Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

Annotazione del testo

Rocco Tripodi

rocco@unive.it

Ricostruzione

Filologia

Ricostruire la forma originaria dei testi tramite l'analisi critica delle fonti

Ecdotica

Avvicinarsi il più possibile alla forma originaria del testo
Edizione diplomatica
Edizione critica

Esegesi

Disciplina nata durante XI sec. Incentrata sull'interpretazione critica dei testi per la comprensione del loro significato
Glossa: nota esplicativa annotata all'interno del testo

Annotazione

Annotazione, markup, mercatura
- Nasce in ambito tipografico
- Evidenziazioni delle parti “speciali” del testo

Codifica di informazione linguistica associata al dato testuale

Permette di rendere esplicita, interpretabile ed esplorabile la struttura linguistica del testo

Oltre ai dati linguistici vengono codificati anche gli elementi del paratesto
- struttura editoriale costituita da componenti organizzati in modo gerarchico (Frontespizio, Capitoli, Titoli, Paragrafi, versi, battute, ecc)

Rappresentare i diversi livelli del testo

Linguaggi di annotazione

Testo digitale

Linguaggi di marcatura

Tag: indicano la funzione astratta della porzione di testo che delimitano
Grammatica che regola l'uso dei tag

Linguaggi procedurali

Nei linguaggi procedurali il mark-up specifica quali operazioni un dato programma deve compiere su un documento elettronico per ottenere un determinato effetto presentazionale (font, dimensione,…)

TeX - LaTeX (Link)

Linguaggio usato per testi scientifici. Nato per rappresentare in modo professionale simboli ed espressioni matematiche

```
E &= mc^2
m &= \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}
```

$$E = mc^2$$

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Linguaggi di annotazione

Linguaggi referenziali

Si fa riferimento ad entità esterne che il programma che processa il file può richiamare.

Es: nel mark-up si inserisce una sigla e il browser la visualizza in forma estesa

Linguaggi dichiarativi

il mark-up descrive la struttura di un testo identificandone i componenti.

Gli elementi del testo vengono associati a determinate classi di elementi testuali.

Sono inserite informazioni riguardanti la struttura del testo e vengono esplicitate le proprietà di determinate unità.

Separazione forma – contenuto

SGML – HTML - XML

Proprietà delle annotazioni

Copertura
: Corrispondenza esatta tra le categorie e strutture dello schema e i fenomeni indagati. Quanti aspetti del fenomeno lo schema riesce a cogliere e quanti ne lascia in ombra

Riproducibilità
: Possibilità di applicare lo schema allo stesso modo a tutti i fenomeni indagati

Espressività
: Traduzione dello schema in un linguaggio di marcatura

Livelli
: Generalmente si effettuano annotazioni seguendo i livelli di studio del linguaggio (morfologia, sintassi, semantica, pragmatica) e la struttura presentazionale del testo

Interazione
: Comunicare con gli altri livelli dell'annotazione

Requisiti 1

Potenza espressiva

capacità di rappresentare il maggior numero di tipologie testuali
capacità di rappresentare adeguatamente il maggior numero di livelli strutturali e di caratteristiche
impiego di diverse prospettive metodologiche
possibilità di metadati descrittivi e gestionali

Portabilità e preservazione

fruibilità senza limitazioni di spazio e di tempo
accessibile su diverse piattaforme e dispositivi informatici (portabilità)
riusabilità in archi temporali ampi (conservazione)

Versatilità

testo disponibile in differenti formati di fruizione (testuale, audio, diagramma, stampa, smart – screen, ecc)

Requisiti 2

Standardizzazione e apertura
utilizzo dello stesso formato da parte della comunità di utenti
di dominio pubblico (open source)

Standard formale
insieme di norme relative ad una particolare tecnologia emesse da un ente
istituzionale nazionale o internazionale (UNI, ANSI, ISO)

Standard informale
insieme di norme e linee guida relative ad una particolare tecnologia
adottate da una comunità di utenti o produttori, eventualmente
rappresentata da enti associativi

Standard di fatto
standard che si impongono per la diffusione commerciale

Linguaggi dichiarativi: SGML

Standard Generalized Markup Language

Nel 1986 diventa il primo metalinguaggio di markup per la rappresentazione di testi digitali ad essere standardizzato

Il markup si concentra sulla struttura rappresentazionale del testo

Serve ad automatizzare il processo di interscambio di grandi quantità di documenti (machine – readable) in particolare in ambito militare e industriale

Definizione della struttura dei documenti

Elementi strutturali

Relazioni e occorrenze degli elementi

Posizionamento degli elementi

Sintassi

SGML declaration: <!SGML "ISO 8879:1986 (WWW)" … >

Prologo (DTD o riferimento)

Contenuto (struttura ad albero)

Linguaggi dichiarativi: HTML

Hypertext Markup Language

File di testo .html -> Browser

Tag annidati

<html>

<head> Contiene informazioni non visualizzate, che riguardano il modo in cui il documento deve essere letto e interpretato da un browser o da un agente esterno. Contiene i meta-tag (alcuni sono ideati per i motori di ricerca), script, fogli di stile, ecc.. **</head>**

<body> Contenuto vero e proprio del documento **</body>**

</html>

Tag del body: a = link, div = sezione, p = paragrafo, ecc)

CSS: Cascading Style Sheets (rappresentazione degli elementi)