Costruzione di un dataset di clip e metadati per la *knowledge extraction* della piattaforma Città Educante

OBIETTIVI DELLA RICERCA

PREMESSA

Affinché le nuove soluzioni tecnologiche che verranno sperimentate nell'ambito del progetto di ricerca Città Educante permettano l'accesso alla conoscenza e la sua condivisione, è necessario che questa venga estratta dai materiali multimediali che la contengono.

Nonostante l'ampia disponibilità di sorgenti di contenuti utili a progetti e attività educative, il loro utilizzo risulta problematico a causa delle difficoltà nell'individuazione di quelli più appropriati e del fatto che necessitano di adattamenti. Inoltre le piattaforme attualmente proposte per la fruizione di contenuti audiovisivi in ambito educativo sono piuttosto eterogenee e poco evolute.

OBIETTIVI

Innanzitutto si è inteso verificare la possibilità di impostare e sfruttare la ricerca sul Catalogo Multimediale in un'ottica appropriata a selezionare contenuti adatti a supportare attività di insegnamento e apprendimento.

È da appurare l'idoneità dei formati tecnici del materiale proxy per l'estrazione di frammenti di contenuti utilizzabili dai servizi della piattaforma Città Educante.

Si intende inoltre arricchire i metadati già ottenuti dal Catalogo Multimediale per mezzo di strumenti di documentazione automatica, opportunamente adeguati al contesto del progetto.

Infine sono da individuare le modalità di messa a disposizione dei materiali e dei metadati estratti verso la piattaforma e gli altri servizi del progetto.

CATALOGO MULTIMEDIALE

La sorgente utilizzata per selezionare i contenuti multimediali è stata il Catalogo Multimediale della RAI (CMM), un sistema di accesso all'archivio dei contenuti audio-visivi RAI in formato file che consente di fare ricerche testuali sui metadati associati, esaminare i contenuti in qualità proxy ed eventualmente richiederne una copia.

CLIP

Al fine di costruire un dataset di clip di contenuto significativo per il progetto, sono stati individuati alcuni argomenti ritenuti adatti a supportare l'attività di insegnamento e apprendimento all'interno di un contesto educativo qual è quello della Città Educante.

Tabella 1 - Argomenti dataset clip

Argomenti
Arte
Filosofia
Letteratura
Scienze
Storia

Dopodiché all'interno di ciascun argomento sono stati individuati uno o più soggetti da cercare nel Catalogo Multimediale.

Inoltre sono state sfruttate alcune delle 100 pillole di cultura del progetto Rai BIGnomi, la web

series dedicata a chi vuole ripassare le principali tematiche di storia e letteratura italiana con l'aiuto dei personaggi famosi.

Tabella 2 - Soggetti dataset clip

Soggetti	
Espressionismo	Petrarca
Impressionisti	Astronomia
Macchiaioli	Corpo Umano
Picasso	Archeologia
Platone	Emancipazione
Socrate	Etruschi
Dante	Resistenza
Nievo	Shoah

Nella lista dei risultati della ricerca sono stati selezionati alcuni dei materiali presenti in archivio, in base all'appropriatezza e idoneità nel contesto di utilizzo.

Infine per ciascun materiale ne è stata ritagliata una porzione più o meno ampia. In alcuni casi perché il soggetto di interesse era presente solo in una porzione del materiale selezionato. In altri casi, specie quelli di materiali piuttosto lunghi, ne è stata ricavata una porzione più piccola perché la clip risultante fosse maggiormente usabile.

Il risultato di questa fase è un primo insieme di 83 clip, già analizzate con gli strumenti a nostra disposizione e pronte per ulteriori elaborazioni da parte dei partner al fine di ricavarne tutti i metadati di interesse.

STRUMENTI DI ANALISI

MediaInfo è un programma open-source che estrae le informazioni tecniche di un file multimediale. Supporta i formati più comuni (ad esempio AVI, WMV, QuickTime, Real, DivX, XviD) ma anche quelli meno noti o emergenti.

Avprobe è un programma che estrae informazioni tecniche da flussi multimediali e le fornisce in output in formato leggibile. Le principali informazioni fornite sono formato, durata, dimensione, bit rate, codec, frame rate.

ANTS è un sistema integrato per l'annotazione automatica di contenuti telegiornalistici. Le principali funzionalità del sistema sono le seguenti: trascrizione automatica del parlato in testo, segmentazione del contenuto in unità informative elementari, classificazione per contenuto delle unità informative elementari ed estrazione di elementi semantici dalla trascrizione del parlato.

METADATI

Il set di clip è stato sottoposto agli strumenti appena illustrati, e ne sono stati ricavati i seguenti tipi di metadati: informazioni tecniche sul file (durata, wrapping, codec video, codec audio), "speech-to-text" e identificazione delle scene.

Inoltre fanno parte dei metadati associati al set di clip quelli già presenti sul Catalogo Multimediale, ovvero tipo di programma e tipo di prodotto, titolo del programma, canale, data, ora di messa in onda, descrizione testuale del contenuto audio e video, e nominativi di conduttori ed eventuali ospiti del programma.

PROSEGUIMENTO ATTIVITÀ

Le aree di prosecuzione ipotizzate per l'attività includono l'affinamento della configurazione di ricerca in base ai risultati della trascrizione, la creazione di mappe dei documenti audiovisivi analizzati rispetto alle configurazioni di ricerca, la segmentazione e categorizzazione dei contenuti, adattando la categorizzazione ora in uso in ambito "news" ai diversi ambiti di "insegnamento e apprendimento". Inoltre potranno risultare utili l'esplorazione delle possibilità fornite dalle tecnologie CDVS, cui si contribuisce nella standardizzazione MPEG, e l'identificazione di standard di metadatazione per le risorse pedagogiche e di apprendimento da utilizzarsi come riferimenti per la progettazione di tecnologie automatiche di classificazione.

IPOTESI DI MODELLO DI SFRUTTAMENTO

Errore. L'origine riferimento non è stata trovata. mostra un diagramma del possibile modello di sfruttamento dei contenuti dell'archivio RAI sulla piattaforma di Città Educante. Le attività svolte da RAI internamente dovranno includere la digitalizzazione e la documentazione dei materiale, la verifica della disponibilità dei diritti per la RAI e la gestione della "messa a disposizione.

Il catalogo così offerto agli utilizzatori della piattaforma potrà essere oggetto di ricerche e di richieste di permessi (diritti) secondo le intenzioni di utilizzo. La "messa a disposizione" potrà essere accordata per (1) la fruizione semplice, (2) riutilizzo, anche per la creazione di nuovi contenuti, ma limitatamente ad ambiti educativi, (3) il riutilizzo senza limitazioni.

Figura 1 - Possibile modello di sfruttamento