

Laboratorio di IU Dati

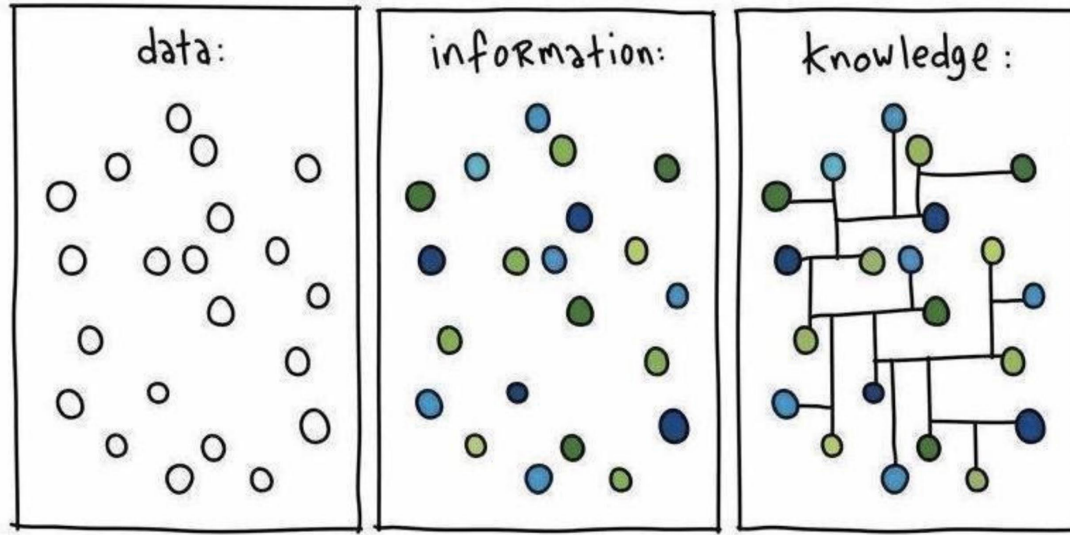
2 marzo 2021

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

DATI, INFORMAZIONE, CONOSCENZA











- DATI: rappresentazioni di oggetti, eventi, entità che possono essere misurabili
- INFORMAZIONE: insieme di dati collocati in un contesto
- CONOSCENZA: mix di esperienze, competenze e informazioni

DATI NELLE DISCIPLINE UMANISTICHE

- DATA = CULTURAL DATA
- Tipi di dati
 - qualitativi
 - quantitativi
 - geospaziali
- Livello di strutturazione
 - strutturati
 - non strutturati
 - semi-strutturati
- In DH: DATA = DIGITAL DATA

ESEMPI: DATI NON STRUTTURATI

 <p>Text files and documents</p>	 <p>Server, website and application logs</p>	 <p>Sensor data</p>	 <p>Images</p>
 <p>Video files</p>	 <p>Audio files</p>	 <p>Emails</p>	 <p>Social media data</p>

ESEMPI

NON STRUTTURATO

DESIDERIO

Il giorno della prova è giunto;
Figlio, sei tu con me?

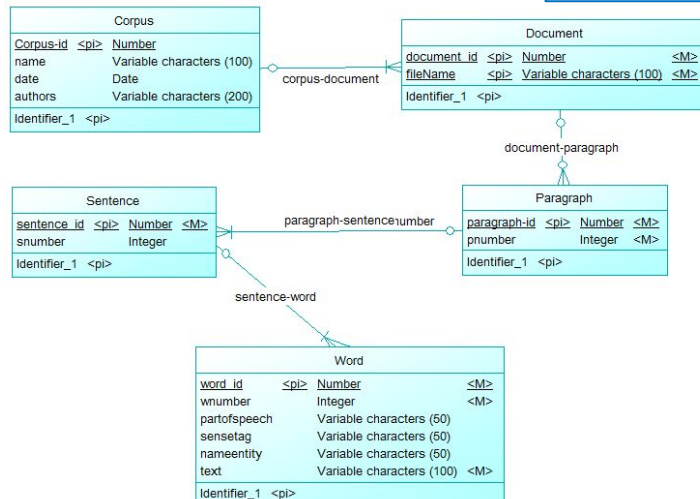
ADELCHI

Sì dura inchiesta
Quando, o padre, mertai?

SEMI-STRUTTURATO

```
<sp>
<speaker>DESIDERIO</speaker>
<lg>
<l>Il giorno della prova è giunto;</l>
<l>Figlio, sei tu con me?</l>
</lg>
</sp>
<sp>
<speaker>ADELCHI</speaker>
<lg>
<l>Sì dura inchiesta</l>
<l>Quando, o padre, mertai?</l>
</lg>
</sp>
```

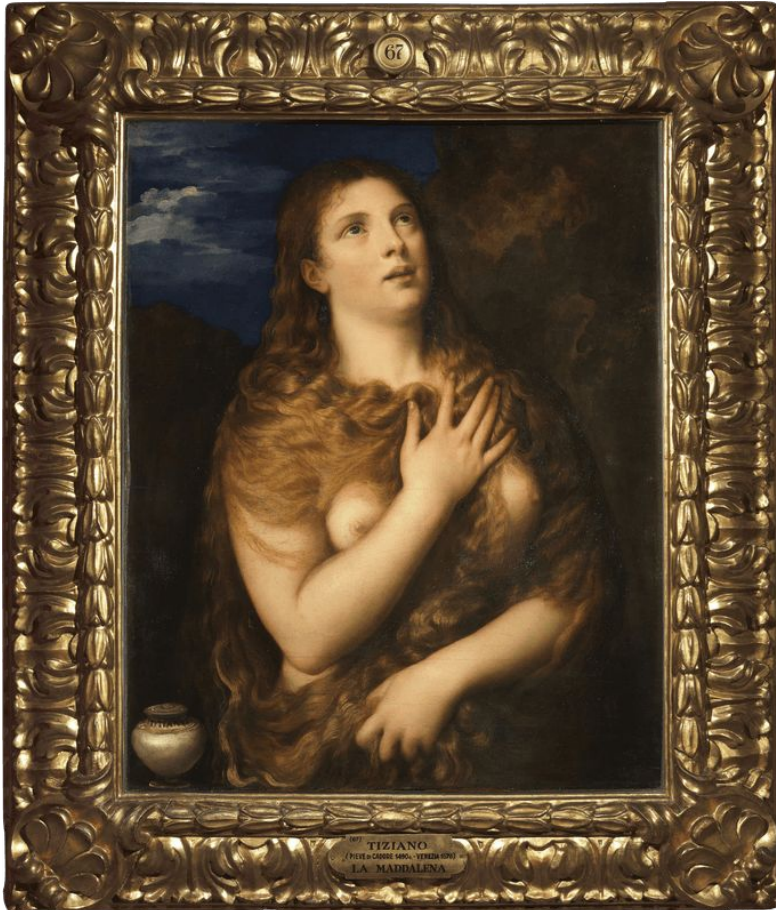
STRUTTURATO



	A	B	C	D
1	Source	Type	Target	Weight
2	vermondo	Undirected	desiderio	3
3	vermondo	Undirected	adelchi	2
4	vermondo	Undirected	ermengarda	1
5	desiderio	Undirected	adelchi	10
6	desiderio	Undirected	ermengarda	2

ESEMPI

NON STRUTTURATO



STRUTTURATO

Autore	Tiziano Vecellio (Pieve di Cadore 1488/90 – Venezia 1576)
Data	1531-1535
Museo	Palazzo Pitti
Collezione	Galleria Palatina
Collocazione	Sala di Apollo
Tecnica	Olio su tavola
Dimensioni	85,8 x 69,5 cm
Iscrizioni	"TITIANUS" sul profilo dell'imboccatura del vasetto in basso a sinistra
Inventario	Inv. 1912 n. 67

QUESTIONI IMPORTANTI

- Importanza del contesto
- Formato digitale versus analogico
- Scelta dei metadati: DublinCore / Manus
- Stato di conservazione
- Copyright
- Archiviazione a lungo termine
- Metodi per trovare/scaricare dati

METADATI: Dublin Core

DC Element Name	Definition
1. Title	A name given to the resource.
2. Creator	An entity primarily responsible for making the resource.
3. Subject	The topic of the resource.
4. Description	An account of the resource.
5. Publisher	An entity responsible for making the resource available.
6. Contributor	An entity responsible for making contributions to the resource.
7. Date	A point or period of time associated with an event in the lifecycle of the resource.
8. Type	The nature or genre of the resource.
9. Format	The file format, physical medium, or dimensions of the resource.
10. Identifier	An unambiguous reference to the resource within a given context.
11. Source	A related resource from which the described resource is derived.
12. Language	A language of the resource.
13. Relation	A related resource.
14. Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
15. Rights	Information about rights held in and over the resource.

COLLEZIONI DIGITALI

TESTI

COLLEZIONI GENERALI:

- 68 lingue: <https://it.wikisource.org/>
- 67 lingue: <http://www.gutenberg.org/>
- 40 lingue: <http://www.intratext.com/>
- solo letteratura italiana: <https://www.liberliber.it/> / <http://www.bibliotecaitaliana.it/>
- latino, greco, arabo: <http://www.perseus.tufts.edu/>

COLLEZIONI TESTI SPECIFICI

- Decameron Web: http://www.brown.edu/Departments/Italian_Studies/dweb/
- Commedia: <http://www.worldofdante.org/>
- Digital Archives of PHilosophical Texts: <http://www.daphnet.org/>
- Articoli di giornale: <https://archivio.unita.news/> / <https://github.com/swapUniba/unita/>

COLLEZIONI DIGITALI

IMMAGINI

- Pharos: <http://images.pharosartresearch.org/>
- Open Art Images: <https://openartimages.com/>
- Biblioteche italiane: <http://www.internetculturale.it/>
- David Rumsey Collection (mappe): <https://www.davidrumsey.com/>
- Biblioteca digitale ambrosiana:
<https://ambrosiana.comperio.it/biblioteca-digitale/>

COLLEZIONI DIGITALI

MULTIMEDIALE

- Europeana: <https://www.europeana.eu/>
- Centro virtuale per la conoscenza dell'Europa: <https://www.cvce.eu/>
- Internet Archive: <https://archive.org/>
- Library of Congress (US): <https://www.loc.gov/>
- New York Public Library: <https://digitalcollections.nypl.org/>
- Shoah Visual History: <http://vhaonline.usc.edu/>

STRUTTURATI

- Dati aperti PA: <https://dati.gov.it/>
- Coding Dürer: <http://codingdurer.de/data.html>
- Met: <https://github.com/metmuseum/openaccess>

STRUMENTI

1. OCR - Tesseract → convertire da pdf a text
2. GutenTag → creare corpora di testi
3. BootCaT: <https://bootcat.dipintra.it/?section=download>
→ scaricare testi dal web

1. OCR - Tesseract

- Codice: <https://github.com/tesseract-ocr/tesseract>
 - Demo online: <http://195.148.30.97/cgi-bin/ocr.py>
1. Scaricare un pdf dall'archivio dell'Unità:
<https://archivio.unita.news/>
 2. Nella demo online di Tesseract scegliere la lingua "Italian"
 3. Caricare il pdf cliccando su "Scegli file"
 4. Cliccare su "Submit"

2. GutenTag (parte 1)

- Versione Web: <http://www.cs.toronto.edu/~jbrooke/gutentag/>
- Scaricare testi da Gutenberg project:
 1. Cliccare su “Web version”
 2. Scegliere “Genre: Poetry”
 3. Sotto “Author”, scrivere “Ada Negri” nel campo “Author Name” e scegliere “Italian” nel campo “Author Nationality”
 4. Sotto “Text”, scegliere “Italian” nel campo “Language”
 5. Cliccare su “Export”
 6. Sotto “Format” scegliere il formato “Plain Text” o “XML”
 7. Sotto “Output” indicare il nome che avrà il file da scaricare
 8. Cliccare su “Export”

2. GutenTag (parte 2)

- Versione Web: <http://www.cs.toronto.edu/~jbrooke/gutentag/>
- Analizzare testi da Gutenberg project:
 1. Ripetere le indicazioni da 1 a 4 della slide precedente
 2. Cliccare su “Analyze”
 3. Sotto “Add Textual Measure”, scegliere le metriche da usare: se ne può aggiungere più di una cliccando su “+”
 4. Sotto “Output” indicare il nome che avrà il file da scaricare
 5. Cliccare su “Analyze”

3. BootCaT (parte 1)

- Tutorial:

https://docs.sslmit.unibo.it/doku.php?id=bootcat:tutorials:basic_1

1. Aprire il software con un doppio click
2. Dare un nome al corpus, “Caravaggio”, selezionare la lingua “Italian” e cliccare su “Next”
3. Cliccare su “Simple mode”
4. Cliccare su “Next”
5. Scrivere parole chiave di ricerca, una per riga, almeno 5 e cliccare su “Next”
6. Scegliere la lunghezza delle tuple e il numero massimo poi cliccare su “Generate tuples” e poi su “Next”
7. Scrivere “.com” nel campo “Exclude these Internet domains...” e cliccare su “Generate Queries”

3. BootCaT (parte 2)

1. Cliccare su “Open Queries Folder”
2. Cliccare su ogni bottone “Open in Browser”: salvare ogni pagina (CTRL+S su Windows o CMD+S su Mac) poi cliccare su “Collect URLs”
3. Deselezionare le URL da non includere nel corpus (pagine ridondanti o troppo generiche) e cliccare su “Next”
4. Cliccare su “Build corpus” e aspettare...
5. Cliccare su “Open corpus folder”: controllare i file “report.csv”, “corpus.xml” e la cartella “corpus”



GRAZIE!

Email: rachele.sprugnoli@unicatt.it

Twitter: [@RSprugnoli](https://twitter.com/RSprugnoli)

