

Presentazione del Corso Codifica di Testi a.a. 2018-2019

Angelo Mario Del Grosso

angelo.delgrossos@ilc.cnr.it

CNR-ILC-LicoLab

Istituto di Linguistica Computazionale “A. Zampolli” ,
21st September 2018

Piano della presentazione

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

1 Presentazione

2 Introduzione

3 Codifica dei Caratteri

4 Codifica dei Testi

5 Ecosistema XML

6 Conclusioni

Progress status

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

1 Presentazione

2 Introduzione

3 Codifica dei Caratteri

4 Codifica dei Testi

5 Ecosistema XML

6 Conclusioni

Di cosa mi occupo

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Filologia Computazionale

Analisi, progettazione e sviluppo di componenti software per sistemi Web di linguistica e filologia digitale/computazionale volti al trattamento di testi di tradizione medievale, a stampa e di autori moderni e contemporanei.

Modelli Object Oriented per il Textual Scholarship

Impiego delle nuove tecnologie nell'ambito delle Digital Humanities (DH) per la progettazione object-oriented di strumenti digitali Web-based rispondenti alle esigenze degli utenti accademici, studenti e sviluppatori.

Presentazione del Corso

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia



Progress status

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

1 Presentazione

2 Introduzione

3 Codifica dei Caratteri

4 Codifica dei Testi

5 Ecosistema XML

6 Conclusioni

Introduzione al Corso di Codifica di Testi

Obiettivi, competenze e conoscenze

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Obiettivo

Illustrare i principi di modellazione e le prassi di codifica del testo per una adeguata rappresentazione ed elaborazione digitale di risorse testuali.

Rationale

Fornire gli strumenti e le conoscenze necessarie per progettare e realizzare criticamente una codifica digitale di testi complessi, in particolare testi letterari e di interesse storico-culturale, usando le linee guida della Text Encoding Initiative (TEI).

Argomenti trattati

Obiettivi, competenze e conoscenze

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Cosa ci aspettiamo alla fine del corso

- valutare il metodo di codifica più appropriato allo scenario d'interesse
- creare uno schema di codifica TEI
- usare gli strumenti più idonei per la codifica di una risorsa testuale
- elaborare e visualizzare il testo codificato

Principali Argomenti

Obiettivi, competenze e conoscenze

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni
Bibliografia

- Codifica dei caratteri e di testi
- I linguaggi di markup e introduzione a XML
- Creazione di schemi di codifica
- Le norme TEI (Text Encoding Initiative)
- Alcuni specifici Moduli TEI
- Definizione di schemi di codifica personalizzati
- Introduzione ai fogli di stile XSLT
- Elaborazione documenti XML-TEI (XSLT2.0, DOM)
- Esempi, esercitazioni e seminari

Bozza piano del Corso: Calendario 2018/2019

Lunedì dalle 14.00 alle 18.00

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

- Lezione 1 (24/9): Introduzione al corso e strumenti software utilizzati
- Lezione 2 (1/10): Elementi teorici relativi alla rappresentazione del testo
- Lezione 3 (8/10) Elementi di XML e introduzione alla definizione degli Schemi XML
- Lezione 4 (15/10): Elementi di Codifica TEI
- Lezione 5 (22/10): Modulo Specifico TEI

Bozza piano del Corso: Calendario 2018/2019

Lunedì dalle 14.00 alle 18.00

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

- Lezione 6 (29/10): Moduli Editoriali TEI
- Lezione 7 (5/11): Codifica dei caratteri e Unicode
- Lezione 8 (12/11): Elaborazione documenti XML-TEI (XSLT e DOM)
- Lezione 9 (19/11): Esercitazione e discussione sui progetti
- Lezione 10 (26/11): Esercitazione e Recupero Argomenti
- (03/12-17/12): Eventuali esercitazioni, attività laboratoriali, recupero argomenti

Perché è importante la codifica dei testi

Motivazioni pratiche

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Dove troviamo i testi

Nella nostra cultura tradizionale la quasi totalità dei testi è “registrata”, “trasmessa” e quindi “conservata” attraverso supporti e materiali fisici di varia natura e forma (iscrizioni su pietra, manoscritti di pergamena, papiri, carta, libri a stampa, incunabula, cinquecentine, etc).

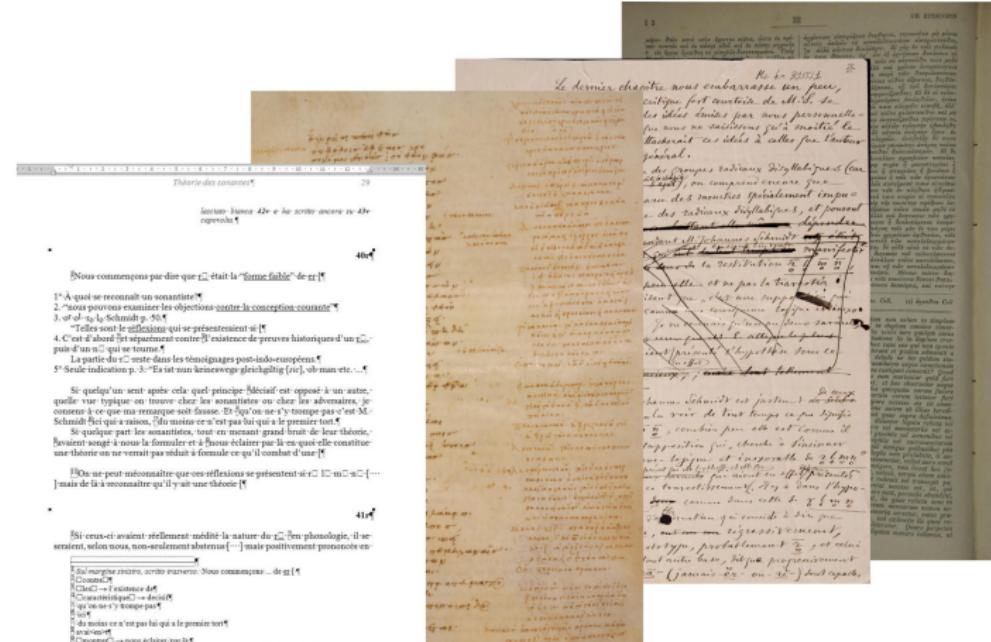
Perché è importante la codifica dei testi

Dove troviamo i testi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Introduzione



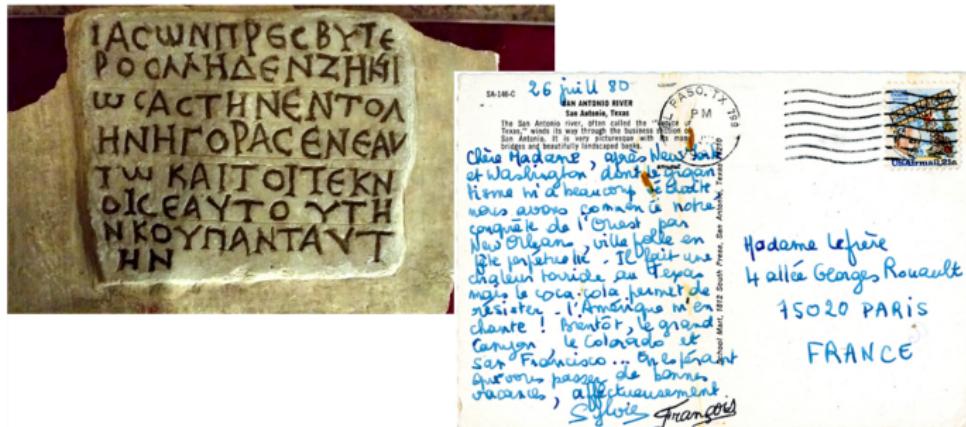
Perché è importante la codifica dei testi

Dove troviamo i testi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia



Perché è importante la codifica dei testi

Motivazioni pratiche

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Perché codificare i testi

Per rendere disponibile questo patrimonio attraverso i sistemi per la gestione dell'informazione digitali e computazionali è necessario effettuare una trasposizione/transcodifica* dei testi dal loro supporto originario verso il nuovo supporto elettronico (*Machine Readable Form*).

* *procedimento di conversione dei dati codificati secondo un sistema verso un sistema diverso*

Perché è importante la codifica dei testi

Motivazioni teoriche

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Scienze umane vs Scienze informatiche

Il rapporto tra sapere umanistico e informatica non è solo una questione meramente strumentale:

l'informatica non è solo un utensile dalle notevoli capacità.

Salto di paradigma sia da un punto di vista teorico sia metodologico.

Perché è importante la codifica dei testi

Motivazioni teoriche

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

La codifica come metodologia

L'attività di codifica diviene metodo di analisi della conoscenza nell'ambito delle discipline che si occupano del testo.

La codifica come analisi teorica

Il linguaggio di codifica adottato può essere considerato come un linguaggio teorico:

Esplorare e formalizzare le ipotesi interpretative su un certo oggetto di studio

Perché è importante la codifica dei testi

In sintesi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019
A.M. Del
Grosso

Rappresentare il testo

Il focus del corso sarà incentrato sulla rappresentazione digitale del testo.

Esistono dibattiti e controversie

Per ottenere una rappresentazione digitale del testo ci sono diversi formati, formalismi e prassi:
la nostra scelta ricade sulle norme suggerite dal consorzio TEI.

Molte questioni non sono risolte altre sono controverse, sia dal punto di vista teorico-metodologico, sia pratico-tecnologico.

Perché è importante la codifica dei testi

Ma in definitiva

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Perché codificare

Le differenze di formato sono più che altro estetiche e non sostanziali

Perché codificare

Ma anche l'occhio umano vuole la sua parte

Progress status

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

1 Presentazione

2 Introduzione

3 Codifica dei Caratteri

4 Codifica dei Testi

5 Ecosistema XML

6 Conclusioni

Elementi di Codifica dei Caratteri

Definizioni

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Rappresentare il testo in formato digitale

L'adozione di metodologie informatiche per il trattamento dei testi richiede in primo luogo la disponibilità di un'adeguata rappresentazione dei dati testuali in formato digitale.

Elementi di Codifica dei Caratteri

Problemi di rappresentazione

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

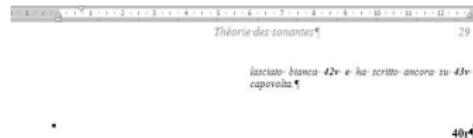
A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia



«Nous commençons par dire que t était la "forme faible" de er.»

1. À quoi se reconnaît un sonantiste?
2. «nous pouvons examiner les objections contre la conception courante»
3. cf. cf. 12, 13, Schmidt p. 50.¶
"Telles sont les réflexions qui se présentent si [¶]
4. C'est d'abord l'opposition contre l'existence de preuves historiques d'un t, puis d'un n qui se tourne.¶
La partie du t reste dans les témoignages post-indo-européens.¶
Seule indication p. 3: "Es ist nur keineswegs gleichgültig [sic], ob man etc...."

Si quelqu'un sent après cela quel principe décisif est opposé à un autre, quelle vue typique on trouve chez les sonantistes ou chez les adversaires, je consens à ce que ma remarque soit fausse. Et qu'on ne s'y trompe pas c'est M. Schmidt qui qui a raison, «du moins ce n'est pas lui qui a le premier tort.¶

Si quelque part les sonantistes, tout-en-menant grand bruit de leur théorie, avaient songé à nous la formuler et à nous éclairer par là en quoi elle constitue une théorie on ne verrait pas réduit à formuler ce qu'il combat d'une.¶

¶On ne peut néanmoins que ces réflexions se présentent si t l m n [...] mais de là à reconnaître qu'il y ait une théorie.¶

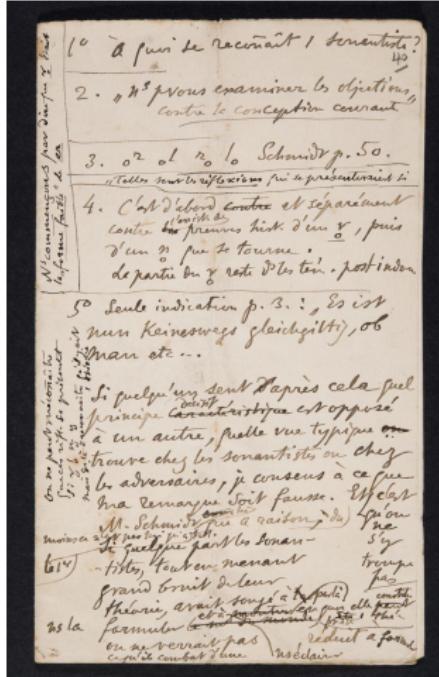
41r

¶Si ceux-ci avaient réellement médité la nature du t [en phonologie, il se seraient, selon nous, non-seulement abstenus [...] mais positivement prononcés en-

¶Sur margine sinistro, scritto traverso: Nous commençons... de gr [¶]

- control
- les t → l'existence de t
- caractéristique → décisif
- que ce n'est pas lui qui a le premier tort
- et
- du moins ce n'est pas lui qui a le premier tort
- au contraire
- montrer t → nous éclairer par là

¶Sur margine sinistro, scritto traverso: On ne peut... qu'il y ait une théorie.



Elementi di Codifica dei Caratteri

Definizioni

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Perché è importante la codifica dei caratteri

La codifica dei caratteri costituisce il grado zero (basso livello) della rappresentazione di testi su supporto digitale.

Le codifiche dei caratteri sono la base di qualsiasi schema di codifica testuale.

Rappresentazione digitale dei caratteri

I caratteri vengono rappresentati all'interno di un elaboratore mediante una sequenza di codici binari formati da opportune disposizioni di cifre composte da 0 e 1: 01100001 *lettera a*

Elementi di Codifica dei Caratteri

Definizioni

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Tabella Code Page ASCII 7 bit

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
10	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
20	SP	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
60	~	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{	}	~	DEL	

7 bit = 128 possibili caratteri; 32 caratteri di controllo; 96 caratteri effettivi

Elementi di Codifica dei Caratteri

Esempio codifica binaria

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

codifica *ciao mondo!* 7 bit ASCII

6369 616F 206D 6F6E 646F 210A

codifica *ciao è mondo!* 8 bit ASCII

6369 616f 20e8 206d 6f6e 646f 210a

codifica **ciao è mondo!** UNICODE UTF-8

6369 616f 20c3 a820 6d6f 6e64 6f21 0a

Elementi di Codifica dei Caratteri

Definizioni

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Character set Per le discipline che studiano i sistemi di scrittura e l'analisi del linguaggio naturale, un insieme di caratteri astratti è detto Character set (unità alfabetiche). Astratto perché non riguarda la rappresentazione materiale della forma sul supporto, ma è relativo alla forma mentale, fatta di simboli di codifica (referenti).

Coded Char Set Per poter trattare un insieme di unità alfabetiche in formato digitale bisogna assegnare a ciascun carattere un numero intero non negativo detto code point.

Elementi di Codifica dei Caratteri

Definizioni

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Character encoding Il fine ultimo della codifica è quello di rappresentare una sequenza di caratteri in una sequenza di byte. La codifica di un carattere utilizza uno “encoding schema” che a sua volta mappa o trasforma ciascun code point in una sequenza di byte e quindi in ultima istanza in una sequenza di bit.

Tabella del code page Generalmente i code points sono espressi attraverso un sistema numerico esadecimale e disposti in una tabella di associazione.

Elementi di Codifica dei Caratteri

In sintesi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Codifica dei caratteri

Quindi trasformare una sequenza di caratteri appartenenti ad un char set in una sequenza di byte (bit) significa prima di tutto trasformare/mappare ciascun carattere nel proprio corrispettivo code point e successivamente codificare/serializzare questo code point nella relativa sequenza di byte (bit).

Elementi di Codifica dei Caratteri

Complessità e rappresentazione

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Complessità di rappresentazione universale dei caratteri

Se si considerano tutti i possibili alfabeti del mondo e le molteplici esigenze poste dalla scrittura delle fonti manoscritte antiche e medievali, ci si accorge che la realizzazione di un sistema universale per la codifica dei caratteri è un progetto molto complesso con svariate sfide da affrontare.

Complessità e rappresentazione di Codifica dei Caratteri

Un Esempio

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

ost̄p aut̄ compleuit ser̄ suū et uotū redidit ap̄lo sup̄tō. re-
duuit ī italiam. q̄ curitate quā p̄us incoluerat. uisitauit. Accep-
e m̄ asp̄u sancto ī uia. ut nō deberet se m̄ aliquo exaltare. Et ac-
cepto confirmatoq; ī corte p̄cepto. sp̄reuit artem suā. uelicit sub-
ture ul̄ cerdonū. et adhesit se uiliori offiō quod inuenit. Cum ergo
uoluerit p̄ceptum dñi obſuare nō ſolum telecīt uultum ſuū uilissimo
elemento ſer̄ equauit ſe aſinis et mulis. horum a honora deportantib;.
p̄ceptū dñi hoc eſt. Qui ſe huiliat exaltabitur. Et iterum ſi q̄
nō iniquatus fuerit et efficiatur ut parvulus nō intrabit regnum
celor̄. Inclinauit aut̄ oculos ad terram. corte ū humili. ſpeculabat
ad celum. honora corpus portabat. ſer̄ aīa ſemp in dno gloriaſ. lucra-
batur vir tei pecuniam magnam. qua mediante minimo panes eme-
bat ul̄ placentulas. et illis īt̄r̄ p̄tes duuallis interrogabat de duabz
p̄tib; que eſſent meliores. et ip̄as paupibus errogabat. Cernā p̄te;
ſibi obtinebat. et ip̄a aduertitionē corporis ſuū fruebat. Audieſ
aut̄ quidam de ſur mentis q̄ hec fatiebat. et q̄ ſepe placentulū uiceret
motus iuidia. et ad dyabolo iuigatus. ad impleuit. placentula unā
excungia porqua. et ip̄am illi tribuit ad edendū. Atenſ aut̄ sanctus
Thebaudus illusiones de ſe ip̄o factas. mirabiliter motus fuit. q̄ uis ſuū
non eſſ iraſci. et maledixit illi dicens. maledictus ſis inter hoīes.
et generatio tua nō exaltetur. nec ſemen tuū multiplicet iemuz.

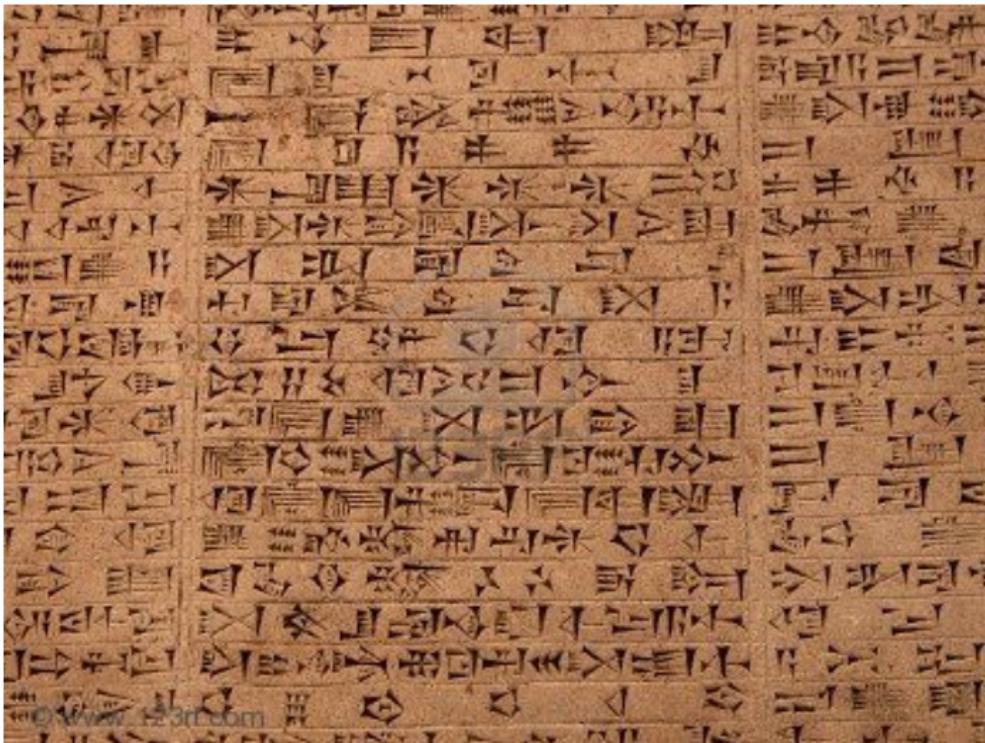
Complessità e rappresentazione di Codifica dei Caratteri

Un Esempio

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Codifica dei Caratteri



Elementi di Codifica dei Caratteri

Unicode

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Complessità di rappresentazione universale

Ad oggi, lo standard de facto per la codifica dei caratteri è lo UNICODE. Esso è in grado di codificare più di un milione di differenti unità alfabetiche, segni di interpunkzione e diacritici, appartenenti a centinaia di diverse lingue.

Complessità di rappresentazione universale

Unicode assegna i propri code point in un range che va da 0x0 a 0x10FFFF. In Unicode il code point viene indicato con una "U" seguita da un segno "+" seguito a sua volta dall'esadecimale con padding del codice (es: U+0041 lettera a).

Elementi di Codifica dei Caratteri

Unicode

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Unicode Transformation Format

Lo Unicode è un Coded Char Set e per essere concretamente serializzato su un supporto elettronico deve essere trasformato attraverso qualche tipo di schema di codifica. L'UTF (Unicode Transformation Format) mappa i code point Unicode in sequenze di byte (bit).

UTF standards

Esistono tre tipi di schemi di codifica che vanno sotto il nome di UTF, ciascuno è identificato dal minimo numero di bit necessario a codificare ciascun code point: UTF-8; UTF-16; UTF-32.

Progress status

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

1 Presentazione

2 Introduzione

3 Codifica dei Caratteri

4 Codifica dei Testi

5 Ecosistema XML

6 Conclusioni

Rappresentazione digitale dei testi

basso e alto livello di codifica

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Codificare un testo

La codifica dei caratteri evidentemente non esaurisce i problemi per una opportuna rappresentazione delle caratteristiche interne ed esterne di un testo.

Codificare un testo

Difatti la codifica del testo è una questione molto più complessa di una semplice riproduzione meccanica di un dato.

Rappresentazione digitale dei testi

basso e alto livello di codifica

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Rappresentare un testo

La rappresentazione digitale di un testo è una operazione intrinsecamente assai difficile perché coinvolge una pletora di aspetti, a varie dimensioni, a varie granularità e a vari livelli di astrazione sia teorici, sia metodologici, sia tecnologici e sia pratici.

Rappresentazione digitale dei testi

basso e alto livello di codifica

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Rappresentare un testo

Prima di poter fare qualsiasi ipotesi su come compiere una codifica di un testo e su come rappresentarlo digitalmente, bisogna stabilire cosa si intende per testo.

Rappresentazione digitale dei testi

Modello dati di un testo

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Un testo non ha una struttura rigida, predefinita:

- Non è rappresentabile solo come un insieme di record di un archivio elettronico.
- Non è rappresentabile solo come un insieme di tabelle di una banca dati.
- Non è rappresentabile solo come un albero o un insieme di sotto-alberi
- Non è rappresentabile solo come un grafo o come un insieme di sotto grafi

Molteplici modelli per diverse esigenze

Strutture dato e testo

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

La rappresentazione di un testo

- modello lineare: sequenza di dati non strutturati
- modello a record: enumerazione delle proprietà
- modello tabulare: insieme di dati omogenei
- modello ad albero: gerarchie di dati e insiemi di dati
- modello grafo: rete di strutture informative interconnesse tra loro

Elementi di Codifica del testo

Formalismi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Formati di rappresentazione

Un formato è un insieme di regole e convenzioni formali per rappresentare un insieme di dati, nel nostro caso un testo.

Importanza dei formati

Seppur isomorfi la scelta dei formati condiziona molto l'efficienza delle operazioni e l'efficacia delle dichiarazioni.

Elementi di Codifica del testo

Tabella Formalismi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Formalismi

	Data	Text	Hierarchies	Presentation	Validation	References	Annotations	Overlapping
CSV								
JSON								
RDF								
Markdown								
HTML								
HTML+RDFa								
XML								
Overlapping formats								

courtesy of *Fabio Vitali*

Elementi di Codifica del testo

Varietà di rappresentazione

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

In literature, there are **four main models** to represent a text:

1. Text as **linear data** model (typical in Natural Language Processing)
 - ✓ Plain Text and Regular Expression
 - *String Manipulation*
 - ✗ Limit: nor physical nor logical structure among textual elements
2. Text as **tabular data** model (typical in document processing)
 - ✓ Comma separated values (CSV)
 - *Indexing, retrieval*
 - ✗ Limit: no hierarchies inherently described
3. Text as **hierarchical data** model (typical in scholarly editing - OHCO)
 - ✓ XML data model
 - *TEI/XML Encoding and Serialization*
 - ✗ Limit: no overlapping hierarchies allowed
4. Text as **graph data** model (typical in knowledge representation)
 - ✓ Multi-dimensional Texts
 - ✓ Multi-versioned Texts
 - *RDF as Data Model*
 - *OWL as Language for Knowledge Description*
 - ✗ Limit: No type systems for textual scholarship

Elementi di Codifica del testo

Esempio di codifica del testo utilizzando CSV

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Why it's worth giving up .
The human_body is remarkably resilient and will start to repair the damage caused by smoking almost as soon as you start giving_up .
The benefits to your health can be seen in as little as 20 minutes from the time you put_down your last cigarette .
The whole process remains only marginally_safe on average for women_and_children .
Both men_and_women want the process to be safer .
Both men_and_women are involved in developing and exploiting the sciences and technologies to make_this_happen .
The gender_dynamics of public_perceptions on DNA_paternity_testing are likely to become more important as the tests become more widespread .
Some perceptions of paternity_testing will obviously be dependent on the inherently_different reproductive_roles of men_and_women .

Assume	VB	assume	"'
that	IN	that	"'
a	DT	a	"'
natural	JJ	natural	"'
cyclic	JJ	cyclic	"'
phenomenon	NN	phenomenon	"'
has	VBN	have	"'
been	VBN	be	"'
measured	VBN	measure	"'
,	,	,	,
but	CC	but	,
the	DT	the	,
data	NN	data	,
is	VBN	be	,
corrupted	VBN	corrupt	,
by	IN	by	,
errors	NNS	error	,
.	SENT	.	,

```
<? s_id="1"?>
<? surface="Cathy" lemma="Cathy" pos="N" syn="su:2" />
<? surface="Zag" lemma="zile" pos="V" syn="ROOT:0" />
<? surface="hen" lemma="hen" pos="Pron" syn="obj1:2" />
<? surface="wild" lemma="wild" pos="Adj" syn="mod:5" />
<? surface="zwaaien" lemma="zwaai" pos="N" syn="vc:2" />
<? surface="." lemma="." pos="Punc" syn="punct:5" />
```

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://574heritago.com/annotations/2/",
  "type": "Annotation",
  "created": "2017-03-12T12:03:45Z",
  "body": {
    "type": "TextualBody",
    "purpose": "tagging",
    "format": "text/plain",
    "language": "en",
    "value": "Where is the origin of this statue?"
  },
  "creator": {
    "id": "http://heritago.com/users/1/",
    "type": "Person",
    "email_shai": "58bad08927902ff9307b621c54716dcc5083e339"
  },
  "selector": {
    "type": "TextPositionSelector",
    "start": 15,
    "end": 25
  }
}
```

<<http://example1.com/citingwork>>
cito:cites <<http://example2.com/citedwork>>;
cito:inTextCitationFrequency [
 a **cito:IntTextCitationCount** ;
 cito:inTextCountValue "10"^^xsd:integer ;

Elementi di Codifica del testo

Formalismi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Formati come formalismi

Data l'importanza metodologica il formato del dato diviene un vero e proprio formalismo, si parla cioè di linguaggi di codifica in quanto questi sistemi si basano su un insieme di istruzioni rigorose di codifica.

Elementi di Codifica del testo

Formalismi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Formati e formalismi di codifica

Quindi ogni pezzo di informazione aggiunta ad un testo grezzo attraverso l'inserimento di dati metatestuali (markup, annotazione, codifica), constituisce il risultato di una analisi e di una interpretazione che è stata condotta (da un umano o da una macchina) al fine di esplicitare e rappresentare nel modo più accurato e completo possibile le informazioni da veicolare attraverso il formato digitale prescelto (anche in modo incrementale).

Progress status

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

1 Presentazione

2 Introduzione

3 Codifica dei Caratteri

4 Codifica dei Testi

5 Ecosistema XML

6 Conclusioni

Codifica ad alto livello

Sistemi di marcatura

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Metodi e tecniche per la codifica di testi

La riflessione sui metodi e le pratiche migliori per la codifica elettronica dei testi è stato uno dei temi fondamentali della ricerca e della sperimentazione nel dominio dell'Informatica umanistica per molti anni.

Markup language e XML

soluzione corrente per la codifica dei testi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

XML per la descrizione e la codifica

Ad oggi la soluzione considerata ottimale per una corretta rappresentazione del testo è l'adozione dei markup language descrittivi basati su XML.

TEI-XML

Standard de facto per la codifica dei testi è considerato lo schema XML messo a punto dalla Text Encoding Initiative (TEI-XML).

Lo schema TEI-XML

Estratto di documento TEI-XML

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

```
1  <?xml version="1.0" encoding="utf-8" standalone="yes"?>
2  <!DOCTYPE div SYSTEM "divTEIsnippet.dtd">
3  <divTEIsnippet>
4      <div type="destination">
5          <ab>
6              <stamp type="postmark">
7                  <placeName>El Paso</placeName>
8                  - TX 799 -
9                  <date notBefore="1980-07-26">
10                     <unclear>PM JUL</unclear>
11                     </date>
12             </stamp>
13             <stamp type="postage">
14                 Profil masculin, avec un avion et un radar au second
15                 plan:
16                 <mentioned>US Airmail 21 c.</mentioned>
17             </stamp>
18         </ab>
19         <ab>
20             <address>
21                 <addrLine>
22                     Madame
23                     <name>Lefrère</name>
24                 </addrLine>
25                 <addrLine>4, allée George Rouault</addrLine>
26                 <addrLine>75020 Paris</addrLine>
27                 <addrLine>France</addrLine>
28             </address>
29         </ab>
30     </div>
31 </divTEIsnippet>
```

Perché TEI

La Text Encoding Initiative (TEI) è un autorevole progetto internazionale, a cui afferiscono varie organizzazioni e università, il cui scopo è fornire agli studiosi di informatica umanistica uno strumento il più espressivo e flessibile possibile per rappresentare qualsiasi aspetto di interesse relativo alla risorsa testuale da rappresentare digitalmente.

Impiego di XML

Benefici

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Perché XML

- separazione dei dati dall'applicativo di authoring/editing
- separazione della rappresentazione dei dati dalla presentazione dei dati
- possibilità di trasformare i dati in qualsiasi altro formato compatibile
- leggibilità dei documenti XML da parte di esseri umani

Impiego di XML

Benefici

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Perché XML

- standard w3c testuale, aperto, personalizzabile e liberamente utilizzabile
- semplicità di condivisione e scambio dati (interoperabilità e portabilità)
- adatto per codificare dati semistrutturati oltre che a dati strutturati
- validazione del documento attraverso specificazioni formali

Ecosistema XML

Le tecnologie XML per la definizione ed elaborazione di documenti XML

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi

Ecosistema
XML

Conclusioni
Bibliografia

- XSD: XML Schema Definition Language
- XPath: XML Path Language
- XSL: eXtensible Stylesheet Language
- XSL-T: XSL – Transformations
- XSL-FO: XSL – Formatting Objects
- XQuery: XML Query Language for XML Databases
- XInclude: XML inclusion Language
- DTD: Document Type Definition Language
- RelaxNG: Regular Expression Language for XML (New Generation)

Linguaggio di marcatura XML

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Perché XML

Adottando la tecnologia e il linguaggio XML abbiamo la possibilità di creare linguaggi di marcatura personalizzati e specifici per ogni esigenza e dominio.

XML come linguaggio per la codifica di testi

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Vantaggi

Attraverso XML è possibile strutturare i dati, gestire in modo nativo strutture gerarchiche, elaborare e presentare i dati con strumenti XML nativi, validare i tipi di strutture e i tipi di dati consentiti, gestire riferimenti incrociati tramite opportuni meccanismi di dereferenziazione, aggiungere e gestire annotazioni a vari livelli di granularità.

Progress status

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

1 Presentazione

2 Introduzione

3 Codifica dei Caratteri

4 Codifica dei Testi

5 Ecosistema XML

6 Conclusioni

Prerequisiti

Cosa è bisogna sapere

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

- Buona conoscenza dell'inglese
- Nozioni di HTML / CSS / JS
- Computer personale o a disposizione per svolgere gli esercizi assegnati
- Uso di editor di testo
- Uso di editor per programmatore o editor XML
- Eseguire comandi da shell

Strumenti

Cosa è consigliato usare

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

- Editor di testo professionali (syntax highlighting per XML, workspace)
- Editor XML + processore XSLT (normalmente è integrato nell'editor XML)
- Navigatore web
- Manuali di codifica (Guidelines TEI P5)
- Materiale didattico (slide, esempi, esercizi)

Editor di testo

Visual Studio Code

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Overview

SETUP

GET STARTED

Intro Videos

Tips and Tricks

User Interface

Themes

Settings

Key Bindings

Display Language

USER GUIDE

LANGUAGES

NODEJS /

JAVASCRIPT

PYTHON

JAVA

AZURE

EXTENSION

AUTHORING

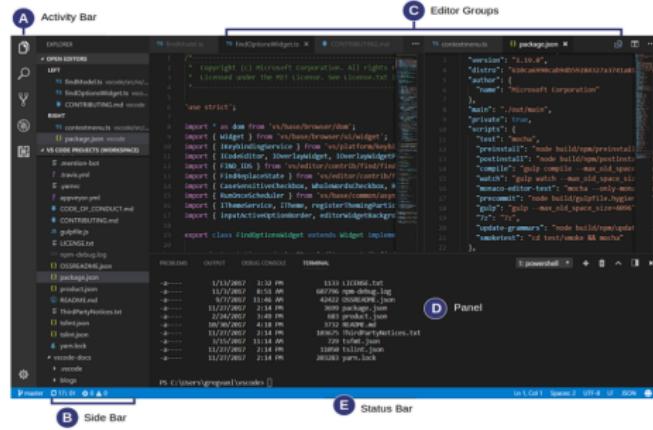
EXTENSIBILITY

REFERENCE

OTHER

User Interface

At its heart, Visual Studio Code is a code editor. Like many other code editors, VS Code adopts a common user interface and layout of an explorer on the left, showing all of the files and folders you have access to, and an editor on the right, showing the content of the files you have opened.



IN THIS ARTICLE

Basic Layout

Side by Side Editing

Minimap - outline view

Explorer

Open Editors

Views

Command Palette

Configuring the Editor

Tabs

Preview mode

Editor Groups

Grid editor layout

Working without Tabs

Window Management

Next Steps

Common Questions

Tweet

Subscribe

Ask questions

Follow @code

Request features

Report issues

Watch videos

Basic Layout

Home page del tool: <https://code.visualstudio.com/>

Validatore XML

XMLlint

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

```
XMLINT(1)                               xmllint Manual                               XMLINT(1)

NAME
    xmllint - command line XML tool

SYNOPSIS
    xmllint [--version | --debug | --shell | --xpath "XPath_expression" | --debugent | --copy | --recover | --noent | --mount | --nomet |
    --path "PATH(S)" | --load-trace | --htsout | --nowrap | --valid | --postvalid | --ddtvalid URL | --ddtvalidfpi FILE | --timing |
    --output FILE | --repeat | --insert | --compress | --html | --xsltout | --push | --memory | --maxmem NBYTES | --nowarning | --noblanks |
    --nodata | --format | --encode ENCODING | --dropdtd | --nsclean | --testFO | --catalogs | --nocatalogs | --auto | --xinclude |
    --noxmlincludenode | --loaddtd | --ddtaddr | --stream | --walker | --pattern PATTERNVALUE | --chkregister | --relaxng SCHEMA |
    --schema SCHEMA | --c14n {XML-FILE(S)... | -}

    xmllint --help

DESCRIPTION
    The xmllint program parses one or more XML files, specified on the command line as XML-FILE (or the standard input if the filename provided is -
    ). It prints various types of output, depending upon the options selected. It is useful for detecting errors both in XML code and in the XML
    parser itself.

    xmllint is included in libxml(3).

OPTIONS
    xmllint accepts the following options (in alphabetical order):

    --auto
        Generate a small document for testing purposes.

    --catalogs
        Use the SGML Catalog(s) from SGML_CATALOG_FILES. Otherwise XML catalogs starting from /etc/xml/catalog are used by default.

    --chkregister
        Turn on node registration. Useful for developers testing libxml(3) node tracking code.

    --compress
        Turn on gzip(1) compression of output.

    --copy
        Test the internal copy implementation.

    --c14n
        Use the W3C XML Canonicalisation (C14N) to serialize the result of parsing to stdout. It keeps comments in the result.

    --ddtvalid URL
        Manual page xmllint(1) line 1 (press h for help or q to quit)
```

Home page del tool: <http://www.xmlsoft.org/>

Editor XML

Cosa è consigliato usare

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

- un buon editor open source: XML Copy Editor (<http://xmlcopy-editor.sourceforge.net/>)
- per Mac: XMLSpear, Textmate, Eclipse, IntelliJIDEA
- un ottimo editor: Oxygen (<http://www.oxygenxml.com/>)
 - multi piattaforma, ma non gratuito (in prova gratuita per un mese)
- altri editor: funzioni fondamentali sono la validazione, l'autocompletamento, l'esecuzione di fogli di stile

Editor XML

EditiX

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

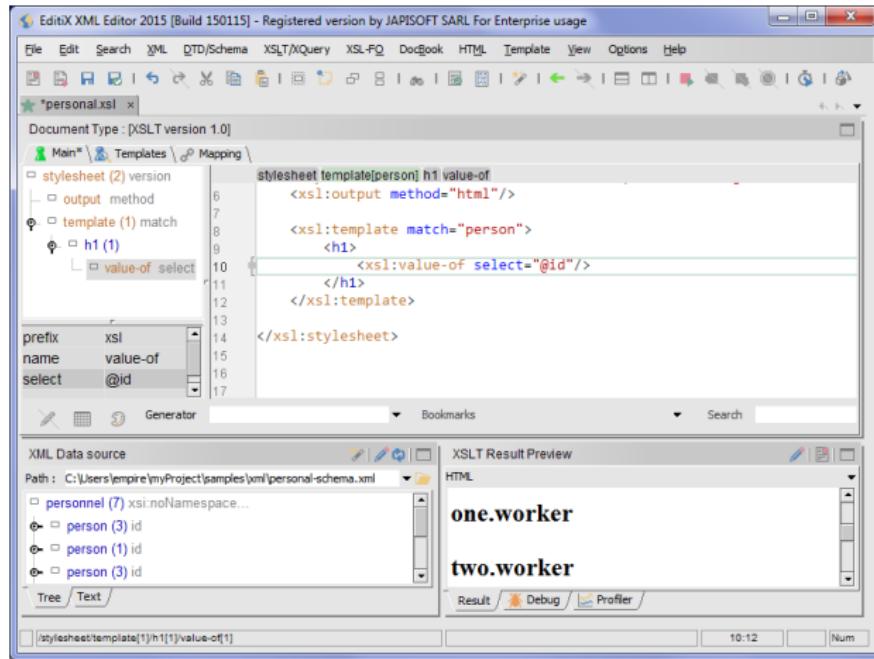
Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia



Home page del tool: <http://www.editix.com>

Programma d'esame

Cosa bisogna fare per superare l'esame

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione

Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

- Studiare i lucidi delle lezioni
- Padroneggiare gli esercizi svolti durante il corso
- Studiare i testi indicati dal docente
- Studiare i capitoli delle Guidelines TEI che riguardano il corso e il progetto
- Realizzare il progetto di codifica concordato con il docente

Modalità d'esame

In cosa consiste l'esame

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

- Invio del progetto (obbligatoriamente tramite github)
- Colloquio
 - Discussione del progetto
 - Verifica delle conoscenze di base XML, XSD, XSL
 - Verifica delle basi teoriche
 - Conoscenza di TEI P5 (moduli principali, parti spiegate a lezione, moduli particolari utilizzati nel progetto)

Materiale didattico

Riferimenti per studiare

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

Slide

- Slide delle lezioni corso a.a. 2018-2019
- Materiale integrativo fornito dal docente
- Repo github del corso
<https://github.com/angelodel80/corsoCodifica>

Libri

- Burnard, L. (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources.* Marseille: OpenEdition Press.
- Ciotti, F. (2007). *Il testo e l'automa: saggi di teoria e critica computazionale dei testi letterari.* Aracne.
- Goldberg, K. H. (2010). *XML: Visual QuickStart Guide.* Pearson Education.
- Carey, P., and Vodnik, S. (2014). *New Perspectives on XML, Comprehensive.* Cengage Learning.

Materiale didattico

Riferimenti per studiare

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione

Codifica dei
Caratteri

Codifica dei
Testi

Ecosistema
XML

Conclusioni

Bibliografia

Siti Web

- <http://www.tei-c.org/>
- <http://teibyexample.org/>
- <https://www.w3.org/standards/xml/>

References

Presentazione
del Corso
Codifica di
Testi
a.a.
2018-2019

A.M. Del
Grosso

Presentazione
Introduzione
Codifica dei
Caratteri
Codifica dei
Testi
Ecosistema
XML
Conclusioni
Bibliografia

-  Ciotti F., e Crupi G, a c. di. 2012. Dall'Informatica umanistica alle culture digitali. ROMA : Casa Editrice Università La Sapienza. open access: <http://www.editricesapienza.it/node/7688>
-  Orlandi, T. (2010). Informatica testuale: teoria e prassi. Laterza.
-  Pierazzo, E. (2015). Digital Scholarly Editing : Theories, Models and Methods. Farnham Surrey: Ashgate.
-  Driscoll, M. J., and Pierazzo, E. (Eds.). (2016). Digital Scholarly Editing: Theories and Practices (Vol. 4). Open Book Publishers.
-  Williams, I. (2009). Beginning XSLT and XPath: Transforming XML Documents and Data. Wiley.
-  Kay, M. (2011). XSLT 2.0 and XPath 2.0 Programmer's Reference. Wiley.
-  **XML Standards Reference**, MSDN.
[https://msdn.microsoft.com/en-us/library/ms256177\(v=vs.110\).aspx](https://msdn.microsoft.com/en-us/library/ms256177(v=vs.110).aspx)
-  **IBM XML Tutorial**,
<https://www.ibm.com/developerworks/xml/tutorials/xmlintro/xmlintro.html>