
Taipei, Taiwan, 19-20 October 2006

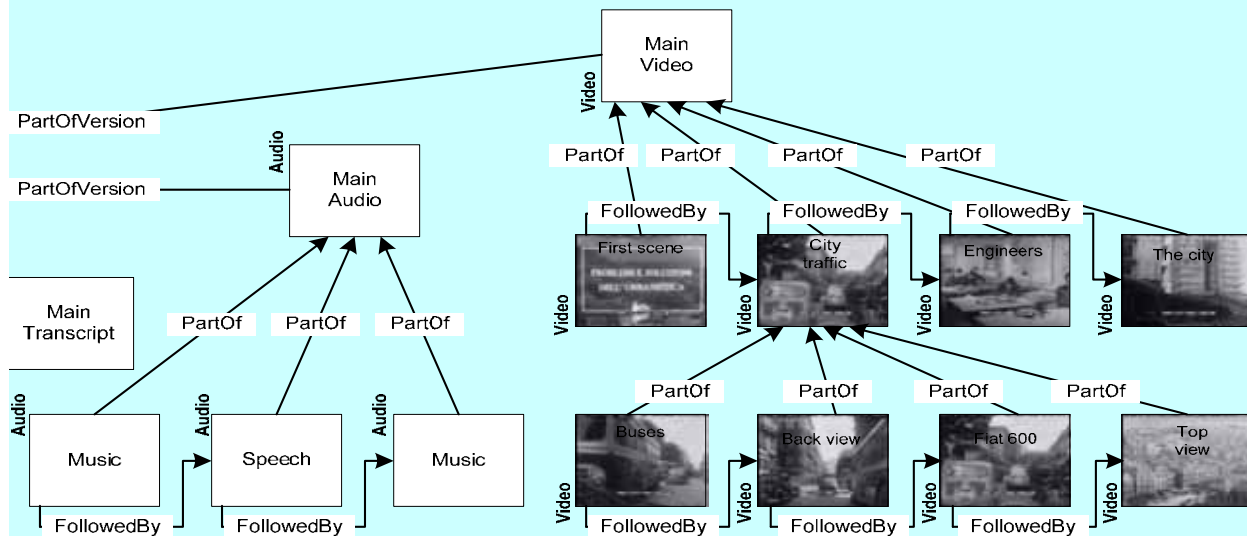
Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

An example



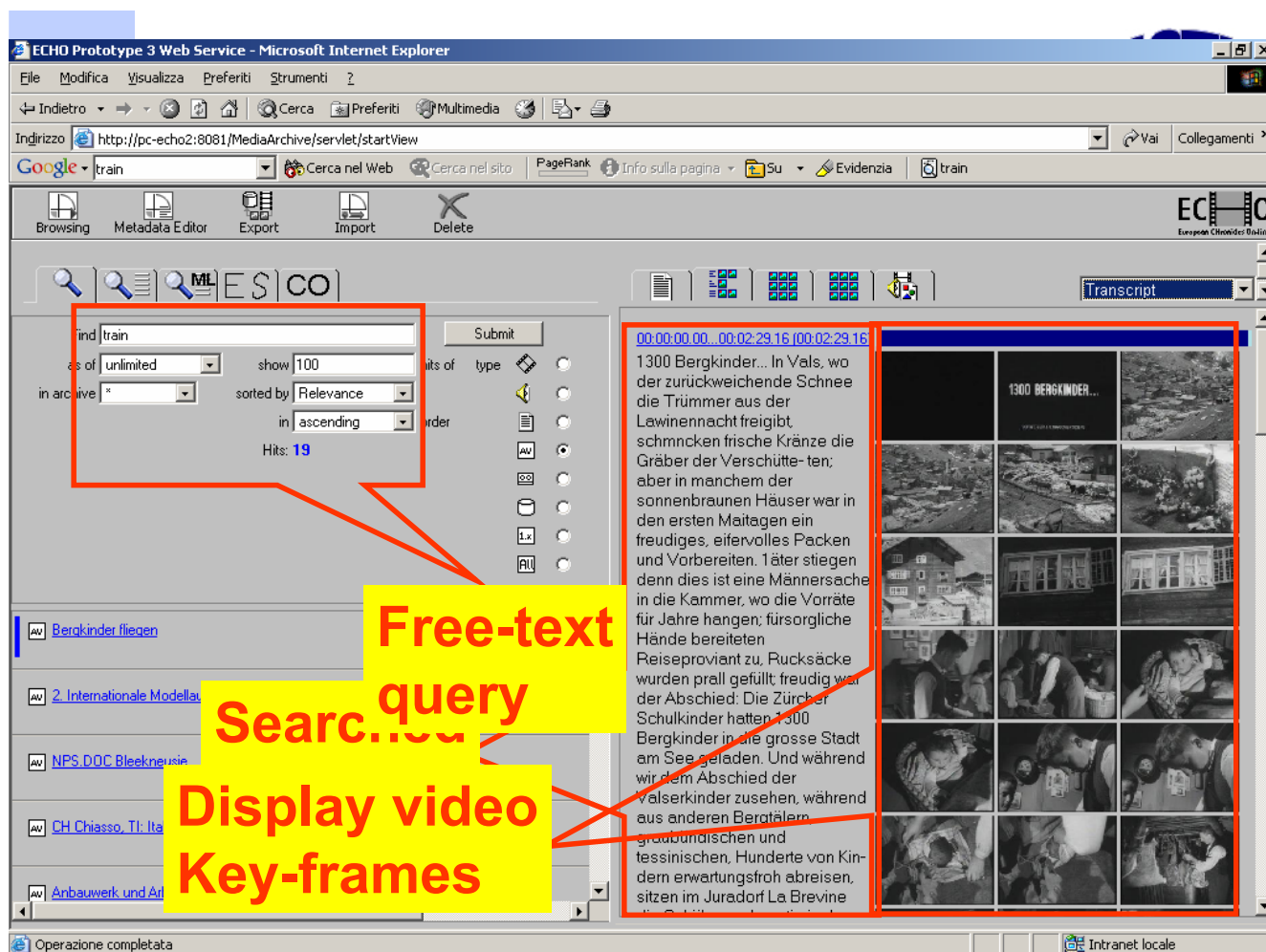
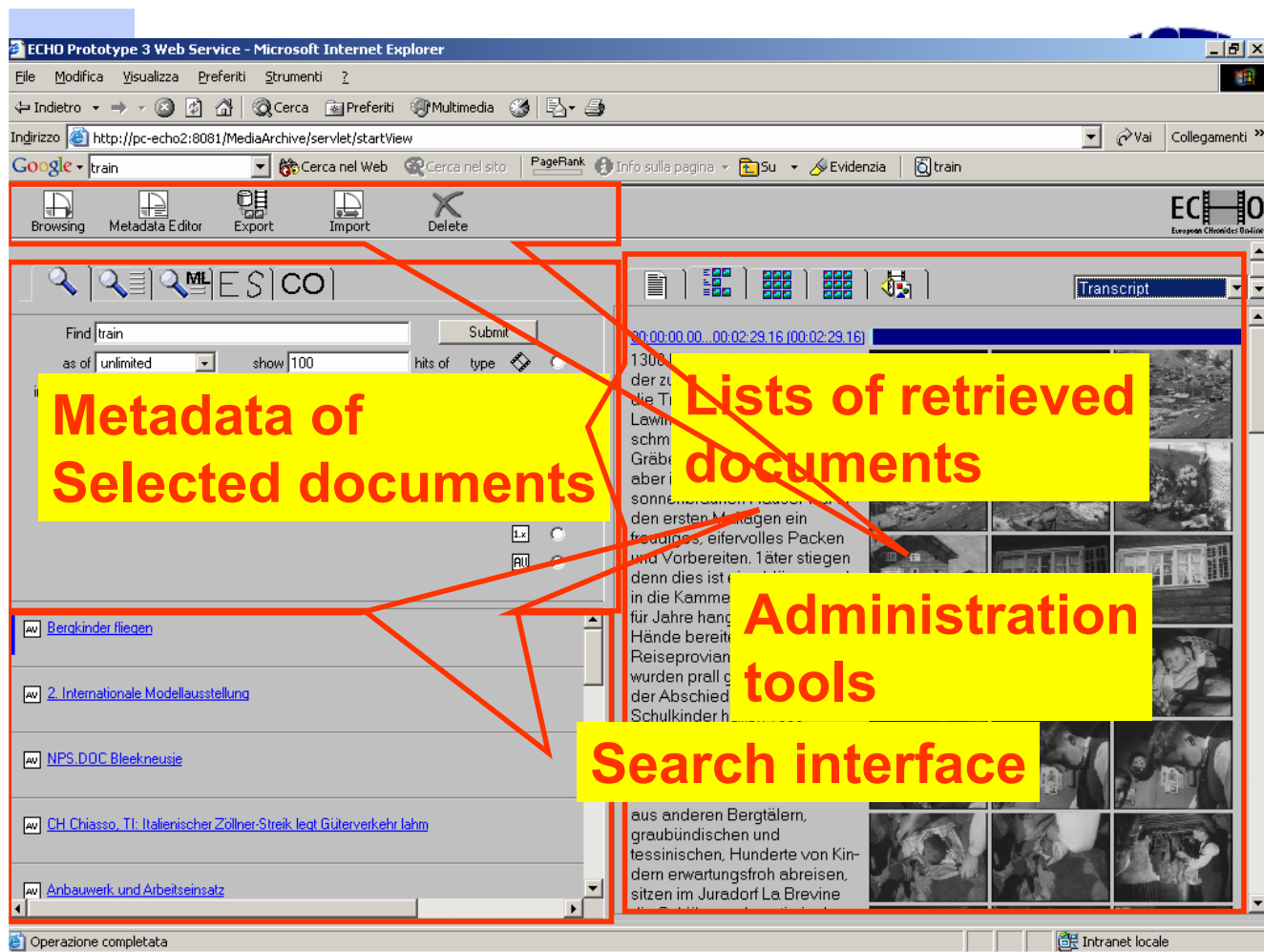
An example

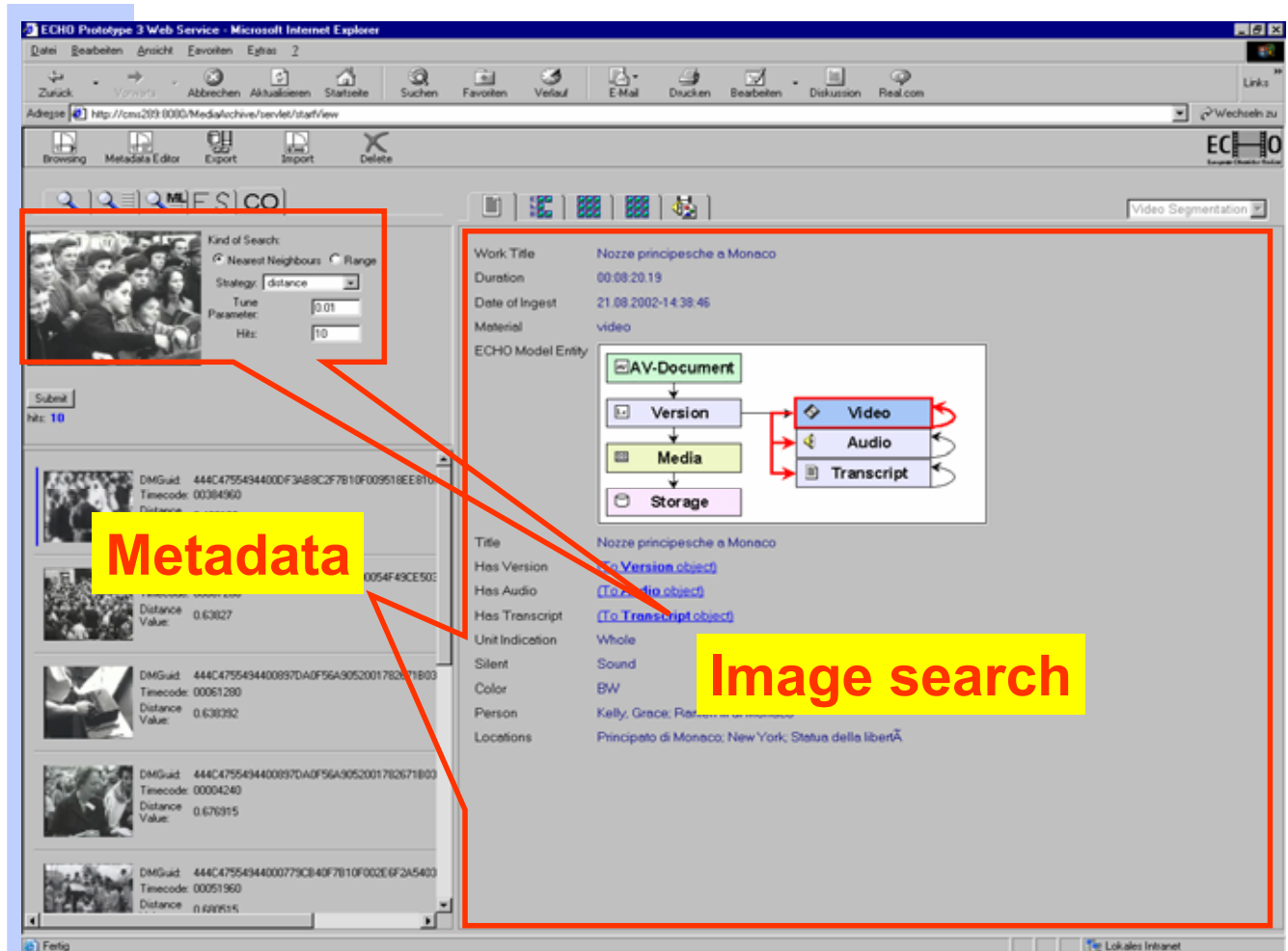
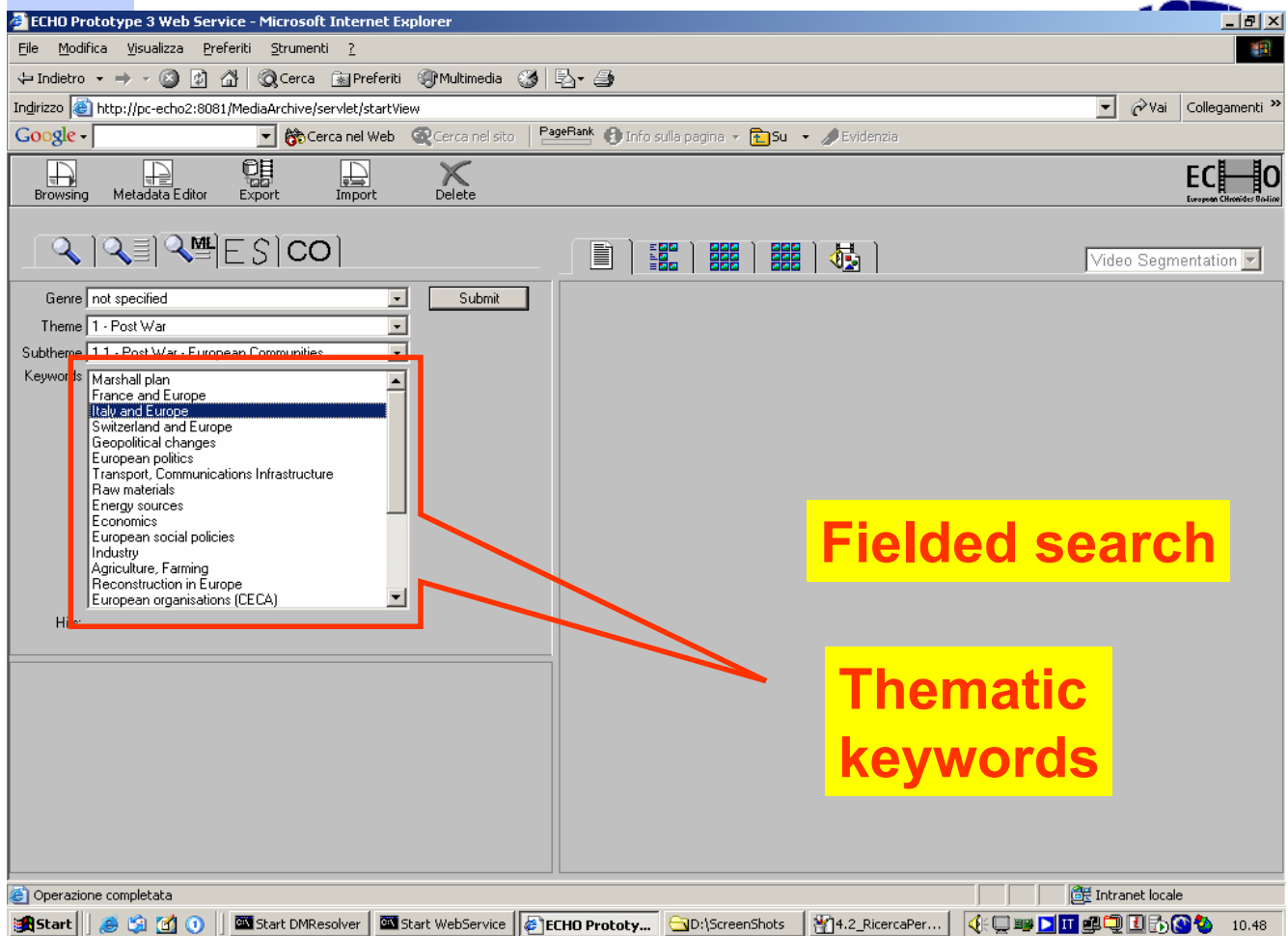


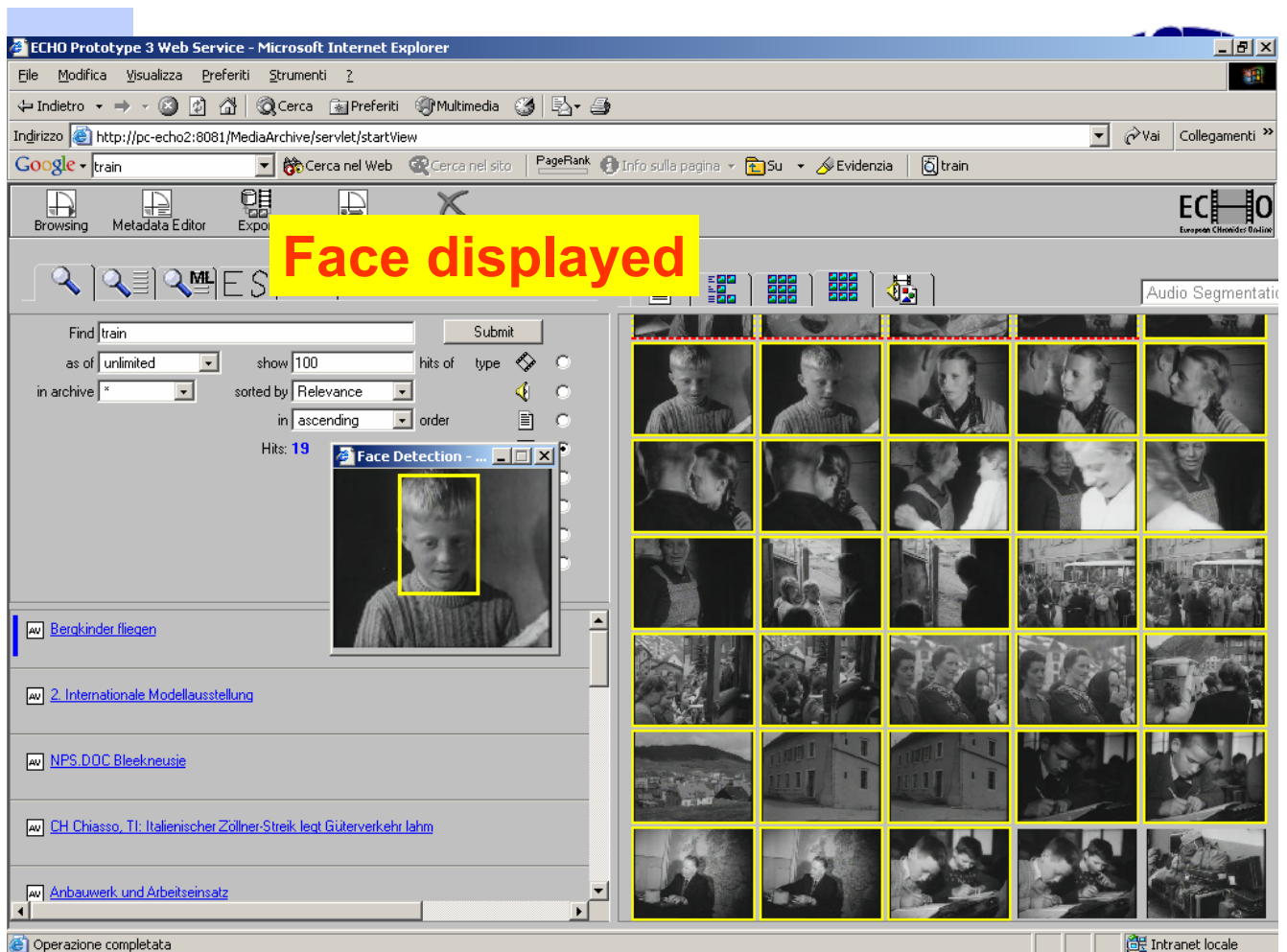
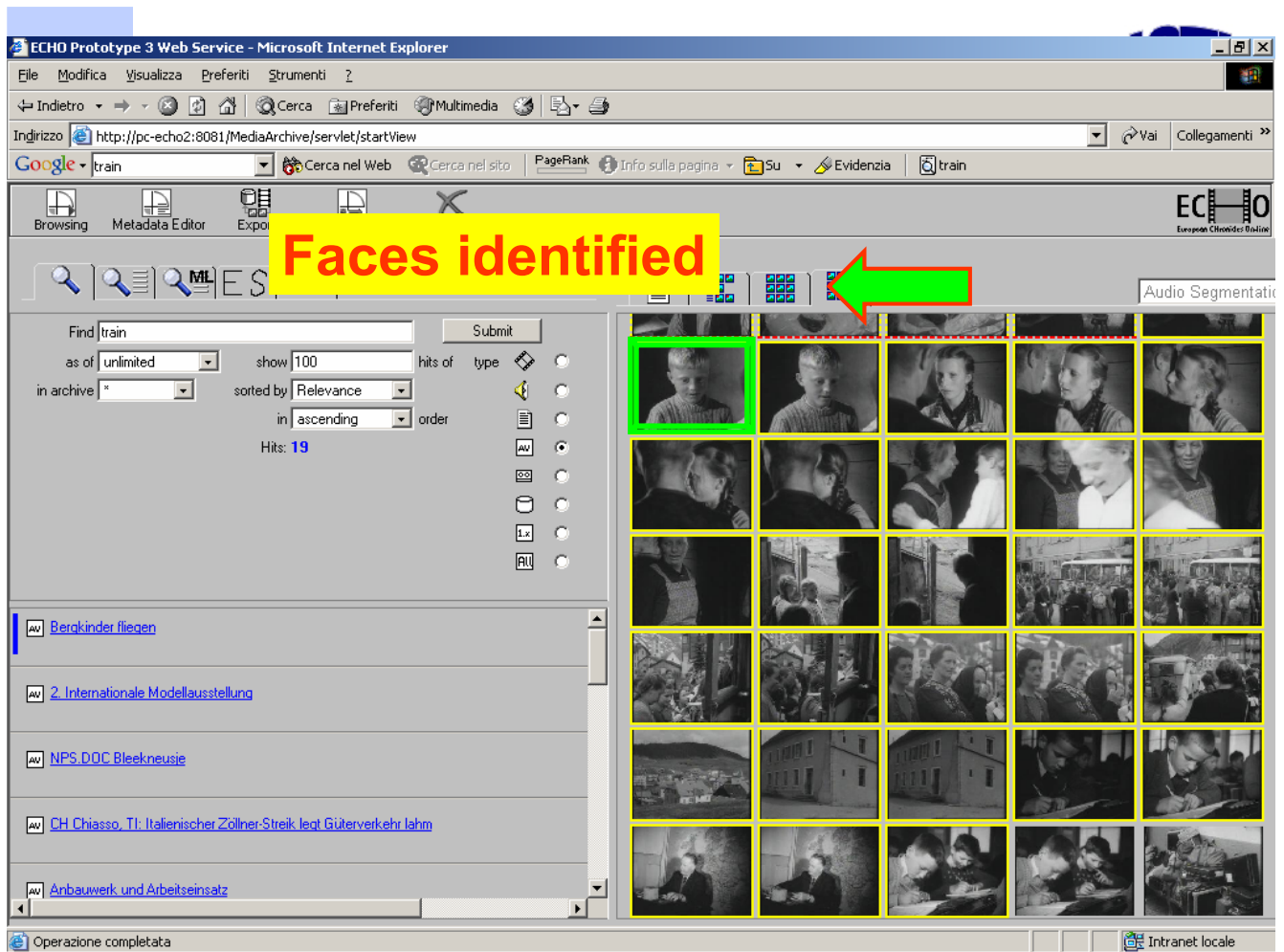


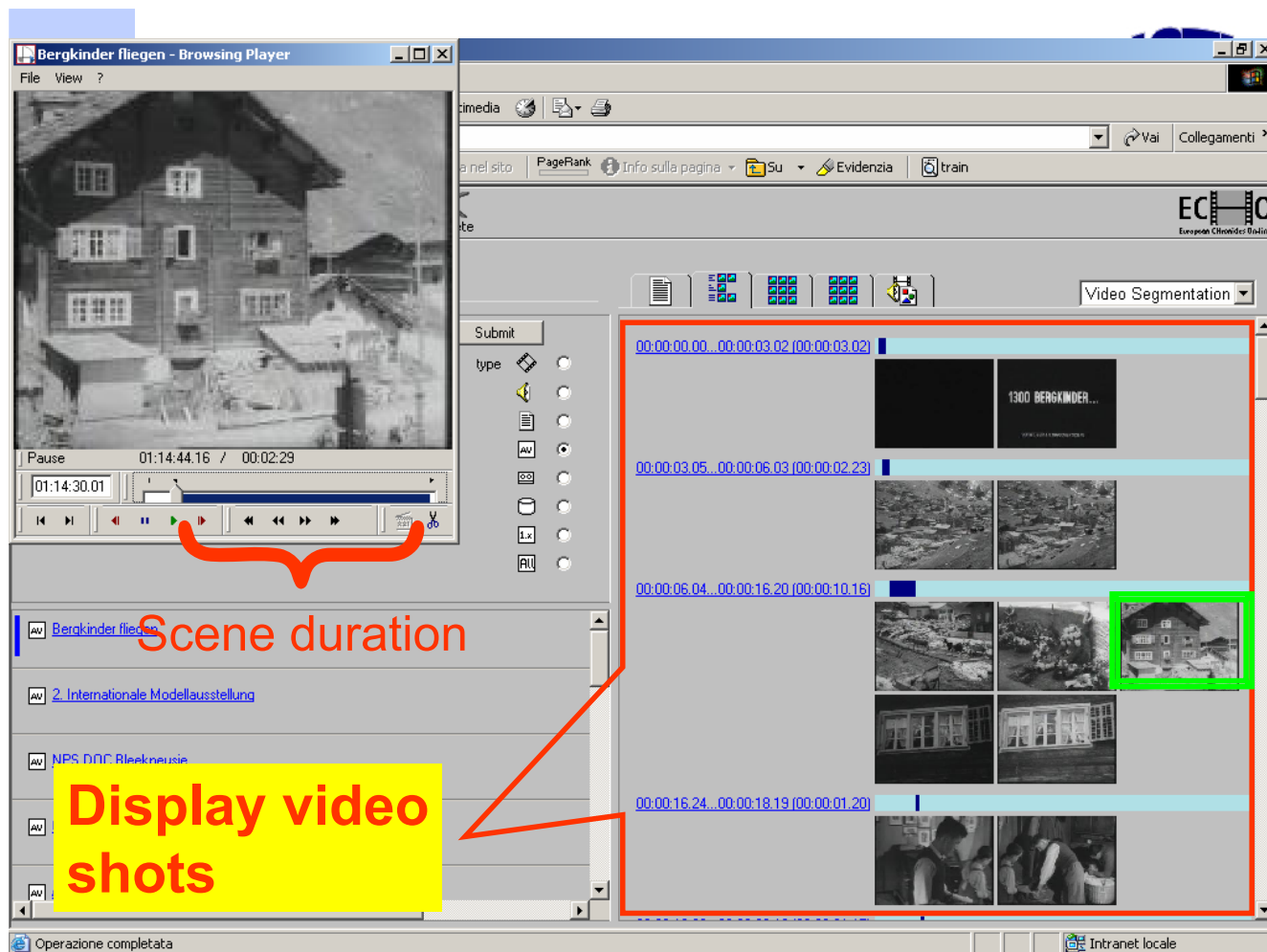
ECHO System Functionality

- **Indexing**
 - Automatic indexing: key frames, object recognition, text extraction, face detection
 - Manual semantic indexing through a specialized video metadata editor
 - Automatic speech recognition for multiple languages (italian, dutch, german, english)
 - Automatic generation of video abstracts
- **Retrieval**
 - Content-based retrieval based on speech transcript, image similarity search, metadata
 - Multilingual retrieval support for free text, and selected metadata fields
- **Metadata model**
 - The model supports the video description at different levels and includes the management of metadata automatically extracted or associated manually
 - Support of multilingual indexing
- **Result presentation**
 - Support of multiple visualization modes of retrieved videos
 - Metadata only
 - Key frames
 - Speech transcripts









ECHO limitations

- Supports the management of **video data only**
- Based on a **proprietary metadata model**
- **Efficiency** limitations
- Complex system architecture



Evolution

- We built a general purpose **Multimedia Content Management System (MILOS)** that supports the archiving, indexing and content based retrieval of many different types of data
- MILOS supports the **interoperability** of different metadata models
- We built, on top of MILOS, an A/V Digital Library with the same functionality offered by ECHO

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

The MILOS MCMS

- Search capabilities:
 - **Traditional** fielded search capabilities
 - **Full text search** (e.g. on video transcripts)
 - Search on automatically associated **classification categories**
 - **Visual content** similarity search
- The system is **not tied to a specific metadata schema**
 - **Any XML encoded metadata** can be managed by the system (e.g. DC, MPEG-7, ECHO, proprietary model)
 - **Metadata mapping techniques** are used to provide users with a homogeneous view
 - Several **different and heterogeneous applications** can be supported

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR

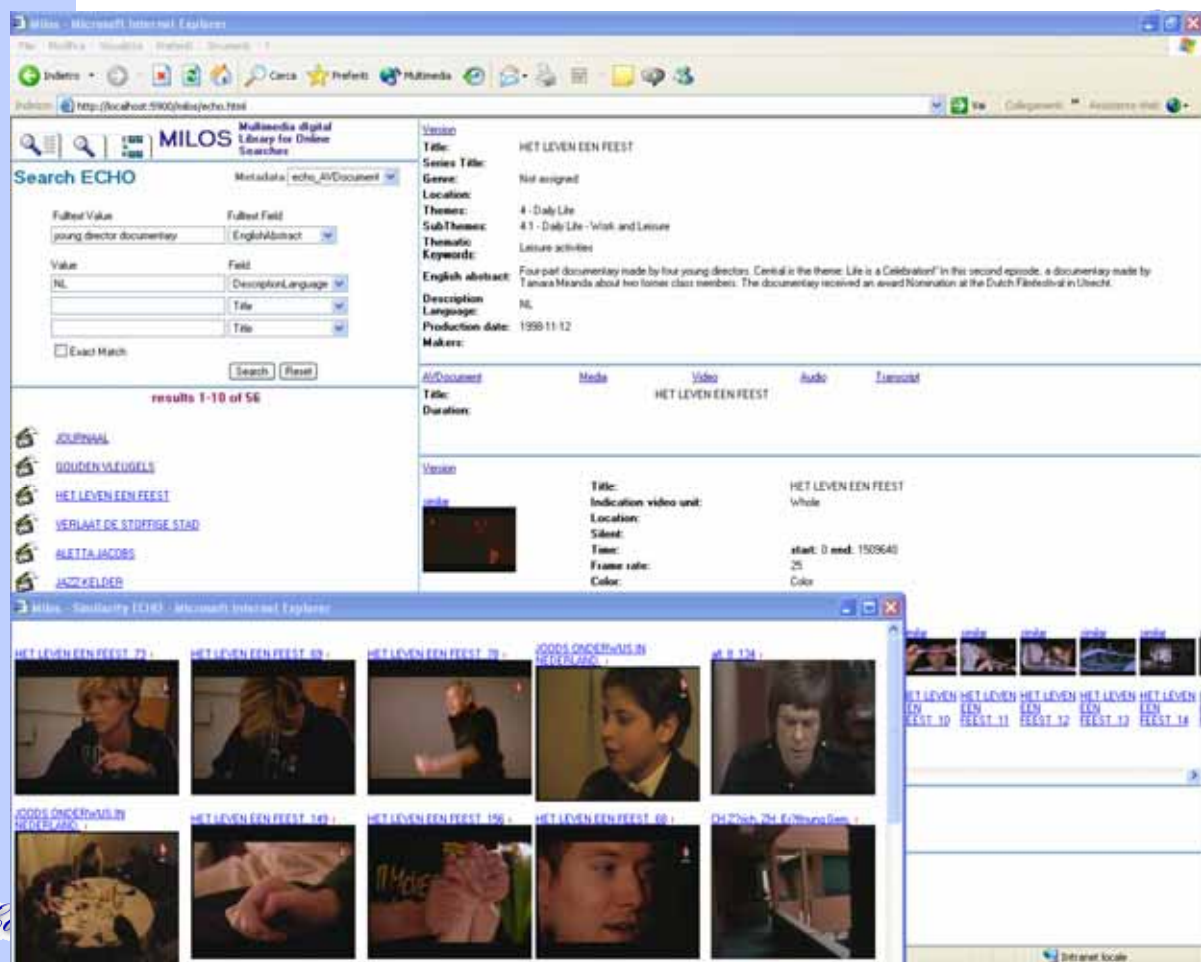
Digital Library Examples

- We have tested MILOS building the following DL applications
 - **ECHO**
 - 50 hour of A/V data with IFLA-FRBR and MPEG-7 metadata (21 Gb of MPEG-1, 43.000 XML files)
 - **REUTERS**
 - 810.000 XML encoded, news agencies (2.6 Gb)
 - **DBLP and SIGMOD Records**
 - 187 Mb of XML files
 - **Image archive**
 - 1000 Color images, with MPEG-7 visual descriptor metadata
 - Each application was developed on average in a week
 - User queries are processed in less than a second on average

Consiglio Nazionale delle Ricerche

International Conference on Digital Archive Technologies (ICDAT 2006)
Taipei, Taiwan, 19-20 October 2006

Pasquale Savino
(pasquale.savino@isti.cnr.it)
ISTI-CNR





References

- <http://pc-erato2.iei.pi.cnr.it/echo/>
- <http://milos.isti.cnr.it/>
- My e-mail address:
 - pasquale.savino@isti.cnr.it