

Capitolo 5

I metadati per documenti audiovisivi

5.1 Introduzione

La produzione di dati multimediali è in continua crescita. Con l'avvento della TV digitale, migliaia di canali sono oggi disponibili all'utente. In pratica si può dire che tutto il materiale multimediale prodotto è oggi registrato in formato digitale; si pensi ai CD Audio, ai DVD etc. Inoltre, la produzione di contenuti multimediali è diventata molto più facile. Usando videocamere digitali, personal computer, e Internet, virtualmente ogni persona nel mondo è oggi un potenziale produttore di contenuti digitali.

Senza strumenti adeguati di ricerca, tutto questo patrimonio umano rischia di rimanere poco o per niente utilizzato. Infatti, la necessità di disporre di meccanismi universali che consentano prima una rapida ed agevole catalogazione dei contenuti, e successivamente l'utilizzo di strumenti di ricerca sofisticata per accedere ai contributi multimediali complessivi, oppure ai singoli componenti elementari di cui essi sono composti, diventa una esigenza sempre più imprescindibile.

Un primo passo verso la soluzione di questo problema è sicuramente la realizzazione di standard di metadati per la descrizione dei contenuti multimediali, tali da permettere un'agevole descrizione dei contenuti audiovisivi, sui quali poter fare ricerche, selezioni, sincronizzazioni temporali e simili.

Il problema principale del contenuto audiovisivo è quello di essere molto più ricco e complesso del testo. Infatti, quest'ultimo è tipicamente già organizzato dagli autori in una struttura in capitoli o sezioni, ed ha il vantaggio di essere più facile da ricercare se già presente in forma elettronica.

Nell'Audio/Video la strutturazione può essere sì presente, tuttavia essa è implicita nel contenuto stesso del materiale. Ad esempio, il palinsesto di un canale televisivo è diviso in film, news, varietà, pubblicità, etc. A loro volta le singole unità possono essere suddivise in sottoparti, ad esempio un telegiornale è diviso in servizi. In teoria il contenuto audiovisivo prodotto ad esempio dai canali televisivi potrebbe essere senza grossi sforzi arricchito da metadati testuali utili per l'identificazione delle parti. Questo in parte avviene già in alcuni canali nelle trasmissioni satellitari, dove insieme allo segnale video vengono inviate informazioni sui programmi in corso (la guida DVB - Digital Video Broadcast).

I metadati possono essere usati per descrivere caratteristiche di basso livello sul contenuto audiovisivo, come ad esempio la distribuzione dei colori di un'immagine oppure informazioni ancora più sintetiche ma a maggior contenuto semantico (es: la Reggia di Caserta). In ogni caso, il livello di astrazione dei metadati è influenzato dal modo con cui tali informazioni sono ottenute dal sistema: nel primo caso un programma di elaborazione di immagini può estrarle automaticamente dall'analisi della scena, nel secondo caso è indispensabile l'intervento di un operatore umano.

In questo capitolo vedremo due importanti esempi di metadati per contenuti audiovisivi: MPEG-7 e ECHO. Il primo è uno standard internazionale sviluppato e promosso dal gruppo MPEG nel 2001, il secondo è un modello realizzando nel progetto europeo ECHO per la gestione di filmati storici.

5.2 Tipologie di metadati per documenti audiovisivi

Esistono in letteratura diverse classificazioni per i metadati che possono essere adottate per strutturare i metadati per audiovisivi. In questo capitolo utilizzeremo una classificazione che divide i metadati audiovisivi in due grandi gruppi: metadati per la *gestione del contenuto* e metadati per la *descrizione del contenuto*. Si noti come questo primo livello di raggruppamento è valido anche per metadati non solo nell'ambito del multimediale.

5.2.1 Metadati per la gestione del contenuto

Questo gruppo comprende tutti quei metadati che descrivono un oggetto generico (in teoria anche astratto) nella sua interezza, come se si trattasse di una scatola nera. Questi metadati possono essere a loro volta suddivisi in tre categorie:

- metadati di produzione
- metadati d'uso
- metadati di archiviazione

I metadati di produzione contengono informazioni sul documento audiovisivo legate alla sua creazione e produzione. Questi metadati comprendono metadati tipici quali: il titolo, la casa di produzione, i nomi degli attori, il genere, il soggetto, la lingua, etc.

I metadati d'uso contengono informazioni relative ai diritti d'autore, il tipo d'uso che ne può essere fatto (vendita, noleggio, trasmissione, etc.).

I metadati di archiviazione permettono di fornire informazioni su come è memorizzato l'oggetto multimediale (che in teoria può non essere necessariamente digitale, come ad esempio una videocassetta). In particolare questi metadati comprendono informazioni sulla codifica adottata (per esempio MPEG-1, DivX, etc.). Nel caso in cui siano disponibili più versioni dello stesso oggetto multimediale, è possibile descrivere le varie versioni, raggruppando insieme informazioni comuni a tali versioni. Per fare un esempio lo stesso video può essere presente sia in versione MPEG-1 che in DivX con risoluzioni diverse, a loro volta i file associati potrebbero trovarsi in più copie su archivi differenti.

5.2.2 Metadati per la descrizione del contenuto

Questi metadati contengono informazioni relative alla tipologia dei dati a cui sono associati e ne descrivono il loro contenuto. Ad esempio nel caso di documenti testuali i metadati della descrizione del contenuto potrebbero comprendere la struttura del testo in capitoli, sezioni, etc. In generale qualsiasi informazione utile ad identificare elementi o parti che compongono un documento può essere definita come metadato per la descrizione del contenuto.

Nel caso dei dati audiovisivi questi metadati potrebbero corrispondere alla decomposizione del video o dell'audio in segmenti (brani o scene), a caratteristiche di basso livello sull'audio o sul video, oppure a metadati di alto livello come il commento ad una scena. Questi metadati possono essere utilizzati per effettuare ricerche più approfondite su di un archivio multimediale. Ad esempio in un archivio di video contenenti servizi giornalistici televisivi potrebbe essere utile, cercare le scene (o i servizi giornalistici) che mostrano un particolare personaggio politico, oppure cercare il logo di

un marchio commerciale, individuare immagini protette da diritti di autore come fotografie vendute dall'ANSA.

Grazie all'avvento del digitale, questi metadati possono essere facilmente associati in forma binaria alla sorgente di dati che tipicamente risiede in un file di un disco rigido o un DVD. Tuttavia, il grosso vantaggio non risiede solo nella facilità di associazione, ma quanto nella possibilità di utilizzare strumenti automatici in grado, se non di generare completamente quanto meno di aiutare l'uomo nella produzione di questi metadati. In questa sezione daremo una breve panoramica delle tecniche utilizzate per estrarre metadati per la descrizione del contenuto.

Caratteristiche di basso livello per il video e per l'audio

Il video può essere visto come una serie di immagini che si susseguono in rapida sequenza, ed un sonoro ad esso sincronizzato. Le immagini contengono molte informazioni utili, non a caso la ricerca su come recuperarle basandosi sul loro contenuto informativo è da anni oggetto di ricerca. Un tipo di approccio molto utilizzato con le immagini è quello di effettuare ricerche per similarità. In generale, verificare se un'immagine è uguale ad un'altra o analizzare quanto è simile ad un'altra, è un problema complesso e difficile anche da definire. Sicuramente confrontare ogni singolo punto (*pixel*) di un'immagine con un'altra non è un modo intelligente di affrontare il problema. È preferibile estrarre delle informazioni sintetiche utili per effettuare ricerche per similarità. Queste informazioni sono chiamate *features*. Dietro una specifica feature esiste un modello matematico in grado di sintetizzare uno specifico aspetto di un'immagine: la distribuzione dei colori, la tessitura¹, le forme, etc.

La distribuzione dei colori è sicuramente il metodo più semplice ed intuitivo per analizzare la similarità di due immagini. Un concetto molto utilizzato e facile da comprendere è l'*istogramma dei colori*. Lo spettro dei colori dell'immagine viene diviso in porzioni e vengono analizzati quanti pixel con un certo intervallo di colori compaiono nell'immagine. La Figura 5.1 mostra un esempio di un istogramma dei livelli di grigio di un'immagine in bianco e nero. Ogni elemento dell'istogramma indica quanto è frequente il livello di grigio corrispondente. Naturalmente se l'immagine fosse a colori questa potrebbe essere scomposta per esempio nei tre colori fondamentali (rosso, verde, blu), e verrebbero creati tre istogrammi per ognuno dei colori.

¹La tessitura rappresenta un mezzo utile per mettere in luce regolarità nell'intensità dei colori in immagine.

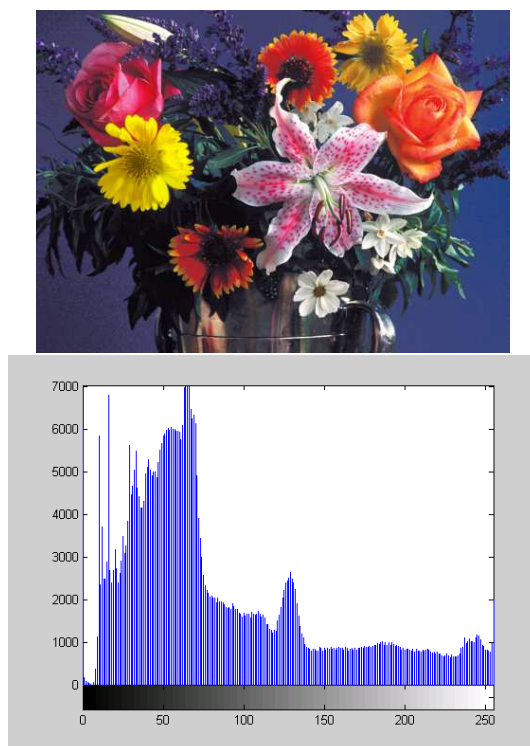


Figura 5.1: Istogramma dei livelli di un'immagine campione in bianco e nero.

La similarità di due immagini può essere in questo modo valutata misurando la differenza fra i due istogrammi corrispondenti, vale a dire sommando i valori assoluti delle differenze della frequenza di ogni colore. Il risultato rappresenta la dis-similarità in quanto una differenza nulla indicherebbe una similarità perfetta. Si noti come due immagini risultanti perfettamente simili non sono necessariamente uguali. Per convincersene basta pensare di confrontare un'immagine qualsiasi con la stessa capovolta, i loro istogrammi dei colori risulterebbero identici.

È importante notare come queste features se confrontate singolarmente possono portare a risultati inaspettati, ad esempio un cetriolo verde può essere tranquillamente essere riconosciuto simile ad una bottiglia verde, se si analizzano solo i colori. Tuttavia, se la similarità viene calcolata analizzando contemporaneamente ai colori anche la forma e la tessitura, sarà possibile discriminare meglio i due oggetti.

In realtà realizzare un sistema che permetta di cercare su tutti i fotogrammi di un filmato è poco vantaggioso e dispendioso. Difatti un video realizzato nello standard televisivo europeo PAL è composto da 25 fotogrammi al secondo. Quindi un'ora di video in PAL è composto da $25 \times 60 \times 60 = 90.000$ fotogrammi, una quantità enorme se si pensa che un sistema di archiviazione può contenere migliaia di ore di video, inoltre molti di questi fotogrammi sono simili tra loro. Per ovviare a questo problema si utilizza solo un sotto insieme dei fotogrammi del video, i keyframes (fotogrammi chiave). Un keyframe è un fotogramma particolare che rappresenta adeguatamente una scena di un video. I keyframes vengono generati durante la fase di identificazione delle scene discussa nel seguente paragrafo.

Considerare il video come una semplice sequenza di immagini può essere riduttivo, infatti i fotogrammi di un video sono tra loro correlati temporalmente. L'estrazione di features del moto forniscono un modo efficace per effettuare ricerche attraverso la dimensione temporale. Queste features permettono di descrivere in modo sintetico informazioni sul movimento di oggetti nel video o della cinepresa.

Un esempio tipico è la feature che descrive la quantità di moto presente in una data scena. Dato che la quantità di moto è semplicemente uno scalare (vale a dire un numero) è possibile ridurre il numero di fotogrammi su cui effettuare la ricerca usando la similarità per immagini. Ad esempio volendo cercare delle scene di calcio da programmi televisivi può essere utile ridurre l'insieme di fotogrammi da cercare a quelli che possiedono una quantità di moto superiore ad una certa soglia. Dopodichè è possibile selezionare le immagini che contengono ad esempio un prato verde.

Un altro elemento importante di un documento multimediale che può essere analizzato è la traccia audio. Oltre al parlato di cui si discuterà più avanti, può essere utile estrarre altre informazioni riguardanti la traccia audio come ad esempio, effetti sonori, esplosioni, musica, etc. Per quanto riguarda la musica esistono sistemi automatici per l'estrazione di features che descrivono la melodia, il tempo, il timbro, di una traccia audio contenente musica. Sono stati anche realizzati strumenti per la ricerca di melodie musicali sulla base di un motivo fischiato o cantato dall'utente.

Oggetti

Il riconoscimento del contenuto di un video è senz'altro la sfida più importante che vede impegnati molti ricercatori sia del mondo accademico che industriale. Il miglioramento di queste tecniche di riconoscimento potrebbe un giorno colmare quello che è chiamato in inglese semantic gap, ossia il

divario semantico, che rappresenta oggi il più grosso ostacolo nella ricerca sui dati multimediali. Il semantic gap è in pratica la differenza tra quello che l'utente percepisce e quello il sistema automatico riconosce. Questo divario probabilmente non potrà mai essere del tutto colmato in quanto la percezione umana è qualcosa di soggettivo; inoltre, attualmente i sistemi automatici riescono a fare ben poco. Ad esempio riescono a riconoscere automaticamente abbastanza bene scritte all'interno di un'immagine (e quindi in un video), e si riesce ad identificare un certo insieme di oggetti bene definiti come automobili, animali, volti, etc. Attenzione però a non confondere il concetto di identificazione con riconoscimento. Il primo implica semplicemente l'individuazione all'interno di un'immagine di un'area in cui probabilmente è contenuto un oggetto noto (ad esempio un volto); il secondo, oltre l'identificazione del volto implica anche, appunto, il riconoscimento della persona.

Il riconoscimento di scritte all'interno di un'immagine non è un'operazione complessa. In una prima fase si identificano le regioni che contengono testo, isolando parti dell'immagine che hanno elementi peculiari dei caratteri tipografici. In seguito i segmenti individuati vengono elaborati in modo da accentuare i caratteri rispetto al resto, aumentando ad esempio il contrasto. In fine la parte estratta viene analizzata da un programma OCR (Optical character recognition) che estrae il testo.

Algoritmi per l'identificazione e il riconoscimento di volti sono sicuramente richiesti in svariate applicazioni pratiche come ad esempio nel giornalismo o nella sicurezza. Naturalmente, per procedere al riconoscimento di un volto è necessario prima indentificarlo. Esistono decine di metodi di identificazione di volti, la maggior parte si basano sui colori o sulla ricerca di schemi predefiniti come ad esempio occhi-naso-bocca. I più promettenti utilizzano tecniche che si basano sull'approccio delle reti neurali. I problemi da affrontare nell'identificazione di un volto sono svariati. Un volto può essere illuminato in modo insufficiente, trovarsi di profilo, ci possono essere altri volti nella stessa inquadratura o esserci altri elementi di disturbo come animali o cose. Sono quindi frequenti falsi positivi e falsi negativi.

Un volto, una volta identificato (tipicamente viene racchiuso in un rettangolo), può essere inviato ad una procedura per il suo riconoscimento. Per questa fase è necessario avere uno o più modelli della persona da riconoscere, non è pensabile realizzare un sistema che riesce a riconoscere chiunque. I modelli possono essere utilizzati per creare un base di conoscenza con la quale è possibile annotare un video con i personaggi riconosciuti. Ad esempio in video di un telegiornale, può essere utile identificare il volto di personaggi politici, personaggi dello sport, etc.

L'identificazione di oggetti è una problematica ancora più ampia della precedente (se vogliamo comprende anche quello dei volti). In generale nel caso di oggetti si parla semplicemente riconoscimento piuttosto che identificazione in quanto, quasi sempre, il secondo implica necessariamente il primo. Se devo identificare un cavallo necessariamente lo avrò anche riconosciuto come tale. Questa distinzione è però necessaria per oggetti particolari come le scritte e i volti.

Quando si cerca di individuare diversi oggetti in un'immagine generalmente si procede ripartendo l'immagine in segmenti significativi che sono poi confrontati con una base di conoscenza costituita da modelli vari, ad esempio aeroplani, automobili, biciclette, etc.

Segmentazione del video

La segmentazione del video è la decomposizione temporale del contenuto visuale in unità più piccole. Analogamente a quanto accade per il testo, è tipicamente organizzato in capitoli, sezioni, paragrafi, etc., anche un video può essere organizzato temporalmente secondo una gerarchia di segmenti. I segmenti del video sono normalmente (dal più grande al più piccolo) noti come *sequenze*, *scene*, *shots* e fotogrammi. Lo shot è formato da una serie di fotogrammi che identificano una singola azione della telecamera. La scena è una sequenza di shots che appartengono ad uno stesso contesto, ad esempio la scena di un inseguimento d'auto. In fine, una sequenza è un segmento video composto da più scene correlate semanticamente, ad esempio un episodio all'interno di un film.

Il processo di segmentazione si basa sul partizionamento delle sequenze video in shots che sono più facili da identificare. Gli shots vengono di solito rilevati automaticamente e sono rappresentati da uno o più keyframes. Gli algoritmi per l'estrazione degli shots sono utilizzati anche da programmi per l'editing video per facilitare la navigazione nel video e si basano sulla determinazione delle transizioni temporali da uno shot all'altro. Esistono due tipi di transizioni, quelle improvvisate (*cuts*) e quelle graduali, come la dissolvenza. Un caso particolare di dissolvenza sono la *fade-in* e la *fade-out*), che sono rispettivamente delle transizioni del video dal buio a una scena illuminata e viceversa.

Dal punto di vista tecnico le transizioni vengono rilevate analizzando le differenze di due fotogrammi consecutivi. Queste differenze possono essere rilevate osservando la variazione di colore dei singoli pixel, se molti pixel dei due fotogrammi sono cambiati allora esiste una buona probabilità di aver individuato una transizione. Gli svantaggi di questa tecnica sono due: è mol-

to sensibile ai movimenti della cinepresa, e bisogna lavorare sul video già in forma decompressa. Quest'ultimo aspetto può essere di qualche rilevanza, se si ricorda che i documenti audiovisivi sono quasi certamente disponibili in formato compresso (MPEG-1/2, DivX, etc.)² e che la fase di decompressione può avere un certo peso sul tempo di elaborazione totale. A volte, invece di analizzare singoli pixel, si analizzano alcune statistiche di blocchi di pixel (colore medio, varianza del colore, etc.).

Un altro tipo di approccio analizza le differenze dell'istogramma dei colori di due fotogrammi successivi. Se la differenza tra i due istogrammi supera una certa soglia predefinita, il punto di passaggio tra i due fotogrammi viene marcato come transizione. Algoritmi più sofisticati analizzano la variazione su un tempo più lungo (ossia un numero di fotogrammi maggiore di due). In modo da rilevare anche variazioni lente di una scena.

Alcuni algoritmi sono in grado di lavorare sul flusso digitale non compresso, con l'ovvio vantaggio di non dover perdere tempo nella fase di decompressione. Infatti, molte features possono essere estratte direttamente nel video compresso, come ad esempio vettori di movimento.

L'identificazione delle scene, vale a dire il rilevamento di transizioni nel contenuto audiovisivo dal punto di vista semantico, invece che fisico (come quello degli shots), è molto più difficile da ottenere ed è ancora oggetto di ricerca. Le soluzioni adottate richiedono un alto livello di analisi del contenuto audiovisivo, e si basano su tre tipi di strategie:

1. l'identificazione di manifestazioni locali (dal punto di vista temporale) basata sulle regole cinematografiche di produzione, che possano far pensare ad una transizione più macroscopica. Come ad esempio, effetti di transizione, comparsa/scomparsa di musica dalla colonna sonora.
2. il raggruppamento secondo vincoli temporali: si basa sull'idea che gruppi di contenuti correlati semanticamente tendono ad essere localizzati temporalmente. Quindi solo gli shots che cadono all'interno di una predefinita soglia temporale vengono eventualmente aggregati in un'unica scena.
3. l'uso di modelli a priori sulla base soggetto: si affidano alla conoscenza a priori del tipo di contenuto: notizie, sport, etc.

²Fanno eccezione i filmati prodotti dalle videocamere digitali MiniDV, che comprimono solo i singoli fotogrammi.

Estrazione del parlato

La ricerca di parole o frasi all'interno della trascrizione del parlato può essere sorprendentemente efficace nel recupero di informazioni in documenti audiovisivi. L'utilità del testo estratto dipenderà naturalmente dal tipo di documento audiovisivo. Ad esempio il parlato in un documentario o di un telegiornale è fortemente correlato con quello che è mostrato nel video. Viceversa, in un film il parlato caratterizza poco una scena dal punto di vista semantico, in quanto i personaggi in un film normalmente (come d'altra parte succede nella realtà) non commentano le scene in cui sono coinvolti. Inoltre, mentre in un documentario c'è un solo narratore che parla e che scandisce bene le parole, in un film non succede altrettanto e spesso più persone che sono coinvolte nella stessa conversazione parlano insieme.

L'estrazione del parlato è un processo particolarmente complesso, che si basa su principi simili a quelli utilizzati dai programmi di dettatura. A differenza di questi ultimi però l'estrazione del parlato deve fronteggiare diversi problemi come l'ambiente (sovrapposizione di rumori, suoni, musica o riverberi), la qualità di registrazione, le caratteristiche vocali della persona che parla (sesso, età, proprietà di pronuncia, stato emotivo).

Ovviamente il risultato di questi algoritmi è affetto da errori, però è stato dimostrato che anche con una percentuale di errori del 50%, l'efficacia della ricerca sul testo è soddisfacente. Il riconoscimento del parlato utilizza le seguenti tre fasi di riconoscimento:

- *Riconoscimento acustico.* Utilizza un modello che descrive il suono dei singoli fonemi del parlato.
- *Riconoscimento delle parole.* Utilizza un modello del lessico che descrive quali sequenze di fonemi rappresentano parole valide.
- *Riconoscimento di frasi.* Utilizza un modello linguistico che determina la probabilità che una specifica parola sia stata pronunciata, sulla base della lingua utilizzata nel parlato.

5.3 Note sullo standard MPEG-7

5.3.1 Introduzione

MPEG-7 è uno standard realizzato dal gruppo MPEG (Moving Picture coding Experts Group), gli stessi degli standard MPEG-1, MPEG-2 ed MPEG-4. Il gruppo MPEG nasce nel 1988 all'interno di ISO, l'International Stan-

dard Organisation, per sviluppare gli standard di codifica per la rappresentazione di immagini in movimento e audio, nel contesto della memorizzazione di tali informazioni su supporti digitali (Digital Media Storage).

Lo scopo MPEG-1, MPEG-2 ed MPEG-4 è stato quello di realizzare standard per la rappresentazione, memorizzazione e compressione di contenuti digitali di tipo audiovisivi. Nel caso di MPEG-1 e 2 lo scopo principale è quello di permettere la codifica di video in forma digitale compressa. Lo standard MPEG-4 introduce, tra le tante cose, l'interattività (ossia la possibilità di creare contenuti digitali in grado di interagire con l'utente attraverso la tastiera e il mouse) e la possibilità di creare oggetti sintetici 3D. Mentre nel caso di MPEG-1 e 2 il contenuto digitale può essere visto come un unico flusso audio/video con MPEG-4 è possibile creare, mappe interattive, animazioni, interfacce di apprendimento, etc.

Per avere una portata dell'importanza di questi standard si pensi che MPEG-1 viene molto utilizzato per distribuire video nel Web, e il suo formato compressione audio (chiamato MPEG-1 Audio Layer III) è noto come MP3. MPEG-2 viene utilizzato per trasmettere Video attraverso il satellite (la comune TV satellitare digitale) ed è usato come formato di codifica per i video dei DVD. MPEG-4 è soprattutto adottato per il suo elevato livello di compressione video. La incarnazione più famosa di MPEG-4 è il DivX (molto utilizzata per registrare interi film in pochissimo spazio) che è una versione di MPEG-4 realizzata da Microsoft, successivamente "rubata" eliminando alcune limitazioni inserite da Microsoft e resa disponibile al pubblico.

La compressione video è importantissima, per convincersene si pensi che se non si usassero sistemi di compressione per i dati, un solo minuto di video registrato su un hard disk o su un DVD come semplice sequenza di immagini occuperebbe ben 1.270 MB. Questo significa che un comune film di 95 minuti (1h e 35') a 25 fotogrammi/sec occuperebbe 120 GB e un comune DVD da 4.7 GB potrebbe contenere solo 4 minuti di video (circa il 4%).

Comprimendo un video utilizzando MPEG-2 in un DVD, vale a dire 4.7GB, si riescono ad inserire circa 80 minuti di filmato (si ricordi che DVD-ROM commerciali hanno una dimensione doppia in quanto dotati di un doppio strato). Utilizzando DivX si riescono ad inserire nello stesso spazio circa cinque ore di film con una qualità paragonabile al DVD-ROM commerciale.

L'MPEG-3 non è mai esistito in quanto era stato inizialmente pensato per la TV digitale ad alta risoluzione, ma per ottenere tale qualità è stato sufficiente apportare alcuni piccoli miglioramenti allo standard MPEG-2.

L'MPEG-7 nasce con uno scopo totalmente diverso dai i suoi predecessori. Mentre MPEG-1, MPEG-2 ed MPEG-4 servono per rappresentare il

contenuto multimediale (the bits), MPEG-7 serve per rappresentare informazione sul contenuto (the bits about the bits) ossia i bits che parlano di bits, in pratica si parla di metadati.

5.3.2 Il modello di metadati di MPEG-7

Il lavoro di definizione dello standard MPEG-7 è iniziato nel 1996 ed ha portato ad una prima versione nel 2001. Originariamente denominato “Multimedia Content Description Interface” ha lo scopo di introdurre uno standard per la descrizione di contenuti audiovisivi in modo da permetterne la ricerca.

MPEG-7 consente di associare ai contenuti audiovisivi una specie di etichetta costituita da un insieme di “descrittori” che ne rappresentano le caratteristiche principali: nel caso di un brano musicale, ad esempio la sequenza di note oltre naturalmente al nome dell’autore, del cantante, la casa discografica, l’anno di pubblicazione; nel caso di un film, invece, la trama, il nome del regista, degli attori, del produttore, dello sceneggiatore, la colonna sonora ed anche il trailer e le scene principali.

Attraverso interrogazioni (*query*) apposite sarà possibile effettuare ricerche sui contenuti multimediali presenti in Rete, esattamente come già facciamo sui testi attraverso gli appositi motori di ricerca, per trovare esattamente ciò che cerchiamo.

Oltre ad informazioni di tipo testuale, come autore, titolo, data di creazione, etc. MPEG-7 permette di associare al contenuto audiovisivo anche features dell’audio e del video, che sono state descritte nella precedente sezione. Al fotogramma di un video MPEG-7 permette di associare delle features del colore, forma, e tessitura. Ricevendo sul nostro computer un filmato così arricchito sarà possibile effettuare una ricerca all’interno del video utilizzando come interrogazione una nostra immagine e ottenendo come risultato i fotogrammi del video più simili. In modo del tutto analogo esistono features audio che permettono di descrivere, ad esempio, il timbro di uno strumento musicale. Queste features nello standard MPEG-7 vengono chiamati Descrittori (*Descriptors*).

È importante puntualizzare che MPEG-7 definisce solo il formato da usare per creare queste descrizioni e non le metodologie usate per estrarle. Infatti, per lo scambio, ricerca, ecc. è necessario conoscere solo il formato delle descrizioni, non come sono state ottenute. Questo permette di avere uno standard che si “adatta” alle evoluzioni tecnologiche (nuovi algoritmi di estrazione di features possono essere utilizzati senza modificare lo standard).

In particolare lo standard MPEG-7 definisce i seguenti componenti:

Descriptors (D) Rappresentazione di una feature. Definisce la sintassi, il

tipo di dato, i valori permessi e la semantica della rappresentazione di una feature (ad esempio RGB-Color:[integer, integer, integer], vale a dire che un colore è descritto da tre numeri interi che rappresentano le intensità del rosso, verde e blu);

Description Schemas (DS) Insieme di Descriptors e di altri Description Schemas. Definisce anche la struttura e la semantica delle relazioni tra i vari Descriptors e Descriptor Schemes.

Description Definition Language (DDL) Permette di definire nuovi Descriptors e Descriptor Schemes e di estendere quelli esistenti. Meccanismo utilizzato per l'estensione del modello.

Prima di fare qualche esempio che permetta di capire meglio di cosa stiamo parlando, bisogna dire che il linguaggio scelto per descrivere le istanziazioni di Descriptors e Descriptor Schemes di MPEG-7 è XML. Questo perché XML è molto flessibile e potente. Inoltre, esiste un gran numero di strumenti per creare, processare e manipolare documenti scritti in XML. Anche il DDL è basato su XML, in realtà si tratta di una versione estesa dell'XML Schema del W3C di cui si è accennato nel capitolo 2.

Descriptors

I Descriptors sono le features predefinite di MPEG-7 e ne descrivono la sintassi e la semantica. Benché in MPEG-7 non esista una vera e propria definizione formale, generalmente per Descriptors si intende una feature percettiva di basso livello di un oggetto multimediale che può essere estratta automaticamente. Ad esempio un istogramma dei colori di un'immagine. Esistono comunque anche Descriptors per features di alto livello per oggetti semantici, eventi e concetti astratti, etc., i quali sono tipicamente prodotti attraverso l'intervento umano.

In MPEG-7 vi sono sia Descriptors visuali e sia Descriptors per l'audio. Per fare un esempio concreto consideriamo il descrittore visuale *Dominant Color* (colore dominante), che, come si intuisce, fornisce una descrizione compatta dei colori rappresentativi in un'immagine o una regione di un'immagine. Ad esempio un'immagine completamente rossa verrebbe descritta con il descrittore Dominant Color con il seguente spezzone XML:

```
<VisualDescriptor xsi:type="DominantColorType">
  <SpatialCoherency>31</SpatialCoherency>
  <Value>
    <Percentage>31</Percentage>
```

```
<Index>255 0 0</Index>
  <ColorVariance>1 0 0</ColorVariance>
</Value>
</VisualDescriptor>
```

Senza entrare nel dettaglio dei vari campi, l'elemento **Percentage** definisce la percentuale di colore descritto nell'elemento **Index** in RGB che vanno da 0 a 255. Infatti, in **Index** vi è la definizione del colore rosso: 255 per il Rosso, 0 per il Verde e 0 per il Blu. Per quanto riguarda **Percentage** il valore 31 identifica corrisponde al 100% (la scala delle possibili percentuali è divisa in 32 valori da 0 per lo 0% fino a 31 per il 100%.)

Descriptor Schemes

Attraverso un Descriptor Scheme, MPEG-7 permette di combinare insieme più Descriptors o anche altri Descriptor Schemes per creare delle strutture più complesse. Alcuni Descriptor Schemes sono predefiniti nello standard MPEG-7, altri possono essere definiti attraverso il linguaggio DDL.

Con un Descriptor Scheme si può mettere, ad esempio, insieme una descrizione testuale di una scena e alcuni descrittori visuali per un keyframe della scena.

I Descriptors e Descriptor Schemes sono organizzati nello standard in tre gruppi fondamentali: *Multimedia Description Schemes*, *Visual Descriptors* e *Audio Descriptors*.

Il primo gruppo contiene una serie di Descriptors e Descriptor Schemes che permette di descrivere oggetti di tutti i tipi visivi, audio e testuali. Mentre gli altri due gruppi sono usati per descrivere oggetti prettamente visuali e audio, rispettivamente. Per esempio, il Descriptor Dominant Color visto sopra fa parte Visual Descriptors.

A loro volta i tre gruppi visti sopra sono organizzati in sotto-categorie.

Multimedia Description Schemes in particolare è organizzato nelle seguenti sotto categorie: *Basic Elements*, *Content Management*, *Content Description*, *Navigation and Access*, *Content Organization* e *User Interaction*. Senza entrare nel dettaglio di ogni singola categoria, ma per avere un'idea possiamo dire che ad esempio Basic Elements definisce un insieme di elementi basilari che permettono poi di realizzare oggetti più complessi. Ad esempio, vettori di numeri, stringhe, riferimenti, identificatori univoci, etc. Ad esempio, per definire direttamente un'immagine è possibile utilizzare il formato *inline* di **ImageLocator** come segue:

```
<ImageLocator>
```

```

<InlineMedia type="image/jpeg">
  <MediaData>/9j/4AAQSkj2345h234h5k2h34...</MediaData>
</InlineMedia>
</ImageLocator>

```

Content management permette di descrivere informazioni collegate all'amministrazione del contenuto. Dati legati al tipo di compressione, multimediale, proprietario, etc.

Content description permette una rappresentazione delle informazioni percepibili (contenuti del documento multimediale). Ad esempio si possono definire segmenti di un video come scene, oppure aspetti semantici e concettuali come la descrizione concettuale di un'immagine in cui due persone si stringono la mano.

Content organization fornisce Descriptor Schemes per l'organizzazione di collezioni dati multimediali. In questo modo è possibile rappresentare proprietà di collezioni di oggetti in base alle loro proprietà comuni.

Navigation and access fornisce Descriptor Schemes per facilitare la navigazione ed il recupero definendo sommari, viste e versioni dello stesso oggetto multimediale.

In fine User Interaction permette di descrivere le preferenze e la storia dell'uso di un certo oggetto multimediale. Lo scopo è quello di consentire la personalizzazione dell'accesso ai contenuti multimediali.

Per quanto riguarda Visual Descriptors esso è diviso in: *Color*, *Shape*, *Texture* e *Motion*. Queste quattro sotto categorie contengono definizioni di Descriptors del colore, tessitura, forma e movimento.

Audio Descriptors comprende una serie di Descriptors e Descriptor Schemes che riguarda la scala, i mezzi per la descrizione di effetti sonori, i descrittori degli strumenti e il riconoscimento del parlato e di segmenti di silenzio uniforme.

Description Definition Language

DDL è il linguaggio per la definizione di nuovi Descriptor Schemes e nuovi descrittori. Permette l'*importazione* di dichiarazioni di schemi MPEG-7, la *ridefinizione* di Descriptors o Descriptor Schemes esistenti, la *restrizione* di certi aspetti di Descriptors o Descriptor Schemes esistenti, l'*estensione* di Descriptors o Descriptor Schemes.

DDL di fatto è una estensione del linguaggio XML Schema di cui si fatto cenno nel Capitolo 2. Le estensioni riguardano la definizione di tipi di dati

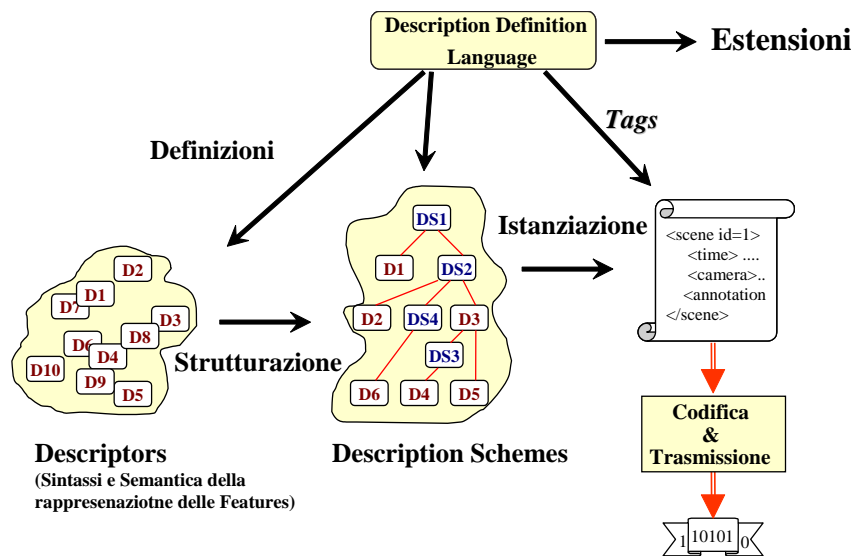


Figura 5.2: Componenti dello standard MPEG-7

non supportati da XML Schema utili per dichiarare array, matrici, istanti e intervalli temporali.

DDL è organizzato in:

- XML Schema Structural Language Components;
- XML Schema Data-Type Language Components;
- MPEG-7 Specific Extensions.

I primi due fanno parte dello standard W3C di XML Schema e concernono la dichiarazione di strutture XML e tipi di dati. L'ultima parte riguarda le specifiche estensioni introdotte da MPEG-7 di cui si è parlato sopra.

La Figura 5.2 mostra uno schema riassuntivo di tutte le componenti dello standard MPEG-7. Alla fine il prodotto di tutte queste parti è un documento XML che descrive uno o più oggetti multimediali. La trasmissione di questo documento avviene normalmente in forma binaria, per cui MPEG-7 standardizza come deve essere trasformato il documento MPEG-7 (XML) in binario.

5.4 Il modello di metadati di ECHO

5.4.1 Introduzione

ECHO (European CHronicle On-line) è un progetto finanziato dall'Unione Europea nell'ambito del V programma quadro, realizzato con lo scopo di fornire una biblioteca digitale per l'accesso remoto a collezioni di documentari storici audiovisivi. L'architettura di ECHO fornisce un'infrastruttura software estendibile ed interoperabile per il supporto ad archivi di video digitali.

5.4.2 Il modello Entità-Relazione

Per affrontare meglio la trattazione del modello di metadati di ECHO in questo paragrafo parleremo del modello descrittivo di tipo Entità-Relazione.

Un modello di metadati definisce in modo chiaro ed univoco i metadati utilizzati da una biblioteca digitale. Esso è associato ad uno schema di una base di dati (dove sono memorizzati i metadati veri e propri). Nel caso di ECHO poiché i metadati sono memorizzati in un base di dati XML, il modello di metadati è realizzato attraverso uno schema XML. Descrivere il modello di ECHO parlando del suo schema XML è tuttavia scomodo, a tale scopo utilizzeremo invece il modello Entità-Relazione più semplice e comprensibile.

Le Entità rappresentano classi di oggetti (cose, fatti ecc.) con proprietà comuni ed esistenza propria, indipendente dall'applicazione. Ogni entità possiede un certo numero di attributi, che sono i campi dei metadati e corrispondono a elementi XML. Un'occorrenza di una entità è un oggetto della classe relativa. Ad esempio, nel modello ECHO esiste un'entità chiamata *Media* che descrive una realizzazione di un documento audiovisivo come ad esempio un file MPEG contenente un filmato. I campi di metadati descrivono ad esempio il formato utilizzato, la compressione, il nome del file, etc.

Le relazioni permettono di collegare tra loro le entità. Nel caso del modello ECHO una relazione è realizzata attraverso uno speciale campo che contiene uno o più URNs (Unique Resource Name) che altro non sono che dei identificatori unici di documenti XML. Date due entità arbitrarie chiamate A e B esistono tre tipi di relazioni che possono legarle:

- *uno-a-uno*, fa corrispondere 1 oggetto di A ad 1 oggetto di B. Non può esistere alcuna oggetto, né di A né di B, che compaia due volte, abbinato ad oggetti diversi.

- *uno-a-molti*, in tal caso ad un oggetto di A possono corrispondere più oggetti di B. Non è, invece possibile che lo stesso oggetto di B corrisponda a oggetti distinti di A.
- *molti-a-molti*, in tal caso a oggetti distinti di A possono corrispondere oggetti distinti di B e viceversa.

Chiaramente la scelta di utilizzare XML è una delle tante possibili. Infatti, si poteva adottare un altro tipo di base di dati per realizzare le relazioni del modello di ECHO come ad esempio una base di dati relazionale. In tal caso le entità sarebbero state istanziate da tabelle invece che da documenti XML.

5.4.3 Il modello di ECHO

Il modello di metadati adottato in ECHO è un'estensione del modello di metadati *IFLA/FRBR*. Questo modello è strutturato in quattro livelli che descrivono diversi aspetti di un lavoro intellettuale o artistico: *Work*, *Expression*, *Manifestation* e *Item*. Le entità del modello sono organizzate secondo una gerarchia che ricalca questa organizzazione a livelli. In particolare, esistono quattro entità *AVDocument*, *Version*, *Media* e *Storage*, che corrispondono rispettivamente ai quattro livelli di *IFLA/FRBR*.

In accordo alla metodologia di *IFLA/FRBR* l'entità *AVDocument* è la più astratta. Essa fornisce la vista generale del lavoro intellettuale o artistico. Per esempio, supponiamo di voler descrivere un documento audiovisivo che ha per tema "Giochi Olimpici di Berlino del 1936". Un oggetto di tipo *AVDocument* rappresenterà l'idea astratta del documentario sui Giochi. Vi saranno un certo numero di oggetti di tipo *Version* che rappresentano diverse versioni dello stesso film, per esempio, una versione in italiano, una in inglese oppure una versione ridotta.

Tuttavia, un oggetto di tipo *Version* non rappresenta una specifica realizzazione di un film. Questo aspetto può essere espresso per mezzo del livello di *Manifestazione*, attraverso l'entità *Media*. Per esempio, la versione dei Giochi in italiano potrebbe possedere due oggetti *Media* una corrispondente alla codifica MPEG ed un secondo oggetto corrispondente alla codifica in DivX.

Ciò nonostante, l'entità *Media* non corrisponde ad un'implementazione fisica vera e propria. Questo aspetto è espresso dall'entità di tipo *Storage*. Ogni *Storage* corrisponde ad una versione fisica veramente esistente, un DVD, un file, etc. Questo è utile per rappresentare copie multiple dello stesso oggetto. Ad esempio, lo stesso *Media* corrispondente alla codifica MPEG

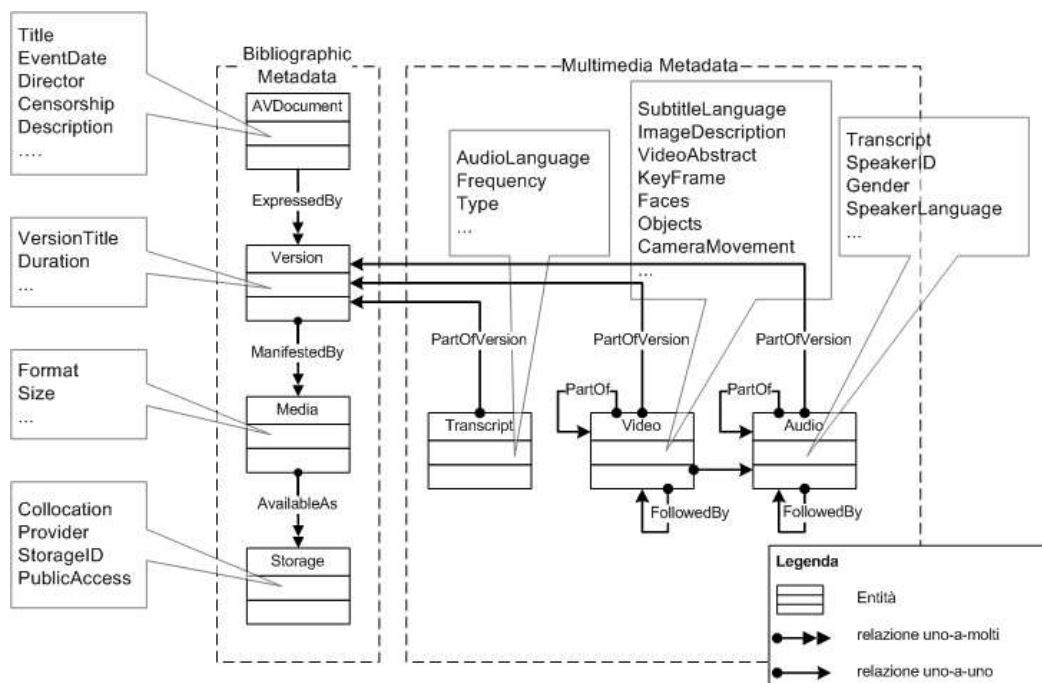


Figura 5.3: Rappresentazione schematica del modello di metadati di ECHO

potrebbe essere disponibile su più elaboratori, oppure essere memorizzato su di un CD-ROM depositato su uno scaffale.

La Figura 5.3 mostra uno schema del modello dei metadati di ECHO, insieme ai nomi di alcuni attributi delle entità del modello.

Si noti, come la relazione fra le entità appena descritte è di tipo uno-a-molti, che significa, ad esempio che un oggetto AVDocument può possedere uno o più oggetti Version, che a sua volta può possedere una o più oggetti di tipo Media, e così via.

Per comodità le entità del modello di ECHO sono state organizzate secondo il raggruppamento visto sopra tra metadati per la *gestione* del contenuto e metadati per *descrizione* del contenuto. In particolare al primo gruppo, chiamato "Bibliographic Metadata" è costituito dalle quattro entità viste sopra. Il secondo gruppo, chiamato "Multimedia Metadata", contiene le entità (*Video*, *Audio* e *Transcript*). Formalmente queste entità appartengono al livello Expression del modello IFLA/FRBR e dispongono di una relazione di tipo *PartOfVersion* all'entità Version cui appartengono di tipo uno-a-uno. Questo significa che ad ogni oggetto di tipo Version possiamo

associare un oggetto Video, un oggetto Audio e un oggetto Transcript. Essi rappresentano rispettivamente la segmentazione della parte visuale, audio e del parlato. Gli oggetti che riferiscono all'oggetto Version attraverso la relazione `PartOfVersion` sono indicati come oggetti principali e rappresentano il segmento temporale che corrisponde al documento audiovisivo nella sua interezza.

Per rappresentare segmenti più piccoli (ad esempio una scena) si utilizza una nuova istanza di Video, Audio o Transcript utilizzando la relazione uno-a-uno `PartOf`. Ad esempio, l'intero filmato sui Giochi Olimpici sarà rappresentata da un oggetto di tipo Video, due attributi temporali indicano l'inizio e la fine della scena, che, in questo caso, coincideranno con l'inizio e la fine del filmato. Una sequenza all'interno del filmato sarà rappresentato da un oggetto video con opportuni limiti temporali, che avrà come `PartOf` l'URN dell'oggetto Video principale. In modo analogo potrà esserci una scena all'interno di quest'ultima sequenza, e così via. I segmenti allo stesso livello sono collegati tra loro attraverso la relazione uno-a-uno `FollowedBy` che permette di esprimere la successione temporale e descrivere il tipo di transizione tra un segmento e l'altro (dissolvenza, taglio, etc.).

La Figura 5.4 mostra un esempio di un'istanza del modello ECHO. Il documentario di breve durata tratta i problemi urbanistici della città di Palermo nel 1955 mostrando alcune scene del traffico cittadino. Nell'esempio si è immaginato di possedere una versione in lingua italiana e una in francese. La versione in italiano è mostrata in dettaglio, ed il filmato è disponibile in formato MPEG e in AVI, rappresentate da altrettanti oggetti di tipo Media. Esistono poi due copie fisiche dei due documenti audiovisivi, ognuna corrispondente ai due oggetti Media incorporati da oggetti di tipo Storage.

Sulla parte destra della Figura dell'esempio è mostrata la segmentazione dell'audio e del video. I tre oggetti intitolati "Video Principale", "Audio Principale" e "Transcript Principale", rappresentano rispettivamente, come detto, l'intera traccia video, audio e transcript. La traccia audio è suddivisa in tre parti che corrispondono a tre segmenti distinti dal punto di vista acustico: all'inizio si sente della musica, poi parla un uomo, e infine si termina con la musica. La traccia video è suddivisa in quattro scene principali, in particolare la seconda scena è suddivisa in quattro shots. Si noti il significato del verso delle frecce corrispondenti alle relazioni: la freccia parte dall'oggetto che possiede il puntatore all'oggetto indirizzato. Questo significa che esisterà, ad esempio, un documento XML corrispondente che rappresenta la scena intitolata "Traffico cittadino" avrà un elemento `PartOf` che contiene un riferimento (URN) all'oggetto Video intitolata "Video Principale".

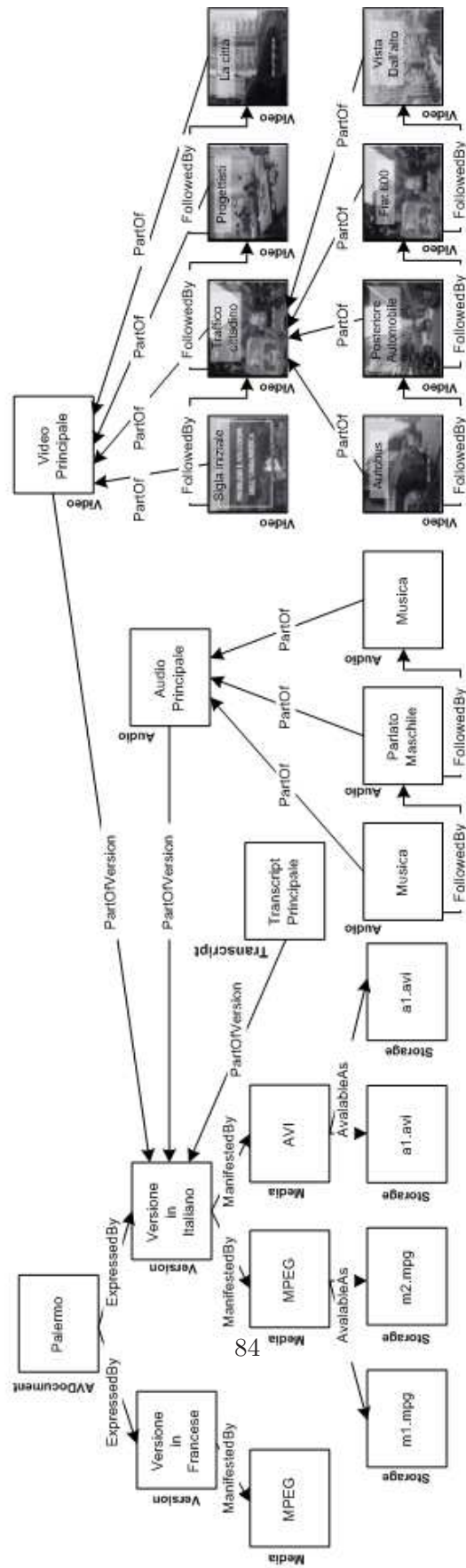


Figura 5.4: Esempio di un'istanza del modello ECHO

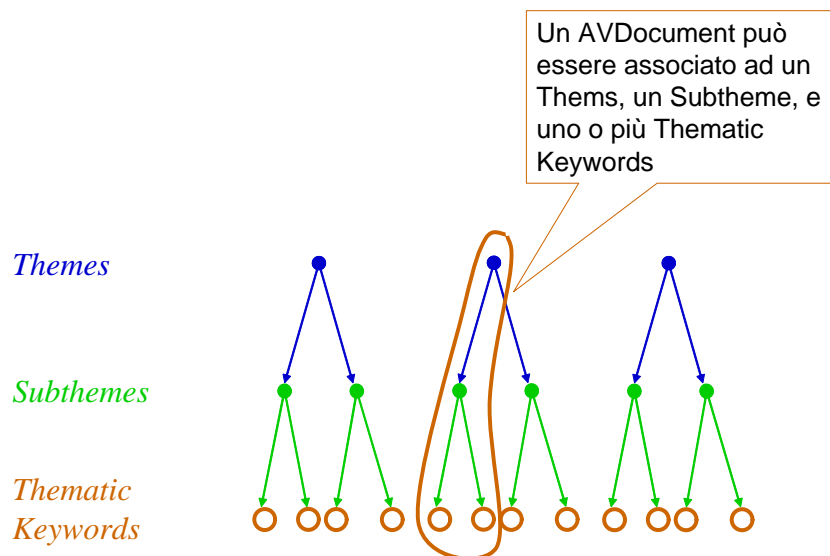


Figura 5.5: Organizzazione del campi multilinguali dei ECHO

5.4.4 Metadati Multilinguali

I documenti di ECHO sono classificati secondo una gerarchia di categorie che permette un facile accesso a gruppi di documenti inerenti lo stesso argomento. In particolare i documenti di ECHO sono stati raggruppati in un certo numero di *Themes* (Temi), corrispondenti a importanti eventi del ventesimo secolo: “Dopo Guerra”, “Le Guerre Mondiali”, “Sport nel 20° secolo”, etc. A loro volta i documenti appartenenti ad un Theme sono divisi in un certo numero di *Subthemes* (Sottotemi); ad esempio il Theme “Dopo Guerra” è suddiviso in Subthemes quali “Comunità Europea”, “Emigrazione”, etc. In fine ad ogni documento appartenente ad un Theme e un Subtheme è associato ad un certo numero di *Thematic Keywords* (Parole chiave tematiche). Ad esempio un documento del Subtheme “Comunità Europea” può essere associato alla Thematic Keyword “Piano Marshall”, “Italia ed Europa”, “Cambiamenti Geopolitici”, etc.

Tutte le parole chiave associate ai Themes, Subthemes e Thematic Keywords sono chiamati campi multilinguali in quanto, sono stati memorizzati in un documento XML (`themes.xml`) nelle quattro lingue usate nel progetto (Italiano, Tedesco, Francese ed Olandese) più l’Inglese. Ogni AVDocument possiede tre campi specifici che permettono di associare un documento ad

un Theme, Subtheme ed uno o più Thematic Keywords, tuttavia i valori contenuti in questi campi sono quelli corrispondenti alla lingua inglese. In questo modo sia l'interfaccia di ricerca di ECHO e sia l'editor dei metadati (utilizzando il documento `themes.xml`) permettono di scegliere quale delle cinque lingue utilizzare, per questi tre campi di metadati, fornendo un tipo di accesso multilinguale.

La Figura 5.5 mostra una schematizzazione dei campi multilinguali di ECHO.