

# Abilità informatiche

A.A. 2023/2024

## 02a - Gestione dei dati

Sebastian Barzaghi

[sebastian.barzaghi2@unibo.it](mailto:sebastian.barzaghi2@unibo.it)

<https://orcid.org/0000-0002-0799-1527>



## 2.0 Una breve digressione

# Ottimo lavoro!

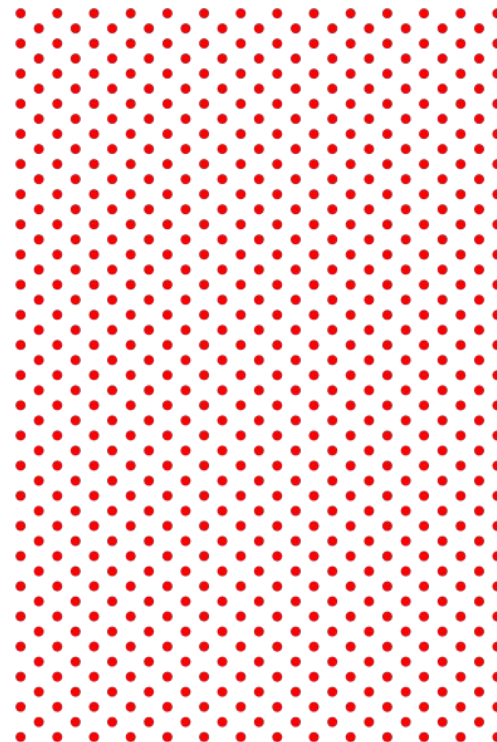
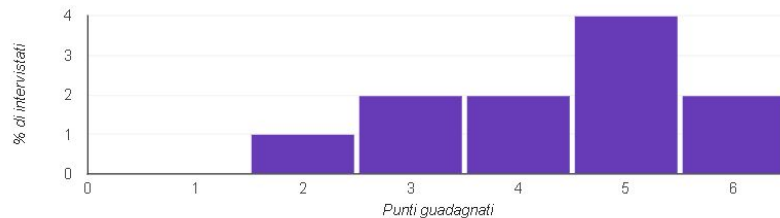
## Statistiche

**Media**  
4,36 / 6 punti

**Mediana**  
5 / 6 punti

**Intervallo**  
2 - 6 punti

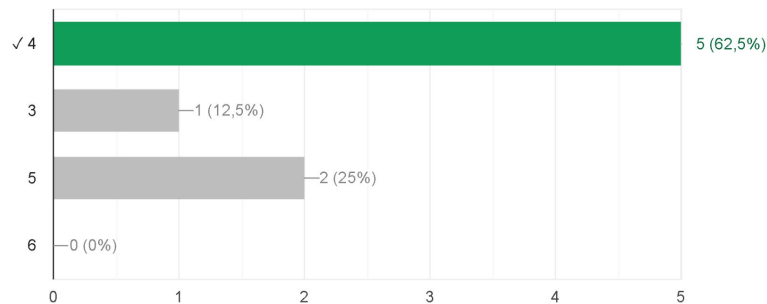
Distribuzione dei punti totali



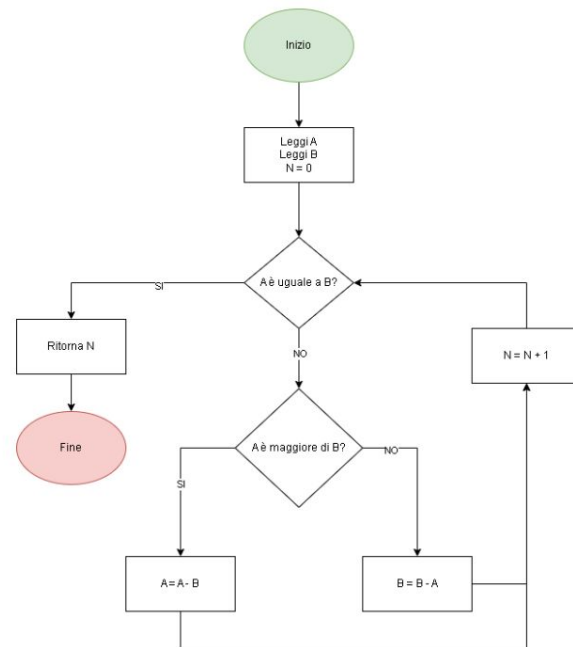
# Ottimo lavoro!

Cosa ritorna l'algoritmo se specifichiamo  $A = 10$  e  $B = 2$ ?

5/8 risposte corrette



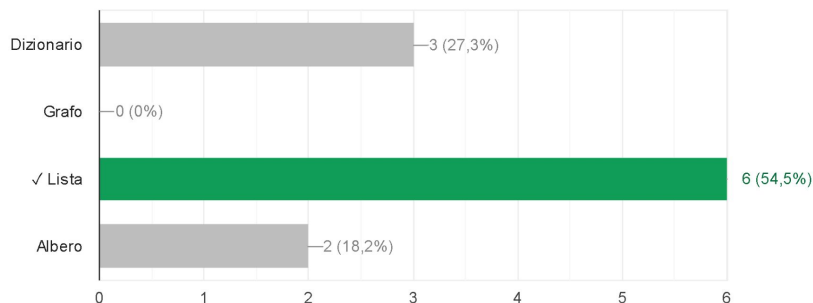
Cosa ritorna l'algoritmo se specifichiamo  $A = 10$  e  $B = 2$ ?



# Ottimo lavoro!

Con quale struttura dati è possibile rappresentare una serie di citazioni bibliografiche?

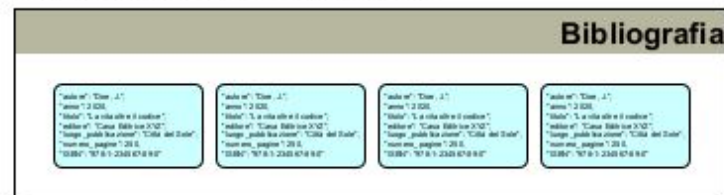
6/11 risposte corrette



## Citazione bibliografica

"autore": "Doe, J.",  
"anno": 2020,  
"titolo": "La vita oltre il codice",  
"editore": "Casa Editrice XYZ",  
"luogo\_publicazione": "Città del Sole",  
"numero\_pagine": 250,  
"ISBN": "978-1-234567-89-0"

## Serie di citazioni bibliografiche





## 2.1 Dati

Cos'è un dato

Cos'è un dataset

# Cos'è un dato?



Fonte:

<https://unsplash.com/it/foto/libro-bianco-e-marrone-su-superficie-intrecciata-marrone-LUGuCtvlk1Q>



Fonte:

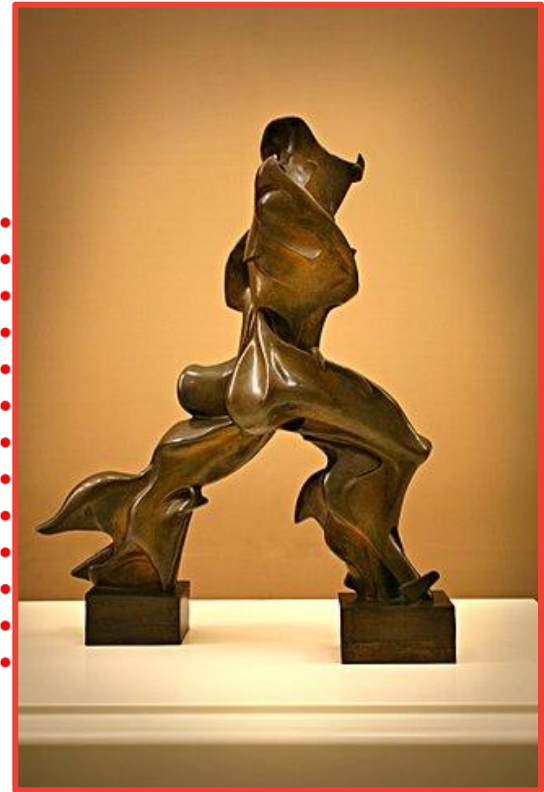
[https://unsplash.com/it/foto/un-vecchio-libro-con-limmagine-di-un-uomo-su-di-esso-U-tDy-NlnTs?utm\\_content=creditShareLink&utm\\_medium=referral&utm\\_source=unsplash](https://unsplash.com/it/foto/un-vecchio-libro-con-limmagine-di-un-uomo-su-di-esso-U-tDy-NlnTs?utm_content=creditShareLink&utm_medium=referral&utm_source=unsplash)



# Cos'è un dato?



Isle of the Dead (first version, May 1880). Oil on canvas, 110.9 x 156.4 cm (43.6 x 61.5 in). Kunstmuseum Basel, Switzerland



Umberto Boccioni, CC BY-SA 4.0  
<<https://creativecommons.org/licenses/by-sa/4.0>>, via Wikimedia Commons



# Cos'è un dato?



Fonte:

[https://it.wikipedia.org/wiki/La\\_casa\\_dalle\\_finestre\\_che\\_ridono#/media/File:Finestra\\_che\\_ride.jpg](https://it.wikipedia.org/wiki/La_casa_dalle_finestre_che_ridono#/media/File:Finestra_che_ride.jpg)



Adam Robinson-Yu - A Short Hike (2019 Video Game). Fonte:

[https://commons.wikimedia.org/wiki/File:A\\_Short\\_Hike\\_Screenshot\\_6.png](https://commons.wikimedia.org/wiki/File:A_Short_Hike_Screenshot_6.png)

# Cos'è un dato?

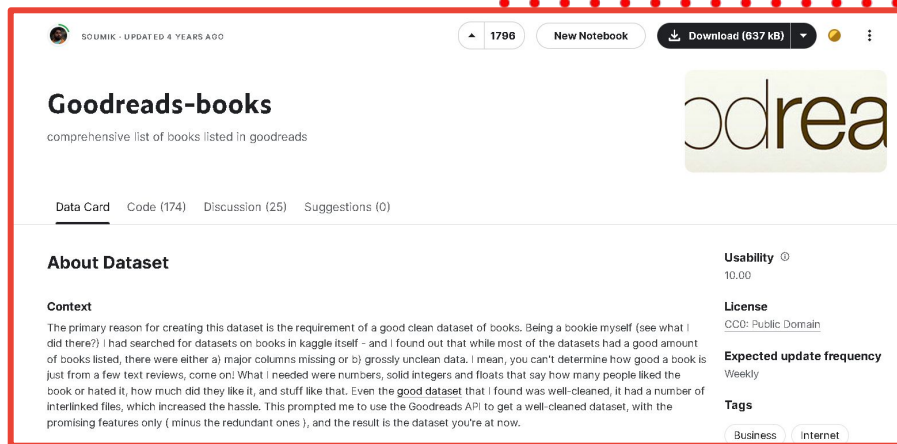


Tabula Quinta d'Europa. Claudio Tolomeo - Herzog August Library, Germany - CC BY-SA.  
[https://www.europeana.eu/item/168/item\\_4HJCJIGGPLEYCMQK5BCNEDQS207MCG5A](https://www.europeana.eu/item/168/item_4HJCJIGGPLEYCMQK5BCNEDQS207MCG5A)



Shrine from El Kab. Claudio Tolomeo - Fitzwilliam Museum, United Kingdom - CC BY.  
[https://www.europeana.eu/it/item/181/share3d\\_1145](https://www.europeana.eu/it/item/181/share3d_1145)

# Cos'è un dato?



The screenshot shows the Kaggle page for the 'Goodreads-books' dataset. At the top, it indicates the dataset was updated 4 years ago and has 1798 rows. The title 'Goodreads-books' is prominently displayed, followed by the description 'comprehensive list of books listed in goodreads'. Below this, there are tabs for 'Data Card', 'Code (174)', 'Discussion (25)', and 'Suggestions (0)'. The 'About Dataset' section explains the dataset's origin and cleaning process. On the right, a 'oodrea' logo is visible. Further down, metadata is provided: 'Usability' is 10.00, the 'License' is 'CC0: Public Domain', and the 'Expected update frequency' is 'Weekly'. At the bottom, there are 'Tags' for 'Business' and 'Internet'.

Goodreads-books  
comprehensive list of books listed in goodreads

Data Card Code (174) Discussion (25) Suggestions (0)

**About Dataset**

**Context**

The primary reason for creating this dataset is the requirement of a good clean dataset of books. Being a bookie myself (see what I did there?) I had searched for datasets on books in kaggle itself - and I found out that while most of the datasets had a good amount of books listed, there were either a) major columns missing or b) grossly unclear data. I mean, you can't determine how good a book is just from a few text reviews, come on! What I needed were numbers, solid integers and floats that say how many people liked the book or hated it, how much did they like it, and stuff like that. Even the good dataset that I found was well-cleaned, it had a number of interlinked files, which increased the hassle. This prompted me to use the Goodreads API to get a well-cleaned dataset, with the promising features only (minus the redundant ones), and the result is the dataset you're at now.

**Usability** ①  
10.00

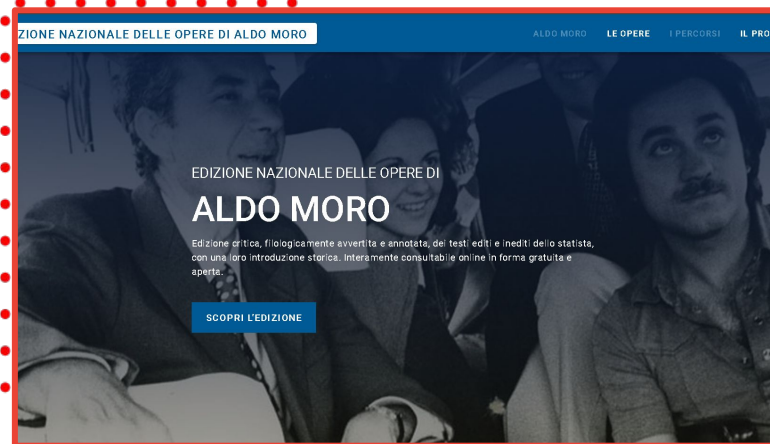
**License**  
CC0: Public Domain

**Expected update frequency**  
Weekly

**Tags**  
Business Internet

Goodreads books dataset.

<https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks>



The screenshot shows the website for the 'Edizione Nazionale delle Opere di Aldo Moro'. The header is blue with the title 'EDIZIONE NAZIONALE DELLE OPERE DI ALDO MORO' and navigation links for 'ALDO MORO', 'LE OPERE', 'I PERCORSI', and 'IL PROGETTO'. The main image is a black and white photograph of Aldo Moro and other people. Overlaid on the image is the text 'EDIZIONE NAZIONALE DELLE OPERE DI ALDO MORO' and a description: 'Edizione critica, filologicamente avvertita e annotata, dei testi editi e inediti dello statista, con una loro introduzione storica. Interamente consultabile online in forma gratuita e aperta.' Below this is a blue button that says 'SCOPRI L'EDIZIONE'.

EDIZIONE NAZIONALE DELLE OPERE DI ALDO MORO

ALDO MORO LE OPERE I PERCORSI IL PROGETTO

EDIZIONE NAZIONALE DELLE OPERE DI  
**ALDO MORO**

Edizione critica, filologicamente avvertita e annotata, dei testi editi e inediti dello statista, con una loro introduzione storica. Interamente consultabile online in forma gratuita e aperta.

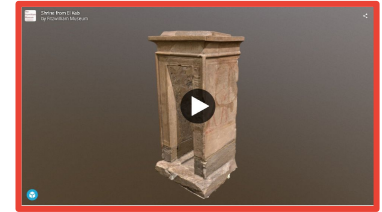
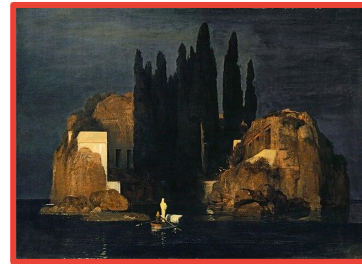
SCOPRI L'EDIZIONE

Moro, Aldo, *Edizione Nazionale delle Opere di Aldo Moro*, voll., Bologna, Università di Bologna, 2021. ISBN: 9788854970496; DOI: <https://doi.org/10.6092/unibo/aldomoro>

# Cos'è un dato?

Premessa: problemi epistemologici

Definizioni innumerevoli e diverse a seconda della disciplina di riferimento, se non addirittura da persona a persona



# Cos'è un dato?

*Datum* (participio passato di *dare*): **“qualcosa che è dato”** (dalla natura allo scienziato)”

Un dato non esiste di per sé, ma deve essere generato o raccolto tramite sensori o attraverso uno sforzo umano

*Datum* → *Captum*: **“qualcosa che viene preso, costruito”**



# Cos'è un dato?

Valore assegnato a qualcosa e che trasmette informazioni

Numeri, parole, immagini, video, fotografie, registrazioni audio, interviste, artefatti, manoscritti, note, fumetti, codici, performance, siti web, esibizioni, ...

Evidenze che possiamo manipolare per trovare pattern e significati

Qualsiasi cosa che può essere quantificata, qualificata o interpretata in qualche modo e usata come evidenza



# Cos'è un dato?

Per essere usati con criterio, i dati necessitano di un contesto

*L'appuntamento con il dottor Watt è martedì alle 14:30 presso la clinica di Heslington Lane.*

Questa informazione contiene i dati seguenti:

- Con chi è l'appuntamento
- La data dell'appuntamento
- L'orario dell'appuntamento
- Il luogo dell'appuntamento



# Cos'è un dato?

## Una “categoria relazionale”

Ciò che è considerabile “dato” dipende da chi lo usa, come, e per quale scopo

Percepire, osservare, raccogliere dati sono tutti atti interpretativi, in un contesto che dà forma ai dati e ne definisce i limiti

**An historical perspective on family violence and child abuse: Comment on Moloney et al, *Allegations of Family Violence*, 12 June 2007**

Es. i tassi di violenza domestica sono stati storicamente sottovalutati perché questi crimini venivano raramente documentati e considerati dal sistema giuridico, dai professionisti di salute mentale e dai ricercatori delle scienze sociali, nonché dalla società in senso lato. Fonte:

<https://doi.org/10.5172/jfs.327.14.2-3.271>

# Cos'è un dataset?

I dati vengono raccolti e conservati in modo da poter essere interrogati e analizzati per spiegare qualcosa

Una collezione di dati (spesso in un formato leggibile da un computer)

id	title	alt	type
1	1 Edipo risolve l'enigma della Sfinxe	Oedipus and the Sphinx	Pittura
2	2 Eracle di Mantinea	statuette	Scultura
3	3 Eracle cattura il cinghiale di Erimanto	amphora	Pittura vascolare
4	4 Eracle brandisce la clava contro Caco	Hercule tuant Cacus	Pittura
5	5 Psiche riceve il primo bacio da Amore	Psyché et l'Amour, dit aussi Psyché recevant le premier baiser de l'Amour	Pittura
6	6 Diana cacciatrice	Statuette: Diane	Scultura
7	7 Eracle giunge all'Olimpo tra gli Dei	olpé	Pittura vascolare
8	8 Odisseo massacrare i pretendenti di Penelope	Cratée des prétendants	Pittura vascolare
9	9 Dioniso, accompagnato dal suo corteo, incontra Arianna	Sarcophage; couvercle de sarcophage	Scultura
10	10 Venere di Milo	Vénus de Milo	Scultura
11	11 Teseo uccide il Minotauro	Skyphos Rayet	Pittura vascolare
12	12 Gigantomachia	Amphore de Milo	Pittura vascolare
13	13 Orfeo incanta gli animali	Orphée charmant les animaux	Disegno
14	14 Clio, Euterpe e Talia	Clio, Euterpe et Thalie	Pittura
15	15 Danzatrici di Ruvo	Danzatrici di Ruvo	Pittura murale
16	16 Nike di Samotracia	Victoire de Samothrace	Scultura
17	17 Ercole Farnese	Ercole Farnese	Scultura
18	18 Atena Mattia	Athéna Mattia	Scultura
19	19 Polifemo e Galatea si baciano		Pittura murale
20	20 Neottolero trasporta il corpo di Astianatte		Scultura
21	21 Ajax and Cassandra		Pittura
22	22 Diana appoggiata a un cervo	Diane appuyée sur un cerf	Scultura
23	23 Eracle iniziato ai Misteri Eleusini		Scultura
24	24 Afrodite Callipige	Afrodite Callipige	Scultura

Parte del dataset *Mythologiae*. Fonte: <https://mythologiae.unibo.it/>

# Cos'è un dataset?

Strumento potente e prezioso, ma anche delicato, parziale, imperfetto

Riflette le circostanze che hanno portato alla sua creazione e gestione (comunità, individui, organizzazioni, ambienti, strumenti, limiti, bias, responsabilità...)

## Researchers' tool finds bias in state-of-the-art generative AI model

August 10, 2023

By Emily Cerf

## Amazon ditched AI recruiting tool that favored men for technical jobs

## Robert was wrongly arrested because of a racist algorithm. Are these the hidden dangers of AI?

By Flint Duxfield and Samantha Hawley

Posted Tue 25 Apr 2023 at 8:58pm

## Finally, We Know What *Minority Report* Would Look Like on the Gulf Coast of Florida

In Pasco County, the sheriff is hunting pre-crime—and engaging in systematic harassment of citizens.

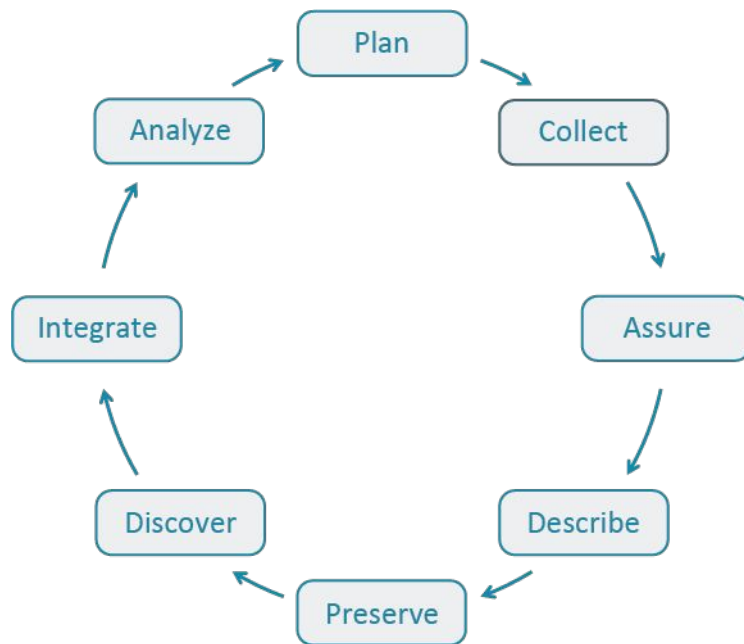
By Charles P. Pierce PUBLISHED: SEP 11, 2020 1:02 PM EST

# Gestione dei dati

Processo **critico** di creazione, raccolta, organizzazione, descrizione, archiviazione, e condivisione dei dati

Obiettivo: produrre dataset autodescritti, sostenibili e utilizzabili

Motivi: trovabilità, usabilità, citabilità dei dati



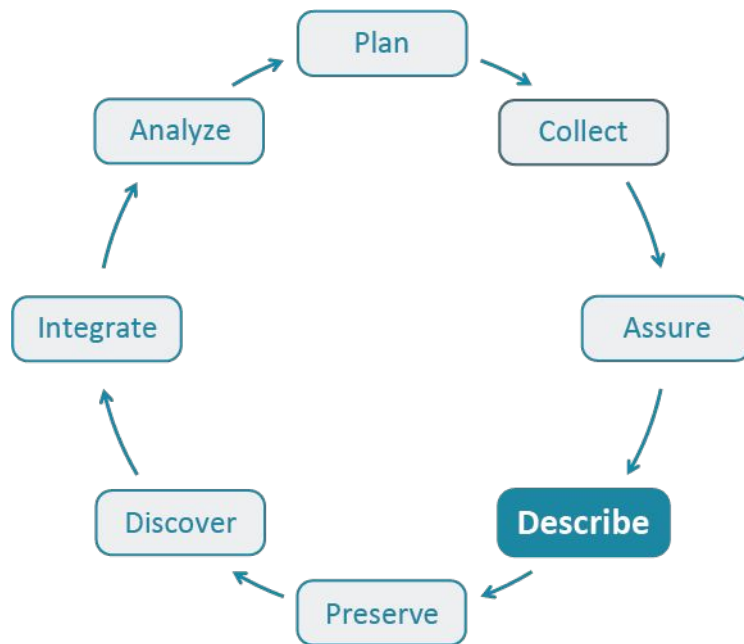
Una possibile rappresentazione del processo di gestione dei dati. Fonte:

<https://dataoneorg.github.io/Education/bestpractices/>

# Descrizione dei dati

Un componente chiave della gestione dei dati è la descrizione dei dati e del loro contesto

Obiettivo: contrastare la naturale tendenza delle informazioni all'entropia



La fase di descrizione dei dati. Fonte:  
<https://dataoneorg.github.io/Education/bestpractices/>



## 2.1 Metadati

Cos'è un metadato  
Tipologie  
Schemi

# Cos'è un metadato?



Data

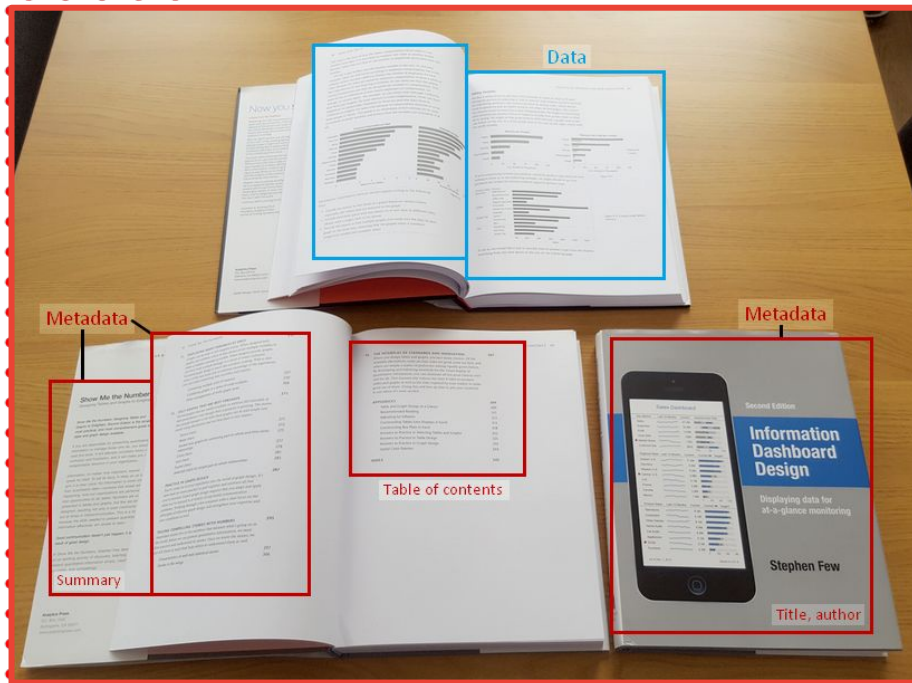
**Filename:** Tadzik.jpg  
**Author:** Piotr Kononow  
**Date:** August 15, 2016 6:40:10PM  
**File:** 5,312 × 2,988 JPEG  
15.9 megapixels  
3,393,448 bytes  
(3.2 megabytes)  
**Camera:** Samsung SM-G920F  
4.3 mm  
**Lens:** Max aperture f/1.9  
(shot wide open)  
Auto exposure  
Program AE  
**Exposure:** 1/402 sec  
f/1.9  
ISO 40  
**Flash:** none



Metadata



# Cos'è un metadato?



Fonte: <https://dataedo.com/kb/data-glossary/what-is-metadata>

# Cos'è un metadato?

**2016 Sales** — Metadata

Month	Forecast	Sales	Variation
Jan 17	42,000	38,532	-3,468
Feb 17	45,000	41,934	-3,066
Mar 17	45,000	42,163	-2,837
Apr 17	45,000	43,050	-1,950
May 17	45,000	45,145	145
Jun 17	48,000	47,745	-255
Jul 17	48,000	49,623	1,623
Aug 17	48,000	52,539	4,539
Sep 17	45,000	47,324	2,324
Oct 17	45,000	44,700	-300
Nov 17	42,000	44,923	
Dec 17	48,000	51,120	
	546,000	548,798	

— Data

**James:**  
Forecast

Fonte: <https://dataedo.com/kb/data-glossary/what-is-metadata>

# Cos'è un metadato?

The screenshot shows a web browser window with the address bar displaying <https://dataedo.com/blog/what-is-metadata-examples>. The page content includes a blue header with the title 'What is Metadata', a byline 'Piotr Kononow', and a brief introduction: 'Metadata is simply data about data. It helps to organize, find and update metadata.' Below this, a section titled 'Typical Metadata' lists five points: 1. Title and description, 2. Tags and categories, 3. Who created and when, 4. Who last modified and when, 5. Who can access or update.

Overlaid on the page is the 'Page Info' window for the same URL. It shows the following details:

- Title: What is Metadata - 9 Examples
- Address: https://dataedo.com/blog/what-is-metadata-examples
- Type: text/html
- Render Mode: Standards compliance mode
- Text Encoding: UTF-8
- Size: 4,05 KB (4 150 bytes)
- Referring URL: https://dataedo.com/blog/posts/what-is-metadata-examples/edit
- Modified: March 23, 2017, 6:24:41 PM

The 'Meta (8 tags)' section is expanded, showing a table:

Name	Content
x-ua-compatible	ie=edge
viewport	width=device-width, initial-scale=1
description	Meaning of metadata and 9 real life examples.
og:url	https://dataedo.com/blog/what-is-metadata-examples
og:title	What is Metadata - 9 Examples
og:description	Meaning of metadata and 9 real life examples.
og:image	https://dataedo.com/asset/img/blog/banners/metadata.png

Red and blue lines with labels point to the 'Page Info' window and the 'Typical Metadata' list respectively. The red line is labeled 'Metadata' and the blue line is labeled 'Data'.

Fonte: <https://dataedo.com/kb/data-glossary/what-is-metadata>

# Cos'è un metadato?

Meta + Datum: “dopo / insieme / oltre il dato”

Dato per descrivere o rappresentare una caratteristica di un altro dato

Informazioni su un oggetto o risorsa che descrive le caratteristiche di tale oggetto, come contenuto, qualità, formato, posizione e diritti di accesso

Possono essere utilizzati per descrivere oggetti fisici (ad esempio frammenti di vaso e campioni) così come oggetti digitali (ad esempio documenti, immagini, set di dati e software)

# Caratteristiche

## Tipo di dato

→ viene creato, gestito e conservato come un qualsiasi altro tipo di dato

## Multiforme

→ può assumere molteplici forme, da testo libero (es. un file README) a contenuto standardizzato e strutturato (es. un file .xml/.json/.rdf)

## Associativo

→ sempre associato al dato che descrive

→ può farlo in due modi:

- internamente (embeddato nel dato)
- esternamente (in un file separato e collegato al dato)

# Tipologie di metadati

## Descrittivi

informazioni riguardanti il contenuto e il contesto di una risorsa: es. titolo, autore, soggetto, descrizione, data di pubblicazione, ecc.

## Strutturali

informazioni riguardanti l'organizzazione di un oggetto: es. ordine dei capitoli di un libro, ordine delle pagine di un capitolo, ecc.

## Amministrativi

informazioni riguardanti l'origine e la gestione del ciclo di vita di una risorsa: es. oggetto utilizzato per la creazione della risorsa, licenza, ecc.

# Livelli di raggruppamento



Componente

metadati assegnati alle sottoparti di oggetti:  
capitoli (di un libro), scene (di un film), ecc.

Oggetto

metadati assegnati agli oggetti di una  
collezione: libri, film, ecc.

Collezione

metadati assegnati ad un insieme di oggetti



# Perché usare i metadati?

## Caso del Mars Space Orbiter (1999)

Sonda spaziale schiantata su Marte a causa di una discrepanza nell'uso delle unità di misura tra due gruppi di scienziati: uno utilizzava le unità imperiali, l'altro quelle metriche.

Ciò ha causato un errore di navigazione e la distruzione della sonda, con una conseguente perdita di milioni di dollari



Remember the Mars Climate Orbiter incident from 1999?

Fonte:

<https://www.simscale.com/blog/nasa-mars-climate-orbiter-metric/>

# Trovare dati

La scopribilità delle informazioni si basa (anche) sulla ricerca dei metadati

I metadati aiutano i ricercatori a trovare dati che, ad esempio:

- si riferiscono a un'area geografica di interesse attraverso i metadati geospaziali
- si riferiscono a una disciplina di ricerca di interesse tramite il campo di ricerca, parole chiave o vocabolario
- sono generati da un altro ricercatore il cui lavoro è di interesse tramite i metadati del ricercatore principale o del contributore

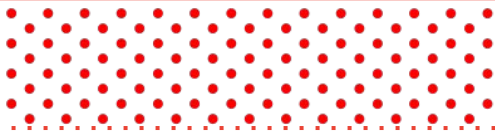
# Determinare il valore dei dati

Per valutare l'utilità, il valore e la qualità di un dataset, è necessario comprendere il contesto attorno ai dati


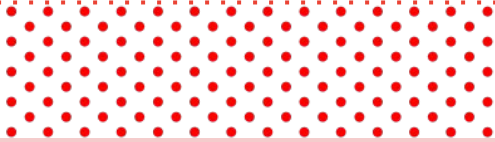
Quindi i metadati aiutano a determinare il valore del dataset perché:

- descrivono i motivi, i metodi e i risultati della raccolta dei dati
- collegano ai ricercatori e alle istituzioni coinvolte
- identificano il programma di ricerca o il finanziamento
- indicano le pubblicazioni derivanti dai dati di ricerca
- forniscono esplicitamente provenienza, licenza, diritti e informazioni tecniche

# Accedere ai dati



Per poter accedere a un dataset, sono necessari alcune informazioni fornite dai metadati

- 
- informazioni che identificano il dataset (es. PID o identificatore permanente)
  - un link di download diretto ai dati online per l'accesso aperto, oppure
  - informazioni di contatto per il gestore dei dati per l'accesso mediato
- 

# (Ri)Usare i dati

Anche per poter utilizzare effettivamente un dataset sono necessari dei metadati (es. astronomi e calibrazione di immagini con filtri)

- come sono strutturati i dati
- cosa descrivono
- come leggerli (ad esempio, intestazioni delle colonne e unità)
- informazioni metodologiche come impostazioni degli strumenti e calibrazioni, reagenti utilizzati o domande del sondaggio
- licenze
- come citare i creatori dei dati

# Perché usare i metadati?

In altri termini, un corretto utilizzo dei metadati assicura che quello che facciamo sia (o almeno cerchi di essere) **FAIR**

- **Trovabile**
- **Accessibile**
- **Interoperabile**
- **Riusabile**

## Findable

Metadata and data should be findable for both humans and computers

## Interoperable

Data needs to work with applications or workflows for analysis, storage and processing

F

A

I

R

## Accessible

Once found, users need to know how the data can be accessed

## Reusable

The goal of FAIR is to optimise data reuse via comprehensive well-described metadata

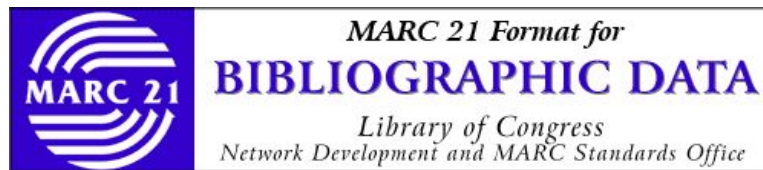
Fonte:

<https://scibite.com/solutions/enterprise-fair-data-mdm/>

# Schemi di metadati

Una struttura concettuale che specifica quali asserzioni (metadati) utilizzare e secondo quali regole

- Insieme di *elementi*
- Definizione di *elementi*
- Relazioni tra *elementi*
- Regole



Fonte: <https://www.loc.gov/marc/bibliographic/>

 Dublin Core™ Metadata Initiative

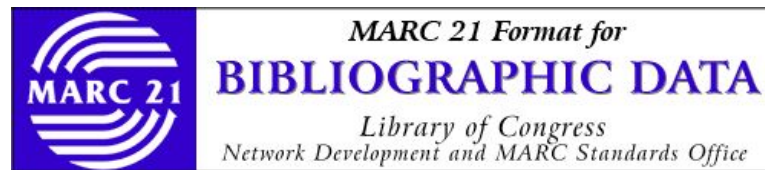
Fonte: <https://www.dublincore.org/>



# Schemi di metadati

Se formalmente validato e riconosciuto, può diventare uno *standard*:

- **MARC** (standard per la rappresentazione di informazioni bibliografiche)
- **Dublin Core** (standard per la rappresentazione di risorse sul Web)



Fonte: <https://www.loc.gov/marc/bibliographic/>

 **Dublin Core™ Metadata Initiative**


Fonte: <https://www.dublincore.org/>

# Dublin Core

Schema di metadati per descrivere risorse pubblicate sul Web

Include quindici elementi (poi estesi ulteriormente) ritenuti fondamentali

→ è un minimo comun denominatore: semplice, economico, facile da imparare e da usare




Dublin Core Elements		
Rights	Contributor	Creator
Subject	Coverage	Title
Publisher	Identifier	Description
Type	Date	Source
Relation	Format	Language


Fonte:

<https://historygonedigital.wordpress.com/2017/10/02/dublin-core-metadata-element-set/>

# Dublin Core



```
Title="Metadata Demystified"  
Creator="Brand, Amy"  
Creator="Meyers, Barbara"  
Subject="metadata"  
Description="Presents an overview of metadata conventions in publishing."  
Publisher="NISO Press"  
Publisher="The Sheridan Press"  
Date="2003-07"  
Type="Text"  
Format="application/pdf"  
Identifier="http://www.niso.org/standards/resources/Metadata_Demystified.pdf"  
Language="en"
```



Esempio di record di metadati basato su Dublin Core. Fonte: NISO Press, Understanding Metadata, <http://www.niso.org>

# Schema di codifica

Schema di codifica == insieme di regole che specificano sintassi o lessico utilizzati nelle asserzioni (metadati) di uno schema di metadati

- **Schema di codifica sintattica:** definisce come rappresentare uno specifico tipo di dati a partire dal formato (es. [ISO 8601](#) per le date)
- **Vocabolario controllato:** definisce come rappresentare uno specifico tipo di dati a partire da un insieme finito e controllato di opzioni (es. [AAT](#))
- **Authority file:** definisce come rappresentare le varianti di un valore stabilito come autoritativo (es. [VIAF](#))

# Vocabolari controllati

Lista strutturata di termini organizzati in un esplicito sistema di relazioni

Metodo consistente per la descrizione dei dati e per controllare i possibili valori applicabili ad un elemento

Esempi:

- [Art & Architecture Thesaurus \(Getty Research Institute\)](#)
- [DCMI Type Vocabulary](#)
- Altri esempi su <https://bartoc.org/>

# Abilità informatiche

A.A. 2023/2024

02a - Fine

Sebastian Barzaghi

[sebastian.barzaghi2@unibo.it](mailto:sebastian.barzaghi2@unibo.it)

<https://orcid.org/0000-0002-0799-1527>