

# Presentazione del Corso Codifica di Testi

Angelo Mario Del Grosso

`angelo.delgrosso@ilc.cnr.it`

*CNR-ILC-LicoLab*

`http://licolab.ilc.cnr.it/`

Istituto di Linguistica Computazionale “A. Zampolli”,  
13th September 2018

# Piano della presentazione

Presentazione  
del Corso  
Codifica di  
Testi

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
- 7 Bibliografia

# Progress status

Presentazione  
del Corso  
Codifica di  
Testi

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.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

slide di presentazione: chi sono, piccola bio, di cosa mi occupo  
prendere dal curriculum alcuni pezzi mettere mail istituzionale  
ed eventualmente telefono

# Presentazione del Corso

Presentazione  
del Corso  
Codifica di  
Testi

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.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.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.

## Razionale

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.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

## Competenze attese

Al termine del corso sarete in grado di valutare il metodo di codifica più appropriato allo scenario d'interesse, di creare uno schema di codifica TEI e di usare gli strumenti più idonei per la codifica e la (semplice) elaborazione e visualizzazione di un testo.



# Principali Argomenti

Obiettivi, competenze e conoscenze

Presentazione  
del Corso  
Codifica di  
Testi

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

# Perché è importante la codifica dei testi

## Motivazioni pratiche

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

## Perché codificare

Nella nostra cultura 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

## Motivazioni pratiche

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

## Perché codificare

Per rendere disponibile questo patrimonio attraverso i sistemi per la gestione dell'informazione digitali e computazionali è necessario effettuare una trasposizione/transcodifica dei testi (procedimento di conversione dei dati codificati secondo un sistema verso un sistema diverso) dal loro supporto originario verso il nuovo supporto elettronico.

# Perché è importante la codifica dei testi

## Motivazioni teoriche

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

### Perché codificare

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.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

### Perché codificare

L'attività di codifica è funzione metodologica nell'ambito delle discipline che si occupano del testo.

### Perché codificare

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

*Esplicitare 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.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

## Rappresentare il testo

Il focus del corso sarà diretto alla rappresentazione digitale del testo.

## Esistono dibattiti e Controversie

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

Molte questioni ancora non sono state risolte altre sono molto controverse, sia teorico-metodologico, sia pratico-tecnologico.

# Perché è importante la codifica dei testi

Ma in definitiva

Presentazione  
del Corso  
Codifica di  
Testi

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.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.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 di fornire un'adeguata rappresentazione dei dati testuali in formato digitale.

# Elementi di Codifica dei Caratteri

## Definizioni

### Perché è importante

La codifica dei caratteri costituisce il grado zero della rappresentazione di testi su supporto digitale. Essa costituisce la base di tutti gli ulteriori sistemi di codifica testuale.

### Codifica dei caratteri

Come qualsiasi altro tipo di dato, anche i caratteri vengono rappresentati all'interno di un elaboratore mediante una codifica numerica binaria. Le codifiche dei caratteri sono la base di qualsiasi schema di codifica testuale.

# Elementi di Codifica dei Caratteri

## Definizioni

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione  
Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

[Character set](#) spiega ...

[Code Set](#) spiega ...

[Character encoding](#) spiega ...

[Tabella del Code page](#) spiega ...

[Unicode](#) spiega

# Elementi di Codifica dei Caratteri

## Definizioni

Presentazione  
del Corso  
Codifica di  
Testi

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.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

## Complessità e rappresentazione

Presentazione  
del Corso  
Codifica di  
Testi

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.

# Elementi di Codifica dei Caratteri

## Unicode

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

Complessità di rappresentazione universale

Cenno al consorzio Unicode

# Progress status

Presentazione  
del Corso  
Codifica di  
Testi

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.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

Di fatti 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.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 plethora di aspetti a vari livelli di astrazione, a varie dimensioni e a varie granularità 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.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, infatti, cosa si intende per testo.**

# Rappresentazione digitale dei testi

## Modello dati di un testo

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.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

## La rappresentazione di un testo

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

[SLIDE DA COMPLETARE]

# Elementi di Codifica del testo

## Formalismi

Presentazione  
del Corso  
Codifica di  
Testi

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 di un formato condiziona molto l'efficienza delle operazioni e l'efficacia delle dichiarazioni.

# Elementi di Codifica del testo

## Tabella Formalismi

### 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

## Formalismi

Presentazione  
del Corso  
Codifica di  
Testi

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.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), costituisce 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 possibile le informazioni da veicolare attraverso il formato digitale prescelto (anche in modo incrementale).

# Progress status

Presentazione  
del Corso  
Codifica di  
Testi

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

# XML titolo

XML sottotitolo

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

## block titolo

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

# XML titolo

## XML sottotitolo

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

### block titolo

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

### block titolo

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

# XML titolo

## XML sottotitolo

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

### Perché XML

importanza dell'XML importanza della definizione di uno  
schema XML importanza della definizione di trasformate XSL e  
manipolazione del DOM

[SLIDE DA COMPLETARE]

- 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)

# XML titolo

## XML sottotitolo

Presentazione  
del Corso  
Codifica di  
Testi

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.

## 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.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

# Conclusioni titolo

## Conclusioni sottotitolo

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione

Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia

block titolo

Conclusioni

block titolo

Strumenti

block titolo

Materiale didattico

# Progress status

Presentazione  
del Corso  
Codifica di  
Testi

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

# References

Presentazione  
del Corso  
Codifica di  
Testi

A.M. Del  
Grosso

Presentazione  
Introduzione

Codifica dei  
Caratteri

Codifica dei  
Testi

Ecosistema  
XML

Conclusioni

Bibliografia



Ciotti 2011



CC Wiki *Best practices for attribution*, CC Wiki 2014,  
[https://wiki.creativecommons.org/wiki/Best\\_practices\\_for\\_attribution](https://wiki.creativecommons.org/wiki/Best_practices_for_attribution)