



Technical Report TR1.1.4

INTEGRAZIONE DEL *CONTLET* NELLO STANDARD MAG

ABSTRACT: Il presente Technical Report approfondisce l'evoluzione dello standard MAG nell'ottica dell'integrazione di un nuovo oggetto di contenuto elementare (Contlet), in grado di gestire la descrizione analitica delle risorse digitali tramite annotazioni semantiche e georeferenziazioni.

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



INDICE GENERALE

1. Evoluzione dello standard MAG	3
2. Possibile estensione del formato MAG 2.0.1.....	7
2.1. Struttura del MAG semantico.....	8
2.2. Il collegamento ontologico.....	9
2.3. La risorsa digitale.....	11

INDICE DELLE ILLUSTRAZIONI

Illustrazione 1: Metadato MAG standard 2.0 – Gestione immagini.....	3
Illustrazione 2: Metadato MAG standard 2.0 – Gestione immagini e audio.....	4
Illustrazione 3: Metadato MAG standard 2.0 – Estensione per Contlet.....	5
Illustrazione 4: Metadato MAG standard 2.0 esteso – Storage standard OAIS.....	6

<i>PROGETTO LC3</i>	<i>Revisione n*</i>	<i>3</i>	<i>Del</i>	<i>07/01/09</i>
---------------------	---------------------	----------	------------	-----------------

1. EVOLUZIONE DELLO STANDARD MAG

Il file MAG, come già evidenziato, si compone di diverse sezioni ciascuna delle quali descrive un particolare tipo di informazione. Escludendo le sezioni <gen>, <bib>, <stru> e <dis> che hanno carattere più generale, , <audio>, <video>, <ocr> e <doc> descrivono le rispettive tipologie di media digitale trattate dal metadato.

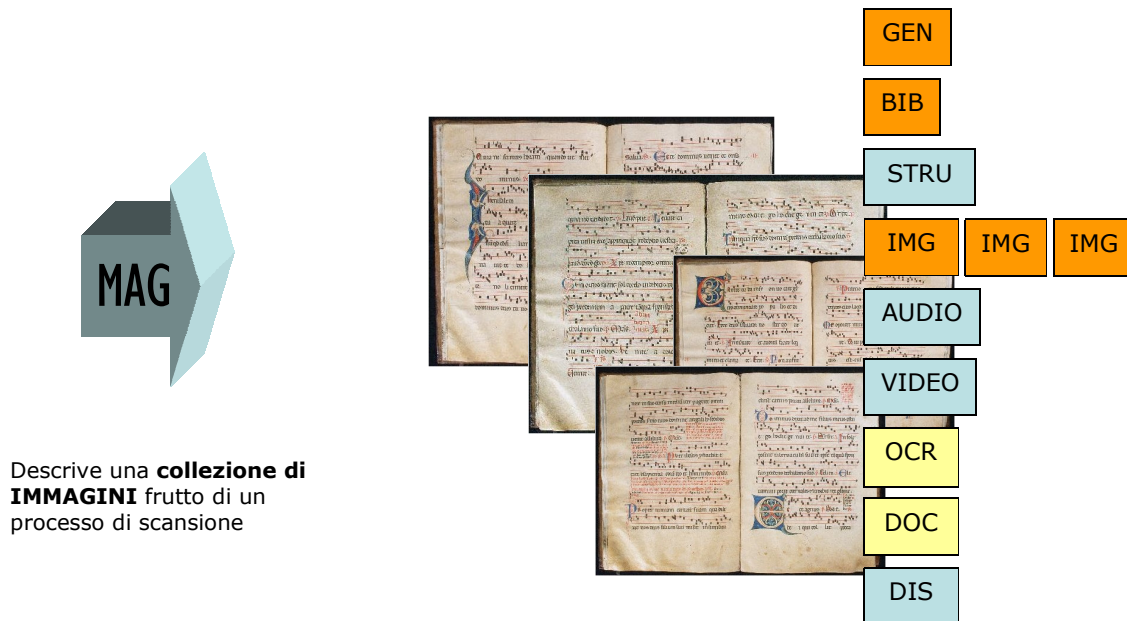


Illustrazione 1: Metadato MAG standard 2.0 – Gestione immagini

Un file MAG solitamente è collegato a un set di immagini digitali, nel caso più frequente a quelle associate all'attività di digitalizzazione di un volume o un documento cartaceo. In questo caso specifico, è possibile rinvenire come, a fronte di un'unica descrizione bibliografica contenuta nella sezione <bib> del metadato, esistano diversi oggetti digitali ad essa collegati.

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------

Per descrivere nel dettaglio la struttura dello stesso volume - ad esempio una miscellanea o un libro composto da più capitoli - si ricorre all'elemento <stru>, che permette di dettagliare le diverse parti dell'oggetto digitalizzato, consentendo anche il collegamento con MAG descrittivi associati alle singole sotto-porzioni del volume in oggetto (ad esempio, gli spogli).

Quanto finora affermato, comunque, risulta valido anche per altre tipologie di media, quali ad esempio la musica, sempre più frequentemente sottoposta ad interventi di digitalizzazione (es. gli spartiti musicali).

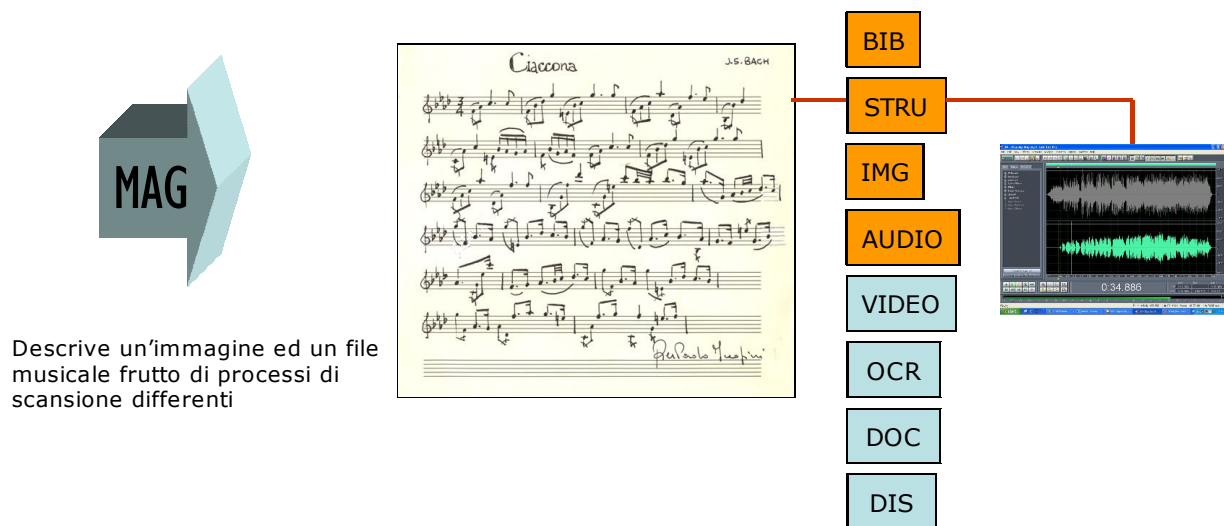


Illustrazione 2: Metadata MAG standard 2.0 – Gestione immagini e audio

In ogni caso, ancora l'elemento minimale non divisibile resta l'immagine singola o il singolo media in genere e l'unica forma di descrizione possibile resta quella limitata alle potenzialità della sezione <bib>, che non contemplano la possibilità di inserire alcuna annotazione semantica.

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------

Frutto dell'attività di ricerca effettuata nell'ambito del Laboratorio LC3 è, dunque, un'estensione del MAG indirizzata, da un lato, a raggiungere anche le sezioni di ciascuna risorsa digitale dello stesso metadato, e dall'altro a poterla collegare ad ontologie di dominio e, dunque, georeferenziali.

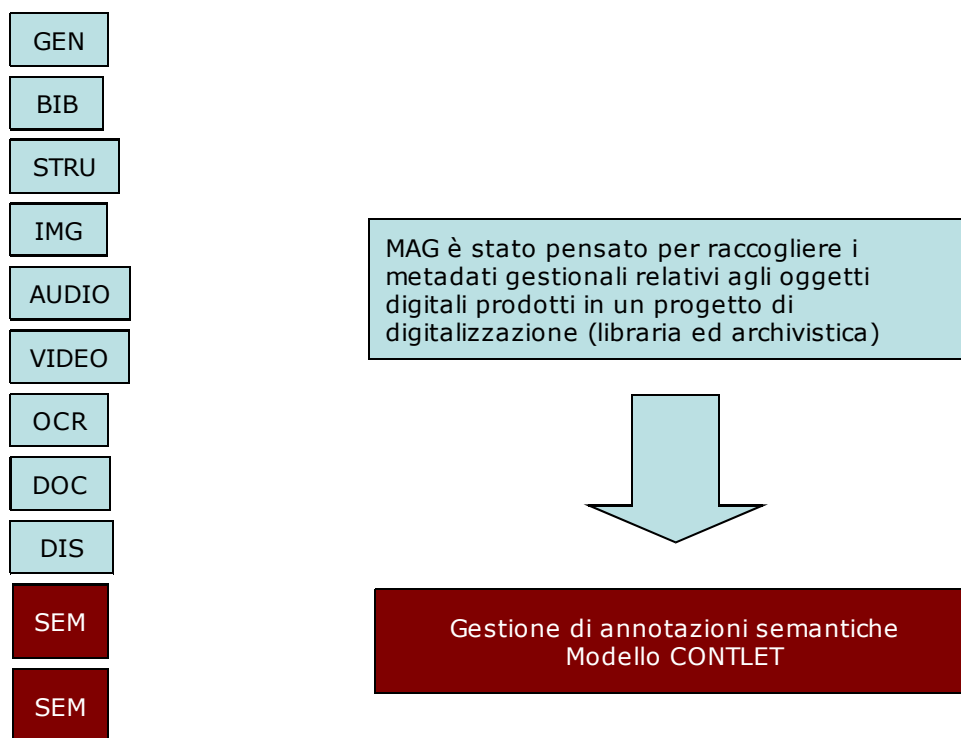


Illustrazione 3: Metadato MAG standard 2.0 – Estensione per Contlet

Per rispondere a queste esigenze, abbiamo introdotto una nuova sezione che contenga le informazioni dei cosiddetti *Contlet*, già descritte nel presente documento. Questa sezione, denominata <sem>, potrà essere ripetibile e, nel caso più semplice, permetterà di puntare a parti delle risorse digitali collegate al MAG di appartenenza e nel contempo a termini di ontologie. Tuttavia, per motivi pratici legati al fatto che solitamente l'annotazione è un'attività successiva alla produzione di metadati MAG, è possibile legare gli oggetti

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------

<sem> a risorse digitali presenti in altri file MAG, permettendo in questo modo di poter creare MAG semplicemente per contenere annotazioni.

Potremo quindi pensare ad un sistema di *storage* di risorse digitali compatibile con lo standard OAIS, contenente esclusivamente files digitali e files MAG come sotto raffigurato:

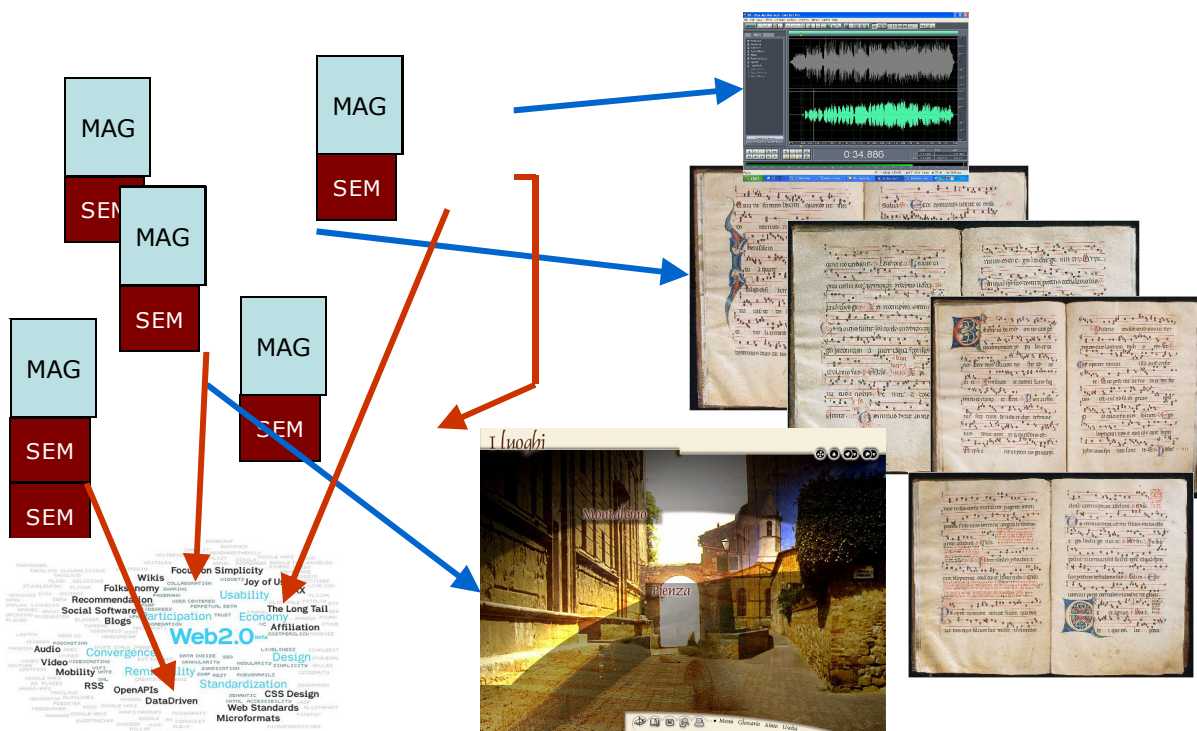


Illustrazione 4: Metadato MAG standard 2.0 esteso – Storage standard OAIS

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



2. POSSIBILE ESTENSIONE DEL FORMATO MAG 2.0.1

Aggiungendo un nuovo elemento (*Contlet*) all'interno del MAG 2.0.1, lo si rende capace di trasportare contenuti semantici. Lo schema della nuova radice “*metadigit*” del MAG modificato diventa dunque:

```
<xsd:sequence>
<xsd:element name="gen" type="gen"/>
<xsd:element name="bib" type="bib"/>
<xsd:element name="stru" type="stru" minOccurs="0" maxOccurs="unbounded"/>
<xsd:element name="img" type="img" minOccurs="0" maxOccurs="unbounded"/>
<xsd:element name="audio" type="audio" minOccurs="0" maxOccurs="unbounded"/>
<xsd:element name="video" type="video" minOccurs="0" maxOccurs="unbounded"/>
<xsd:element name="ocr" type="ocr" minOccurs="0" maxOccurs="unbounded"/>
<xsd:element name="doc" type="doc" minOccurs="0" maxOccurs="unbounded"/>
<xsd:element name="dis" type="dis" minOccurs="0"/>
<xsd:element name="sem" type="sem" minOccurs="0" maxOccurs="unbounded"/> </xsd:sequence>
```

Come risulta dallo schema, è prevista la possibilità di aggiungere 0 o più elementi **<sem>** all'interno di un MAG: ciascuno di questi elementi rappresenta un link che si vuole stabilire fra una risorsa digitale rappresentata nello stesso o in altro MAG ed un concetto espresso in un'ontologia.

Il collegamento con l'ontologia è composto dal link del file OWL, che rappresenta l'ontologia e da una serie di triplette rdf all'interno della stessa ontologia.

Ogni MAG esteso potrà, dunque, contenere sia le vecchie informazioni gestionali e strutturali dello standard 2.0.1, che le informazioni semantiche della sezione **<sem>**, o solo una delle due informazioni. Nel caso un MAG abbia solo le sezioni **<sem>** (oltre a **<gen>** e **<bib>** che sono obbligatori) e non si riferisca a nessuna risorsa digitale, si può considerare che la risorsa digitale rappresentata dal metadato sia l'annotazione semantica stessa.

Oltre all'annotazione vera e propria, la sezione **<sem>** contiene una serie di dati che servono ad identificarla: questo facilita il suo utilizzo multi-utente e distribuito nel tempo.

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



Per fare ciò, si è optato per l'utilizzo dei campi standard Dublin Core, con l'aggiunta di un campo **<replaced_by>**, che realizza un semplice sistema di *versioning*.

2.1. Struttura del MAG semantico

La radice dell'elemento **<sem>** risulta così formalizzata:

```
<xsd:complexType name="sem">
  <xsd:sequence>
    <xsd:element ref="dc:identifier"/>
    <xsd:element ref="dc:title" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:creator" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:publisher" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:subject" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:description" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:contributor" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:date" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:type" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:format" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:source" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:language" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:relation" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:coverage" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element ref="dc:rights" minOccurs="0" maxOccurs="unbounded"/>
    <xsd:element name="replaced_by" type="semantic_element" minOccurs="0"/>
    <xsd:element name="ontologic_link" type="ontologic_link" minOccurs="0"/>
    <xsd:element name="resource" type="resource" maxOccurs="unbounded"/>
  </xsd:sequence>
</xsd:complexType>
```

I campi Dublin Core sono stati inseriti per fornire informazioni sull'annotazione semantica stessa.

L'unico campo obbligatorio è **<dc:identifier>**, che serve per identificare univocamente il *Contlet*.

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



Altro metadato del *Contlet* è il campo `<replaced_by>`, di tipo ***semantic_element***, che, se presente, segnala che l'annotazione è stata sostituita da una versione più recente:

```
<xsd:complexType name="semantic_element">
  <xsd:sequence>
    <xsd:element name="referred_mag" type="xsd:anyURI" minOccurs="0"/>
    <xsd:element name="sematic_element_id" type="dc:identifier"/>
  </xsd:sequence>
</xsd:complexType>
```

Il collegamento al *Contlet* aggiornato viene espresso con l'URI del MAG in cui è contenuto e con il suo identificativo (si ricorda che l'URI non verrà inserita nel caso in cui si voglia puntare una sezione `<sem>` nello stesso MAG). Vedremo in seguito che il collegamento attraverso il tipo ***semantic_element*** verrà usato anche quando ci si intende riferire in modo semantico ad un'annotazione diversa.

2.2. Il collegamento ontologico

Il campo ***ontologic_link*** è il cuore del MAG semantico, in quanto esprime la relazione fra le risorse digitali del MAG-*Schema* classico e i concetti ontologici.

E' definito in questo modo:

```
<xsd:complexType name="ontologic_link">
  <xsd:sequence>
    <xsd:element name="ontology" type="xsd:anyURI"/>
    <xsd:element name="instance" type="ontology_instance" maxOccurs="unbounded"/>
  </xsd:sequence>
</xsd:complexType>
```

Una stessa annotazione può riferirsi ad una sola ontologia, e dunque è in questo campo che qui si deve specificare la sua URI.

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



Questo URI, in realtà, potrebbe essere omesso visto che all'interno di **instance**, i collegamenti ai singoli nodi ontologici (classi, istanze, relazioni.....) vengono espressi attraverso il tipo **rdf:resource**, che permette di anteporre all'identificativo del nodo una URI. Tuttavia, in questo modo, si renderebbe necessario integrare ogni nodo ontologico con l'URI dell'ontologia.

Inoltre, in questo modo, si sottolinea che ci si può collegare ad una ed una sola ontologia.

Il campo **instance** rappresenta la specifica del collegamento all'interno dell'ontologia:

```
<xsd:complexType name="ontology_instance">  
<xsd:sequence>  
<xsd:element name="base_instance" type="rdf:resource"/>  
<xsd:element name="property" type="ontology_property" minOccurs="0"  
maxOccurs="unbounded"/>  
</xsd:sequence>  
</xsd:complexType>
```

Oltre al riferimento al nodo semantico protagonista dell'annotazione, si possono specificare triplette rdf per sottolineare meglio da quale ottica è rilevante il soggetto nella risorsa digitale.

Delle triplette si specifica predicato e oggetto, visto che il soggetto è già stato espresso:

```
<xsd:complexType name="ontology_property">  
<xsd:sequence>  
<xsd:element name="predicate" type="rdf:resource"/>  
<xsd:element name="object" type="rdf:resource"/>  
</xsd:sequence>  
</xsd:complexType>
```

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



2.3. La risorsa digitale

Il campo **resource** stabilisce a quale risorsa digitale è riferita l'annotazione. Ci si può collegare alle risorse espresse dallo standard MAG oppure ad altri *Contlet*, per collegarli semanticamente. All'interno del campo **resource** si può specificare una ed una sola risorsa digitale; nel caso se ne vogliono esprimere più di una nello stesso oggetto <sem> occorrerà ripetere il campo.

Di seguito è riportata un'esemplificazione del campo **resource**:

```
<xsd:complexType name="resource">
  <xsd:sequence>
    <xsd:element name="referred_mag" type="xsd:anyURI" minOccurs="0"/>
    <xsd:element name="matching" type="percentage" minOccurs="0"/>
    <xsd:choice>
      <xsd:element name="img_element" type="img_element"/>
      <xsd:element name="video_element" type="video_element"/>
      <xsd:element name="txt_element" type="txt_element"/>
      <xsd:element name="ocr_element" type="ocr_element"/>
      <xsd:element name="audio_element" type="audio_element"/>
      <xsd:element name="semantic_element" type="semantic_link"/>
    </xsd:choice>
  </xsd:sequence>
</xsd:complexType>
```

Il campo **referred_MAG**, se presente, indica in quale metadato compare la risorsa a cui si fa riferimento. Se il campo è assente, significa che la risorsa è presente nello stesso MAG dell'oggetto <sem>.

Il campo opzionale **matching** esprime in percentuale la pertinenza della risorsa digitale rispetto al collegamento semantico.

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



La risorsa digitale vera e propria è rappresentata da un elemento a scelta che rappresenta il contenuto audio, video, etc. Sarà possibile inserire un unico contenuto per ogni elemento *resource*; nel caso in cui si vogliono più risorse occorrerà usare più oggetti *resource*.

I contenuti a cui è possibile collegarsi sono rappresentati da tutte quelle sezioni di risorse digitali possibili presenti nello standard MAG 2.0.1, che andremo ora ad elencare ed esplicitare in maniera più diffusa.

IMG_ELEMENT

IMG_element contiene un riferimento a una o più sezioni `` presenti all'interno di un MAG, oppure a parti di esse: infatti, oltre a poter esprimere diversi *sequence number* o più intervalli di *sequence number* (si ricorda che questi stessi *sequence number* saranno relativi al MAG indicato precedentemente nell'elemento **referred_MAG** oppure al MAG corrente in assenza di questo), sarà potrà "ritagliare" da questi gruppi di immagini dei poligoni per annotare solo alcune parti dell'immagine.

Di seguito, la definizione:

```
<xsd:complexType name="img_element">
<xsd:sequence>
<xsd:element name="img_sequence_numbers" type="sequence_numbers" maxOccurs="unbounded"/>
<xsd:element name="img_section" type="img_section" minOccurs="0" maxOccurs="unbounded"/>
</xsd:sequence>
</xsd:complexType>
```

Il ***sequence_number***, ripetibile, può rappresentare un unico *sequence number* oppure un intervallo di *sequence number*:

```
<xsd:complexType name="sequence_numbers">
<xsd:choice>
<xsd:element name="sequence_number" type="xsd:positiveInteger"/>
<xsd:sequence>
<xsd:element name="start_sequence_number" type="xsd:positiveInteger"/>
```

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



```
<xsd:element name="end_sequence_number" type="xsd:positiveInteger"/>
</xsd:sequence>
</xsd:choice>
</xsd:complexType>
<xsd:complexType name="img_section">
```

IMG_section, invece, permette di definire le sottosezioni presenti all'interno delle immagini ora specificate con i diversi *sequence number*.

```
<xsd:complexType name="img_section">
<!-- la zona rilevante di un'immagine puo' essere espressa come poligono o come ellissi -->
<xsd:choice>
<xsd:element name="polygonal_section" type="polygonal_section"/>
<xsd:element name="elliptical_section" type="elliptical_section"/>
</xsd:choice>
</xsd:complexType>
```

Come si può notare, la selezione può essere lineare (una sequenza di almeno tre punti), o curva (un'ellisse definita attraverso la posizione dei vertici dei suoi assi).

La posizione di un punto è definita in modo relativo (percentuale) rispetto alla dimensione dell'immagine, partendo dall'angolo inferiore destro.

```
<xsd:complexType name="point">
<xsd:sequence>
<xsd:element name="horizontale_position" type="percentage"/>
<xsd:element name="vertical_position" type="percentage"/>
</xsd:sequence>
</xsd:complexType>
```

Fornire le posizioni sottoforma di percentuale piuttosto che come pixel, consente di applicare gli stessi dati geometrici a risoluzioni differenti delle stesse immagini.

La selezione poligonale appare così definita:

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



```
<xsd:complexType name="polygonal_section">
<xsd:element name="vertex" type="point" minOccurs="3" maxOccurs="unbounded"/>
</xsd:complexType>
```

Mentre questa è la definizione della selezione ellittica:

```
<xsd:complexType name="elliptical_section">
<xsd:sequence>
<xsd:element name="first_axis_vertex" type="point" minOccurs="2" maxOccurs="2"/>
<xsd:element name="second_axis_vertex" type="point" minOccurs="2" maxOccurs="2"/>
</xsd:sequence>
```

VIDEO_ELEMENT

Video_element è assolutamente equivalente a *IMG_element*, ma oltre alla selezione spaziale delle risorse permette la selezione temporale:

```
<xsd:complexType name="video_element">
<xsd:sequence>
<xsd:element name="video_sequence_numbers" type="sequence_numbers" maxOccurs="unbounded"/>
<xsd:element name="video_section" type="video_section" minOccurs="0"
maxOccurs="unbounded"/>
</xsd:sequence>
</xsd:complexType>
<xsd:complexType name="video_section">
<xsd:sequence>
<xsd:element name="spatial_section" type="img_section" minOccurs="0"
maxOccurs="unbounded"/>
<xsd:element name="temporal_section" type="temporal_section" minOccurs="0"
maxOccurs="unbounded"/>
</xsd:sequence>
</xsd:complexType>
<xsd:complexType name="temporal_section">
<xsd:sequence>
<xsd:element name="start" type="xsd:time"/>
<xsd:element name="end" type="xsd:time"/>
</xsd:sequence>
</xsd:complexType>
```



E' obbligatorio specificare i *sequence_numbers* per individuare l'immagine, ai quali eventualmente si possono aggiungere restrizioni spaziali e/o temporali. Come si può notare, la selezione temporale è espressa come semplice tempo di inizio e tempo di fine.

AUDIO ELEMENT

Audio_element individua una sezione temporale all'interno di contenuti audio descritti dalle sezioni audio presenti in un MAG.

```
<xsd:complexType name="audio_element">
<xsd:sequence>
<xsd:element name="audio_sequence_numbers" type="sequence_numbers" maxOccurs="unbounded"/>
<xsd:element name="audio_section" type="temporal_section" minOccurs="0" maxOccurs="unbounded"/>
</xsd:sequence>
</xsd:complexType>
```

L'intervallo di interesse è espresso in modo analogo alla sezione temporale presente in un contenuto audio.

TXT ELEMENT E OCR ELEMENT

Questi due tipi di link sono esattamente uguali ed è necessario distinguere fra loro solo per farli puntare automaticamente alle due corrispettive sezioni del MAG (<doc> e <ocr>), che esistono per rappresentare documenti testuali trascritti o acquisiti tramite riconoscimento ottico della scrittura. Anche se lo standard MAG permette il contemporaneo inserimento di diversi tipi di documenti testuali, per ora questa proposta di sua estensione semantica permette solo il riferimento a file di testo semplice, riferiti ad un'unica sezione del documento. Ciò è documentato dal fatto che il riferimento alla parte di testo viene espresso semplicemente come intervallo dei caratteri da considerare, secondo quanto sotto riportato:

PROGETTO LC3	Revisione n*	3	Del	07/01/09
--------------	--------------	---	-----	----------



```
<xsd:complexType name="txt_element">
<xsd:sequence>
<xsd:element name="txt_sequence_numbers" type="sequence_numbers" maxOccurs="unbounded"/>
<xsd:element name="txt_section" type="txt_section" minOccurs="0" maxOccurs="unbounded"/>
</xsd:sequence>
</xsd:complexType>
<xsd:complexType name="txt_section">
<xsd:sequence>
<xsd:element name="start_character" type="xsd:integer"/>
<xsd:element name="end_character" type="xsd:integer"/>
</xsd:sequence>
</xsd:complexType>
<xsd:complexType name="ocr_element">
```

<i>PROGETTO LC3</i>	<i>Revisione n*</i>	<i>3</i>	<i>Del</i>	<i>07/01/09</i>
---------------------	---------------------	----------	------------	-----------------