UNIVERSITÀ DI TRENTO

CIMeC Center for Mind/Brain Sciences

Community over Commercialization

Trento, 27 Ottobre 2023

THINK OPEN @CIMeC

RN IT Italian Reproducibility Network

Preparare i dati per un riuso efficace

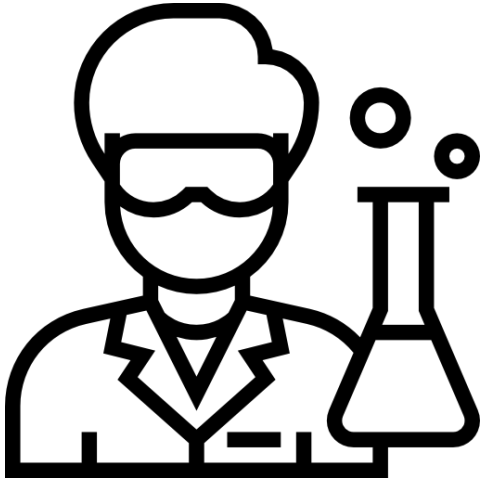
appunti sulla condivisione responsabile di informazioni

Vittorio Iacovella
CIMeC - Center for Mind / Brain Sciences
The University of Trento
@V_Iacovella @v_iacovella@qoto.org @viacovella.bsky.social
vittorio.iacovella@unitn.it

Perché costruire collezioni di dati?

Perché costruire collezioni di dati?

per sviluppare
credibili e rigorosi contributi
al progresso del sapere

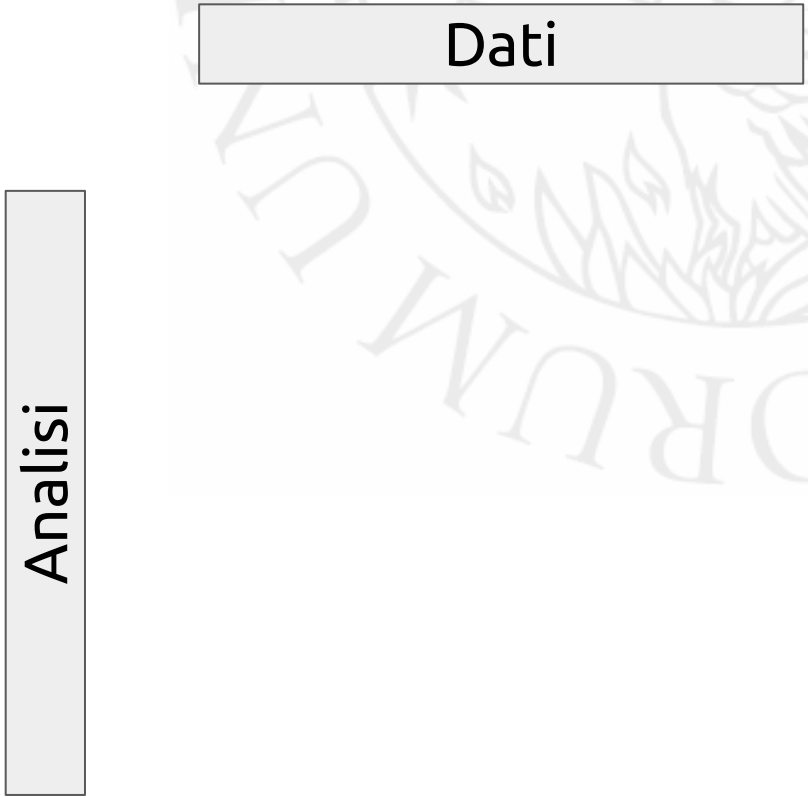


Perché costruire collezioni di dati?

Un contributo mette insieme

un approccio scientifico, rappresentabile come “Analisi”

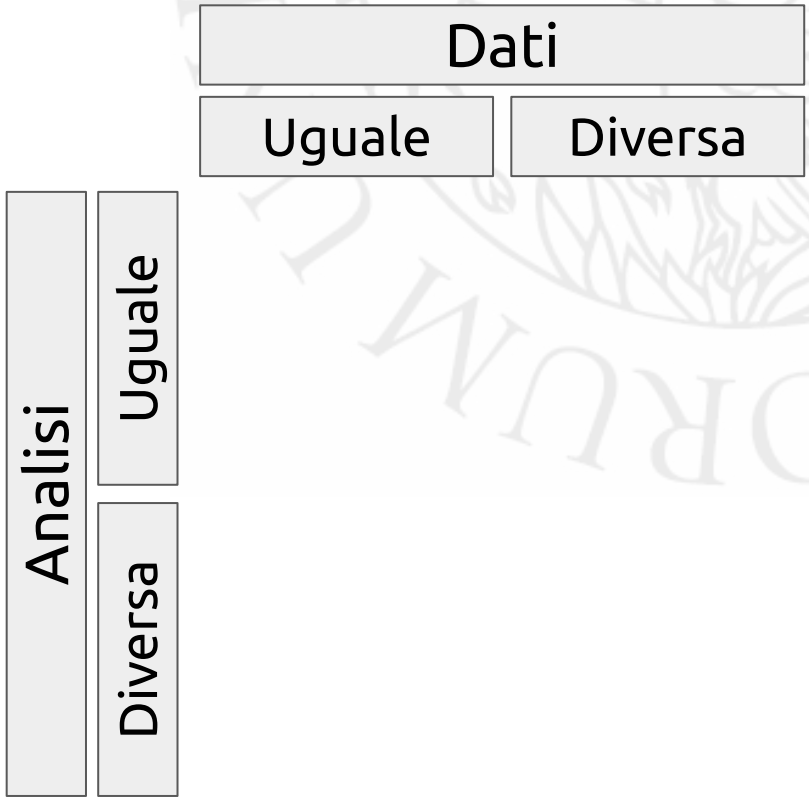
l’informazione acquisita, rappresentabile come “Dati”



Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation. https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

Perché costruire collezioni di dati?

Un contributo si può sempre posizionare in relazione a quello che è stato fatto in precedenza, per Analisi e Dati, che possono essere uguali o diverse.

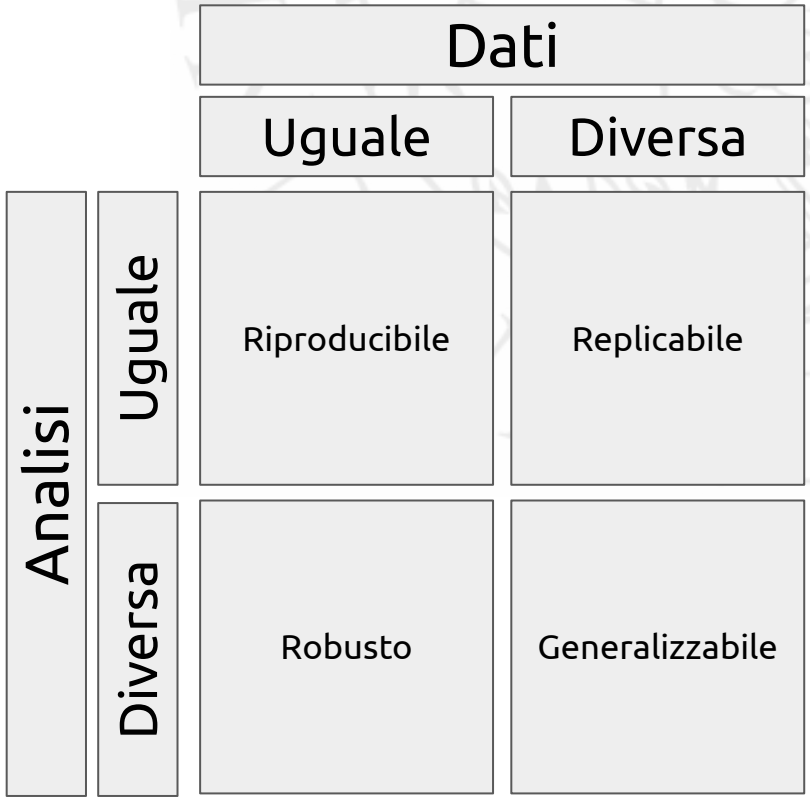


Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation.
https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

Perché costruire collezioni di dati?

Questo divide lo spazio dei contributi in quattro partizioni.

La ricerca di contributi generalizzabili è di gran lunga la più difficile da sviluppare, ma è la più comune finora.

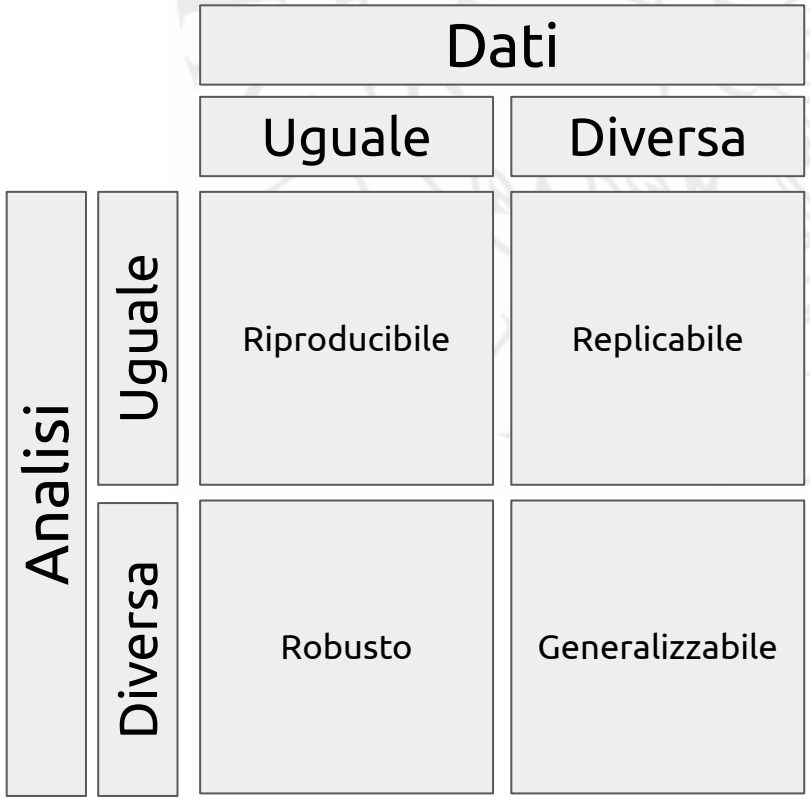


Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation.
https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

Perché costruire collezioni di dati?

La regione meno battuta invece è quella superiore.

Applicare uno stesso metodo di analisi a un altro dataset è una definizione basilare di Replicabilità.



Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation.
https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

Perché costruire collezioni di dati?

Nel 2015 c'è stato un tentativo coordinato di replicare 100 esperimenti pubblicati su 3 riviste del 2008.

Più di ⅔ delle repliche non sono riuscite a replicare l'effetto originale.

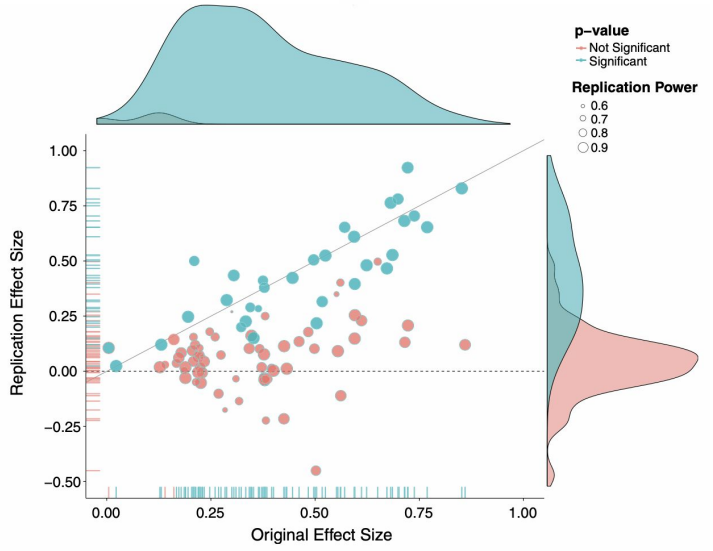
RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

replicability

Estimating the ~~reproducibility~~ of psychological science

Open Science Collaboration*



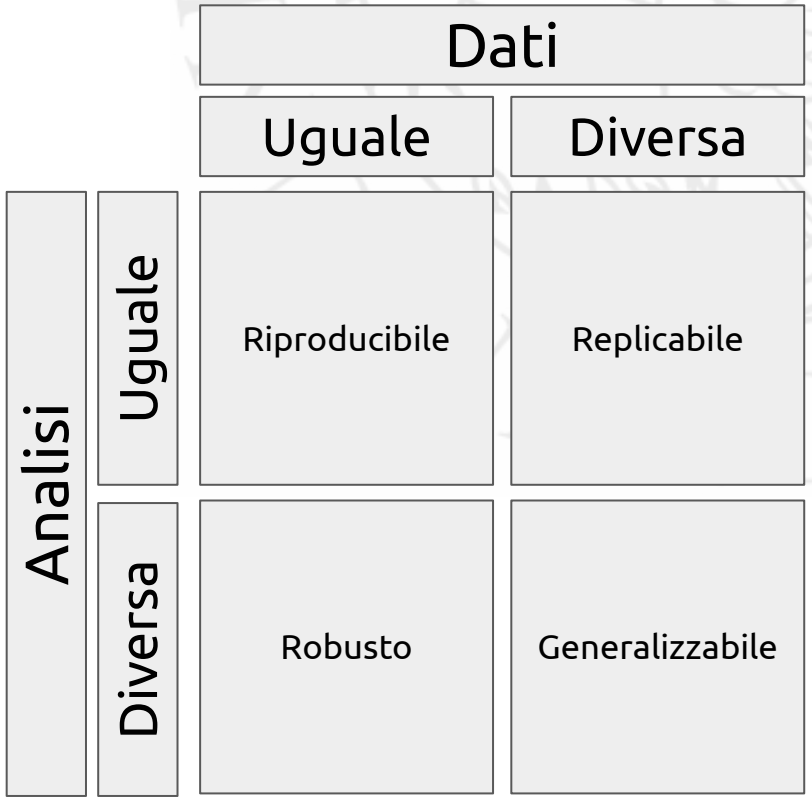
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), 1–8.

Perché costruire collezioni di dati?

Ecco perché uno dei pilastri del movimento open science è quello di favorire il riutilizzo dello stesso codice sugli stessi dati, per garantire la riproducibilità.

A manifesto for reproducible science

Marcus R. Munafò[1,2*], Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]



Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation. https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

Perché costruire collezioni di dati?

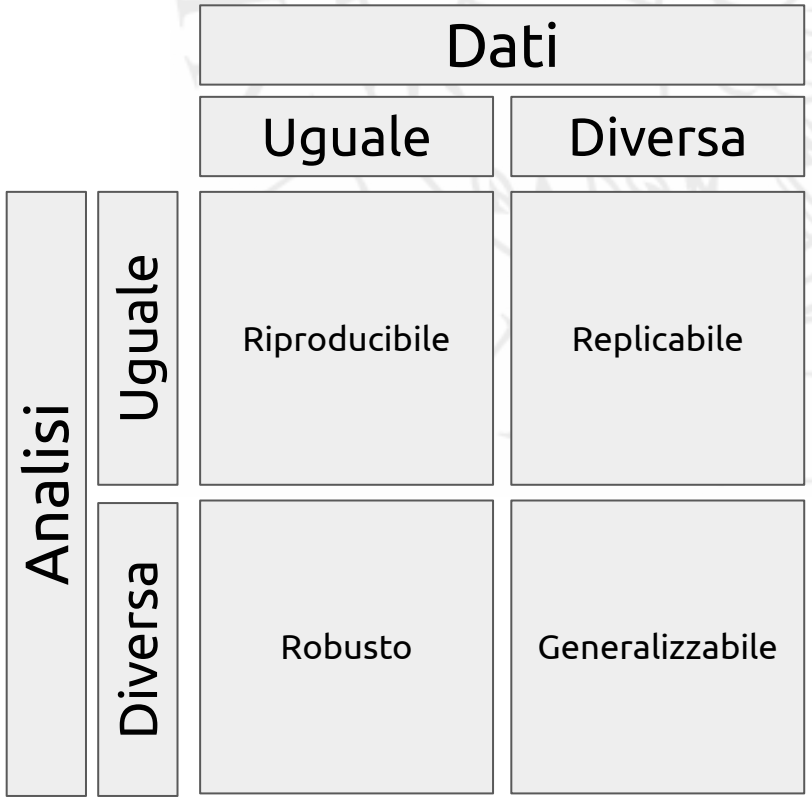
Contributi riproducibili richiedono la condivisione di collezioni, che vanno costruite in maniera rigorosa.

Il compito della condivisione porta con sé innumerevoli altre questioni, spesso (finora) sottovalutate.

Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation. https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

A manifesto for reproducible science

Marcus R. Munafò[1,2*], Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]



Chi potrebbe riutilizzare la tua collezione?

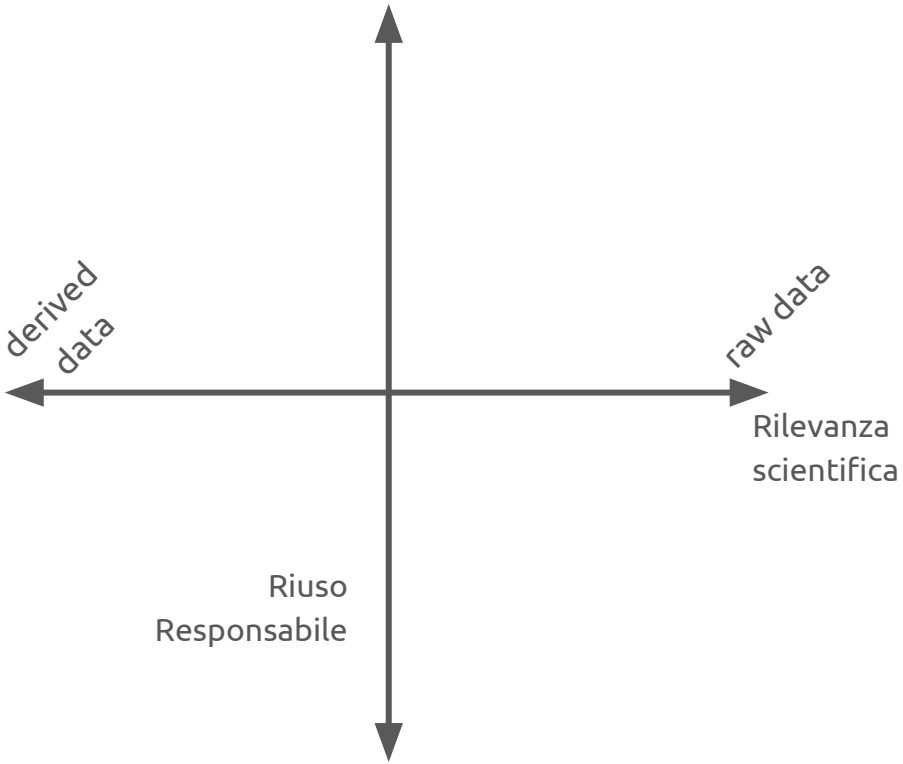
Chi potrebbe riutilizzare la tua collezione?

Ogni raccolta può essere perfezionata eliminando dati personali, dettagli non rilevanti e così via.

In questo modo si trasforma da un insieme di dati "grezzi" a uno derivato.

La rilevanza scientifica non è univocamente correlata alla ricchezza dei dati, anche se un insieme di dati grezzi può prestarsi a un riutilizzo maggiore rispetto a uno derivato.

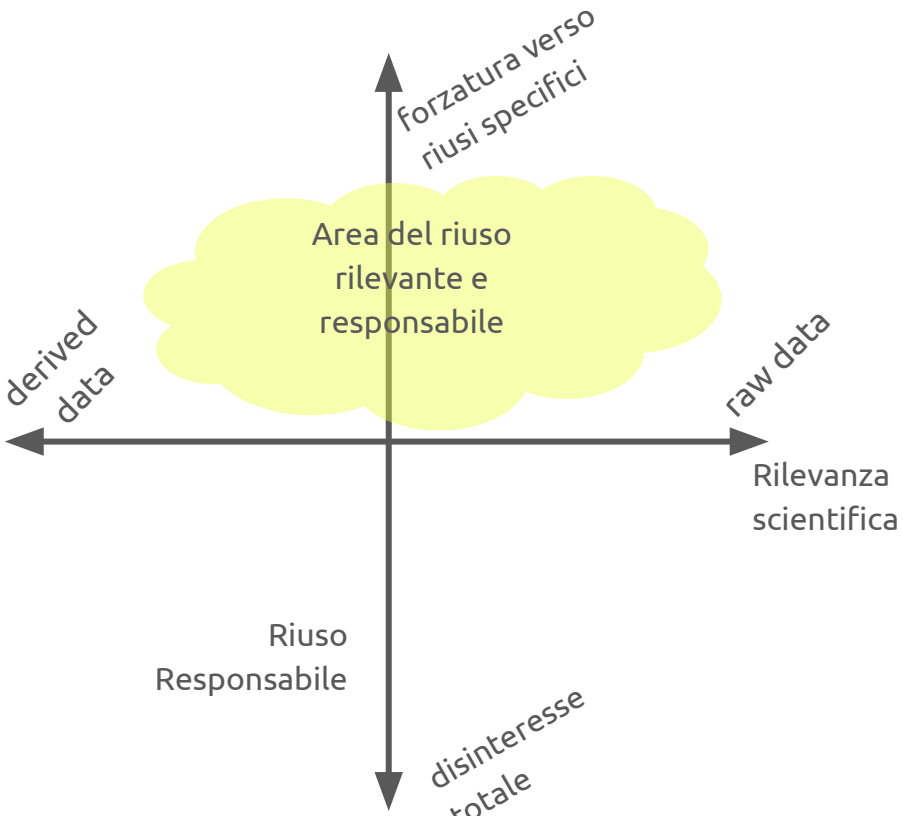
Chi potrebbe riutilizzare la tua collezione?



Il riutilizzo responsabile si riferisce allo sviluppo di un approccio secondario scientificamente rilevante sui dati condivisi, che allo stesso tempo sia conforme ai principi di integrità ed etica della ricerca.

Ciò non dipende dal contenuto dei dati. È possibile ottenere collezioni di dati derivati e riutilizzarli per scopi malevoli.

Chi potrebbe riutilizzare la tua collezione?



Gli autori originali fanno parte del processo di riutilizzo responsabile.

La collezione esposta dovrebbe essere posizionata all'interno di una regione sia di rilevanza scientifica, sia di riutilizzo responsabile.

Allo stesso tempo, gli autori originali non dovrebbero guidare riusi specifici nascondendo parti scientificamente rilevanti della raccolta.

Come scoraggiare riusi sgradevoli?

F - findable

A - accessible

I - interoperable

R - reusable

I - Interoperable

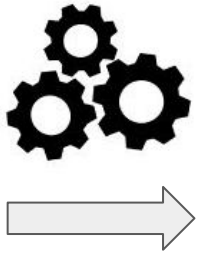
Experimental data

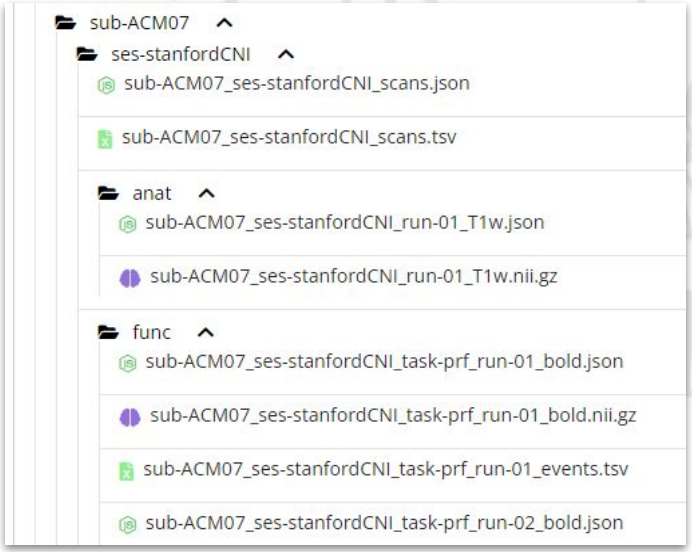
File Modifica Visualizza Inserisci Formato Dati Str

100% € % .0 .00 123 Predefinit

G8

	A	B	E	F
1	Participant ID	Label	Reaction Time	Accuracy
2	001	3382A1A8	0,34	46,24%
3	002	179881BE	0,14	44,11%
4	003	76AE93B1	0,05	57,59%
5	004	B824ADC2	0,46	75,58%
6	005	129147CA	0,03	41,52%
7	006	440FEF25	0,51	84,90%
8	007	824AC0BB	0,36	32,08%
9	008	8341B25F	0,78	42,42%
10	009	CFFFE870	0,23	57,36%
11	010	B32BF2AC	0,46	67,05%
12	011	197296CF	0,17	33,51%
13	012	18DFE4C2	0,84	79,19%
14	013	54B3A12D	0,80	6,75%
15	014	5A2CF043	0,95	47,60%
16	015	56FE4507	0,78	72,19%
17	016	8F5B6442	0,18	72,07%
18	017	60ED878E	0,13	24,08%
19	018	0237A1BF	0,68	44,87%
20	019	1D315730	0,96	51,90%
21	020	D939ABCD	0,78	6,88%
22	021	4AF28762	0,15	19,45%
23	022	79E0557F	0,90	31,41%
24	023	C871FBB9	0,42	79,10%





I - Interoperable

L'informazione è *valida*?

è autoesplicativa?

Cosa posso imparare senza scaricare nulla?

> Metadata and data use a formal, accessible, shared, and broadly applicable language for knowledge representation

F - findable

A - accessible

I - interoperable

R - reusable

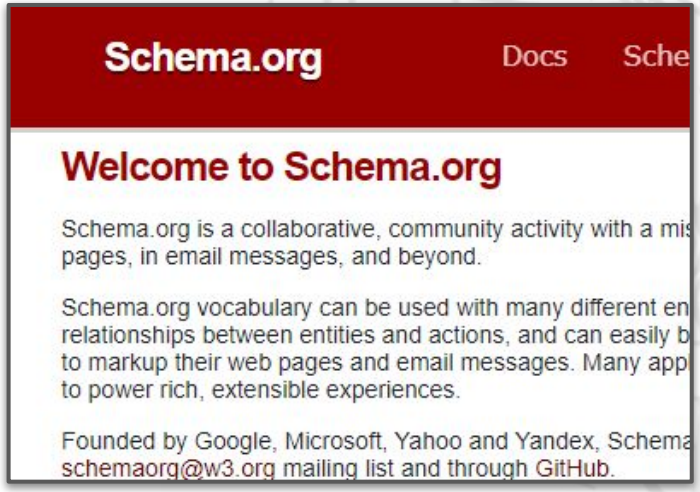
I - Interoperable

L'informazione è *valida*?

è autoesplicativa?

Cosa posso imparare senza scaricare nulla?

> Metadata and data use a formal, accessible, shared, and broadly applicable language for knowledge representation



I - Interoperability - caratteristiche di uno schema

Struttura gerarchica autoesplicativa e informativa (ad esempio, file-system, mark-up, ecc.)

L'etichettatura dei file e/o delle sezioni è costruita in modo modulare per comunicare informazioni senza esplorare il contenuto interno.

evitare la proliferazione superflua
disseminare solo informazione valida

I - Interoperability - caratteristiche di uno schema

Le informazioni descrittive, qualitative e leggibili dall'uomo devono essere posizionate all'interno della directory principale del dataset, per facilitare una rapida e facile sintesi del dataset per l'utente finale.

limitare le comunicazioni private a riguardo

favorire le citazioni

```
+--- BIDS
|   +--- dataset_description.json
|   +--- participants.tsv
|   +--- sub-01
|   |   +--- ses-imaging
|   |   |   +--- anat
|   |   |   |   +--- sub-01_ses-imaging_T1w.json
|   |   |   |   +--- sub-01_ses-imaging_T1w.nii.gz
|   |   |   +--- func
|   |   |   |   +--- sub-01_ses-imaging_task-predict_run-09_bold.json
|   |   |   |   +--- sub-01_ses-imaging_task-predict_run-09_bold.nii.gz
|   |   |   |   +--- sub-01_ses-imaging_task-predict_run-09_events.tsv
|   |   |   +--- sub-01_ses-imaging_scans.tsv
|   |   +--- ses-beh
|   |   |   +--- beh
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-09_beh.json
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-09_beh.tsv
|   |   +--- sessions.tsv
```

```
{
    "Acknowledgements": "",
    "Authors": [
        "Marc Himmelberg",
        "Ekin Tuncok",
        "Jesse Gomez",
        "Kalanit Grill-Spector",
        "Marisa Carrasco",
        "Jonathan Winawer"
    ],
    "BIDSVersion": "1.0.1",
    "DatasetDOI": "doi:10.18112/openneuro.ds004440.v1.0.1",
    "Funding": [
        "R01-EY023915",
        "R01-EY027401",
        "P30-EY013079",
        "R01-EY022318"
    ],
    "HowToAcknowledge": "Cite the publication",
    "License": "CC0",
    "Name": "Stanford Child and Adult Checkerboard Retinotopy Dataset",
    "ReferencesAndLinks": [
        "Comparing visual cortex of children and adults reveals a late-stage change in how V1 samples the visual field",
        "Development differentially sculpts receptive fields across early and high-level human visual cortex"
    ]
}
```

I - Interoperability - caratteristiche di uno schema

I formati di file non devono essere proprietari, facili da compilare e da portare su tutte le piattaforme, senza strumenti specifici (o proprietari).

La serializzazione e la de-serializzazione devono essere implementabili in modo nativo e facilmente automatizzabili.

favorire l'automazione del lavoro ripetitivo sulle informazioni


```
"trial_id":{
"LongName":"Identifier of a specific trial",
"Description":"It identifies specific trials throughout the experiment",
"Units" : "ordinal number"
},

"trial_type":{
"LongName":"Type of a specific trial",
"Description":"A number identifying the category of a trial",
"Units" : "integer number"
},

"RT":{
"LongName":"Reaction Time",
"Description":"It identifies the time when the participant responded to the instruction",
"Units" : "seconds"
},

"total_time": {
"LongName":"Trial full duration",
```


trial_id	trial_type	RT	total_time	grey_c	yellow_c
1	4	0.12983	2232.2	971	321
2	4	0.14882	2190.4	1868	272
3	3	0.049064		3286.1	64	111
4	3	0.13208	2040.3	69	176
5	5	0.1157	2786.7	1360	263
6	4	0.14808	2273.8	511	203
7	5	0.16499	3273.3	488	372
8	4	0.049056		2492	1100	367
9	5	0.18212	2558	260	228
10	3	0.26536	2758	1058	67
11	5	0.13204	3406.2	1652	216
12	5	0.14826	3822.1	800	104
13	5	0.098874		3406.2	290	342

Data minimization - identificatori indiretti

È probabile che i dati contengano molte informazioni che non sono rilevanti per gli scopi scientifici

Non si tratta solo di dati personali (nome, cognome ecc.)

(0008,0070)	Manufacturer	Keep	SIEMENS
(0008,0080)	Institution Name	Keep	Degli Studi Di Trento
(0008,0081)	Institution Address	Keep	Via Delle Regole 101,Trento,Trento,IT,38122
(0008,0090)	Referring Physician's N...	Assign	ABC01234
(0008,1010)	Station Name	Assign	STATION10
(0008,1030)	Study Description	Assign	CLINICAL STUDY 2023
(0008,103E)	Series Description	Keep	AAHead_Scout
(0008,1040)	Institutional Departme...	Assign	Cognitive Science Department
(0008,1050)	Performing Physician's ...	Assign	FAKE MARIA ROSSI
(0008,1070)	Operators' Name	Assign	FAKE MARIO BIANCHI
(0008,1090)	Manufacturer's Model ...	Keep	Prisma
(0010,0010)	Patient's Name	Assign	19830729VTIC_202309221330
(0010,0020)	Patient ID	Assign	19830729VTIC
(0010,0030)	Patient's Birth Date	Assign	19830729
(0010,0040)	Patient's Sex	Keep	M
(0010,1010)	Patient's Age	Assign	040Y
(0010,1020)	Patient's Size	Keep	1.62
(0010,1030)	Patient's Weight	Keep	55

NB: questa è un header DICOM reale in cui ho modificato i campi inserendo informazioni FALSE. NON si tratta di informazioni reali dei partecipanti.

Inspired by E. Glerean's seminar "Handling of Personal Data", 16/03/2022
Anonymization and Personal Data by Finnish Social Science Data Archive, url checked on 21/9/2023
https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

Data minimization - identificatori indiretti

È probabile che i dati contengano molte informazioni che non sono rilevanti per gli scopi scientifici

Non si tratta solo di dati personali (nome, cognome ecc.)

Gli identificatori indiretti potrebbero comunque contenere informazioni rilevanti che possono essere utilizzate per individuare il partecipante.

(0008,0070)	Manufacturer	Keep	SIEMENS
(0008,0080)	Institution Name	Keep	Degli Studi Di Trento
(0008,0081)	Institution Address	Keep	Via Delle Regole 101,Trento,Trento,IT,38122
(0008,0090)	Referring Physician's N...	Assign	ABC01234
(0008,1010)	Station Name	Assign	STATION10
(0008,1030)	Study Description	Assign	CLINICAL STUDY 2023
(0008,103E)	Series Description	Keep	AAHead_Scout
(0008,1040)	Institutional Departme...	Assign	Cognitive Science Department
(0008,1050)	Performing Physician's ...	Assign	FAKE MARIA ROSSI
(0008,1070)	Operators' Name	Assign	FAKE MARIO BIANCHI
(0008,1090)	Manufacturer's Model ...	Keep	Prisma
(0010,0010)	Patient's Name	Assign	19830729VTIC_202309221330
(0010,0020)	Patient ID	Assign	19830729VTIC
(0010,0030)	Patient's Birth Date	Assign	19830729
(0010,0040)	Patient's Sex	Keep	M
(0010,1010)	Patient's Age	Assign	040Y
(0010,1020)	Patient's Size	Keep	1.62
(0010,1030)	Patient's Weight	Keep	55

NB: questa è un header DICOM reale in cui ho modificato i campi inserendo informazioni FALSE. NON si tratta di informazioni reali dei partecipanti.

Inspired by E. Glerean's seminar "Handling of Personal Data", 16/03/2022
Anonymization and Personal Data by Finnish Social Science Data Archive, url checked on 21/9/2023
https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

Data minimization - identificatori indiretti

Non si tratta solo di dati personali (nome, cognome ecc.)

Gli identificatori indiretti potrebbero comunque contenere informazioni rilevanti che possono essere utilizzate per individuare il partecipante.

Possono essere usati per scopi malevoli.

(0008,0070)	Manufacturer	Keep	SIEMENS
(0008,0080)	Institution Name	Keep	Degli Studi Di Trento
(0008,0081)	Institution Address	Keep	Via Delle Regole 101,Trento,Trento,IT,38122
(0008,0090)	Referring Physician's N...	Assign	ABC01234
(0008,1010)	Station Name	Assign	STATION10
(0008,1030)	Study Description	Assign	CLINICAL STUDY 2023
(0008,103E)	Series Description	Keep	AAHead_Scout
(0008,1040)	Institutional Departme...	Assign	Cognitive Science Department
(0008,1050)	Performing Physician's ...	Assign	FAKE MARIA ROSSI
(0008,1070)	Operators' Name	Assign	FAKE MARIO BIANCHI
(0008,1090)	Manufacturer's Model ...	Keep	Prisma
(0010,0010)	Patient's Name	Assign	19830729VTIC_202309221330
(0010,0020)	Patient ID	Assign	19830729VTIC
(0010,0030)	Patient's Birth Date	Assign	19830729
(0010,0040)	Patient's Sex	Keep	M
(0010,1010)	Patient's Age	Assign	040Y
(0010,1020)	Patient's Size	Keep	1.62
(0010,1030)	Patient's Weight	Keep	55

NB: questa è un header DICOM reale in cui ho modificato i campi inserendo informazioni FALSE. NON si tratta di informazioni reali dei partecipanti.

Inspired by E. Glerean's seminar "Handling of Personal Data", 16/03/2022
Anonymization and Personal Data by Finnish Social Science Data Archive, url checked on 21/9/2023
https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

Data minimization - cancella!

Cancella ciò che è completamente irrilevante per un riuso secondario, ad esempio l'indirizzo dell'istituzione, gli esecutori di esperimenti, ecc.

(0008,0070)	Manufacturer	Keep	SIEMENS
(0008,0080)	Institution Name	Keep	
(0008,0081)	Institution Address	Keep	
(0008,0090)	Referring Physician's N...	Assign	ABC01234
(0008,1010)	Station Name	Assign	STATION10
(0008,1030)	Study Description	Assign	CLINICAL STUDY 2023
(0008,103E)	Series Description	Keep	AAHead_Scout
(0008,1040)	Institutional Departme...	Assign	
(0008,1050)	Performing Physician's ...	Assign	
(0008,1070)	Operators' Name	Assign	
(0008,1090)	Manufacturer's Model ...	Keep	Prisma
(0010,0010)	Patient's Name	Assign	
(0010,0020)	Patient ID	Assign	
(0010,0030)	Patient's Birth Date	Assign	
(0010,0040)	Patient's Sex	Keep	M
(0010,1010)	Patient's Age	Assign	040Y
(0010,1020)	Patient's Size	Keep	1.62
(0010,1030)	Patient's Weight	Keep	55

NB: questa è un header DICOM reale in cui ho modificato i campi inserendo informazioni FALSE. NON si tratta di informazioni reali dei partecipanti.

Inspired by E. Glerean's seminar "Handling of Personal Data", 16/03/2022
Anonymization and Personal Data by Finnish Social Science Data Archive, url checked on 21/9/2023
https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

Data minimization - categorizza!

Si può prendere in considerazione la possibilità di classificare i dati che sono scientificamente rilevanti ma che non richiedono uno specifico "approfondimento informativo".

(0010,0040)	Patient's Sex	Keep	M
(0010,1010)	Patient's Age	Assign	040Y
(0010,1020)	Patient's Size	Keep	1.62
(0010,1030)	Patient's Weight	Keep	55

40 Years old -> **35 ÷ 45 years old**
162 cm -> **160 ÷ 170 cm**
55 kg -> **50 ÷ 60 kg**

NB: questa è un header DICOM reale in cui ho modificato i campi inserendo informazioni FALSE. NON si tratta di informazioni reali dei partecipanti.

Inspired by E. Glerean's seminar "Handling of Personal Data", 16/03/2022
Anonymization and Personal Data by Finnish Social Science Data Archive, url checked on 21/9/2023
https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

Data minimization - categorizza!

40 Years old -> **35 ÷ 45 years old**
162 cm -> **160 ÷ 170 cm**
55 kg -> **50 ÷ 60 kg**

(0010,0040)	Patient's Sex	Keep	M
(0010,1010)	Patient's Age	Assign	040Y
(0010,1020)	Patient's Size	Keep	1.62
(0010,1030)	Patient's Weight	Keep	55

Schemi di dati ben congegnati guidano l'utente verso l'eliminazione e la categorizzazione, introducendo formati di file adatti e strutture informative specifiche.

NB: questa è un header DICOM reale in cui ho modificato i campi inserendo informazioni FALSE. NON si tratta di informazioni reali dei partecipanti.

Inspired by E. Glerean's seminar "Handling of Personal Data", 16/03/2022
Anonymization and Personal Data by Finnish Social Science Data Archive, url checked on 21/9/2023
https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

Take home messages

- La condivisione dei dati per favorire la riproducibilità è un compito scientifico, per pianificazione, rigore e finalità.
- Le persone che costruiscono i contributi fanno parte del processo di riutilizzo responsabile di ciò che hanno originariamente raccolto.
- L'interoperabilità tramite schemi di dati offre una soluzione solida, scientificamente valida e facile da usare per questo problema, suggerendo come redistribuire e sistemare le informazioni più rilevanti da esporre.

“As structured as possible, as derived as necessary”

“As structured as possible, as derived as necessary”

vittorio.iacovella@unitn.it