Indexing A/V Documents

Giuseppe Amato
Claudio Gennaro
Pasquale Savino

ISTI-CNR
Pisa, Italy
g.amato@isti.cnr.it

Overview

- Metadata
  - Dublin core
  - MPEG-7
  - IFLA-FRBR / ECHO
- Editing Metadata
- Automatic indexing
  - Text, speech, images, moving pictures

Metadata

- Metadata: data about data
  - Structured information about data
- Types of metadata
  - Resource discovery
  - Right management
  - Content rating
  - Archival status
  - Etc…

Metadata

- Manual generation
  - Time consuming: high cost
  - Detailed metadata, if generated by experts
- Automatic generation
  - Fast: Reduced cost
  - Metadata contain noise
    - Imprecision, uncertainty

Metadata Models

- Dublin core
- MPEG – 7
- IFLA – FRBR / ECHO

Dublin core

- Flat model of 15 base elements:

  Title
  Creator
  Contributor
  Publisher
  Subject
  Description
  Identifier
  Date
  Language
  Type
  Format
  Coverage
  Source
  Relation
  Rights

Dublin core

- Additional detail through qualifiers
  - Element refinements
    - Es.: date.created, relation.isPartOf
- Extensions
  - Es.: audience element (Education, libraries, government)

Dublin core

- Core vocabulary of terms useful for description
- Cross domain discovery
  - It is not designed for a specific domain
- Interoperability
  - Different digital libraries can talk each other
- Known implementations
  - Open archive initiative
  - Many digital libraries projects
  - Open source and commercial tools
- http://dublincore.org/

Dublin core RDF/XML

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF PUBLIC "-//DUBLIN CORE//DCMES DTD
    2002/07/31//EN" "http://dublincore.org/documents/2002/07/31/dcmes-
    xml/dcmes-xml-dtd.dtd">

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
      <rdf:Description rdf:about="http://pc-erato2.iei.pi.cnr.it/amato">
            <dc:title>Giuseppe Amato's Home Page</dc:title>
            <dc:creator>Giuseppe Amato</dc:creator>
            <dc:publisher>ISTI-CNR</dc:publisher>
            <dc:date>2002-11-18</dc:date>
      </rdf:Description>
</rdf:RDF>
```

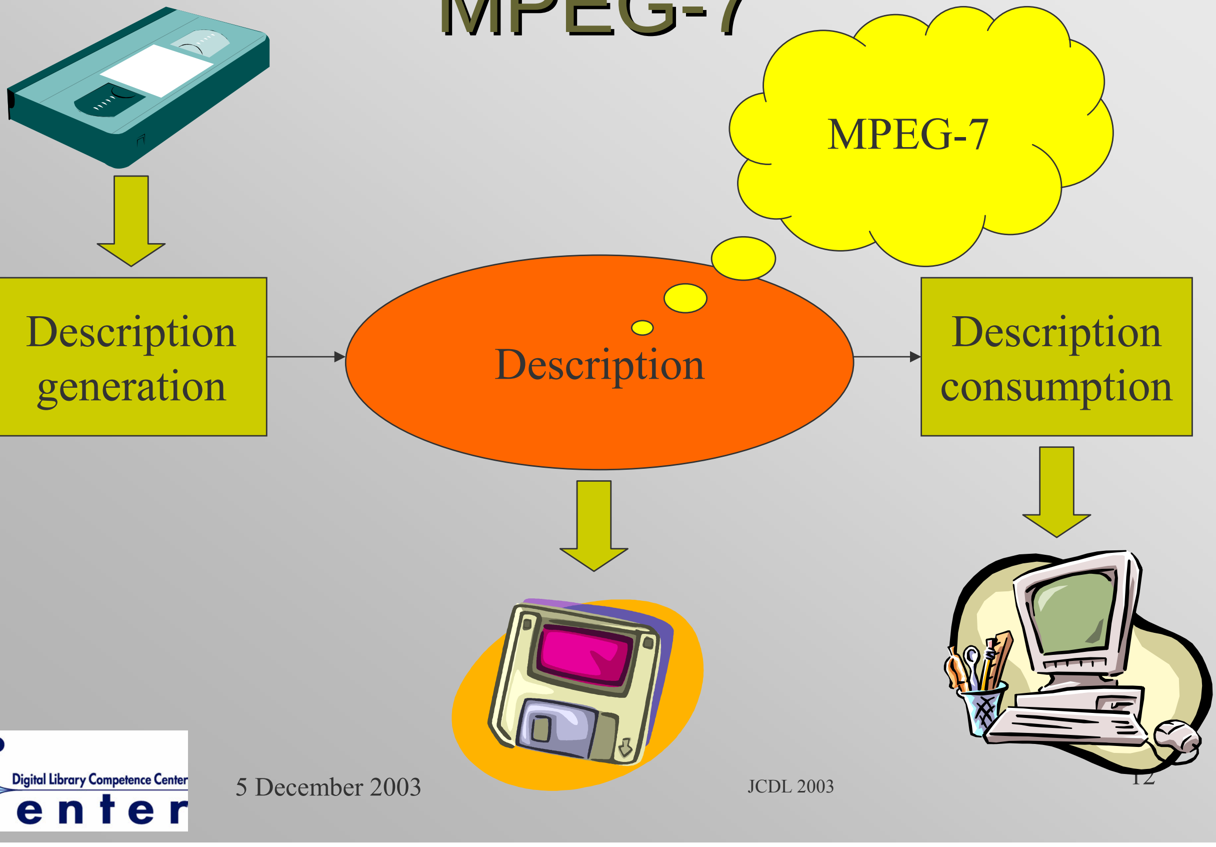
MPEG-7

- MPEG-7: standard developed by MPEG
- It is named “Multimedia content description interface”
- Describes multimedia content data
- A broad range of applications are supported
- It has been developed by experts representing
  - Broadcasters, electronic manufacturers, content creators, publishers, right managers, telecommunication service providers, and academia

MPEG-7

- Application scenarios:
  - Image understanding
  - Intelligent vision
  - Smart cameras/VCRs
  - Information retrieval
  - Information filtering
  - Digital libraries
  - Computer based training

MPEG-7



MPEG-7

- MPEG-7 components:
  - Descriptors (Ds)
    - Semantics and syntax of feature representation
  - Description schemas (DSs)
    - Structure and semantics of relations between Ds and other DSs
  - Description Tools
    - Set of Ds and DSs
  - Description Definition Language (DDL)
    - Defines new Ds and DSs and extends exiting ones

MPEG-7

- Standard description tools
  - MPEG-7 Visual
  - MPEG-7 Audio
  - MPEG-7 Multimedia Description Schemes

MPEG-7

- MPEG-7 Visual:
  - Visual description tools covering the following visual features:
    - Colour, texture, shape, motion, localisation, faces
    - There are elementary and sophisticated Descriptors

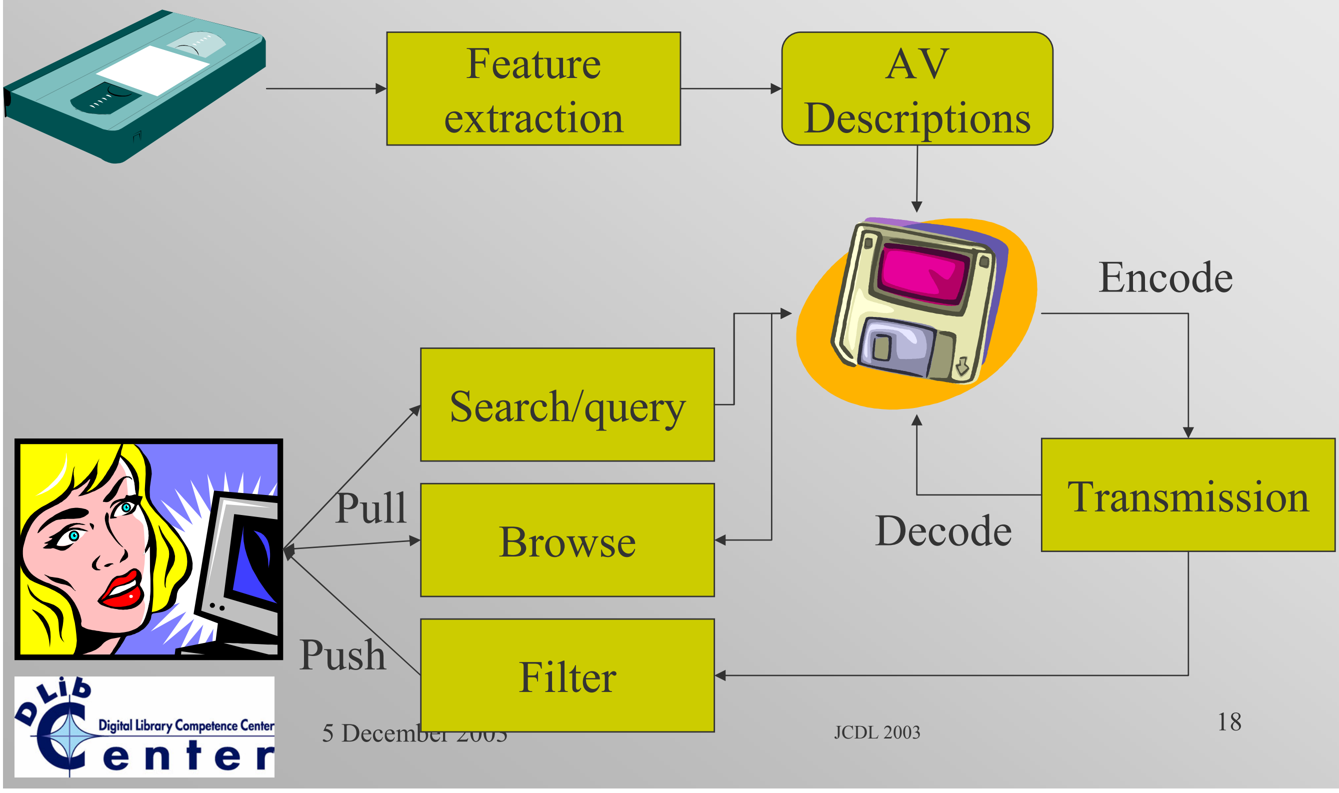
MPEG-7

- MPEG-7 Audio:
  - Audio description tools covering the following:
    - Descriptors:
      - spectral, parametric, temporal features
    - Description Tools:
      - sound recognition, instrumental timber, spoken content, audio signature, melody

MPEG-7

- MPEG-7 Multimedia Description Schemes:
  - Metadata generic structures for annotating audio-visual content:
    - Vector, time, textual, controlled vocabularies
    - Content description: perceivable information
    - Content management: creation, coding, usage
    - Content organisation: collections
    - Navigation and access: summaries, partitions, etc.
    - User interaction: user preferences, usage history

MPEG-7



Feature extraction

AV Descriptions

Encode

Transmission

Decode

Search/query

Browse

Filter

Pull

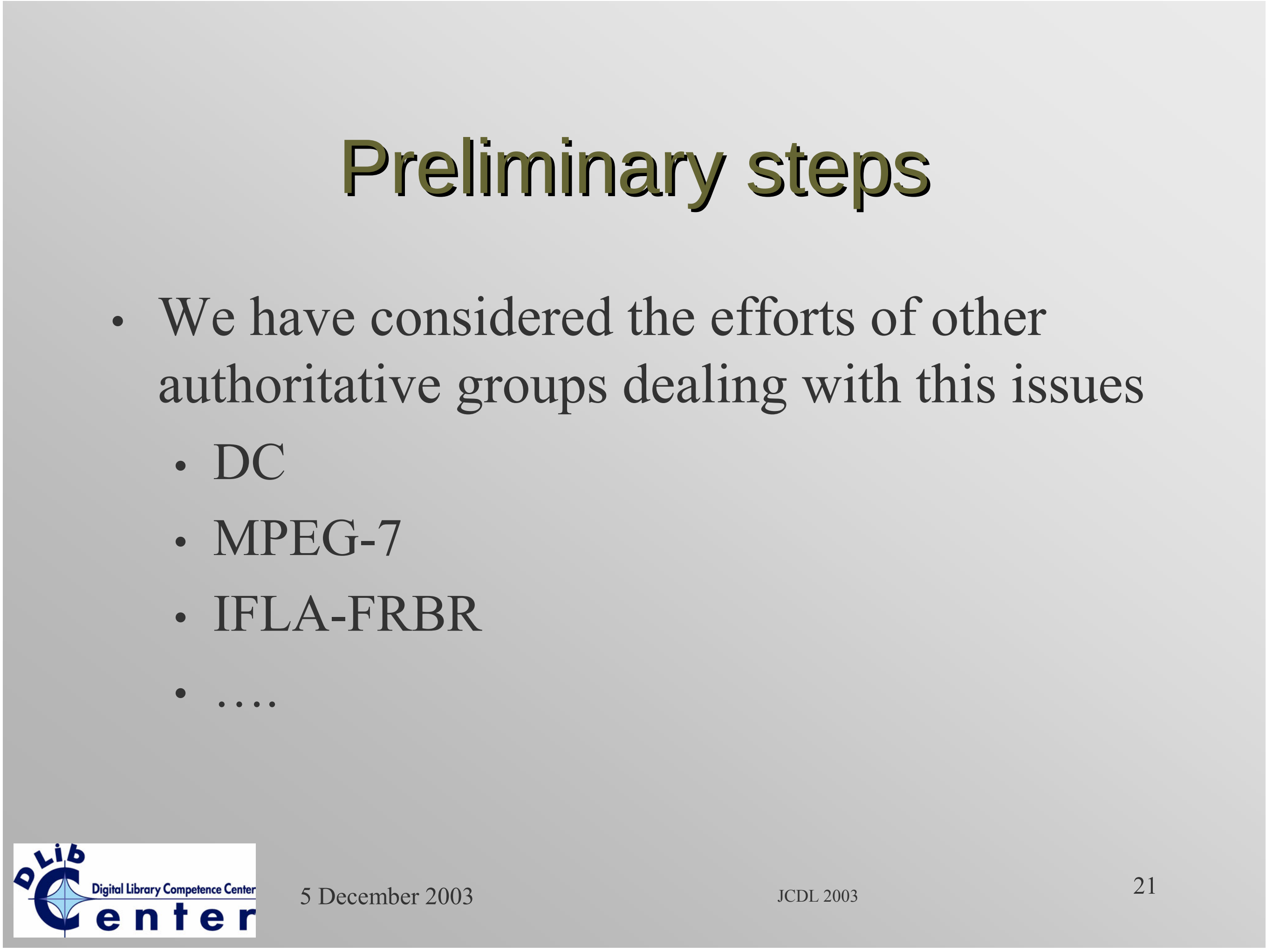
Push

ECHO Metadata Model

- This model originated from our experience in the **ECHO** project (**E**uropean **CH**ronicle **O**n-line)
  - ECHO is an EC funded **IST** project
  - ECHO aims at providing
    - remote access to collection of **historical documentary** audio-video resources
    - a software infrastructure to support digital video archives
    - an extensible and interoperable architecture

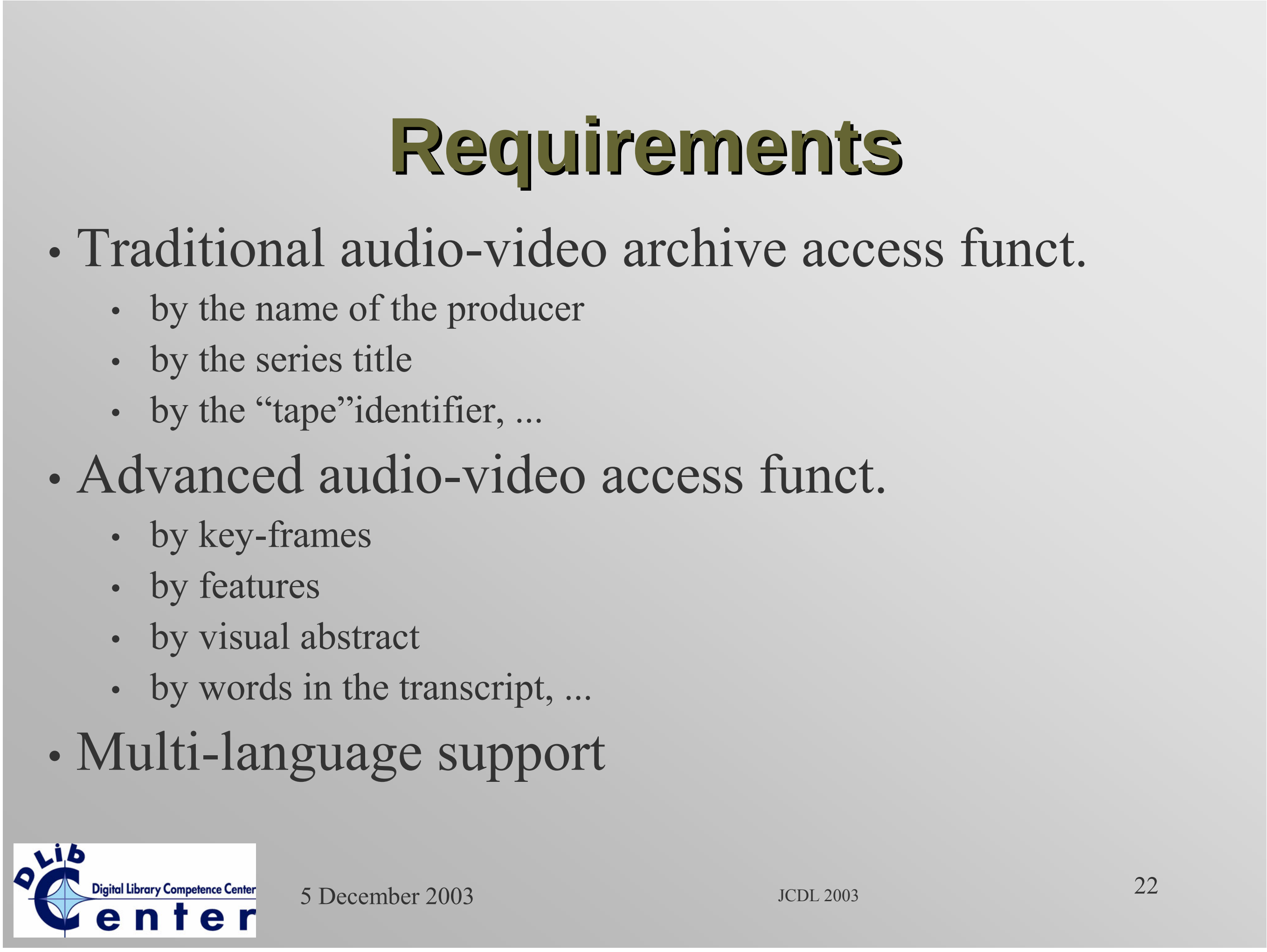
Preliminary steps

- We have interviewed
  - Content providers
    - Audio/visual archives
  - Technology providers
    - feature extraction, speech recognition, indexing, ...
  - End-Users
    - teachers, researchers, cultural heritage institutions...
- demand for a more detailed content description and advanced search capabilities



Preliminary steps

- We have considered the efforts of other authoritative groups dealing with this issues
  - DC
  - MPEG-7
  - IFLA-FRBR
  - ….



Requirements

- Traditional audio-video archive access funct.
  - by the name of the producer
  - by the series title
  - by the “tape”identifier, ...
- Advanced audio-video access funct.
  - by key-frames
  - by features
  - by visual abstract
  - by words in the transcript, ...
- Multi-language support

Requirements

- Specific metadata “fields” for
  - speech recognition processing
  - image/video processing
  - digital video abstracting

to provide advanced search facilities

Requirements

- Each data provider has its own way of cataloguing
- Must be harmonised
  - in general they describe:
    - logical content, using free keywords
    - physical content
    - cataloguing information

The approach

- Hierarchical and Multi-level design
  - Provides support for
    - interoperability
      - by using specialisation and generalisation
    - needs of special interest user communities
      - by using multiple view descriptions

The approach

Extends the IFLA-FRBR model

Four entities used to describe different aspect of a resource:

- WORK
  Describes a distinct intellectual or artistic creation
- EXPRESSION
  Intellectual or artistic realization of a work
- MANIFESTATION
  Physical embodiment of an expression
- ITEM
  A single exemplar of a manifestation

Describes a distinct intellectual or artistic creation

It is the abstract idea of a creation

Intellectual or artistic realisation of a work in the form of alphanumeric, musical, or choreographic notation, sound, image, etc..

Physical embodiment of an expression

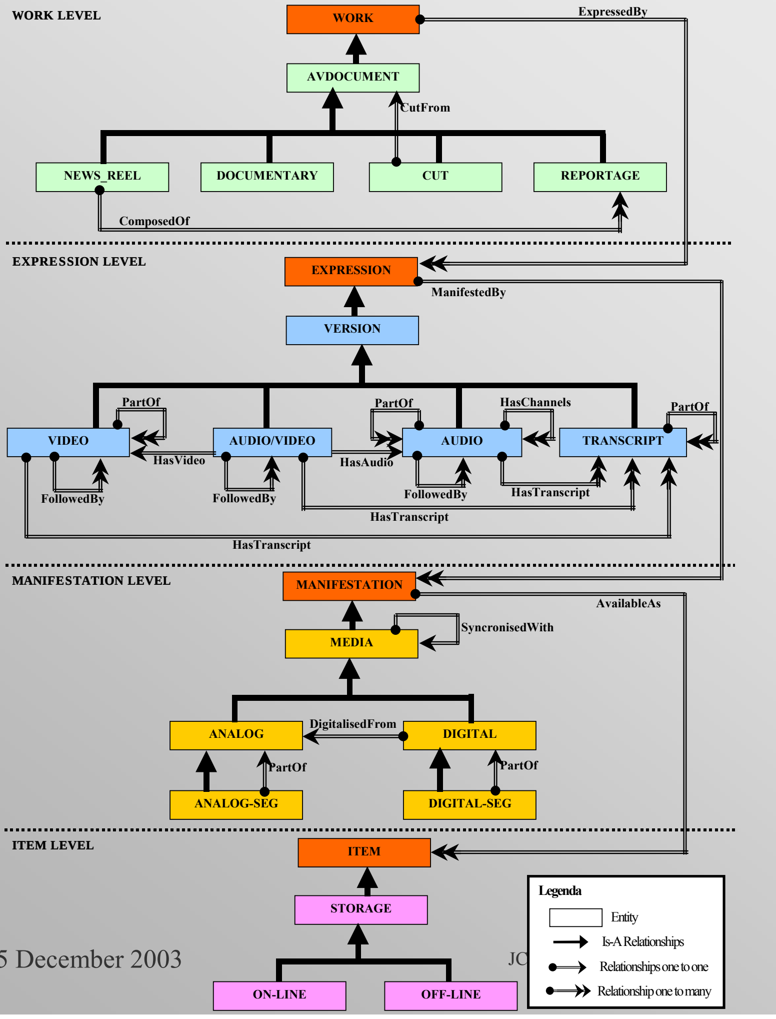
E.g. manuscripts, books, maps, sound, CD_ROM

A single exemplar of a manifestation

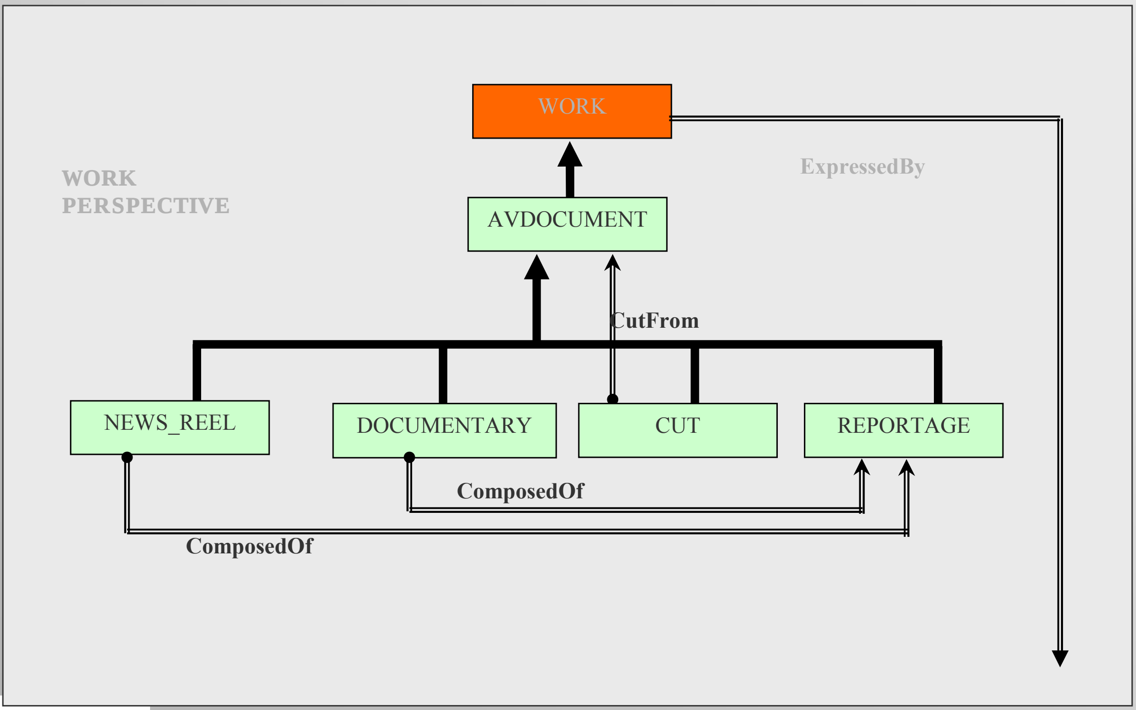
A documentary on the terrorist attack

Interviews on the terrorist attack

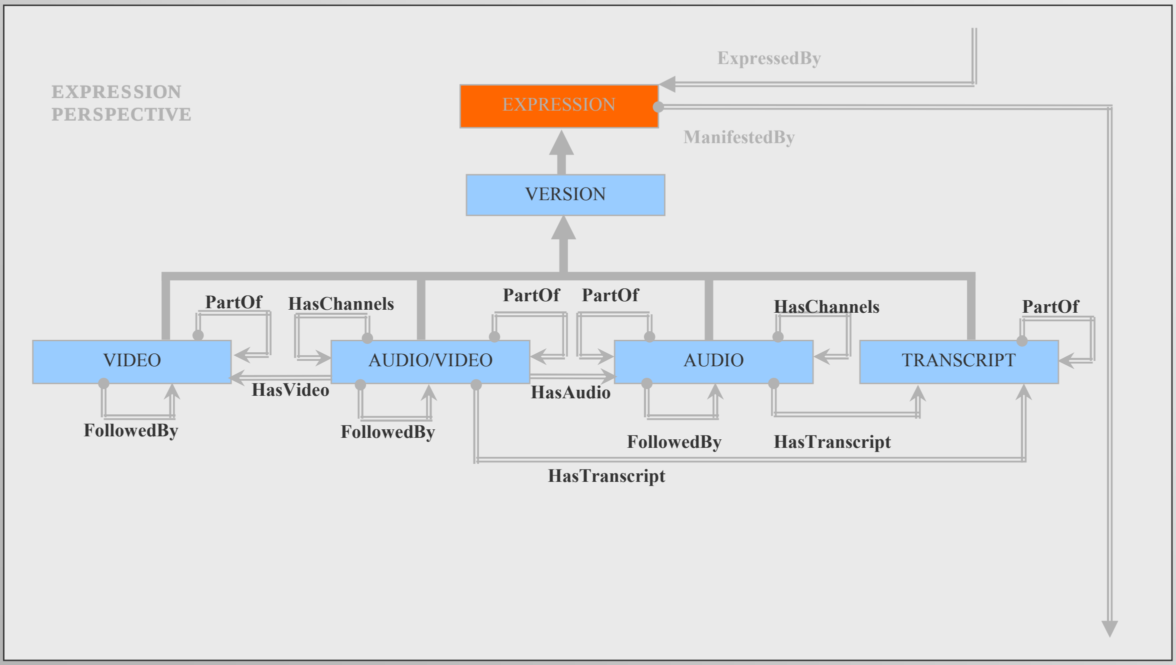
………


WORK LEVEL
WORK
ExpressedBy
AVDOCUMENT
CutFrom
NEWS_REEL
DOCUMENTARY
CUT
REPORTAGE
ComposedOf
EXPRESSION LEVEL
EXPRESSION
ManifestedBy
VERSION
PartOf
PartOf
HasChannels
PartOf
VIDEO
AUDIO/VIDEO
AUDIO
TRANSCRIPT
HasVideo
HasAudio
FollowedBy
FollowedBy
FollowedBy
HasTranscript
HasTranscript
HasTranscript
MANIFESTATION LEVEL
MANIFESTATION
AvailableAs
MEDIA
SyncronisedWith
ANALOG
DigitalisedFrom
DIGITAL
PartOf
PartOf
ANALOG-SEG
DIGITAL-SEG
ITEM LEVEL
ITEM
STORAGE
ON-LINE
OFF-LINE
Legenda
Entity
Is-A Relationships
Relationships one to one
Relationship one to many

Model


WORK
WORK PERSPECTIVE
ExpressedBy
AVDOCUMENT
CutFrom
NEWS_REEL
DOCUMENTARY
CUT
REPORTAGE
ComposedOf
ComposedOf

Proposed Model (cont.)



EXPRESSION PERSPECTIVE

ExpressedBy

EXPRESSION

ManifestedBy

VERSION

PartOf

HasChannels

PartOf

PartOf

HasChannels

PartOf

VIDEO

AUDIO/VIDEO

AUDIO

TRANSCRIPT

HasVideo

HasAudio

FollowedBy

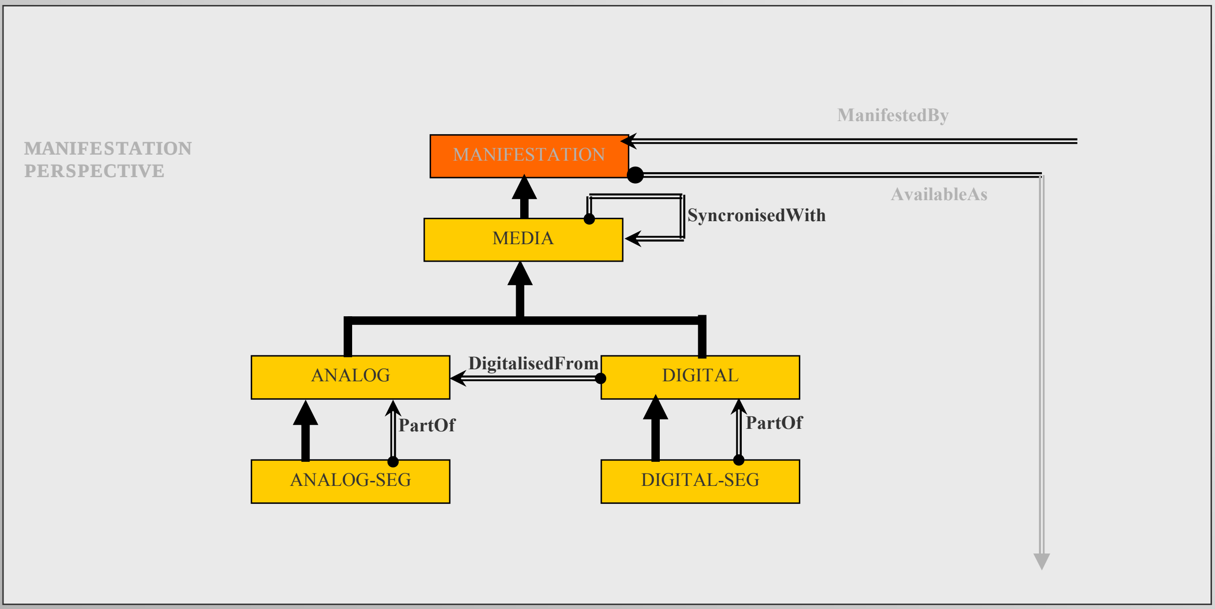
FollowedBy

FollowedBy

HasTranscript

HasTranscript

Proposed Model (cont.)



MANIFESTATION PERSPECTIVE

ManifestedBy

MANIFESTATION

AvailableAs

SyncronisedWith

MEDIA

ANALOG

DigitalisedFrom

DIGITAL

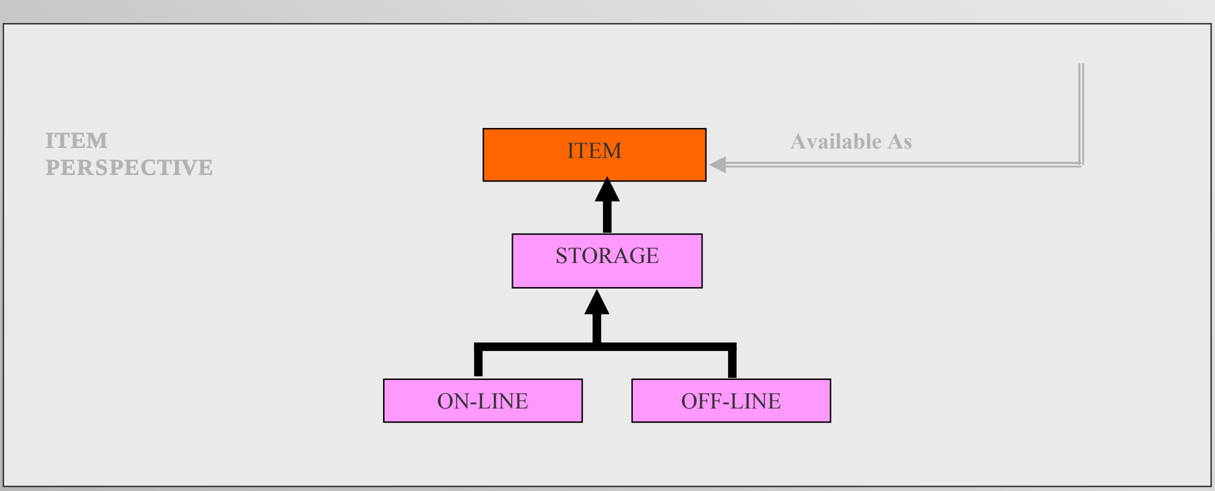
PartOf

PartOf

ANALOG-SEG

DIGITAL-SEG

Proposed Model (cont.)



ITEM PERSPECTIVE

ITEM

Available As

STORAGE

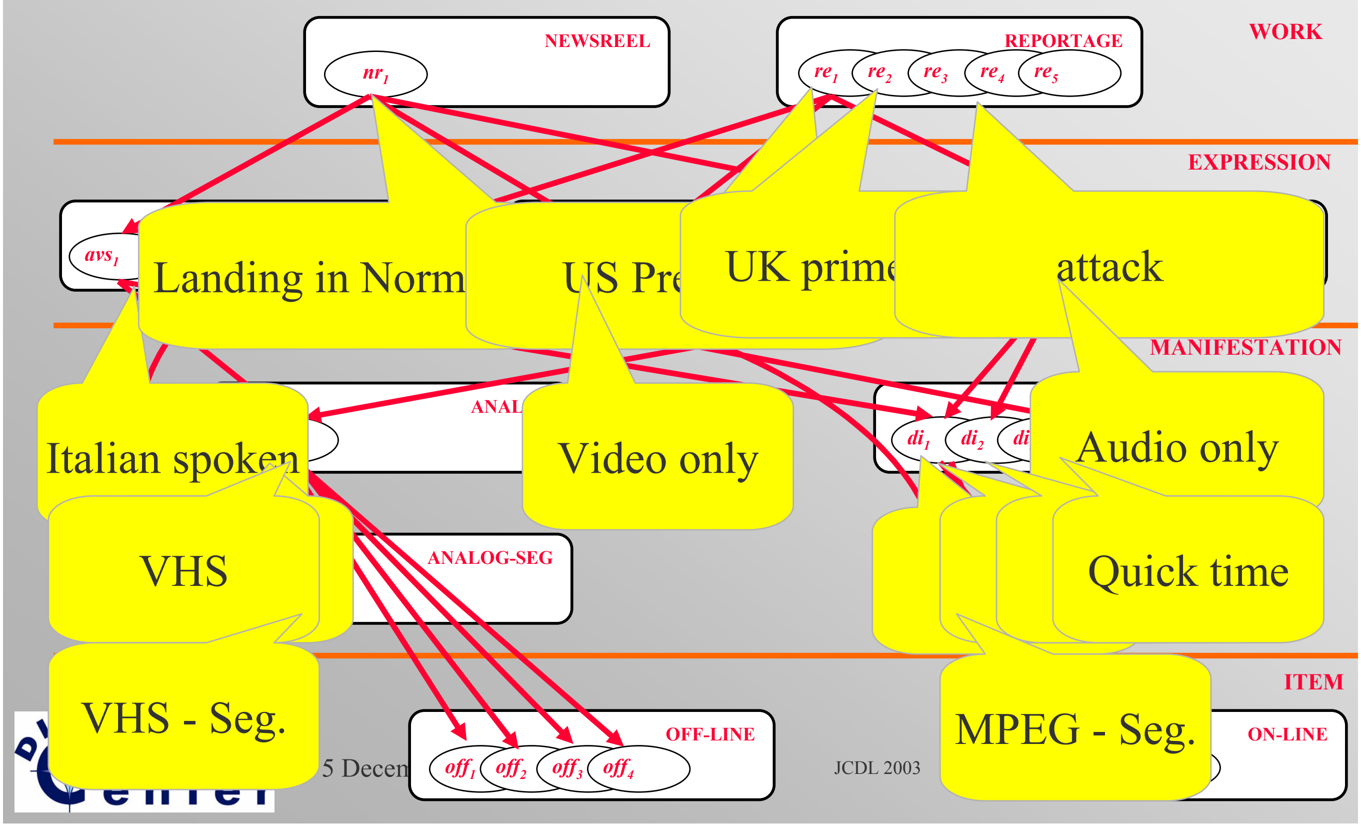
ON-LINE

OFF-LINE

Running example

- Newsreel about the “Landing of the Allied Forces to Normandy”
  - It is composed of several reportages
  - There are several national versions
    - e.g. Italian and French
  - Each version is available on different supports
    - e.g. VHS tapes, MPEG files
  - There are several copies of the VHS tape with different preservation quality
  - There are several copies of the MPEG file with different access speed

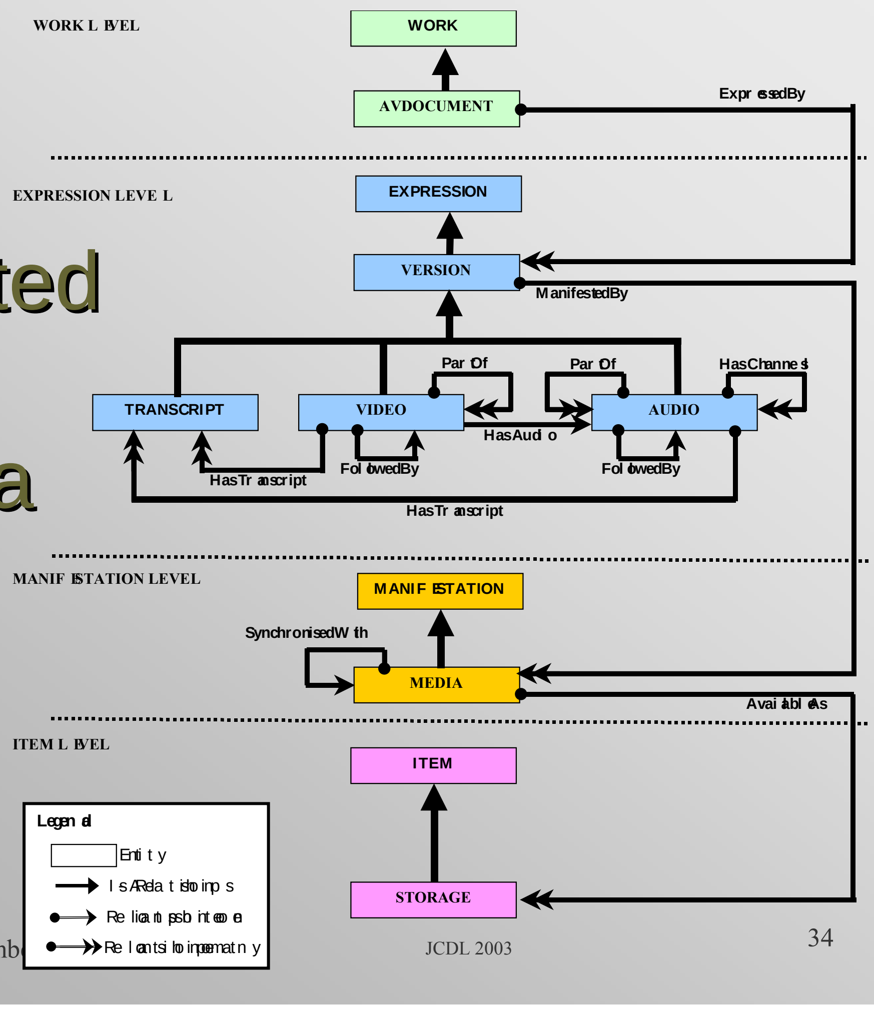
Running example



5 Decem

JCDL 2003

The Implemented ECHO metadata model

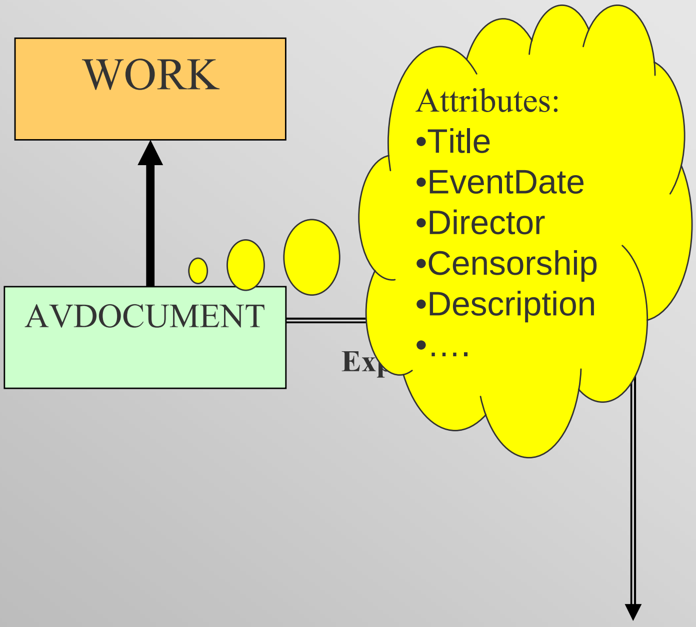


5 Decemb

JCDL 2003

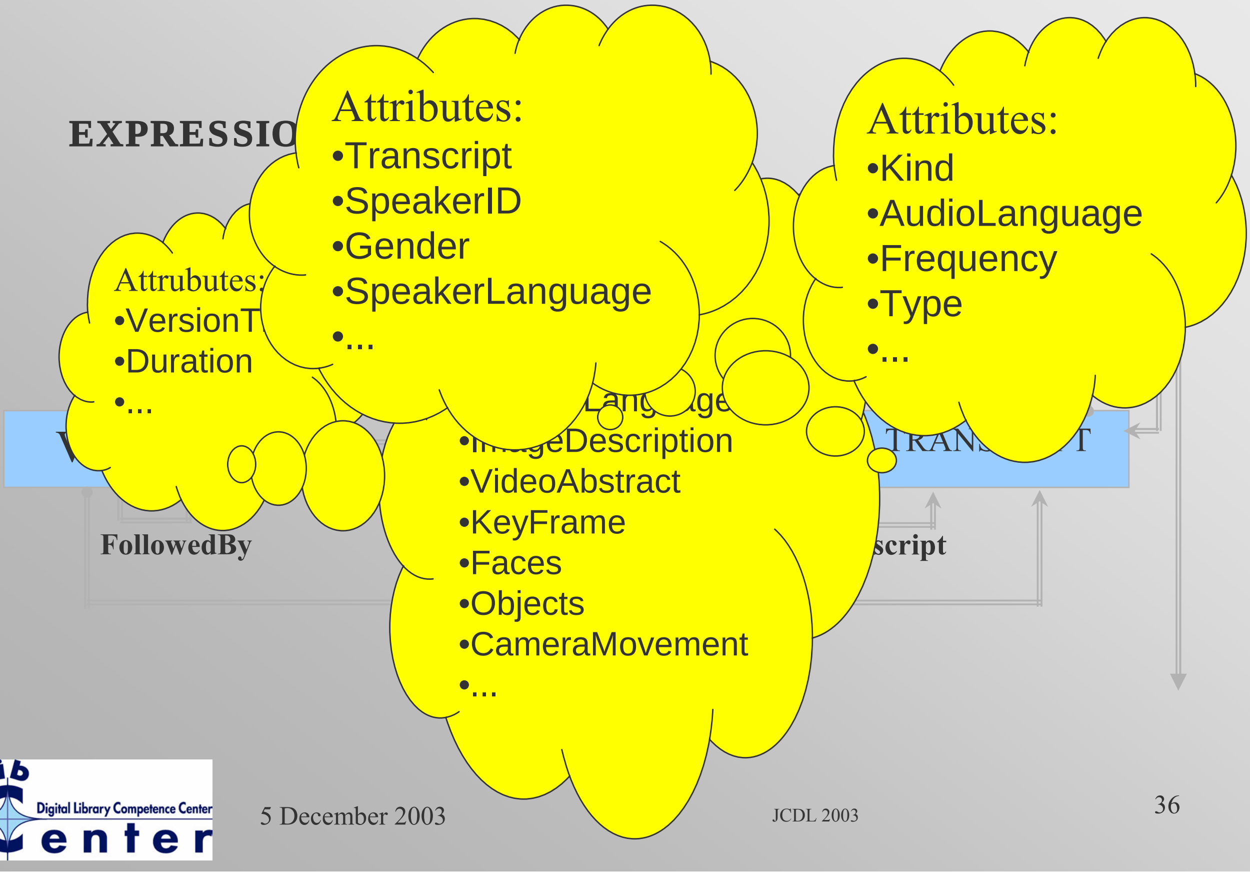
The ECHO Model

WORK



Attributes:
- Title
- EventDate
- Director
- Censorship
- Description
- ….

The ECHO Model (cont.)



Attrubutes:
- VersionT
- Duration
- …

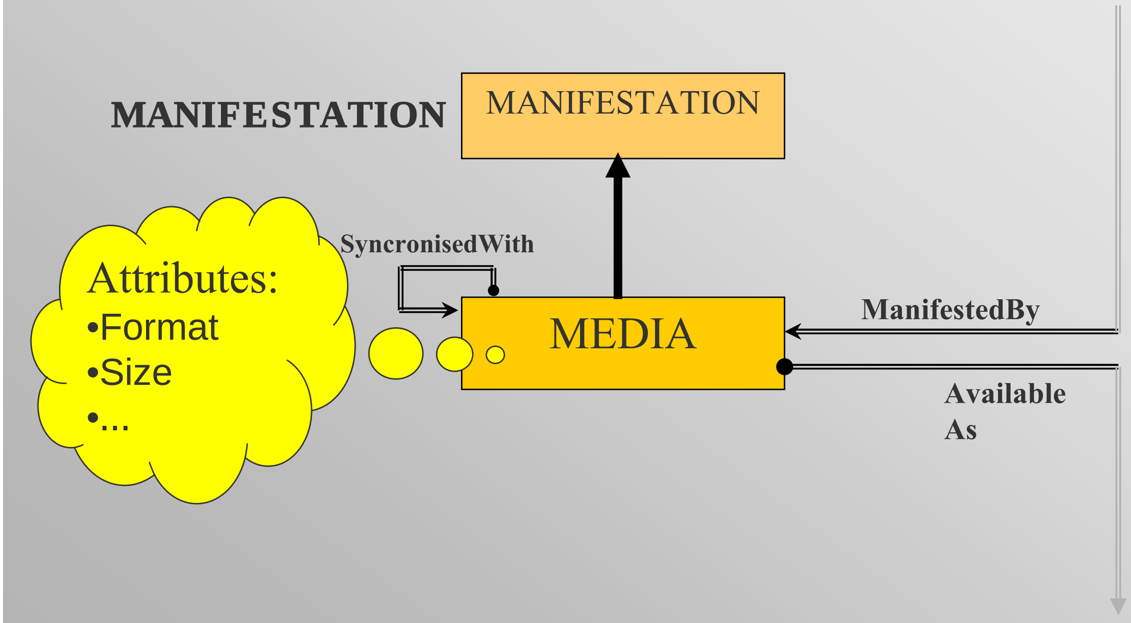
Attributes:
- Transcript
- SpeakerID
- Gender
- SpeakerLanguage
- …

- ImageDescription
- VideoAbstract
- KeyFrame
- Faces
- Objects
- CameraMovement
- …

Attributes:
- Kind
- AudioLanguage
- Frequency
- Type
- …

The ECHO Model (cont.)

MANIFESTATION



MANIFESTATION

MEDIA

SyncronisedWith

ManifestedBy

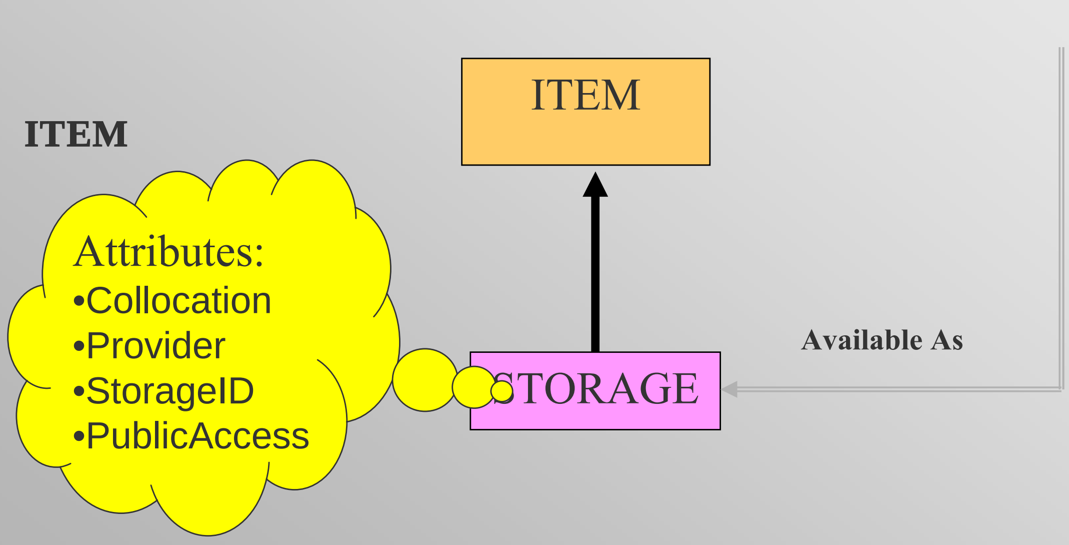
Available As

Attributes:

- Format
- Size
- ...

The ECHO Model (cont.)

ITEM



ITEM

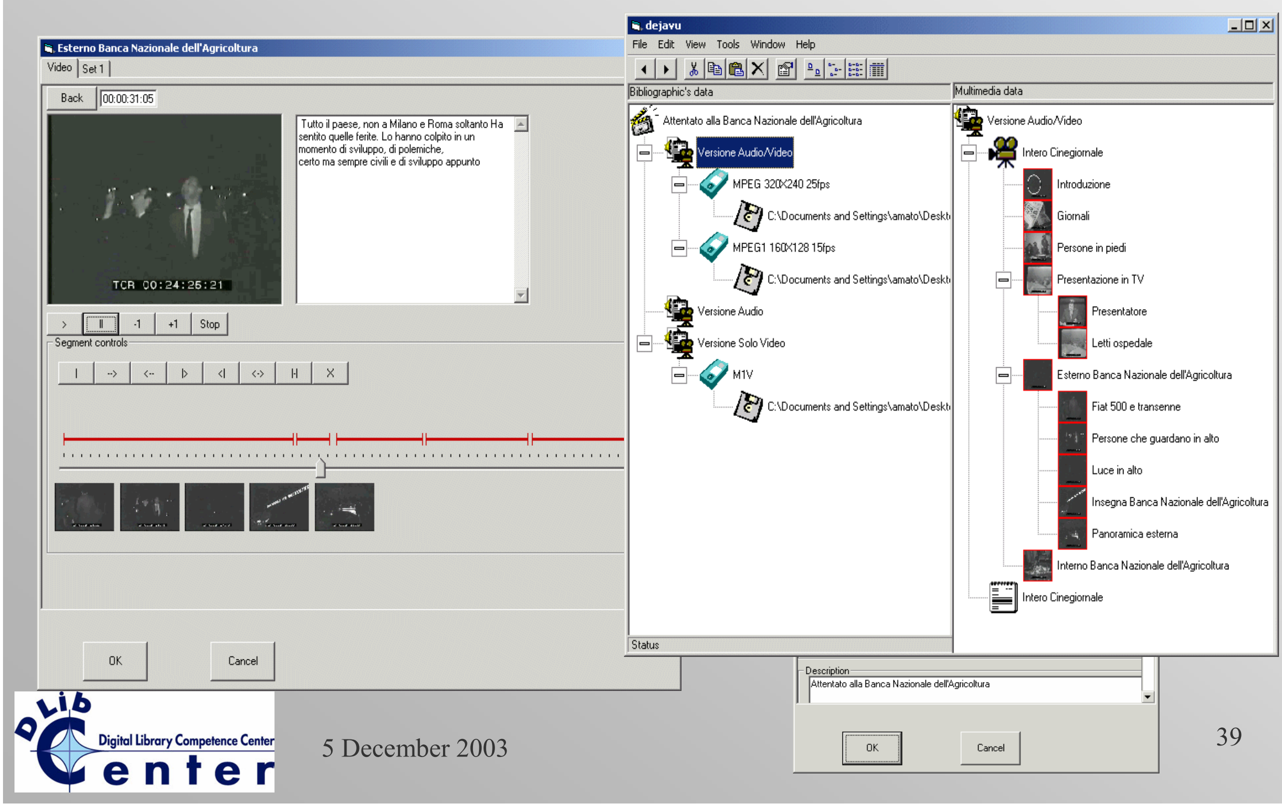
STORAGE

Available As

Attributes:

- Collocation
- Provider
- StorageID
- PublicAccess

Metadata Editor



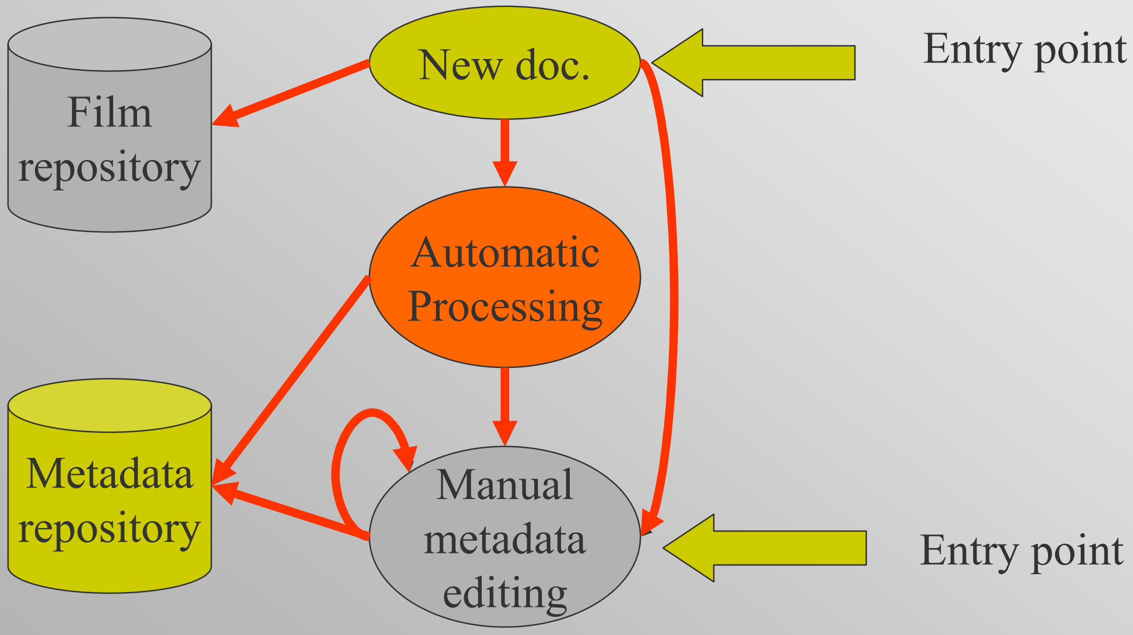
Tutto il paese, non a Milano e Roma soltanto Ha sentito quelle ferite. Lo hanno colpito in un momento di sviluppo, di polemiche, certo ma sempre civili e di sviluppo appunto

Description: Attentato alla Banca Nazionale dell'Agricoltura

Generating metadata: indexing

- Fully Manual indexing:
  - Time consuming and tedious, especially with complex metadata models
- Fully Automatic Indexing:
  - Noise may affect effectiveness
- Manual indexing with automatic support:
  - Could be a good compromise

Work flow



New doc.

Film repository

Automatic Processing

Metadata repository

Manual metadata editing

Entry point

Entry point

Automatic processing tasks

- Cut detection
- Visual features extraction
- Transcript generation
- Object recognition
- Face recognition
- Geospatial information
- Video abstract generation

Automatic Indexing

- Overview
  - Text
  - Speech
  - Images
  - Moving pictures (videos)

Indexing text

- The indexing process associates (weighted) index terms to documents
- Index terms can be
  - Words chosen from a controlled vocabulary
  - Words automatically extracted
  - Steams (e.g. print-)
  - Noun phrases automatically extracted
  - Other metadata

Indexing text

- Experience has shown that using weighted single terms offers the best performance
  - Of course that depends crucially on the choice of the term-weighting system
- Document search is performed by searching for index terms
  - Documents associated with qualifying index terms are retrieved
  - Documents are ranked according to weights of index terms

Indexing text

- The indexing process produces an incidence matrix:

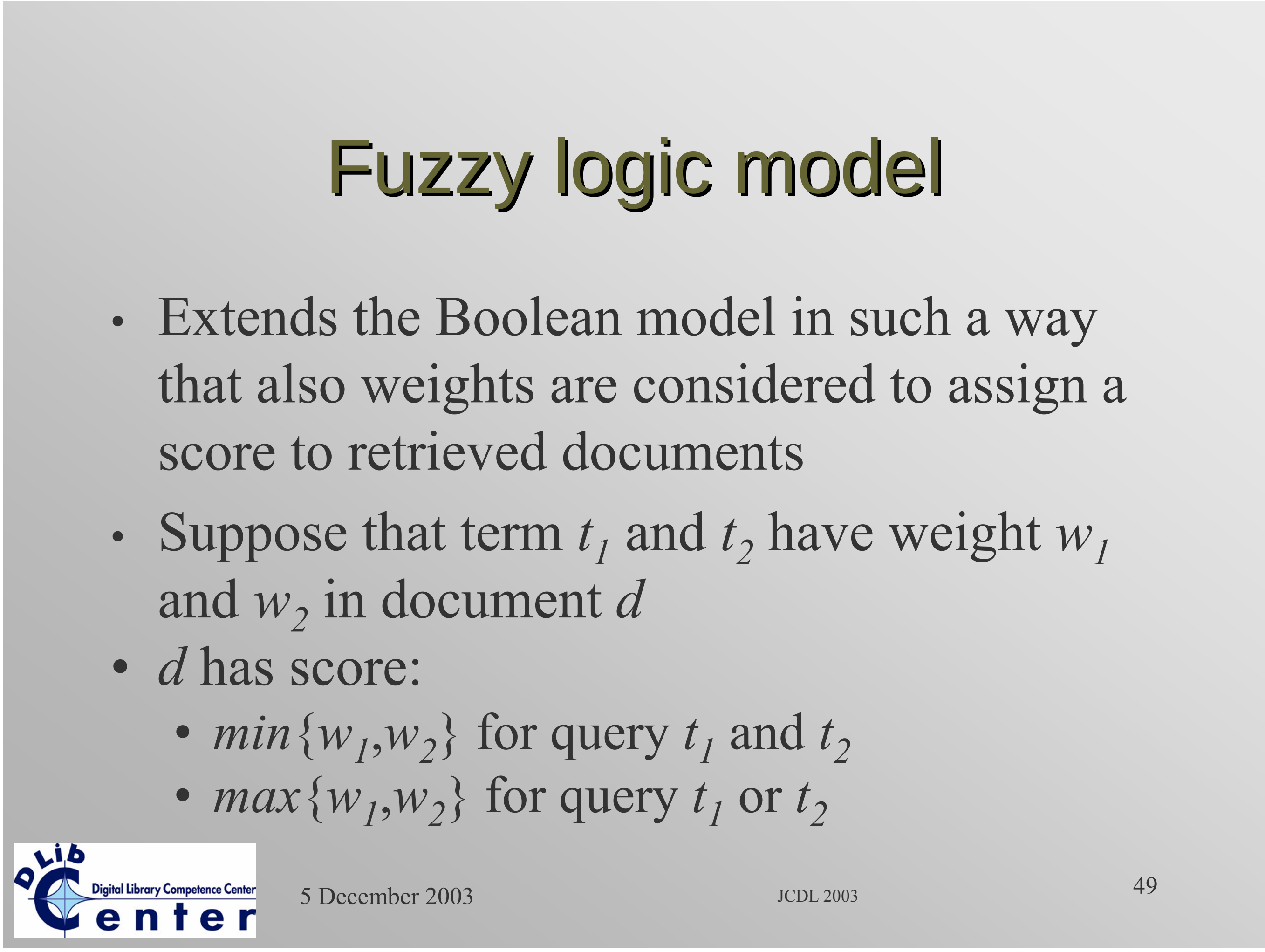
	d_1	…	d_i	…	d_m
t_1	w_{11}	…	w_{1i}	…	w_{1m}
…	…	…	…	…	…
t_k	…	…	w_{ki}	…	…
…	…	…	…	…	…
t_n	w_{n1}	…	w_{ni}	…	w_{nm}

Indexing text

- Models to assess document relevance:
  - Boolean model
  - Fuzzy logic model
  - Vector space model
  - …

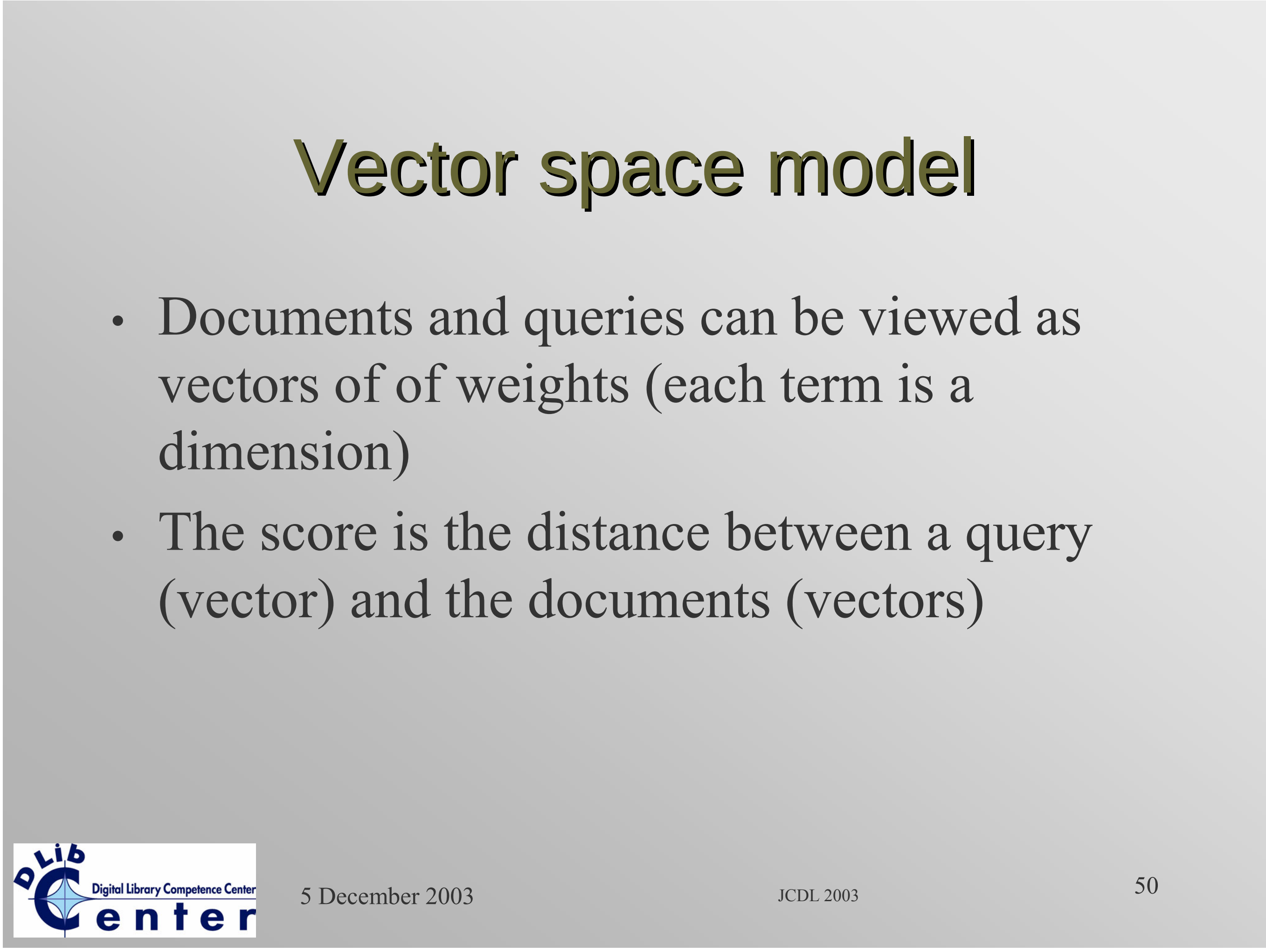
Boolean model

- A query may contain logical operator and/or
  - The query “digital and library” retrieves documents associated with both terms
  - The query “digital or library” retrieves documents associated with at least one of the two terms
- Boolean logic is used to process more complex queries



Fuzzy logic model

- Extends the Boolean model in such a way that also weights are considered to assign a score to retrieved documents
- Suppose that term t_1 and t_2 have weight w_1 and w_2 in document d
- d has score:
  - $min\{w_1,w_2\}$ for query t_1 and t_2
  - $max\{w_1,w_2\}$ for query t_1 or t_2



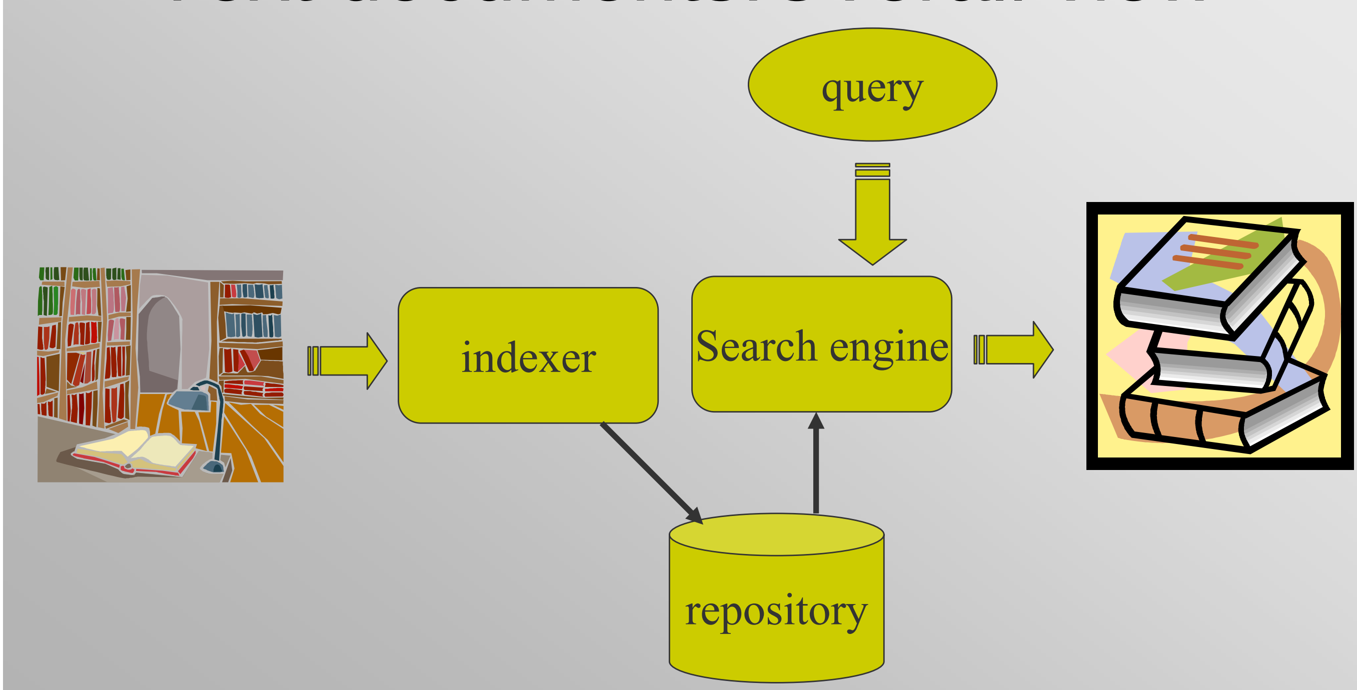
Vector space model

- Documents and queries can be viewed as vectors of of weights (each term is a dimension)
- The score is the distance between a query (vector) and the documents (vectors)

Automatic extraction of weighted index terms

- A widely used technique is the *tfidf* weighting function (term frequency inverse document frequency):
  - The more frequently a term appear in a document the more significant it is for that document: term frequency (*tf*)
  - The more frequently a term occur in the entire collection the less selective it is: document frequency (*df*)
- The weight is directly proportional to the *tf* and inversely proportional to the *df* (*idf*)

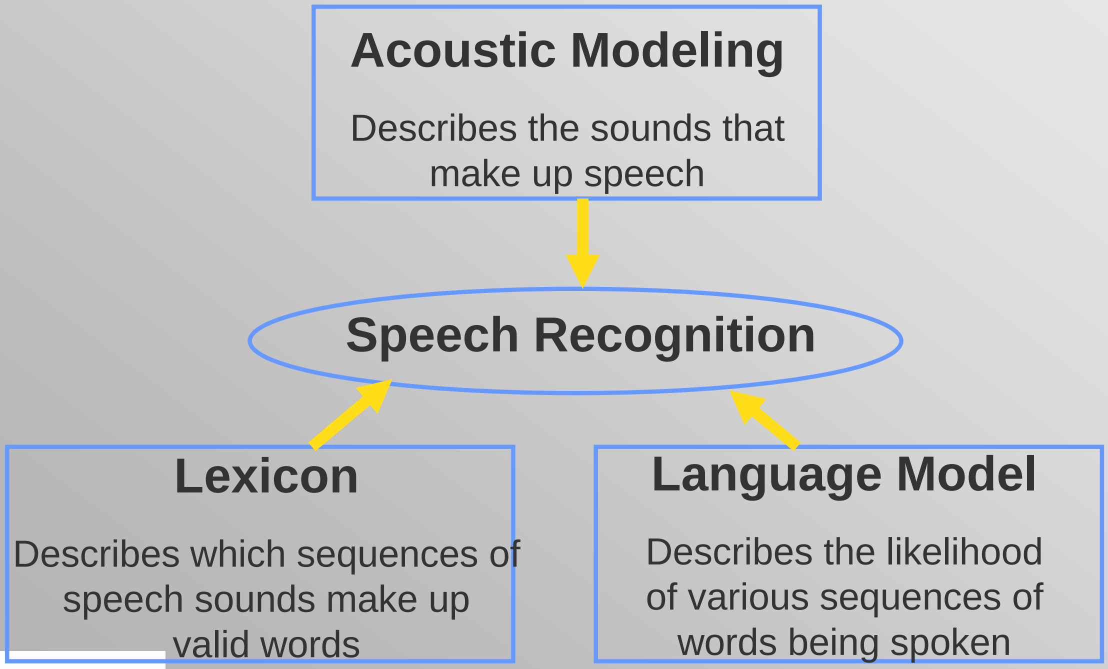
Text documents:Overall view



Indexing speech

- Generates transcript to enable text-based retrieval from spoken language documents
- Improves text synchronization to audio/video in presence of scripts
- Supplies information necessary for library segmentation and multimedia abstractions
- Provides speech interface to digital library

Indexing speech



Acoustic Modeling

Describes the sounds that make up speech

Speech Recognition

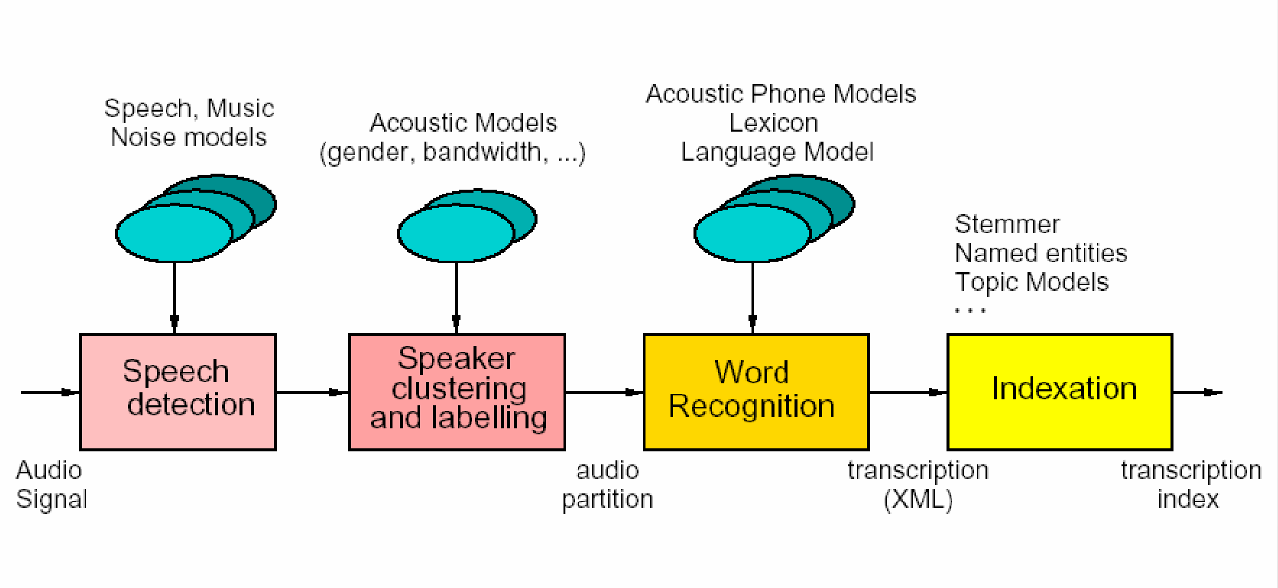
Lexicon

Describes which sequences of speech sounds make up valid words

Language Model

Describes the likelihood of various sequences of words being spoken

Indexing speech



Speech, Music
Noise models

Acoustic Models
(gender, bandwidth, ...)

Acoustic Phone Models
Lexicon
Language Model

Stemmer
Named entities
Topic Models
. . .

Speech detection

Speaker clustering and labelling

Word Recognition

Indexation

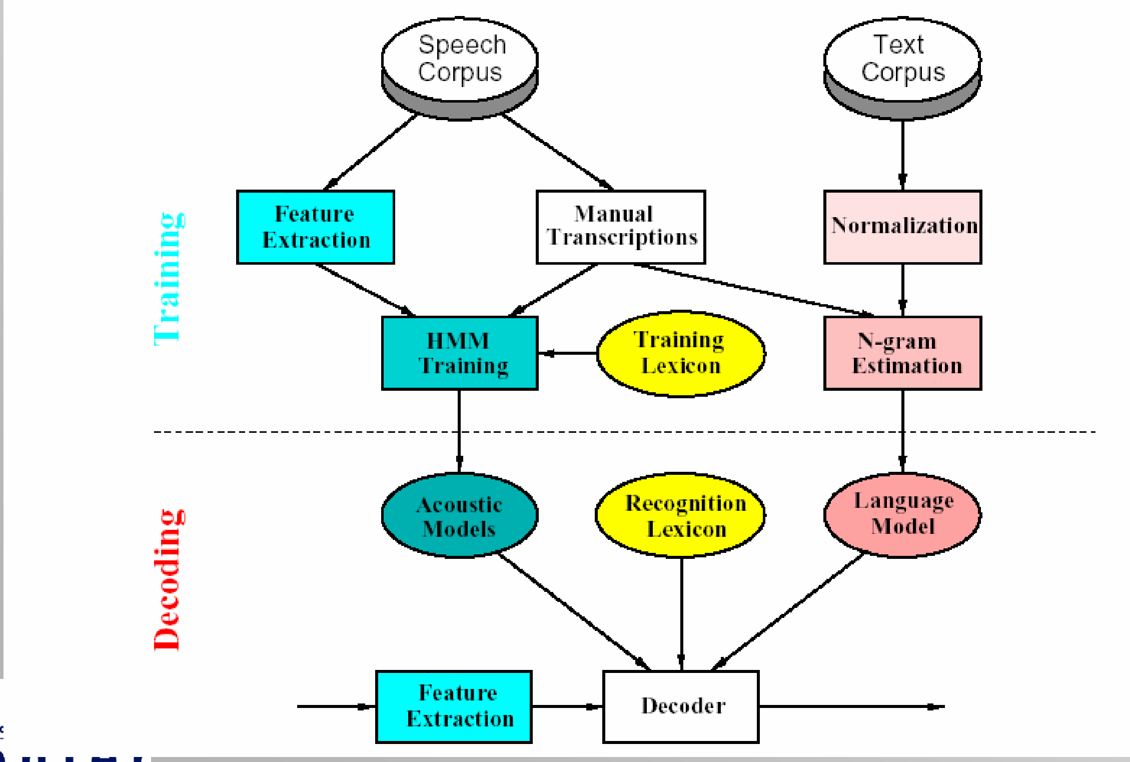
Audio Signal

audio partition

transcription (XML)

transcription index

Indexing speech



Training

Decoding

Speech Corpus

Text Corpus

Feature Extraction

Manual Transcriptions

Normalization

HMM Training

Training Lexicon

N-gram Estimation

Acoustic Models

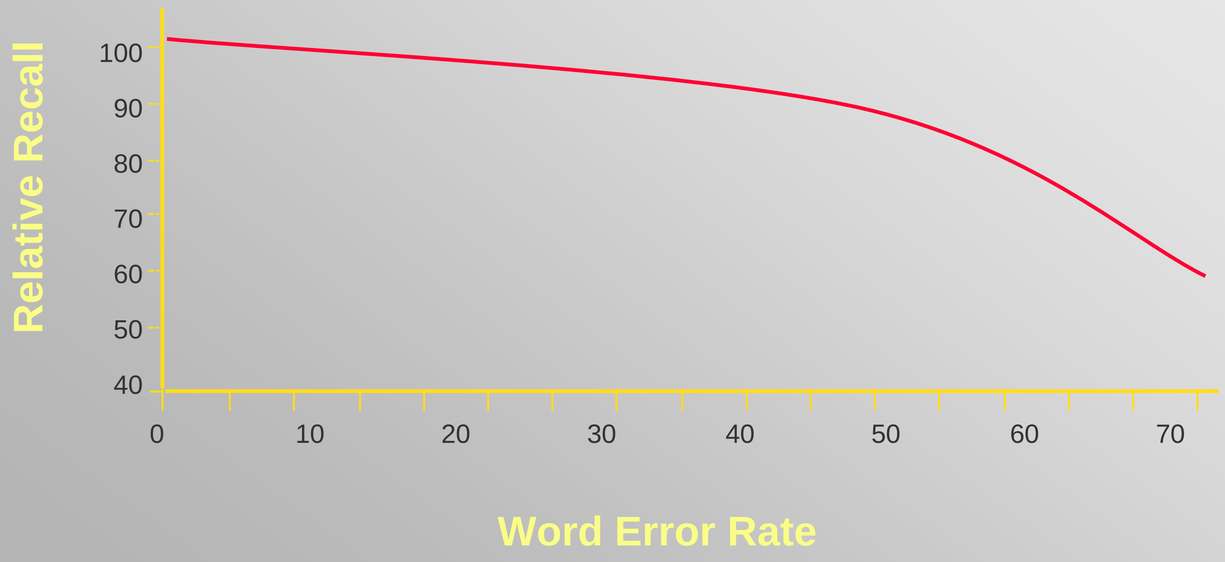
Recognition Lexicon

Language Model

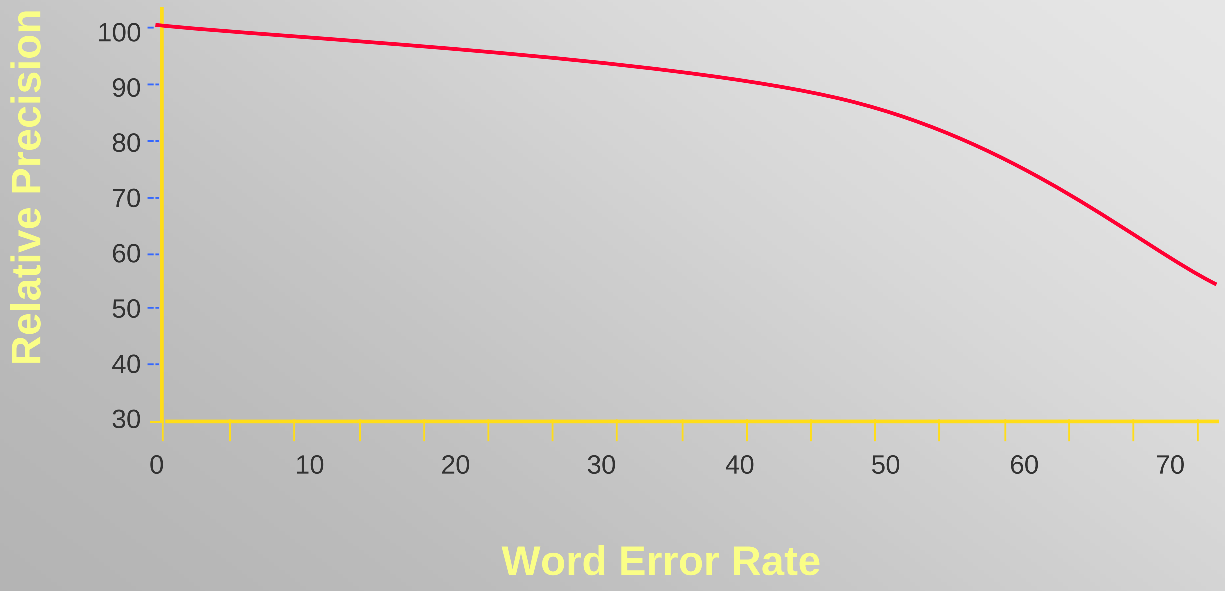
Feature Extraction

Decoder

Text retrieval precision vs. Speech accuracy



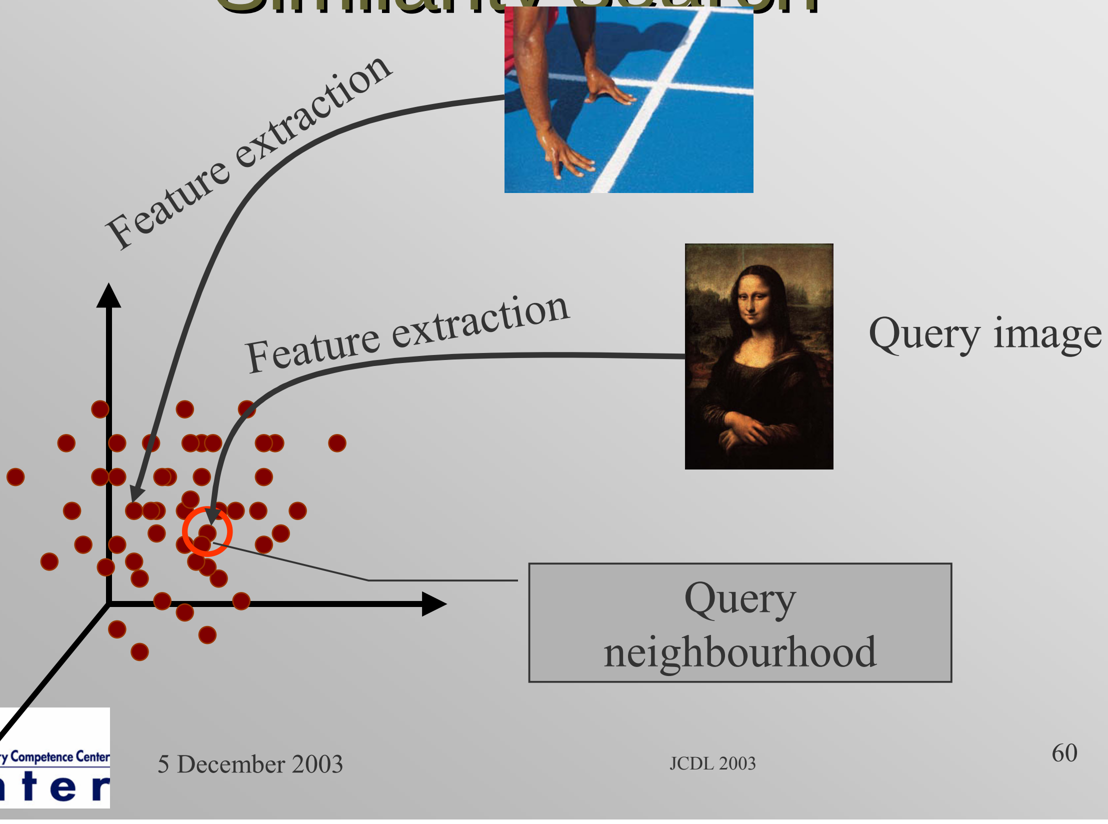
Text retrieval precision vs. Speech accuracy



Indexing images

- The automatic indexing process associates images with features describing their physical content
  - Colour
  - Textures
  - Shapes
  - Spatial organisation
- Image search is performed by using feature similarity

Similarity search



Feature extraction

Feature extraction

Query image

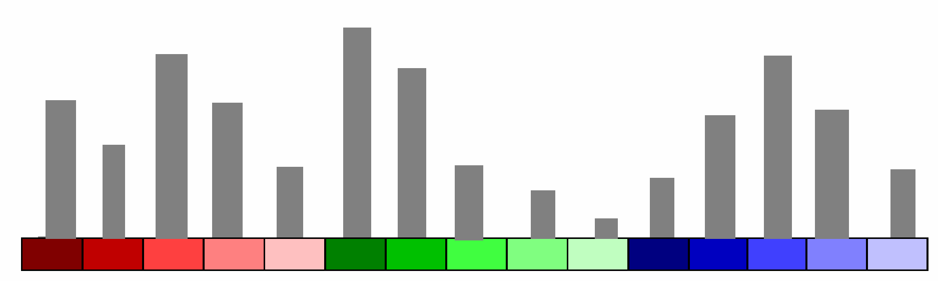
Query neighbourhood

Indexing images

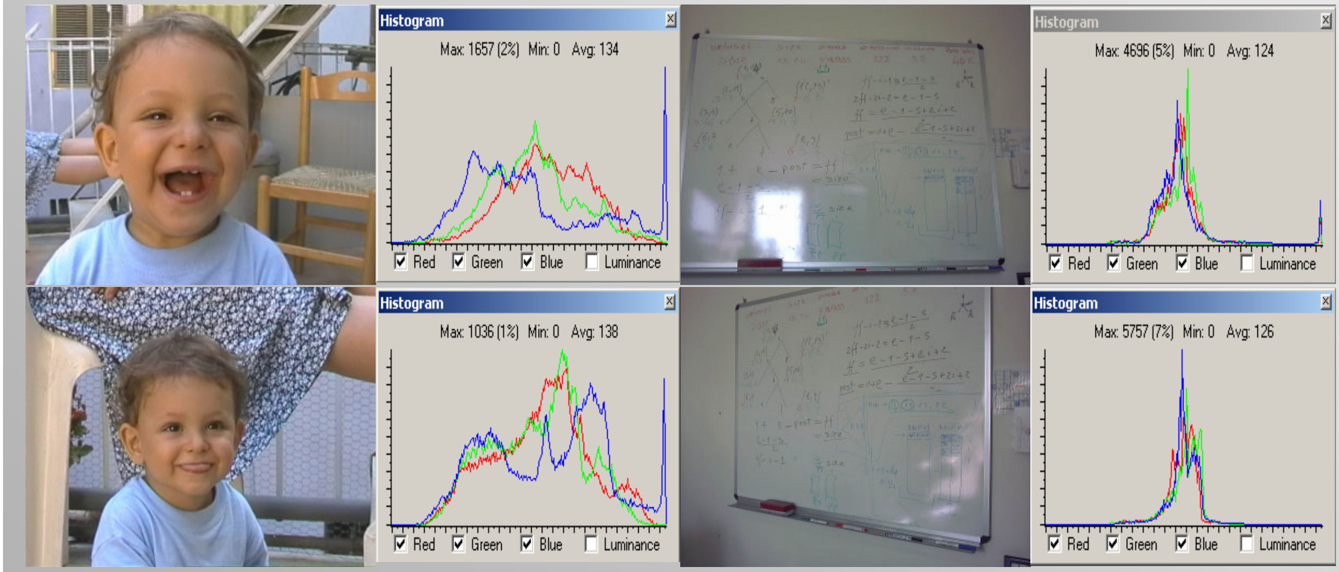
- Colour spaces
  - The most common and intuitive colour space is the RGB (Red Green Blue) colour space
    - Every perceivable colour can be obtained as the sum of three degree of RGB

Image indexing

- Colour histograms
  - The colour spectrum is divided into *n* bins
  - The value contained in each bean is proportional to the amount of pixel having colour of that bean



Indexing images



Indexing images

- Problems with RGB:
  - Colours that are close in the RGB colour space can be distant for the human perception

Indexing images

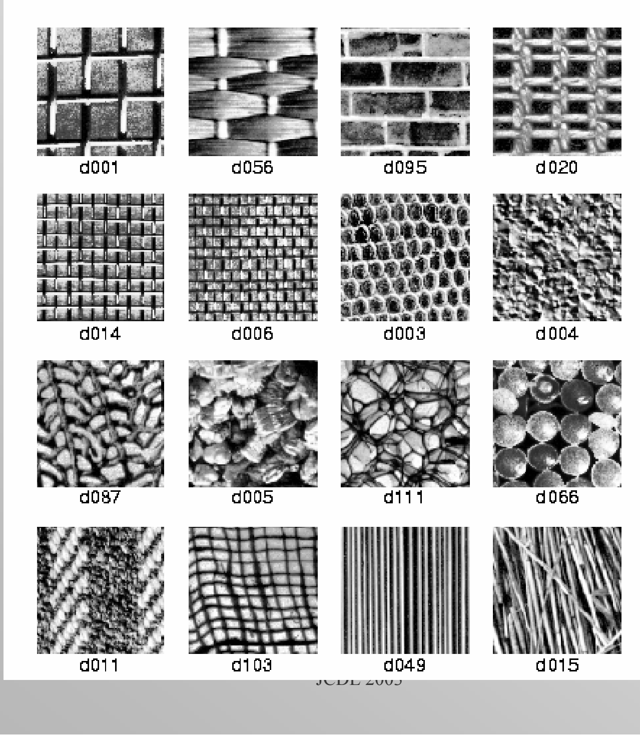
- Wanted properties of colour spaces:
  - Uniformity
    - Close colours are also perceived as similar
  - Completeness
    - All perceivable colours are representable
  - Compactness
    - No redundancy

Indexing images

- Other colour spaces:
  - HSV
    - Hue:Tint of the colour
    - Saturation:Quantity of colour
    - Value (Brightness):Quantity of light
  - YIQ, YUV, YCrCb, etc.

Indexing images

- Textures:



Indexing images

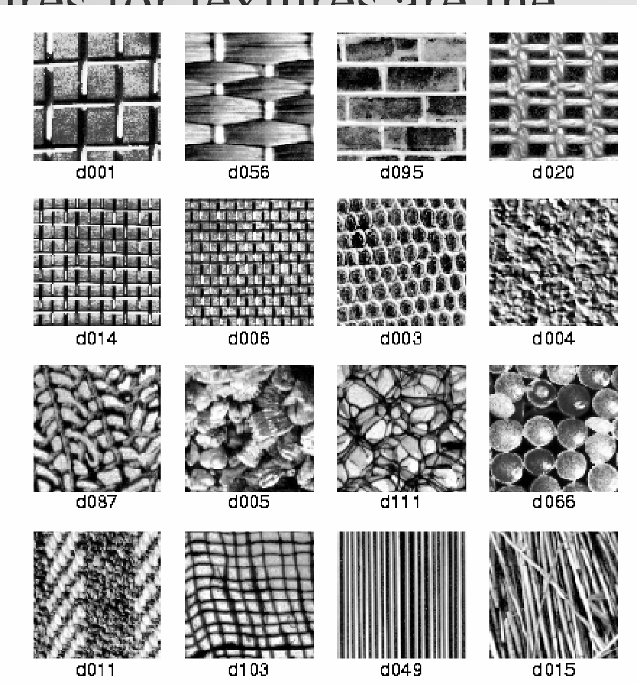
- Textures:
  - Homogeneous patterns
  - Spatial arrangement of pixels
    - Colour is not enough to describe

Indexing images

- Textures descriptions are obtained by using statistical methods
  - Spatial distribution of image intensity
  - Several methods exists
  - Texture descriptions can also be represented as histograms (vectors)

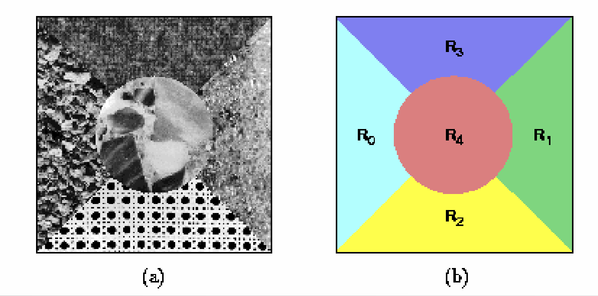
Indexing images

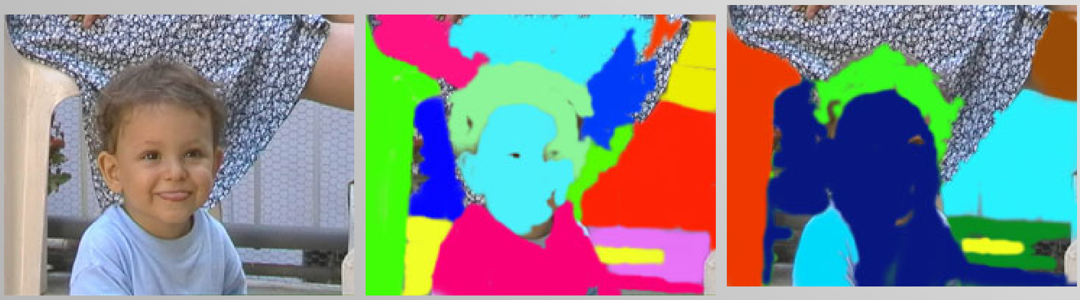
- Widely used features for textures are the Tamura features:
  - Contrast
    - Distribution of
  - Coarseness
    - Granularity of
  - Directionality
    - Dominant dire



Indexing images

- Shapes:
  - Region extraction
  - Segmentation



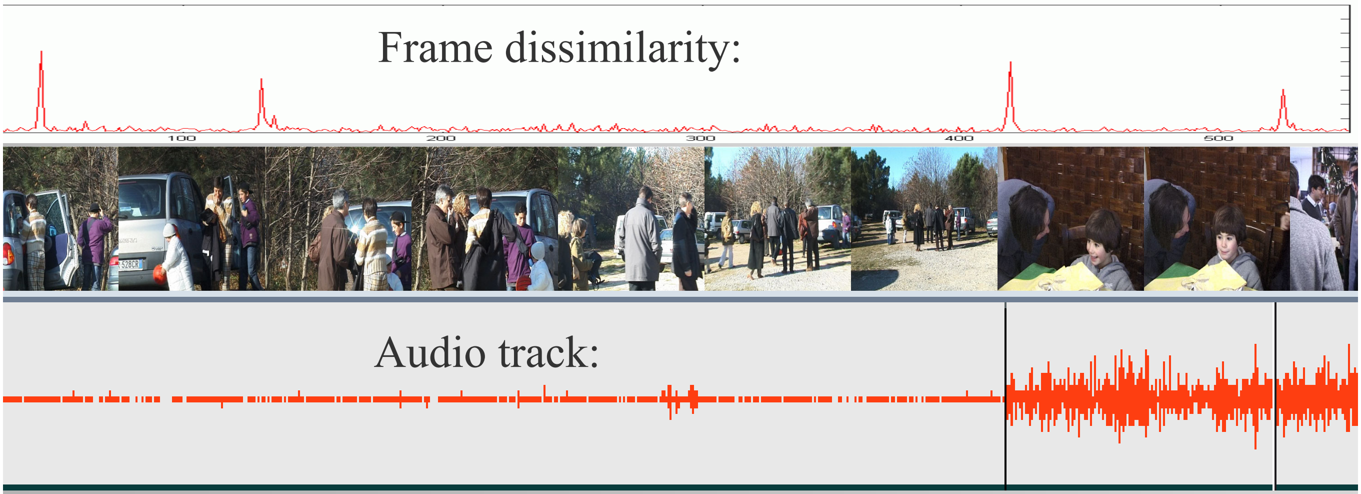


Indexing images

- Colour histograms and textures can be computed for individual regions in addition to entire images
  - Global features
    - Search for images
  - Local features
    - Search for regions in images
- Spatial relationships between region give also additional information
  - Search for images having specific characteristics

Indexing moving pictures

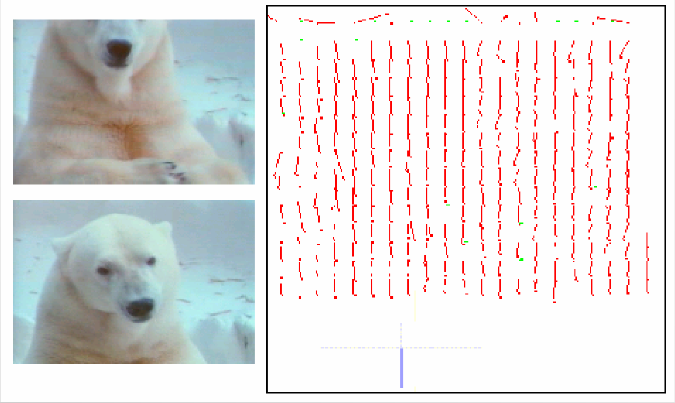
- Cut/scene detection:



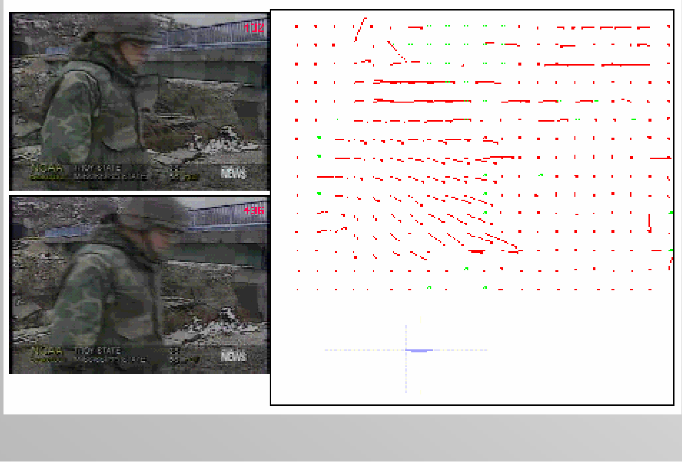
Motion Picture indexing

- Camera and Motion Detection

Pan



Right object motion (not pan left)



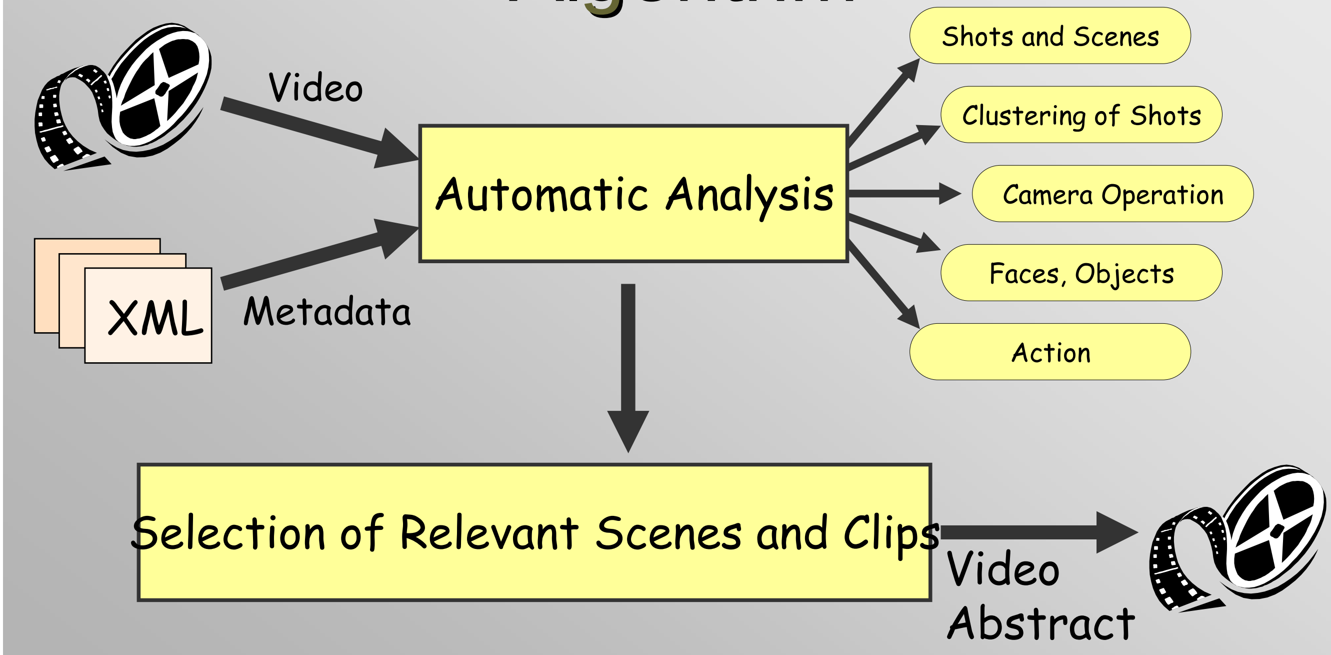
Video Abstract

- 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 user can see the video abstract without any technical knowledge of the application.

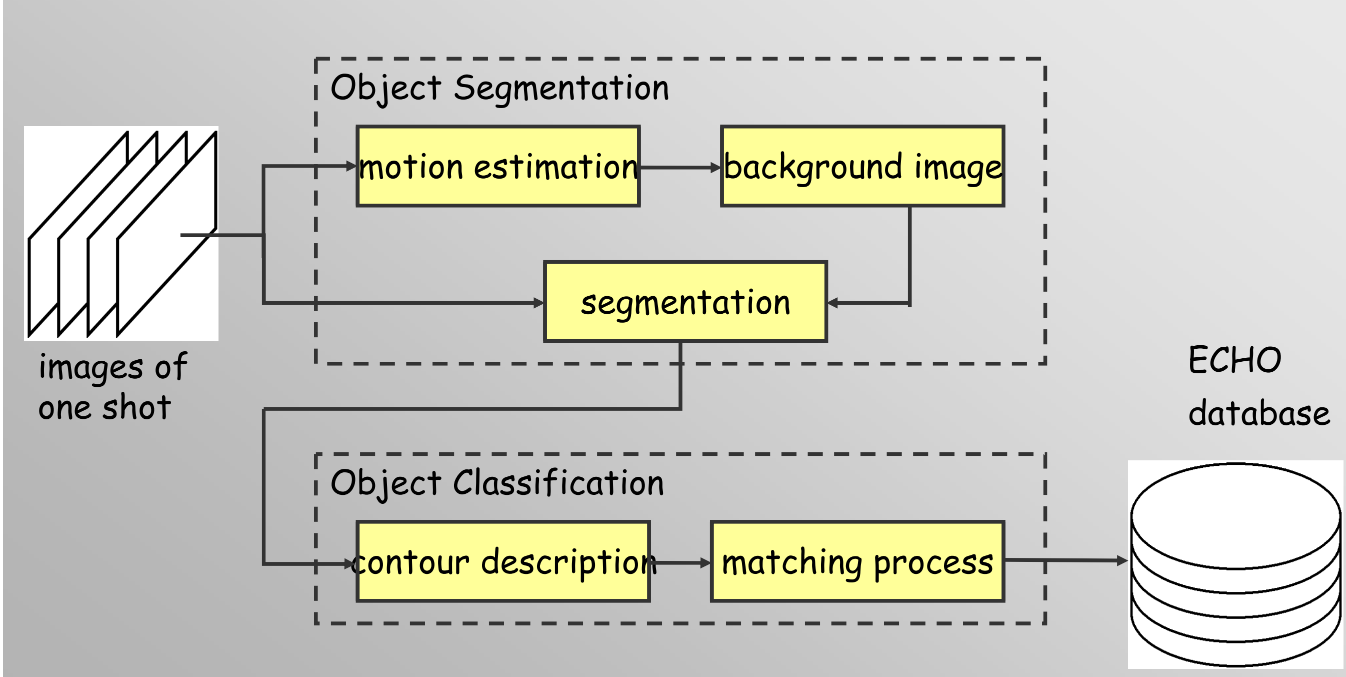
Overview of the Abstracting Algorithm



Moving-Object 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.

Moving-Object Recognition



Object Segmentation

motion estimation

background image

segmentation

images of one shot

Object Classification

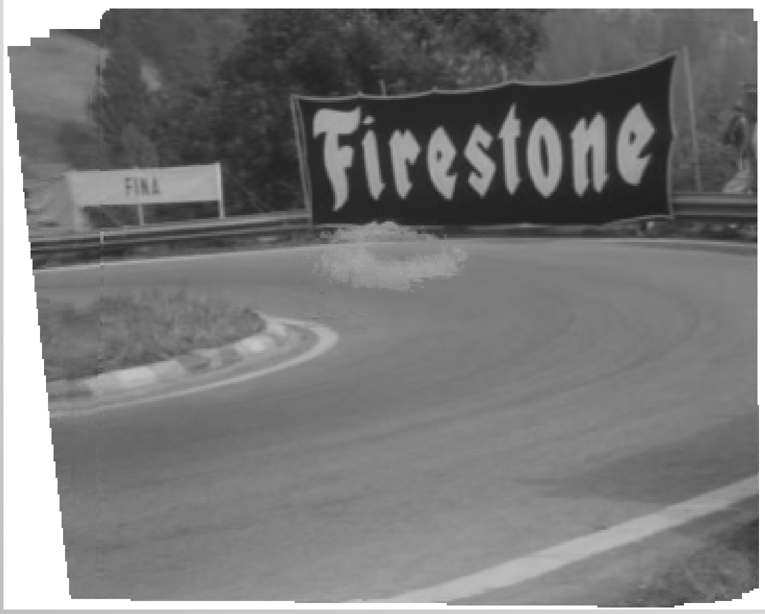
contour description

matching process

ECHO database

Object Segmentation

- The camera model is calculated and all frames are transformed with this camera model.
- The background panorama image is the median of all pixels at the same position.
- A large differences of the frame and the background indicates an object.

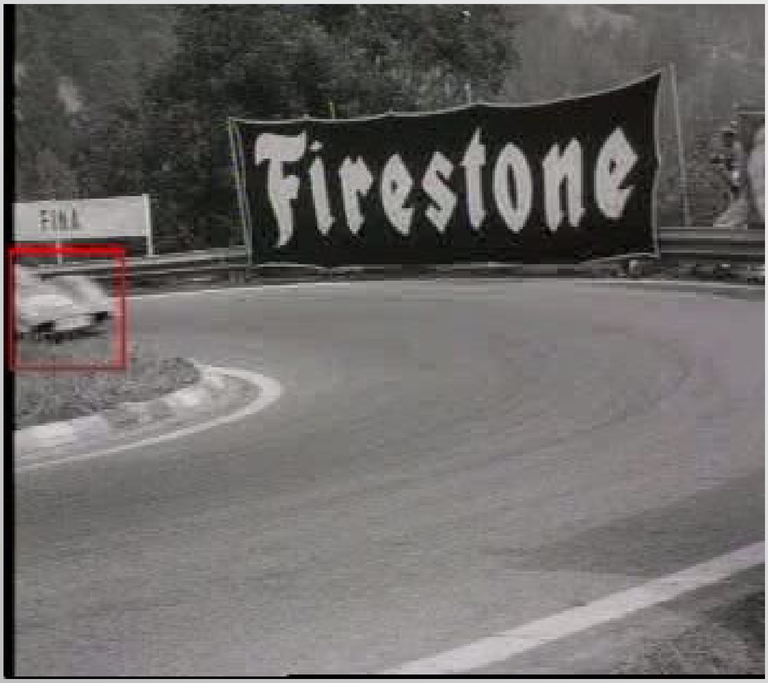


calculated background image

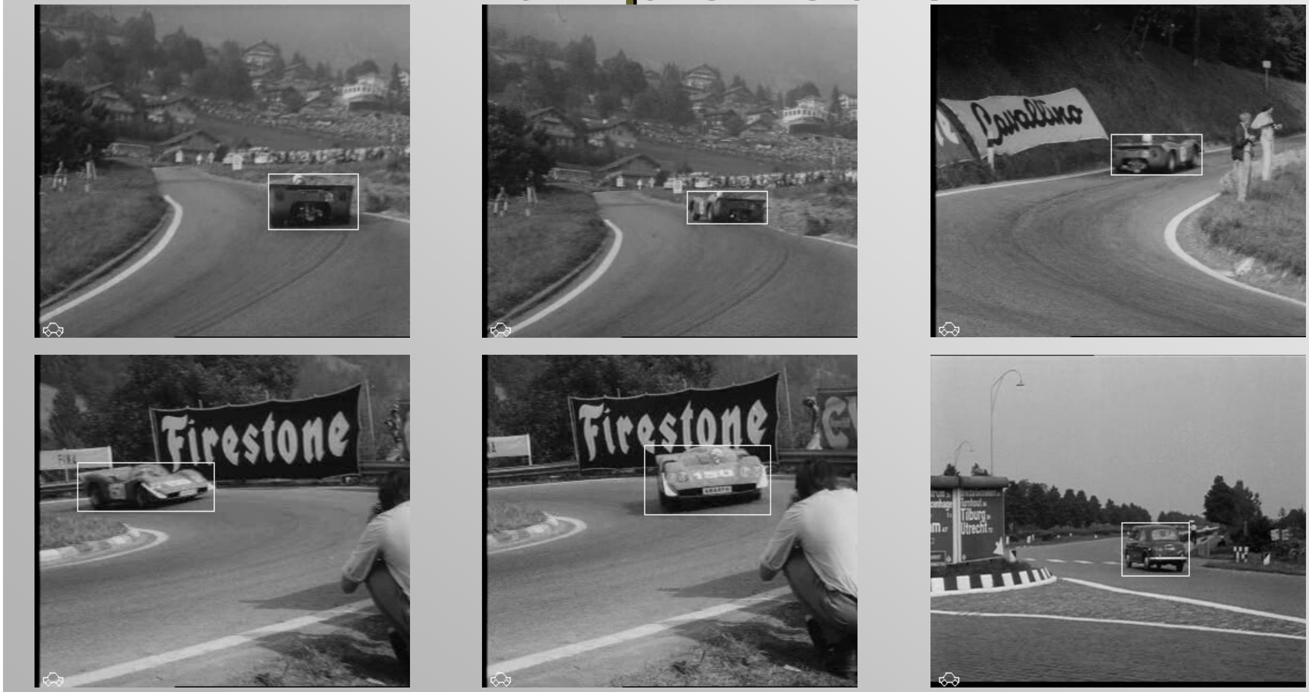
sample video of segmented and recognized objects (cars)

Object Classification

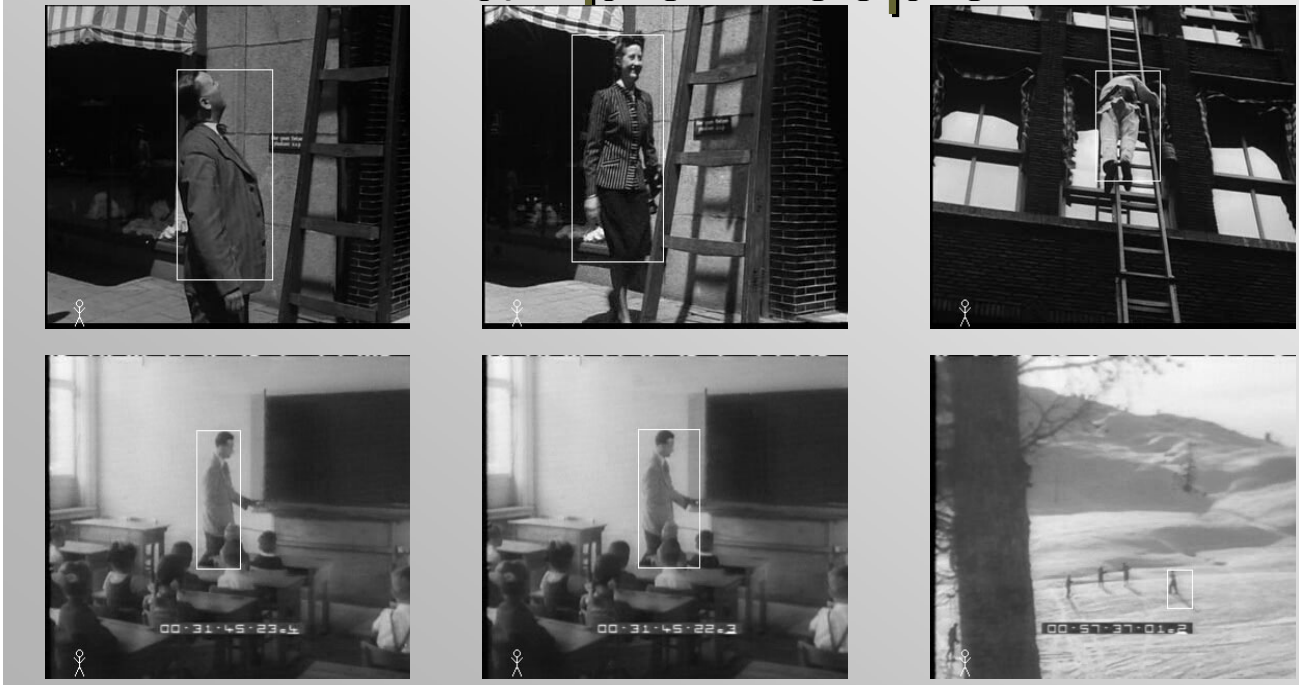
- The classification of the segmented object is based on feature points of the contour.



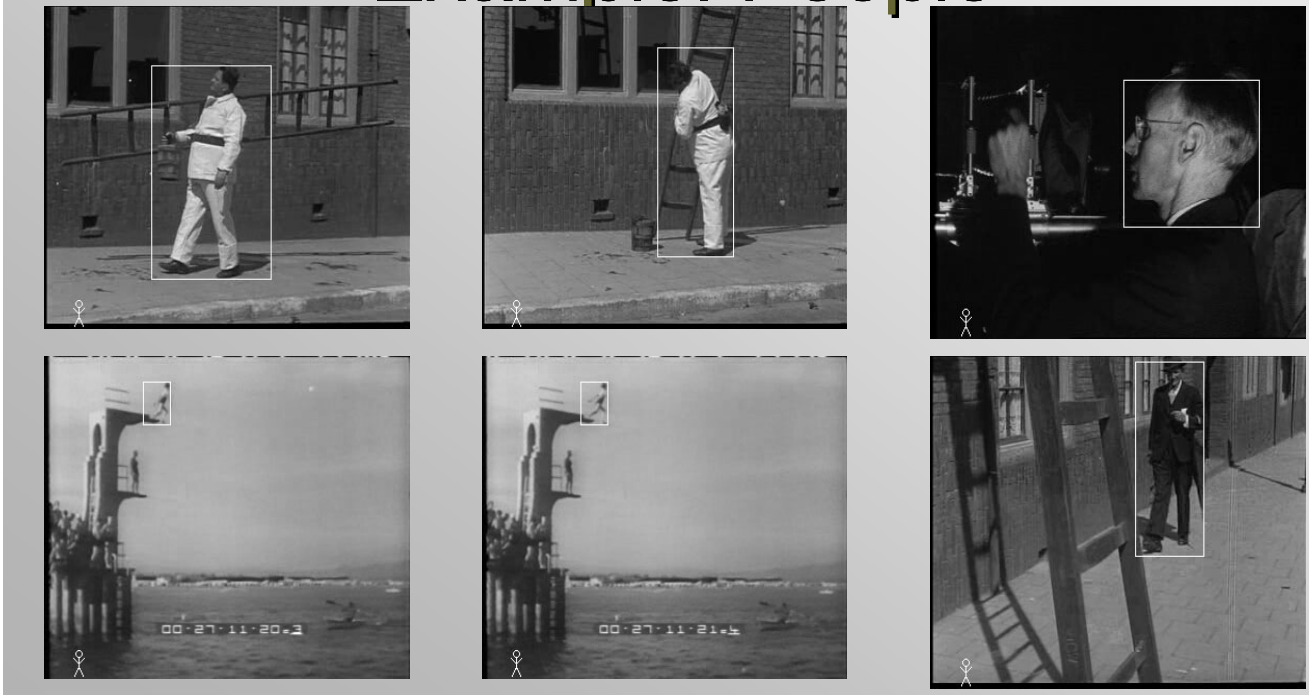
Example: Cars

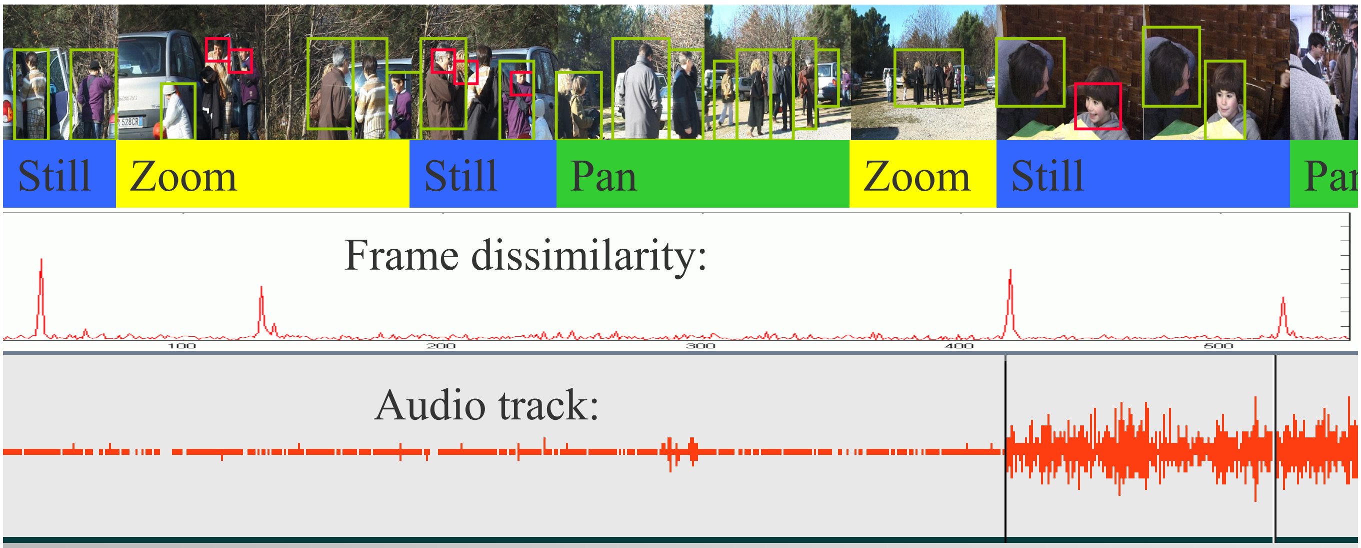


Example: People



Example: People





Face detection

Object detection

Text detection

Speech recognition

Manual editing tasks

- Automatically generated metadata review
  - scene - shot - sequences review
  - face - object review
  - video abstract review
- Descriptive fields input
  - item fields editing
  - manifestation fields editing
  - expression fields editing
  - work fields editing

Questions??