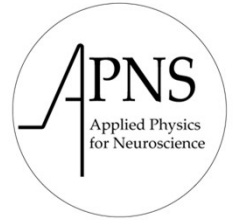




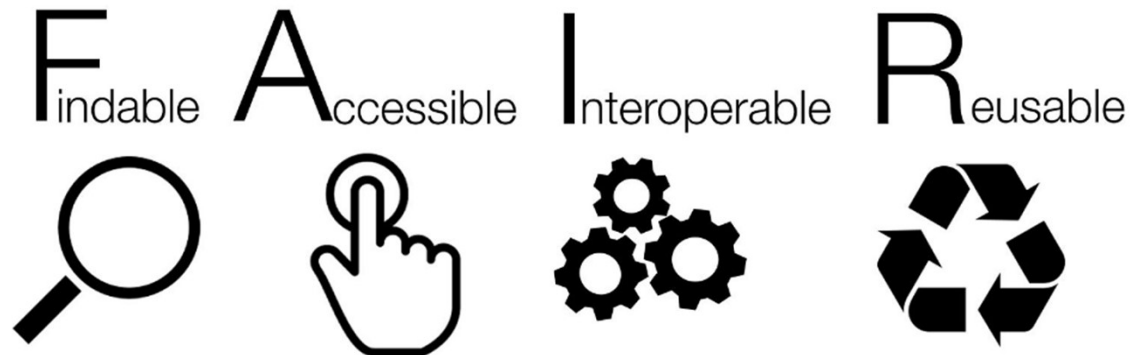
Dipartimento di Neuroscienze
dell'Università di Torino

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



FAIR data: componenti chiave e procedure



Fonte: https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

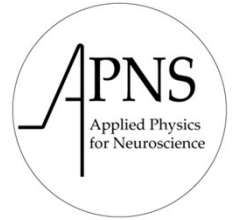
Annamaria Vernone, APNS, UNITO  ORCID 0000-0002-9748-2178
annamaria.vernone@unito.it



Dipartimento di Neuroscienze
dell'Università di Torino

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



FAIR data: componenti chiave e procedure

Cosa vuol dire Open Data?

Qual'è la differenza con FAIR Data?

FAIR data per l'uomo e/o per la macchina?

Come posso rendere i dati FAIR?

Perchè è importante l'analisi strutturale e metodologica dei dati di progetto?

Perchè sono importanti le Infrastrutture?

Chi è il Data Steward?

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



What we do

Through teaching, creating and campaigning, we are working for a fair, free and open future for all.

Open Definition

DEFINING OPEN IN OPEN DATA, OPEN CONTENT AND OPEN KNOWLEDGE

“Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness).”

(Fonte: opendefinition.org)

“Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**”

(Fonte: opendefinition.org)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Base per
l'innovazione della
conoscenza volta
all'integrazione e
al riutilizzo



EUROPEAN COMMISSION
Directorate-General for Research & Innovation

[H2020 Program Guidelines on FAIR Data](#)

Open Access diventa un
obbligo per le
pubblicazioni in Horizon
2020



Please note the distinction between open access to scientific peer-reviewed **publications** and open access to research **data**:

- **publications** – open access is an *obligation* in Horizon 2020.
- **data** – the Commission is running a flexible pilot which has been extended and is described below.

”as open as possible and as closed as necessary”

“open” in order to foster the reusability and to accelerate research, but at the same time they should be “closed” to safeguard the privacy of the subjects

(Fonte: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Gli articoli in preprint medRxiv e bioRxiv pubblicati durante la pandemia COVID-19 “Covid-19 SARS-CoV-2 preprints from medRxiv e bioRxiv” sono cresciuti esponenzialmente



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

I dati collegati alla letteratura non sono però sempre disponibili

I dati guidano le decisioni, permettono rapidità nelle scelte, ma possono fuorviare i risultati se condivisi :

- parzialmente
- con scarsa descrizione
- senza una corretta supervisione nel processo di acquisizione del dato

Anche i **risultati di esperimenti che non supportano l'ipotesi di lavoro** sono da considerarsi dati perché fonte di conoscenza

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Nel **2016** la rivista **Nature** ha pubblicato i principi guida per la creazione dei dati FAIR: “The FAIR Guiding Principles for scientific data management and stewardship” <https://www.nature.com/articles/sdata201618> DOI: 10.1038/sdata.2016.18

(Fonti: <https://www.nature.com/articles/sdata201618>, <https://www.go-fair.org/fair-principles/>)

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

[Comment: The FAIR Guiding Principles for scientific data management and stewardship \(nature.com\)](https://www.nature.com/articles/sdata201618)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Findable

I **Dati** ed i **Metadati** devono essere facilmente **reperibili** “**globally unique and persistent identifier**” sia per l’uomo che per la macchina. I **Metadati** dovrebbero descrivere in modo univoco e facilmente comprensibile i Dati e contenere l’identificativo dei dati che descrivono. Dovrebbero essere indicizzati in una risorsa facilmente rintracciabile attraverso la rete

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

(Fonti: <https://www.nature.com/articles/sdata201618>,
<https://www.go-fair.org/fair-principles/>)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Accessible



I dati, dopo averli rintracciati, devono essere **accessibili** eventualmente anche tramite autenticazione e autorizzazione. I Metadati devono continuare ad essere disponibili anche nel caso in cui i dati non siano più disponibili (A2)

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

(Fonti: <https://www.nature.com/articles/sdata201618>,
<https://www.go-fair.org/fair-principles/>)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Interoperable

I dati devono essere **integrati** con i dati di altri progetti, applicazioni o flussi di lavoro sia per quanto riguarda l'analisi che l'archiviazione e l'elaborazione. Per questo si dovrebbero usare formati condivisi, vocabolari comuni, e riferimenti ad altri dati/metadati:

11. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
12. (Meta)data use vocabularies that follow FAIR principles
13. (Meta)data include qualified references to other (meta)data

((Fonti: <https://www.nature.com/articles/sdata201618>,
<https://www.go-fair.org/fair-principles/>)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Reusable

I **metadati** e i **dati** devono essere **ben descritti** in modo da poter essere replicati e/o combinati in contesti diversi. **I Metadati inoltre dovrebbero far riferimento agli Standard di Dominio della Comunità Scientifica:**

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

((Fonti: <https://www.nature.com/articles/sdata201618>,
<https://www.go-fair.org/fair-principles/>)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Nei Principi FAIR abbiamo parlato di Dati e Metadati. Vediamo cosa sono

Dato L'oggetto di descrizione che dipende dalla realtà di interesse, dal dominio di applicazione, dalla rilevanza che voglio dare ai concetti oggetti di studio

Metadato "Data about data" è una descrizione e un contestualizzazione dei dati. E' utile per organizzare, cercare ma soprattutto comprendere i dati, in particolare se utilizzato da un algoritmo

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Metadato “Data about data”



Dati

Filename: akita.jpg

Proprietà immagine



Attributi file

Ultimo salvataggio: 30/05/2022 16:32

Dimensioni su disco: 394,1 KB

Risoluzione: 120 dpi

Unità

- Pollici
- Centimetri
- Pixel

Colori

- Bianco e nero
- Colore

Larghezza

Altezza:

Metadati

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Metadato “Data about data”



Dati

info = struct with fields:

```
Filename: 'B:\matlab\toolbox\images\imdata\CT-MONO2-16-ankle.dcm'  
FileModDate: '18-Dec-2000 12:06:43'  
FileSize: 525436  
Format: 'DICOM'  
FormatVersion: 3  
Width: 512  
Height: 512  
BitDepth: 16  
ColorType: 'grayscale'  
FileMetaInformationGroupLength: 192  
FileMetaInformationVersion: [2x1 uint8]  
MediaStorageSOPClassUID: '1.2.840.10008.5.1.4.1.1.7'  
MediaStorageSOPInstanceUID: '1.2.840.113619.2.1.2411.1031152382.365.1.736169244'  
TransferSyntaxUID: '1.2.840.10008.1.2'  
ImplementationClassUID: '1.2.840.113619.6.5'  
ImplementationVersionName: '1_2_5'  
SourceApplicationEntityTitle: 'CTN_STORAGE'  
IdentifyingGroupLength: 414  
ImageType: 'DERIVED\SECONDARY\3D'  
SOPClassUID: '1.2.840.10008.5.1.4.1.1.7'  
SOPInstanceUID: '1.2.840.113619.2.1.2411.1031152382.365.1.736169244'  
StudyDate: '1993 04 30'
```

Metadati

<https://www.mathworks.com/help/images/ref/dicomread.html>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

OPEN diverso da FAIR

I dati medici spesso **possono essere FAIR (accessibili) ma non Open**, proprio perché è difficile anonimizzarli completamente. Questo non impedisce che siano riutilizzabili in modo protetto

I dati possono essere resi FAIR con metadati, ontologie ecc ecc e poi depositati **CHIUSI** su Archivi come Zenodo o altrove

I dati Open non sono necessariamente FAIR (dati disponibili pubblicamente ma non FAIR)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Il Ciclo di vita del Dato

Il **dato grezzo**, iniziale, è il primo da considerare nel processo di FAIRificazione per la riusabilità e la ripetibilità dell'esperimento

MA

anche i **dati intermedi** e quelli **finali**, frutto di elaborazione, andrebbero annotati ai fini della riusabilità, ripetibilità degli esperimenti, e archiviazione nel lungo periodo

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Data Management Plan (DMP)

Uno dei primi documenti ad essere prodotto già nella **fase di proposta del Progetto** è il DMP, un documento strutturato e “vivo” in quanto dovrebbe essere costantemente aggiornato sulla base delle nuove necessità

Il DMP descrive il ciclo di vita della gestione dei dati, la metodologia, gli standard che verranno adottati, le politiche di condivisione dei dati, i metodi di conservazione

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Fonte: <https://creativecommons.org/>

Licenze di diritto d'autore che si possono applicare a qualsiasi opera da esso tutelata
(https://it.wikipedia.org/wiki/Licenze_Creative_Commons)

Attraverso le Licenze CC (Creative Common), l'autore, può concedere ad altri il diritto di usare o modificare un'opera che lui stesso (l'autore) ha creato.

CC permette all'autore di scegliere le modalità di utilizzo ma anche di proteggere le persone che usano o diffondono un'opera di altri dalla preoccupazione di infrangere il diritto d'autore, purché siano rispettate le condizioni specificate dall'autore stesso nella licenza

Fonte: <https://www.oa.unito.it/new/come-scrivere-un-data-management-plan/>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Fonte: <https://creativecommons.org/>

Licenze di diritto d'autore che si possono applicare a qualsiasi opera da esso tutelata
(https://it.wikipedia.org/wiki/Licenze_Creative_Commons)

Per scegliere la Licenza più adatta utilizzare il sito Web di Creative Commons all'indirizzo:

<https://creativecommons.org/share-your-work/>

Per un elenco dei tipi di Licenze disponibili:

<https://creativecommons.org/about/clicenses/>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Licenze per i dati di ricerca e come applicarle Guida di Openaire

Fonte: <https://creativecommons.org/>

- If your research data qualifies as a work (literary work such as a journal article or a software), then CC BY 4.0 is usually the best choice. The use of the Share Alike (SA) is also compatible with the Open Access definition and reinforced in Plan S licensing guidance for publications. Non-commercial should be avoided as it is not Open Access compliant. Non-derivative is a tricky issue and should be avoided, especially if you do not know what you are doing. That said, it may not be incompatible with the Open Access definition.
- If your research data is a database or a dataset (unstructured data that do not meet the database definition) usually the best option is a CC0, which waives all your rights in the database.

Fonte: <https://www.openaire.eu/how-do-i-license-my-research-data>

Al seguente indirizzo una sintesi dei requisiti di Horizon Europe per le pubblicazioni e per i dati di ricerca:

<https://www.openaire.eu/horizon-europe-open-science-requirements-in-practice>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

La ricerca è fatta di pubblicazioni, software, annotazioni di laboratorio, progetti, metodologie, procedure ed i **dati li completano**

I dati FAIR condivisi rendono l'**esperimento riproducibile**

I dati FAIR condivisi **velocizzano i processi decisionali**

I dati FAIR condivisi rendono gli **archivi interoperabili** creando **nuova informazione** per la ricerca

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Esempio d'uso di dati da banche dati FAIR per produrre nuova conoscenza

La piattaforma web GPLAB, Gene and Protein Virtual Lab (gplab.diff.org)

Gianpiero Pescarmona, Formerly Department of Oncology, University of Torino, **Francesca Silvagno**, Department of Oncology, University of Torino, **Annamaria Vernone**, Department of Neurosciences "Rita Levi Montalcini", University of Torino

Permette di scaricare, a partire dai databases liberamente accessibili <https://www.uniprot.org/> e <https://www.ensembl.org/index.html> le tabelle utili per l'analisi biochimica e di creare in automatico per ogni cromosoma una nuova tabella canonica con nuovi dati

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

L'applicazione web accede ai database UniProt ed Ensembl in modo programmatico, utilizzando interfacce python per ottenere dati sempre aggiornati

Grazie a questa procedura automatizzata, il contenuto aminoacidico inserito nella tabella dei dati canonici viene sempre calcolato a partire dagli ultimi dati disponibili online dai due database.

Questo è un esempio di costruzione di nuovi Dati a partire da Dati FAIR

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

La piattaforma web GPLAB, Gene and Protein Virtual Lab - gplab.diff.org

Il software è collocato nella sezione “Chromosomes” di GPLAB e permette, selezionando il cromosoma di interesse, di fornire elenchi di proteine ordinate in base alla posizione del loro gene codificante, fornendo la sequenza di AA e la lunghezza della proteina

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Human Chromosomes

Gene/Protein Tables from Uniprot and

Chromosome number: 12

	Gene stable ID	Gene name	Gene start (bp)	Gene end (bp)	id length	entry name	A	C	D	E	
932	ENSG00000120645	IQSEC3	66767	178455	Q9UPP2	1182	IQEC3_HUMAN	127	27	46	93
639	ENSG00000111181	SLC6A12	190077	214570	P48065	614	S6A12_HUMAN	32	23	19	27
534	ENSG00000010379	SLC6A13	220621	262873	Q9NSD5	602	S6A13_HUMAN	40	23	14	28
734	ENSG00000073614	KDM5A	280057	389320	P29375	1690	KDM5A_HUMAN	110	50	87	170
84	ENSG00000120647	CCDC77	389273	442642	Q9BR77	488	CCD77_HUMAN	23	11	20	63
330	ENSG00000139044	B4GALNT3	459939	563509	Q6L9W6	998	B4GN3_HUMAN	58	8	54	74
333	ENSG00000171840	NINJ2	564296	663779	Q9NZG7	142	NINJ2_HUMAN	14	0	3	6

Questo è un esempio di Tabella con nuovi dati prodotta per il cromosoma 12 che contiene le informazioni relative a: geni, posizione genetica, proteine, conteggio degli aminoacidi per ogni proteina, sequenza,

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

EMBL's European Bioinformatics Institute

EMBL-EBI EMBL's European Bioinformatics Institute



Uno studio indipendente di Charles Beagrie Ltd che stima l'impatto e il valore delle risorse di dati gestite dall'EMBL-EBI



Uniprot Protein Data Bank (PDB)

Il consorzio UniProt e le istituzioni ospitanti, tra cui anche EMBL-EBI, sono impegnati nella conservazione a lungo termine delle banche dati UniProt

I databases ed il software di EMBL e di Uniprot sono liberamente accessibili per la comunità scientifica

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Esempio di scheda Uniprot con descrizione di tutte le caratteristiche relative alla proteina P53_HUMAN (<https://www.uniprot.org/uniprot/P04637>)

ID P53_HUMAN Reviewed; 393 AA.

AC P04637; Q15086; Q15087; Q15088; Q16535; Q16807; Q16808; Q16809; Q16810;

AC Q16811; Q16848; Q2XN98; Q3LRW1; Q3LRW2; Q3LRW3; Q3LRW4; Q3LRW5; Q86UG1;

AC Q8J016; Q99659; Q9BTM4; Q9HAQ8; Q9NP68; Q9NPJ2; Q9NZD0; Q9UBI2; Q9UQ61;

DT 13-AUG-1987, integrated into UniProtKB/Swiss-Prot.

Uniprot mette a disposizione un ampio repository relativo alle Proteine, le cui informazioni sono scaricabili in diversi formati open (text, FASTA, XML, GFF, RDF/XML) ed elaborabili con opportuni programmi

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Ogni scheda riporta sempre:

AC (Accession): identificativo primario stabile della proteina, es. P04637

ID (identificativo) della Proteina es. P53_HUMAN, 'Entry name': un identificatore univoco della proteina che può subire modifiche ma fa sempre riferimento ad AC. Può cambiare ad es. perchè la proteina viene promossa da TrEMBL (record annotati ma non completamente approvato) a Swiss-Prot section (record revisionati e approvati)

Fonte: [https://www.uniprot.org/help/difference_accession_entryname#:~:text=An%20accession%20number%20\(AC\)%20is,all%20relevant%20entries%20are%20kept](https://www.uniprot.org/help/difference_accession_entryname#:~:text=An%20accession%20number%20(AC)%20is,all%20relevant%20entries%20are%20kept)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Fonte: <http://www.ensembl.org>

Il progetto Ensembl aggrega, elabora, integra e ridistribuisce insiemi di dati genomici fin dai primi rilasci della bozza del genoma umano, con l'obiettivo di accelerare la ricerca genomica attraverso una rapida distribuzione aperta di dati pubblici

Grandi quantità di dati grezzi vengono così trasformate in conoscenza, resa disponibile attraverso una moltitudine di canali (<http://www.ensembl.org>)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Chromosome Walking: A Novel Approach to Analyse Amino Acid Content of Human Proteins Ordered by Gene Position

by  Annamaria Vernone  ,  Chiara Ricca ,  Gianpiero Pescarmona  and  Francesca Silvagno *  

La composizione aminoacidica delle proteine, ordinate in base alla posizione del corrispondente gene codificante, ha permesso di proporre un nuovo approccio in grado di individuare le regioni di un cromosoma che producono proteine simili per contenuto aminoacidico

L'analisi detta "chromosome walking" può identificare cluster di geni la cui traduzione dipende dalla disponibilità di AA, che potrebbero essere rilevanti in molte patologie umane

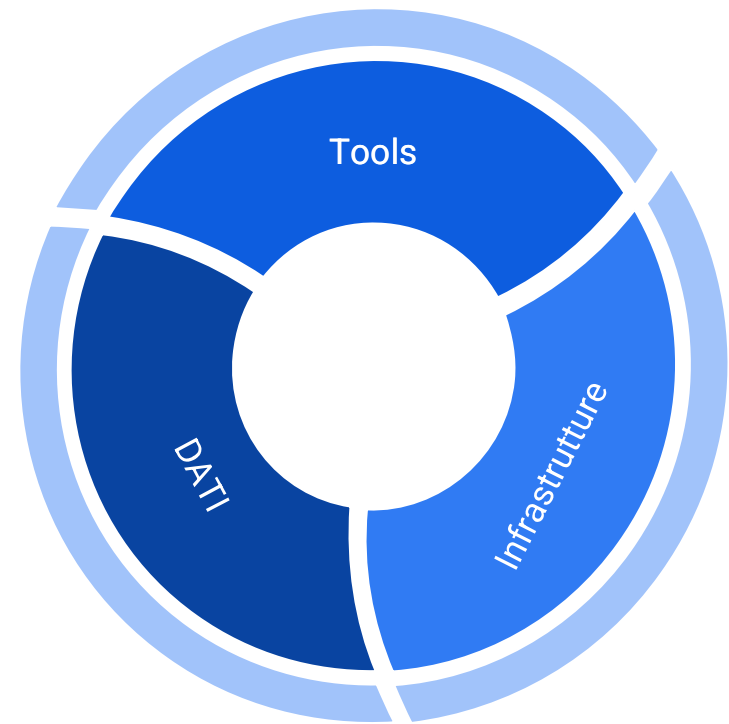
Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Il concetto di FAIR data è multidimensionale
e comprende

Dati, Tools, Infrastrutture

come elementi tra loro interagenti



Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Qualsiasi oggetto digitale di ricerca può beneficiare dei principi FAIR
tools, algoritmi, video, immagini, pipelines,

dal momento che

tutti i componenti del processo di ricerca dovrebbero essere disponibili
per assicurare la trasparenza, la riproducibilità e la riusabilità

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Il **percorso per la costruzione di dati FAIR può essere affrontato in modo graduale, avvicinandosi sempre di più alla costruzione del dato “Machine readable”** cioè accessibile, leggibile, interpretabile, integrabile, attraverso l’uso del software e tale da poter essere riutilizzato, per la corretta ripetibilità di esperimenti e prove di laboratorio, ma anche per **creare in automatico nuova informazione** in modo rapido ed attendibile

Open Science e FAIR Data per le Neuroscience

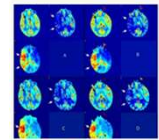
Dipartimento di Neuroscienze, Università di Torino



Digital Object Identifier

10.5281/zenodo.5109415











UNITO BRAIN



IEEE DataPort

Dati e Metadati possono far parte di un Dataset. Un dataset è una collezione di dati che può essere o meno strutturata in base al Dominio di conoscenza

UniToBrain Dataset

 Umberto Gava;  Federico D'Agata;  Edwin Bennink;  Enzo Tartaglione;  Daniele Perlo;  Annamaria Vernone;  Francesca Bertolino;  Eleonora Ficiarà;  Alessandro Cicerale;  Fabrizio Pizzagalli;  Caterina Guiot;  Marco Grangetto;  Mauro Bergui

The University of Turin (UniTO) released the open-access dataset UniTOBrain collected for the homonymous Use Case 3 in the DeepHealth project (<https://deephealth-project.eu/>). UniToBrain is a dataset of Computed Tomography (CT) perfusion images (CTP). The dataset includes 100 training subjects and 15 testing subjects used in a submitted publication for the training and the testing of a Convolutional Neural Network (CNN, see for details: <https://arxiv.org/abs/2101.05992>, <https://paperswithcode.com/paper/neural-network-derived-perfusion-maps-a-model>, <https://www.medrxiv.org/content/10.1101/2021.01.13.21249757v1>). The UniTO team released this dataset publicly. This is a subsample of a greater dataset of 258 subjects that will be soon available for download at <https://ieee-dataport.org/>.

Open Science e FAIR Data per le Neuroscience

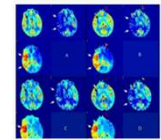
Dipartimento di Neuroscienze, Università di Torino



Digital Object Identifier

10.5281/zenodo.5109415

@UNITOBRAIN



IEEEDataPort

375

views

16,710

downloads

[See more details...](#)

	All versions	This version
Views	375	169
Downloads	16,710	16,519
Data volume	832.7 TB	832.1 TB
Unique views	282	128
Unique downloads	1,274	1,167

Questi sono i dati di accesso al Dataset in numero di visualizzazioni e di downloads dei files

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

UniToBrain Dataset Dataset Open Access

<https://zenodo.org/record/5109415>

UniToBrain è un dataset di immagini di perfusione di Computed Tomography (CT) perfusion images (CTP)

Su Zenodo sono stati pubblicati i dati dell'articolo <https://arxiv.org/abs/2101.05992>

Neural Network-derived perfusion maps: a Model-free approach to computed tomography perfusion in patients with acute ischemic stroke Umberto A. Gava, Federico D'Agata, Enzo Tartaglione, Marco Grangetto, Francesca Bertolino, Ambra Santonocito, Edwin Bennink, Mauro Bergui <https://arxiv.org/abs/2101.05992>, <https://doi.org/10.48550/arXiv.2101.05992>

Il dataset include 100 soggetti di "training" e 15 soggetti di "test" usati per sottomettere la pubblicazione relativa al training ed al test di una Rete Neurale Convolutionale per un totale di circa 51 GB

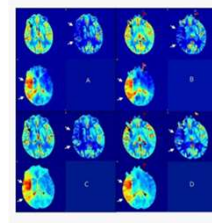
Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Il dataset pubblicato su Zenodo è un sottoinsieme del dataset più grande di 258 soggetti disponibile su <https://iee-dataport.org/> per un totale di circa 80 GB

<https://iee-dataport.org/open-access/unitobrain>

 UNITOBRAIN



DATASET FILES

 UniTOBrain_DeepHealth_IEEE.7z (80.04 GB)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

I programmi con le relative istruzioni sono stati depositati su Github

 EIDOSlab / UC3-UNITOBrain Public

<https://github.com/EIDOSlab/UC3-UNITOBrain>

<https://github.com/EIDOSlab/UC3-UNITOBrain/blob/main/README.md> README.md

<https://github.com/EIDOSlab/UC3-UNITOBrain/tree/main/src>

Es.: [UC3-UNITOBrain](#) / [src](#) / [dataloader.py](#) / <> Jump to ▾

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

The logo for CEDAR, consisting of the word "CEDAR" in a bold, black, serif font, centered within a solid blue rectangular background.

Fonte: <https://metadatacenter.org/>

L'analisi della struttura dei dati ha permesso di identificarne i raggruppamenti, le relazioni e le tipologie. E' stato creato un Dizionario Dati. Sono stati individuati i metadati ed individuato il Vocabolario comune. Le cartelle ed i file caricati sia su Zenodo che su IEEEDataPort seguono uno standard comune di Nomenclatura

Attraverso l'uso della piattaforma Web CEDAR, i dati sono stati quindi trasformati in formato "machine readable"

Fonte: <https://openview.metadatacenter.org/templates/https:%2F%2Frepo.metadatacenter.org%2Ftemplates%2Fe30d8369-6c31-45fa-a10a-2122283a28f2>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

La piattaforma CEDAR è un valido supporto per la creazione di metadati FAIR. Permette di creare metadati strutturati sulla base del modello semantico dei dati. Insieme al vocabolario controllato ed alle Ontologie, CEDAR permette di avvicinarsi sempre di più ai dati Machine Readable

“The CEDAR platform provides an easy-to-use solution for creating and reusing FAIR metadata. CEDAR’s metadata modeling, flexible but rigorous semantics, and ability to quickly produce structured metadata makes it perfect for ongoing Metadata 4 Machines workshops and emerging FAIR training courses

Erik Schultes International Science Coordinator, GO FAIR International Support and Coordination Office (GFISCO) – GO FAIR

Fonte: <https://metadatascenter.org/>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

“Trustworthy Digital Repository” (TDR)

UniToBrain Dataset

Publication date:
June 3, 2021

DOI:
DOI 10.5281/zenodo.5109415



IEEE DataPort™

DOI: 10.21227/x8ea-vh16

Persistent and unique identifier
(PID) come il DOI assegnato al
Dataset

I Metadati devono sempre essere accessibili pubblicamente (anche per i dati con accesso ristretto, es. dati sensibili) e le condizioni per il riuso dei dati es. le Licenze, devono essere definite chiaramente

Il Repository dovrebbe avere un piano di conservazione dei dati a lungo termine

Open Science e FAIR Data per le Neuroscience

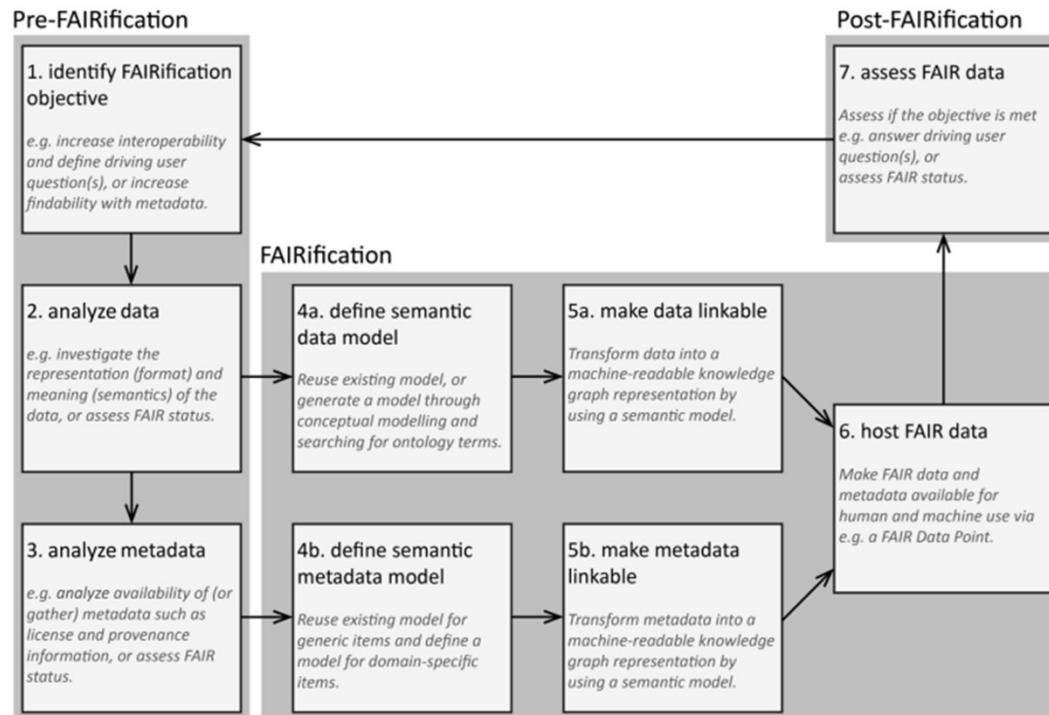
Dipartimento di Neuroscienze, Università di Torino

E' importante decidere il livello di dettaglio ed il raggruppamento in categorie nel deposito del dato e del relativo metadato

La fase preliminare di analisi semantica del Dato e del Metadato è molto importante per poter procedere correttamente a costruire lo schema informatico dei dati ai fini della costruzione dei dati Machine readable

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Questo è un generico flusso di lavoro per la FAIRificazione dei dati volto alla creazione di dati machine readable

Tra i passi troviamo l'**analisi dei dati e dei metadati** intesa a creare dei **modelli semantico dei dati e dei metadati**

Il **collegamento tra modelli** sia interni che esterni all'organizzazione in modo renderli interoperabili
La trasformazione del modello semantico in uno **schema interpretabile dalla macchina**, machine readable

Il deposito dello schema dei dati in formato machine readable su un **Fair Data Point**

Fonte: "A Generic Workflow for the Data FAIRification Process" Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons., Erik Schultes, Marco Roos & Mark Thompson

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Fondamentale è il passo di **analisi e stesura del modello semantico dei dati** che rappresenta la realtà oggetto di studio

Le informazioni a disposizione possono essere:

tabelle, video, file musicali, dati genetici a seconda del dominio di applicazione ma anche dati cartacei, pensiamo ai manoscritti del '700

è possibile renderli completamente FAIR e machine readable attraverso un percorso di analisi per l'estrapolazione dei dati, dei metadati, del modello semantico dei dati, fino alla scrittura dei dati in formato “machine readable” e “linkable”

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

DISPROT database of intrinsically disordered proteins

“DisProt annotations cover both structural and functional aspects of disorder detected by specific experimental methods. Annotation concepts are encoded in the IDPontology. Read more about DisProt” (Fonte: <https://www.disprot.org/>)

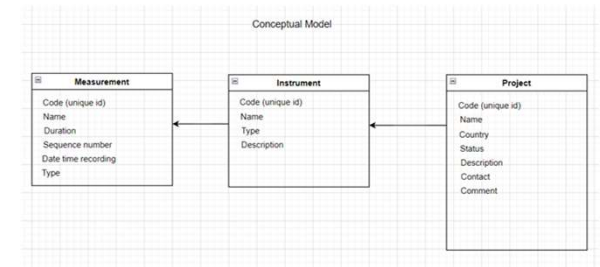


La base di conoscenza biologica è stata estratta da **dati non strutturati**, articoli ed altre fonti di informazione come i dati sperimentali e strutturata in forma digitale machine readable e linkable secondo i principi FAIR ([BIOCURATION](https://www.biocuration.org/)) (Fonte: <https://www.biocuration.org/>)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Realtà -----> Modello Dati

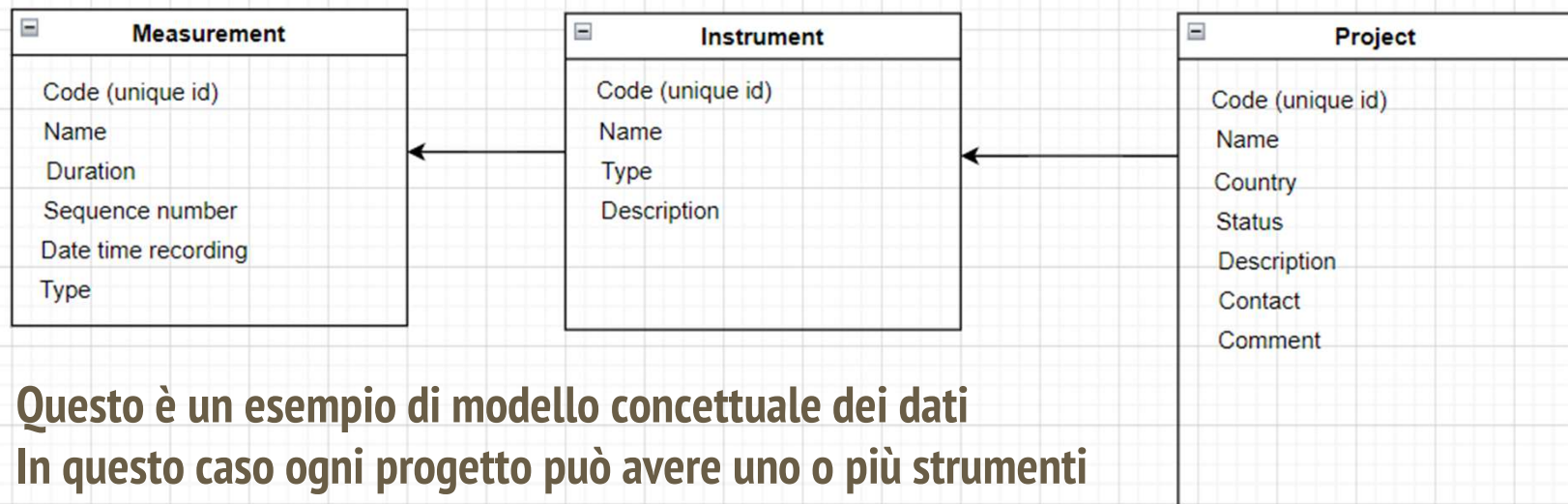


C'è molta teoria e tipi diversi di modelli di dati → Occorre scegliere il più adatto

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Conceptual Model



Questo è un esempio di modello concettuale dei dati

In questo caso ogni progetto può avere uno o più strumenti a disposizione ed ogni strumento potrà compiere una o più misure. Occorre raggruppare correttamente le informazioni, identificarle in modo univoco e correlarle

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Dal modello semantico al formato RDF per i dati riutilizzabili

“I dati della ricerca vengono prodotti in una serie di formati diversi, come fogli di calcolo, file separati da delimitatori (CSV, TSV) o database (relazionali). Per migliorare la riutilizzabilità dei dati e renderli più FAIR, questi set di dati possono essere pubblicati nello standard de facto per i dati riutilizzabili, RDF”

<https://faircookbook.elixir-europe.org/content/recipes/interoperability/rdf-conversion.html>

La conversione dei dati non RDF in RDF è un processo noto come "triplificazione"

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Ontologie

I modelli semantici contengono spesso riferimenti a Ontologie già esistenti interne/esterne all'organizzazione che vengono annotati come relazioni nel modello dati

Un'ontologia è una rappresentazione formale di un insieme di conoscenze all'interno di un determinato dominio

(Fonte: <http://geneontology.org/docs/ontology-documentation/>)

Le ontologie svolgono un ruolo importante per supportare l'interoperabilità semantica
“FAIR principles and Semantics on the Web: where is the meeting point?”

(Fonte: <https://joinup.ec.europa.eu/collection/oeg-upm/news/fair-ontologies>)

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Gene Ontology overview
<http://geneontology.org/docs/ontology-documentation/>



<https://www.uniprot.org/core/RDF Schema Ontology>



<https://proconsortium.org/pro.shtml>



<https://www.ebi.ac.uk/ols/index>



<https://hpo.jax.org/app/> The Human Phenotype Ontology

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



Metadata for Machines
Three-point FAIRification icons

Le tre icone rappresentano tre strumenti necessari per la FAIRificazione dei dati



Workshop Metadata for Machines (M4M)



What is your FIP?

Check out the FIP mini-questionnaire which will lead you through the creation of your own FAIR Implementation Profile: <https://bit.ly/yourFIP> or download the questionnaire in PDF.

FAIR principle	Question	FAIR enabling resource types	Your answers
F1	What globally unique, persistent, resolvable identifiers do you use for metadata records?	Identifier type	e.g. PURL, DOI
F2	What globally unique, persistent, resolvable identifiers do you use for datasets?	Identifier type	
F3	Which metadata schemas do you use for findability?	Metadata schema	
F4	What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata Data linking mechanism	
F4	In which search engines are your metadata records indexed?	Search engines	
F4	In which search engines are your datasets indexed?	Search engines	
A1.1	Which standardized communication protocol do you use for metadata records?	Communication protocol	
A1.1	Which standardized communication protocol do you use for datasets?	Communication protocol	
A1.2	Which authentication & authorization technique do you use for metadata records?	Authentication & authorization technique	
A1.2	Which authentication & authorization technique do you use for datasets?	Authentication & authorization technique	



FAIR Implementation Profile (FIP)



FAIR Data Point (FDP)

Fonte: <https://www.go-fair.org/how-to-go-fair/metadata-for-machines/>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino



FAIR Data Point Africa

22 luglio 2020

Università Internazionale di Kampala



COVID-19 Computer-Readable FAIR Data point of Observational Data
Virus Outbreak Data Network (VODAN-Africa)

<https://www.go-fair.org/2020/07/22/first-fair-data-point-for-covid-19-data-installed-in-africa/>

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Di cosa si occupa il Data Steward?

Il Data Steward si occupa di tutte le tematiche sopra esposte

Collabora con i gruppi di ricerca nell'ambito del processo per rendere i dati FAIR, e può essere specializzato in attività particolari all'interno del flusso volto alla creazione del dato FAIR machine readable

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Come si inserisce nella Comunità Scientifica il Data Steward?

Data Steward per gruppi di ricerca

Data Steward di tipo generale, per tutta l'organizzazione

Data Steward specializzato su un particolare Dominio (genetica, scienze umane, economia, neuroscienze, ...)

Data Steward specializzato sull'Infrastruttura

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Le slide saranno disponibili su

Zenodo

e

sul sito web apns.unito.it

annamaria.vernone@unito.it

Open Science e FAIR Data per le Neuroscience

Dipartimento di Neuroscienze, Università di Torino

Grazie