



**PARTNERS: SECONDA UNIVERSITÀ DI NAPOLI – JM, SPACE**

**RESPONSABILE: ING. FRANCESCO MOSCATO**

Technical Report: **TR1.2.3**

*LC3 – Laboratorio pubblico-privato di ricerca  
sul tema della Comunicazione delle Conoscenze Culturali*

**PAG 1 DI 8**

## Technical Report

### TR1.2.3

#### **Analisi dei Requisiti della Architettura di un Contlet Parser per LC3**

<i>PROGETTO LC3</i>	<i>Revisione n*</i>	<i>6</i>	<i>Del</i>	<i>22/01/2009</i>
---------------------	---------------------	----------	------------	-------------------

## Abstract

*Il seguente Technical report ha lo scopo di definire le specifiche architetture e di interfaccia per quanto riguarda Lo strumento di Contlet Parsing utilizzato all'interno dell'architettura LC3. In particolare è descritta la seguente attività:*

- *Definizione delle specifiche architetture e di interfaccia del Componente COntlet Parser Semantico*

*Le informazioni sui dati e sulle annotazioni associate ad esse sono memorizzate in formato XML, utilizzando grammatiche particolari per la definizione dei linguaggi in cui tali informazioni sono espresse. In questa attività sono state analizzate in linea di massima le grammatiche dei linguaggi definiti ed utilizzati all'interno del progetto LC3 per descrivere le informazioni di interesse, ed è stata definita l'architettura di un Parser per l'analisi di queste informazioni.*

## .1 INTRODUZIONE

Il progetto LC3 prevede la memorizzazione dei documenti testuali e multimediali all'interno di *teche digitali*. Il sistema LC3 deve poter gestire diversi tipi di documenti, da inserire nelle teche digitali, ma anche di associare ai contenuti digitali, delle informazioni volte a descriverle. I Meta-Dati permettono di descrivere queste informazioni. Anche le grammatiche da utilizzare per produrre le annotazioni semantiche e i contenuti all'interno delle teche digitali devono essere descritte in funzione di meta-dati definiti in modo formale. Questo è possibile avviene definendo appositi meta-linguaggi XML per la definizione dei contlet e delle annotazioni. Tali meta-linguaggi, e i contenuti delle teche digitali, andranno analizzati tramite opportuni parser, ovviamente basati sulle tecnologie XML.

L'esigenza innanzitutto è quella di individuare MAG che rappresentano contenuti più o meno equivalenti. Questo può accadere ad esempio quando si vanno ad inserire in MAG differenti diverse edizioni dello stesso libro, o sottoparti di un testo già presente in un altro MAG, magari memorizzate in formati differenti. In questi casi sarebbe opportuno individuare le annotazioni semantiche già prodotte, per annotare anche i MAG appena inseriti.

PROGETTO LC3	Revisione n*	6	Del	22/01/2009
--------------	--------------	---	-----	------------



**PARTNERS: SECONDA UNIVERSITÀ DI NAPOLI – JM, SPACE**

**RESPONSABILE: ING. FRANCESCO MOSCATO**

Technical Report: **TR1.2.3**

*LC3 – Laboratorio pubblico-privato di ricerca  
sul tema della Comunicazione delle Conoscenze Culturali*

**PAG 3 DI 8**

Bisogna quindi mappare annotazioni semantiche già presenti su documenti che hanno eventualmente strutture simili a quelli già annotati.

<i>PROGETTO LC3</i>	<i>Revisione n*</i>	<i>6</i>	<i>Del</i>	<i>22/01/2009</i>
---------------------	---------------------	----------	------------	-------------------

## .2 MODELLO ARCHITETTURALE

Il modello Architetturale di riferimento per la gestione e l'annotazione di contlet all'interno di teche compatibili con lo standard MAG è mostrato in Fig. 1.

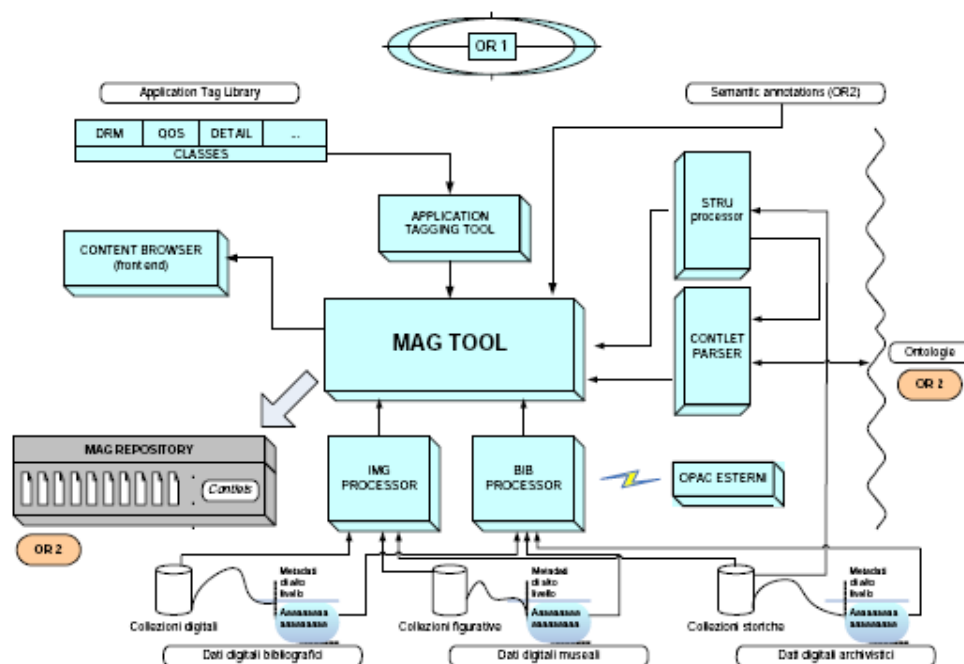


Figura 1: Architettura Sistema LC3

Le parti di interesse per questo documento sono il MAG Tool ed il Contlet Parser. In particolare, il MAG tool è quel componente che permette la creazione dei file MAG a partire dalle risorse digitali e da altri metadati (catalogazione etc.).

Il Contlet Parser (assieme allo STRU Processor) fornisce al MAG Tool le funzionalità per ricavare informazioni di similitudine e strutturali tra documenti contenuti all'interno della teca digitale, al fine di creare nuove annotazioni sulla base di quelle esistenti per MAG simili (per contenuto e struttura) a quello che si sta producendo. Le fasi in cui questo processo viene eseguito sono le seguenti:

1. All'atto della creazione di un nuovo MAG, una volta compilata la parte bibliografica (BIB) e strutturale (STRU), l'autore del MAG richiede l'utilizzo del Contlet Parser

PROGETTO LC3	Revisione n*	6	Del	22/01/2009
--------------	--------------	---	-----	------------

per ottenere una lista di possibili contlet che possano (o meno) costituire una base di annotazione per i contenuti del MAG che si sta producendo.

2. Il Contlet Parser riceve la richiesta del punto 1, e cerca all'interno della teca digitale, tutti i MAG corrispondenti, ricavando una lista di MAG con essa coerenti. La ricerca avviene, in prima istanza, solo rispetto alle informazioni di tipo BIB.
3. Il Contlet Parser utilizza le funzionalità dello STRU Processor, per verificare se esistono relazioni di isomorfismo di struttura tra sotto-alberi della parte STRU di MAG esistenti e del MAG in fase di pubblicazione.
4. Il Contlet Parser effettua delle richieste alla teca digitale per ottenere tutti i Contlet legati ai MAG ricavati al passo precedente.
5. Il Contlet Parser valuta la possibilità di collegare i contlet individuati al passo precedente al nuovo MAG, creando automaticamente dei contlet di annotazione facenti riferimento alle corrette parti strutturali del nuovo MAG.
6. Il Contlet Parser invia i risultati della sua analisi al MAG Tool che suggerisce all'utente i possibili collegamenti a contlet con annotazioni già esistenti per MAG "simili" a quello che si sta producendo.

### **.3 ACCESSO E GESTIONE DELLE INFORMAZIONI CONTENUTE NEI CONTLET E NELLE ANNOTAZIONI**

Come accennato in precedenza, è necessario che le informazioni collegate ai documenti contenuti all'interno delle teche digitali, siano definite per mezzo di grammatiche chiare. Tali informazioni, vengono utilizzate per descrivere documenti o loro parti e prendono il nome in letteratura di meta-dati.

Di particolare rilevanza lo sforzo fatto in questi anni nel campo dei *metadati per la conservazione digitale* da parte di alcune comunità (dai ricercatori in campo scientifico in prima istanza e poi dai bibliotecari stessi) .

Gli sforzi congiunti hanno condotto a un'ipotesi di classificazione ma ormai largamente utilizzata e tradotta nelle norme NISO 2004 anche a seguito della sua adozione da parte dello standard OAIS. A partire da questi standard, sono poi nati una serie di standard per la conservazione dei documenti in teche digitali come lo standard MAG.

Di fatto negli sviluppi implementativi i metadati per la conservazione in quanto informazioni necessarie per archiviare e conservare una risorsa al fine di assicurarne l'autenticità e la possibilità di riproduzione/ricostituzione si limitano a identificare e gestire

PROGETTO LC3	Revisione n*	6	Del	22/01/2009
--------------	--------------	---	-----	------------



informazioni di natura quasi esclusivamente tecnologica e comunque difficilmente riferibili ad archivi digitali complessi.

Scopo del Progetto LC3 è anche quello di permettere la gestione di archivi digitali complessi, grazie all'utilizzo di tecniche semantiche per la classificazione ed il retriwa delle informazioni.

Diventa quindi cruciale, nell'architettura del sistema, utilizzare o definire delle grammatiche per descrivere i metadati di interessi (annotazioni semantiche comprese) e provvedere alla definizione di un componente che permetta di accedere alle informazioni contenute all'interno dei meta-dati.

In letteratura, le grammatiche definite da schema XML sono molto utilizzare per la definizione e la strutturazione dei meta-dati, e in LC3 XML è stato scelto come tecnologia per il trattamento di tali informazioni. Ne consegue che il componente demandato alla gestione dei meta-dati debba essere un parser di documenti XML.

Il parsing consiste nel processo atto ad analizzare uno stream continuo in input (ad esempio descritto da un file XML) in modo da determinare la sua struttura grammaticale grazie ad una data grammatica formale. Nel caso in esame, le grammatiche sono descritte da appositi XML schema (uno XML Schema è un documento XML che utilizza un insieme di tag speciali per definire la struttura di un documento XML).

I parser XML, quindi possono essere costruiti in modo tale da configurare la struttura della grammatica dei linguaggi che andranno ad analizzare, in modo dinamico, leggendo la struttura della grammatica stessa dagli XML Schema.

L'output del processo di parsing, inoltre, costituisce in genere un albero sintattico in cui vengono memorizzate le informazioni associate agli stream XML letti in input, in modo conforme alla definizione delle grammatiche specificate dagli XML Schema.

Il componente dell'architettura in Figura 1, a cui è demandato il compito di analizzare i metadati associati ai documenti nelle teche digitali è il *contlet parser*.

Nella fattispecie, dovendo il contlet parser analizzare informazioni relative alla struttura dei documenti XML e quindi relativi alla struttura dell'albero di parsing, la tecnologia scelta per effettuare il parsing è quella DOM (Document Object Model).

Il Document Object Model (**DOM**), è una forma di rappresentazione dei documenti strutturati come modello orientato agli oggetti.

DOM è lo standard ufficiale del W3C per la rappresentazione di documenti E' inoltre la base per una vasta gamma delle interfacce di programmazione delle applicazioni; alcune

PROGETTO LC3	Revisione n*	6	Del	22/01/2009
--------------	--------------	---	-----	------------



di esse standardizzate dal W3C.

Le specifiche DOM elaborate da W3C sono suddivise in livelli, ciascuno dei quali contiene moduli obbligatori o opzionali. Per sostenere di appartenere ad un certo 'livello', un'applicazione deve soddisfare tutti i requisiti di tale livello e dei livelli inferiori. La specifica attuale di DOM è al Livello 2, tuttavia alcune delle specifiche del Livello 3 ora sono già raccomandazioni del W3C.

- Livello 0: include tutto quello che viene fornito a DOM per la creazione del Livello 1, per esempio: document.images, document.forms, document.layers, e document.all.
- Livello 1: navigazione di un documento DOM e manipolazione del contenuto.
- Livello 2 : supporto al Namespace XML, viste filtrate e Eventi DOM.
- Livello 3 : consiste in 6 specifiche differenti:
  - il nucleo del Livello 3;
  - caricamento e salvataggio del Livello 3;
  - XPath del Livello 3;
  - viste e formattazione del Livello 3;
  - requisiti del Livello 3;
  - validazione del Livello 3.

Un Contlet Parser che utilizzi un Parser DOM è dunque necessario nell'architettura LC3 per i seguenti motivi:

- Dovendo trattare contenuti differenti, il livello 0 offre tutte le funzionalità per distinguere i diversi tipi di documento memorizzati all'interno delle teche digitali. Inoltre, visto che le informazioni facenti riferimento ai metadati possono essere collegati ai documenti nella loro interezza, o a loro parti, avere uno strumento che possa discriminarle ne facilita la definizione.
- I Metadati LC3 rappresentano informazioni complesse, dovendo memorizzare ad esempio, informazioni sulle annotazioni semantiche, oppure informazioni sulla struttura dei documenti. L'analisi dei metadati, quindi, è un processo strutturato che dovrà essere eseguito tenendo conto di tutta la struttura dell'albero di parsing.

PROGETTO LC3	Revisione n*	6	Del	22/01/2009
--------------	--------------	---	-----	------------



Questo è possibile grazie alle funzionalità del livello 1 di DOM.

- Il supporto ai Namespace XML, inoltre, permette ad esempio di definire diversi tipi o versioni di annotazioni, o di fare riferimento a diversi tipi di metadato.
- Nel livello 3, la parte principale è quella collegata alla tecnologia XPATH. Visto che sui documenti e sui metadati sarà necessario effettuare delle ricerche, la tecnologia XPATH permette di definire delle query che verranno eseguite direttamente sull'albero di sintassi in maniera efficiente e costituiscono un modo efficace per l'accesso alle informazioni contenute all'interno delle teche digitali.

Si fa presente che gli standard OAIS e MAG fanno uso di queste tecnologie.

Il Contlet Parser, quindi, nell'Architettura LC3 deve essere implementato come un Parser DOM, riconfigurabile a seconda della grammatica di interesse (per le annotazioni, per i MAG, per i contlets etc.) e permettere in modo nativo di effettuare query XPATH sulla struttura dei documenti parserizzati.

PROGETTO LC3	Revisione n*	6	Del	22/01/2009
--------------	--------------	---	-----	------------