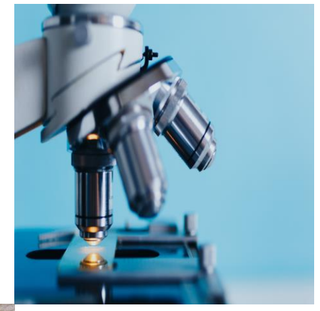
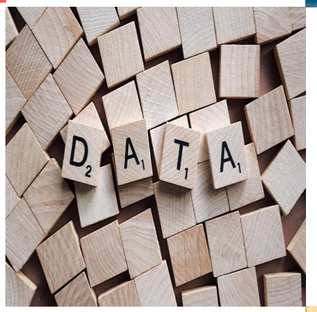
GESTIRE I DATI DELLA RICERCA: tutto ciò che c'è da sapere





LEZIONE 2

Giulia Caldoni e Mario Marino
mario.marino6@unibo.it
Data Steward area Sociale, Research Services Division (ARIC), Alma Mater Studiorum - Università di Bologna

10 Ottobre 2023

DURANTE L’INCONTRO DI OGGI…

- Per favore tenete i microfoni spenti durante la presentazione.
- Sentitevi liberi di accendere microfono e videocamera nei momenti di Q&A.
- Sentitevi liberi di porre le vostre domande in chat in qualsiasi momento, risponderemo durante il Q&A!
- Abbiamo previsto all’interno di questa lezione due pause, che seguiranno i momenti Q&A.

GESTIRE I DATI DELLA RICERCA: TUTTO CIÒ CHE C'È DA SAPERE

Modulo formativo diviso in tre incontri.

Unico pre-requisito è avere una propria esperienza di ricerca.

LE TEMATICHE CHE AFFRONTEREMO OGGI

1. RESEARCH DATA MANAGEMENT
   Cosa e perchè?

2. GESTIRE IL DATO IN PRATICA
   Azioni e strumenti contestualizzati nelle fasi del ciclo di vita del dato.

3. ASPETTI PRIVACY
   Presentati dal dott.Francesco di Tano

CHI SIAMO? DATA STEWARDS @UNIBO

Il progetto Data Stewards @Unibo mira al rafforzamento del supporto alla gestione dei dati della ricerca:

- Chi sono i Data Stewards?

Figure di supporto per le tematiche di gestione FAIR dei dati della ricerca e stesura del Data Management Plan ai team ricerca di UNIBO

- Cosa fanno?

Supportano i ricercatori nella gestione FAIR dei dati (research data management) e nella stesura del Data Management Plan, principalmente nel contesto dei progetti Horizon Europe
Supportano la Governance di Ateneo nella promozione di Open Science

- Qual è il loro profilo?

Background scientifico (esperti di dominio)
Esperienza in data management
FAIR principles e conoscenza di pratiche di OS

MARIO MARINO

Area Sociale

BIANCA GUALANDI

Area Umanistica

GIULIA CALDONI

Area Biomedica

SARA COPPINI

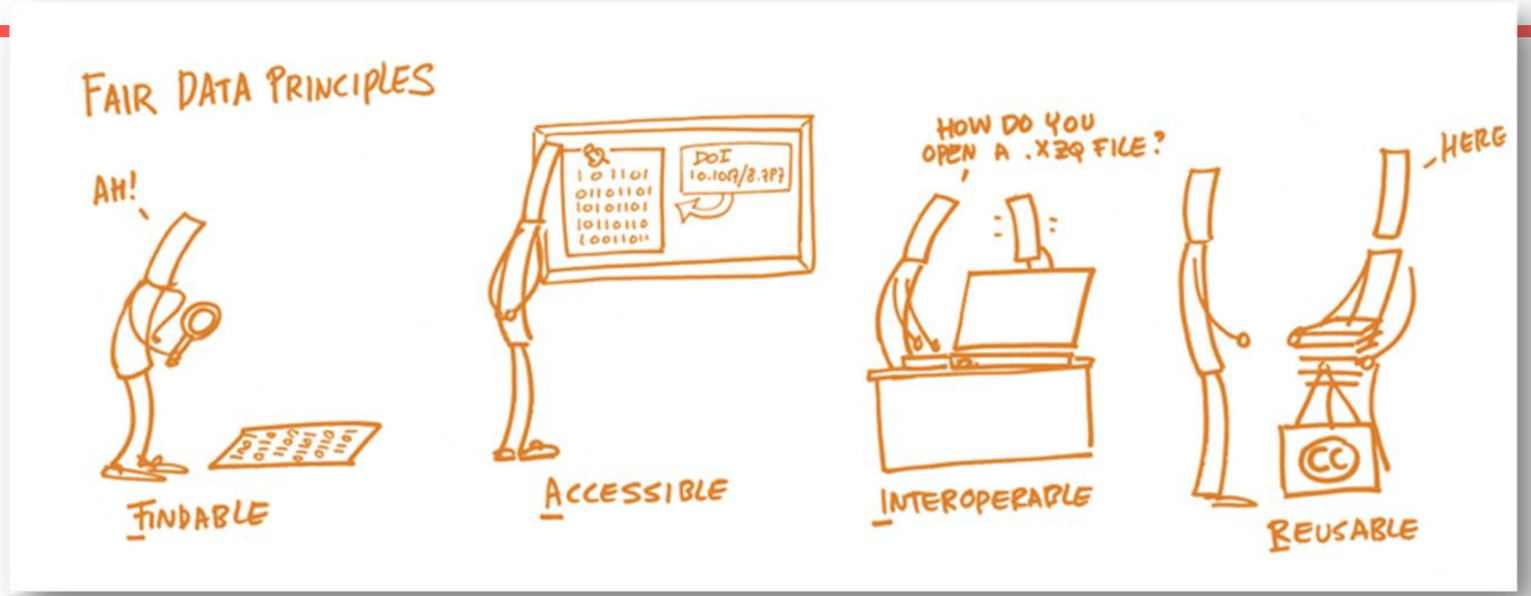
Area Tecnologica

RIPRENDIAMO LE FILA DEL DISCORSO...

Image credit: http://aukeherrema.nl CC-BY

- Esistono diverse definizioni di dato e variano in base alla disciplina di riferimento.
- Dato è tutto ciò che è alla base dei ragionamenti a supporto di una tesi di ricerca.
- Una collezione di dati accomunati dallo stesso obiettivo è chiamata dataset.
- Il dato ha una valore intrinseco come asset della ricerca, per conservarlo deve essere gestito correttamente.
- La gestione corretta è un processo che permea tutte le fasi del ciclo di vita del dato della ricerca.

RIPRENDIAMO LE FILA DEL DISCORSO...



- I dati della ricerca devono essere gestiti in modo trasparente: seguire i principi FAIR e renderli reperibili, accessibili, interoperabili e riutilizzabili.

- I dati FAIR non sono necessariamente dati gestiti correttamente né tantomeno Open Data, " As open as possible, as closed as necessary".

- Un repository per i dati non fa tutto il lavoro per rendere i vostri dati il più FAIR possibile. Tuttavia, offre un'ottima struttura per ottenere le basi giuste.

Image credit: http://aukeherrema.nl CC-BY

RESEARCH DATA MANAGEMENT

RESEARCH DATA MANAGEMENT: UNA DEFINIZIONE

Gestione e organizzazione attenta dei dati di ricerca durante l'intero ciclo di ricerca, con l'obiettivo di rendere il processo di ricerca il più efficiente possibile e di facilitare la cooperazione con gli altri.

Smits, D.A.B. and Teperek, M., 2020. Research Data Management for Master's Students: From Awareness to Action. Data Science Journal, 19(1), p.30. DOI: http://doi.org/10.5334/dsj-2020-030
https://the-turing-way.netlify.app/reproducible-research/overview/overview-benefit.html

QUALI SONO I VANTAGGI DI UNA CORRETTA GESTIONE DEL DATO DI RICERCA?

Organizzare i dati rende il tuo lavoro più efficiente.
In termini di costi/tempo: i dati gestiti una volta restano interpretabili, comprensibili e rintracciabili.

Se li gestisci, potresti non perderli.
L'archiviazione corretta dei dati e il backup regolare prevengono le perdite di dati.

Alcuni dati potrebbero essere unici e non riproducibili.
Questo li rende preziosissimi per la comunità scientifica.

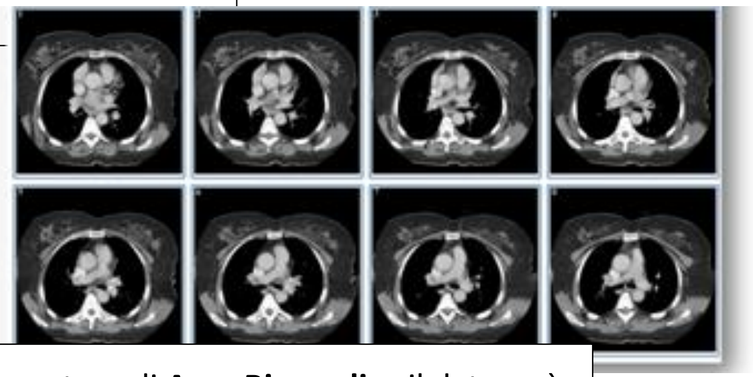
Aumenta l'integrità della ricerca.
Un dato correttamente gestito facilita la validazione e il controllo.

Stimola la collaborazione con altri ricercatori.
Troveranno più facile comprendere e riutilizzare i vostri dati.

QUALI TIPI DI DATI...

... dipende dal dominio di ricerca.

Per un ricercatore di **Area Sociale** il dato può essere un’intervista.

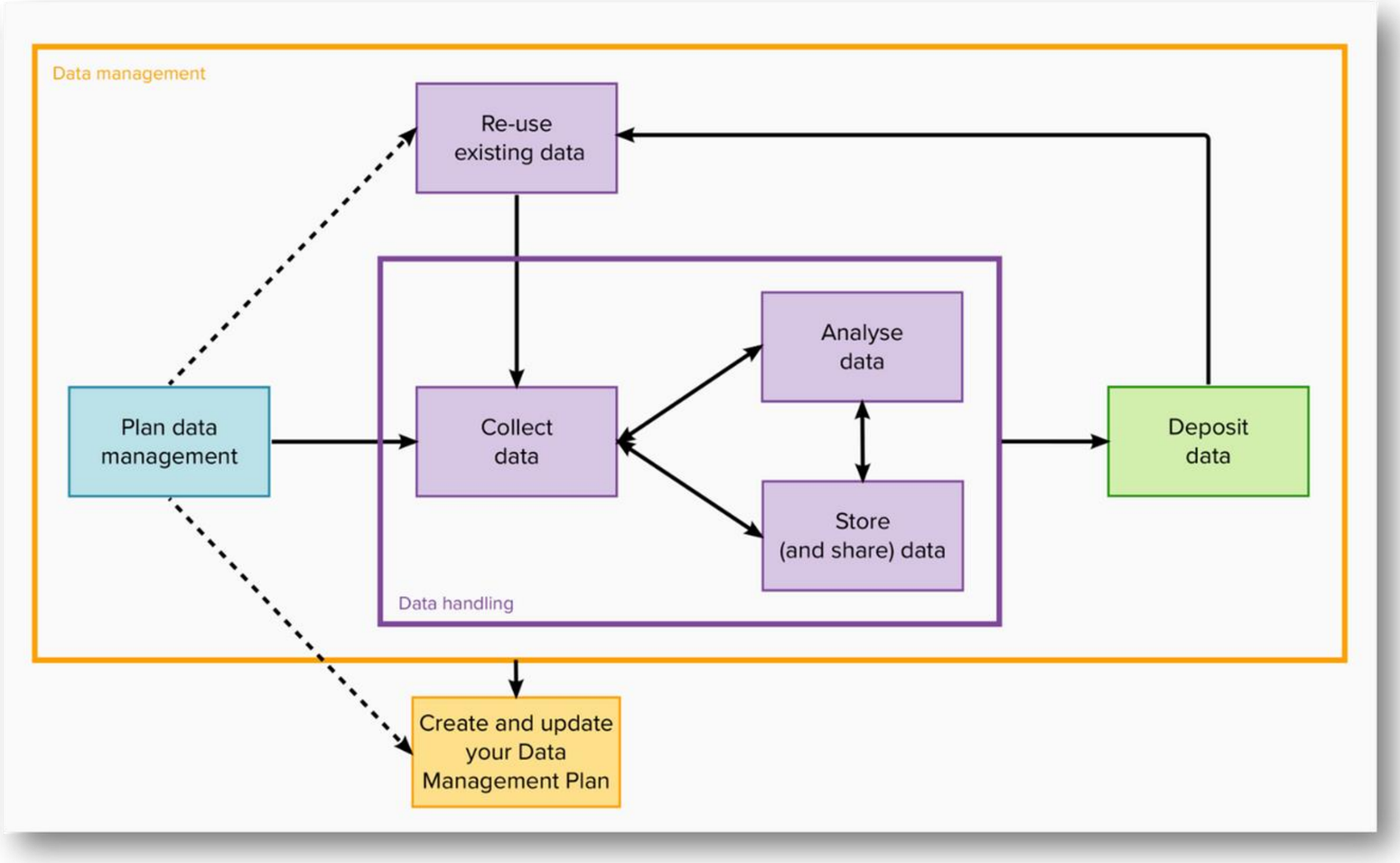


Per un ricercatore di **Area Biomedica** il dato può essere l'immagine di una TAC.

Per un ricercatore di **Area Ingegneristica** il dato può essere il **modello 3D** di un prototipo digitale

Per un ricercatore di **Area Umanistica** il dato può essere la digitalizzazione di una **fonte testuale antica**.

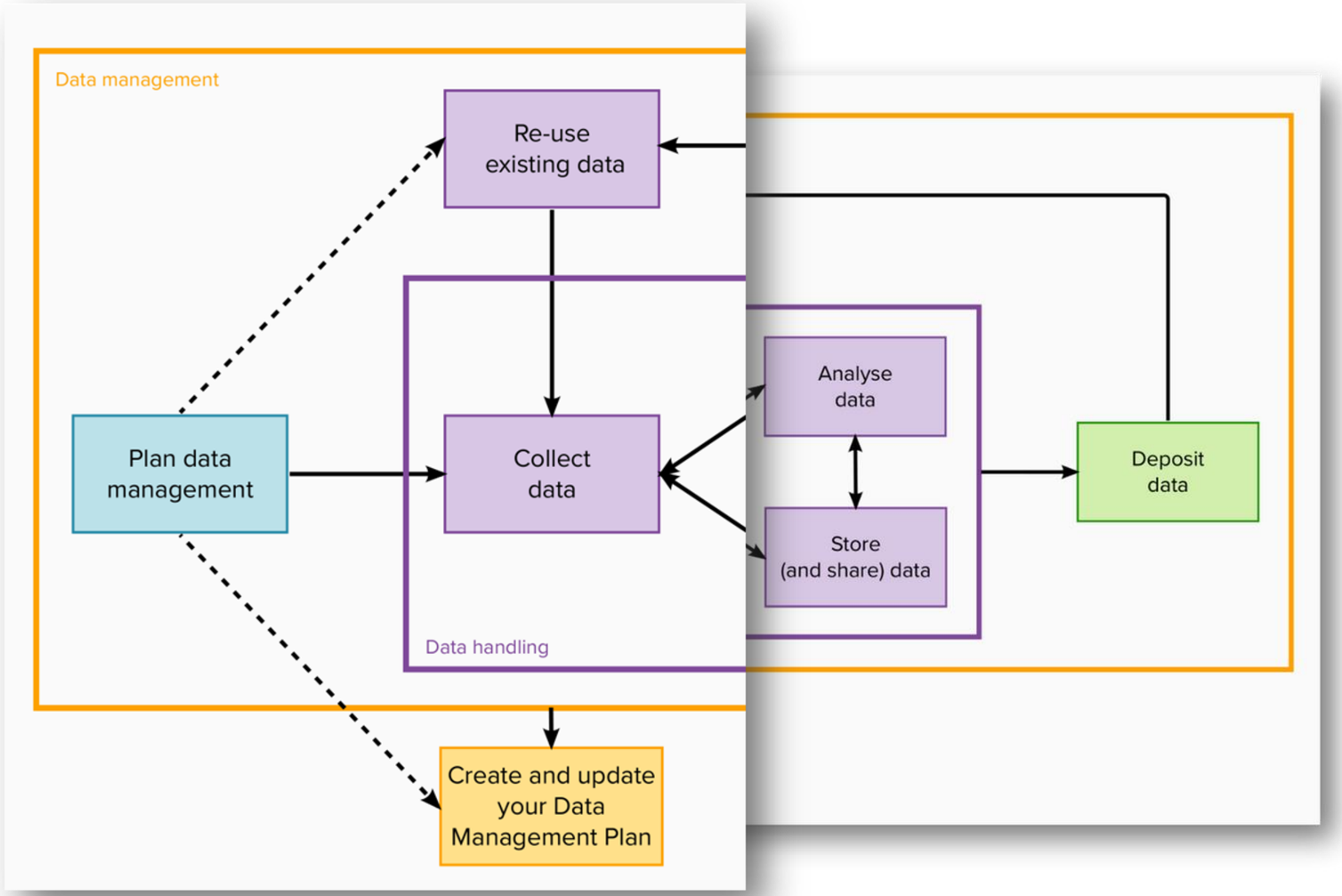
...LE FASI DEL CICLO DI VITA DEL DATO SONO LE STESSE.



Gualandi, Bianca, Caldoni, Giulia, & Marino, Mario. (2022). Research Data Management: Data Lifecycle. Zenodo. https://doi.org/10.5281/zenodo.7249051

GESTIRE IL DATO IN PRATICA

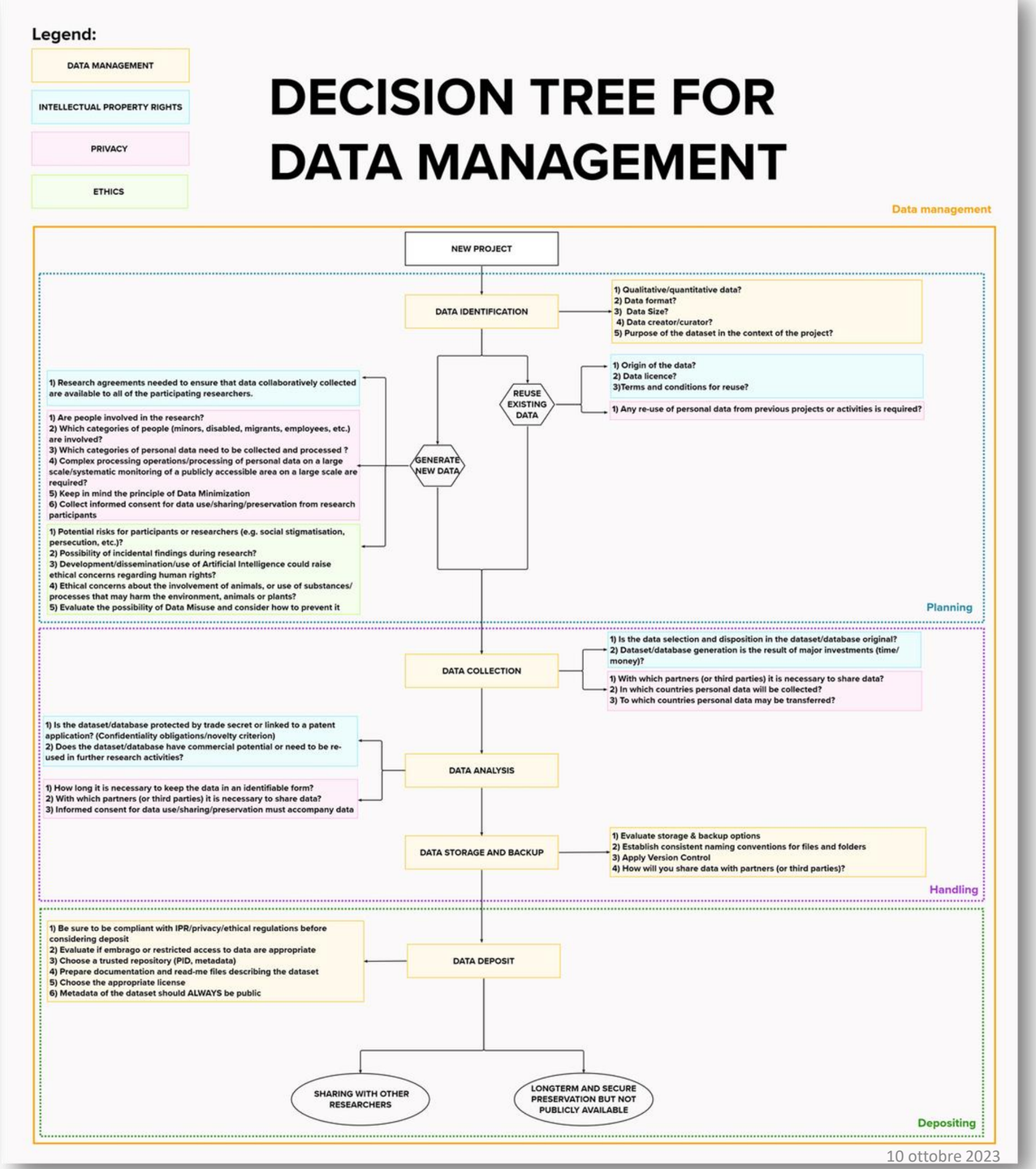
PIANIFICARE LA RICERCA E IL DATA MANAGEMENT



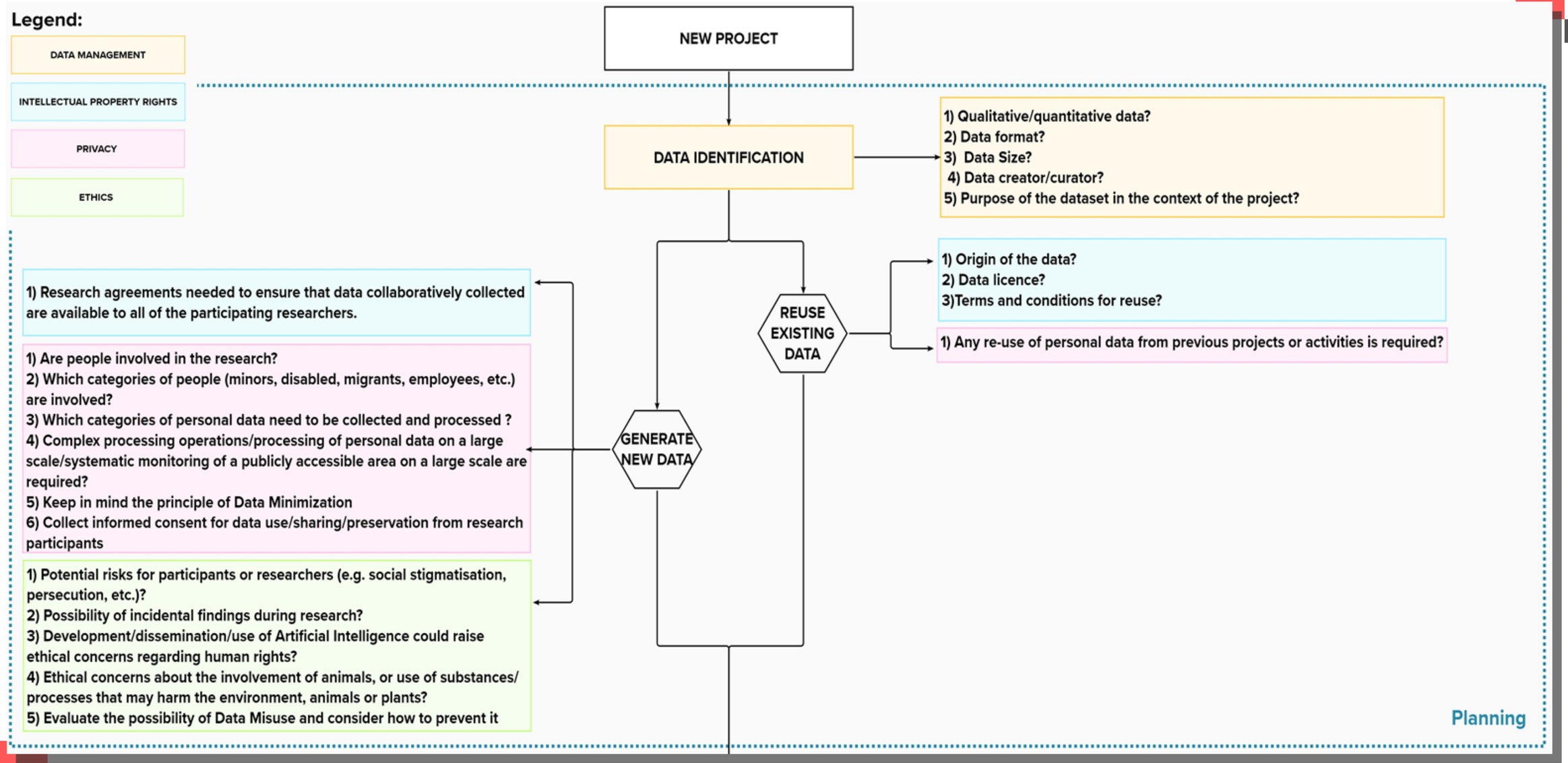
Gualandi, Bianca, Caldoni, Giulia, & Marino, Mario. (2022). Research Data Management: Data Lifecycle. Zenodo. https://doi.org/10.5281/zenodo.7249051

DA DOVE PARTIRE?

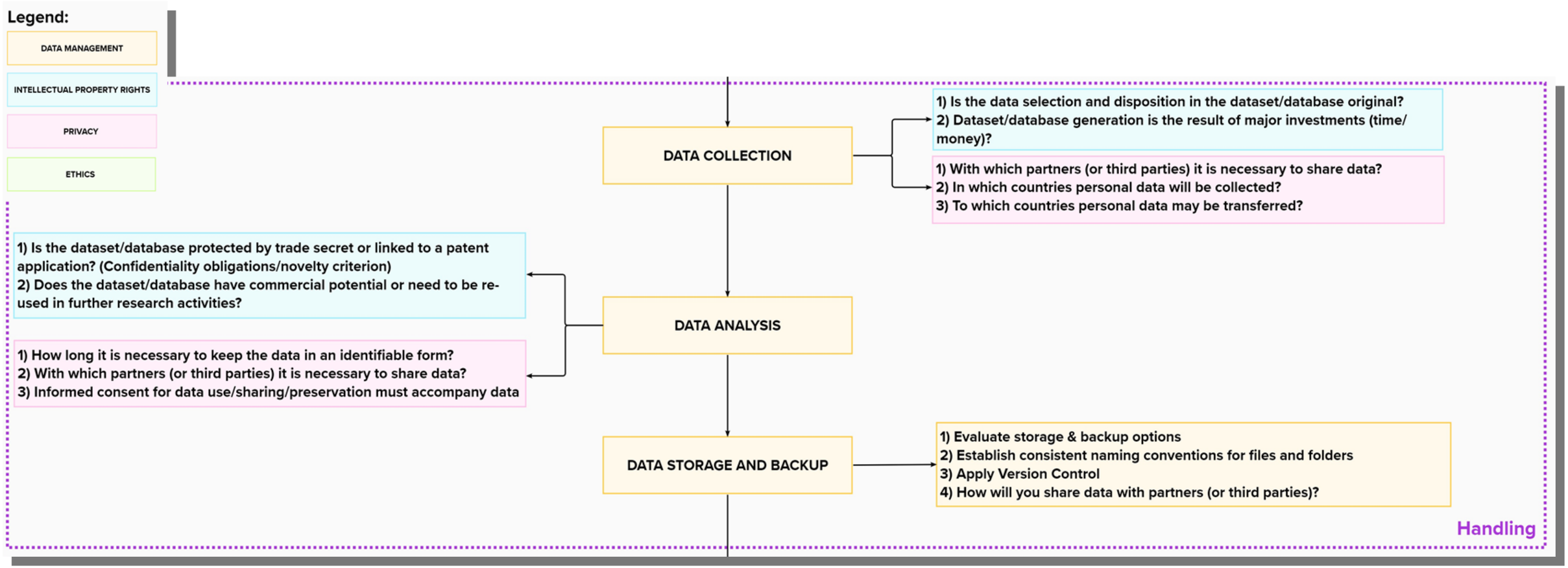
PIANIFICARE
IL DATA MANAGEMENT



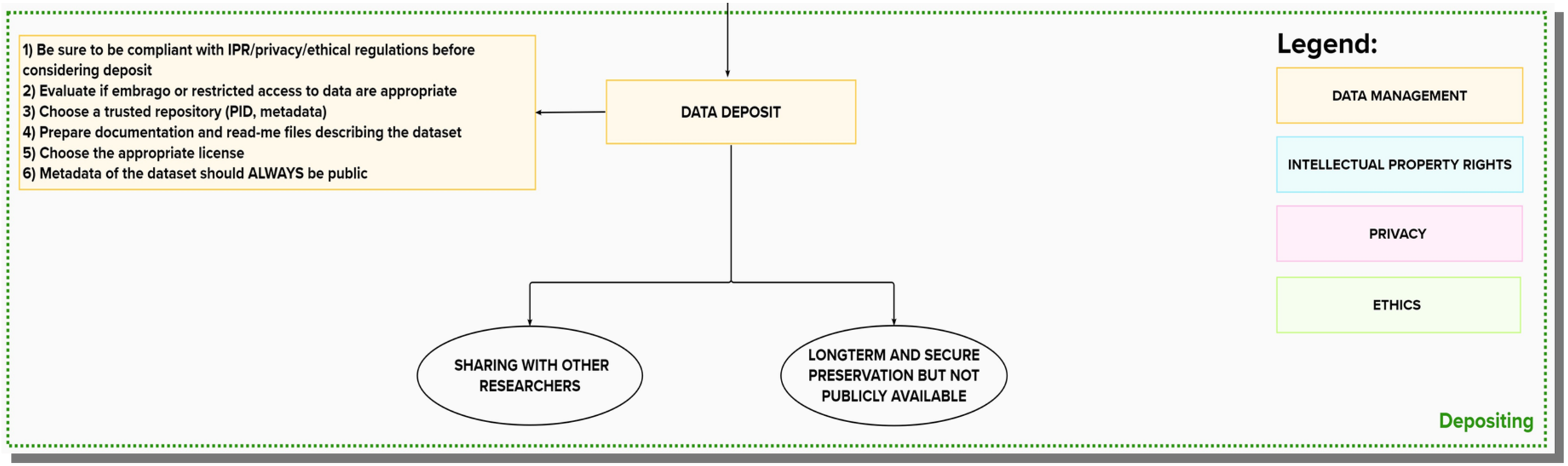
Caldoni, Giulia, Gualandi, Bianca, & Marino, Mario. (2022). Research Data Management Decision Tree. Zenodo. https://doi.org/10.5281/zenodo.7190005



Caldoni, Giulia, Gualandi, Bianca, & Marino, Mario. (2022). Research Data Management Decision Tree. Zenodo. https://doi.org/10.5281/zenodo.7190005



Caldoni, Giulia, Gualandi, Bianca, & Marino, Mario. (2022). Research Data Management Decision Tree. Zenodo. https://doi.org/10.5281/zenodo.7190005



Caldoni, Giulia, Gualandi, Bianca, & Marino, Mario. (2022). Research Data Management Decision Tree. Zenodo. https://doi.org/10.5281/zenodo.7190005

RESEARCH DATA MANAGEMENT DECISION TREE:
PONITI LE DOMANDE GIUSTE!

Partendo dai mattoni fondamentali del ciclo di vita dei dati, integra una serie di domande che mirano a incoraggiare i ricercatori ad affrontare alcuni punti di attenzione principali:

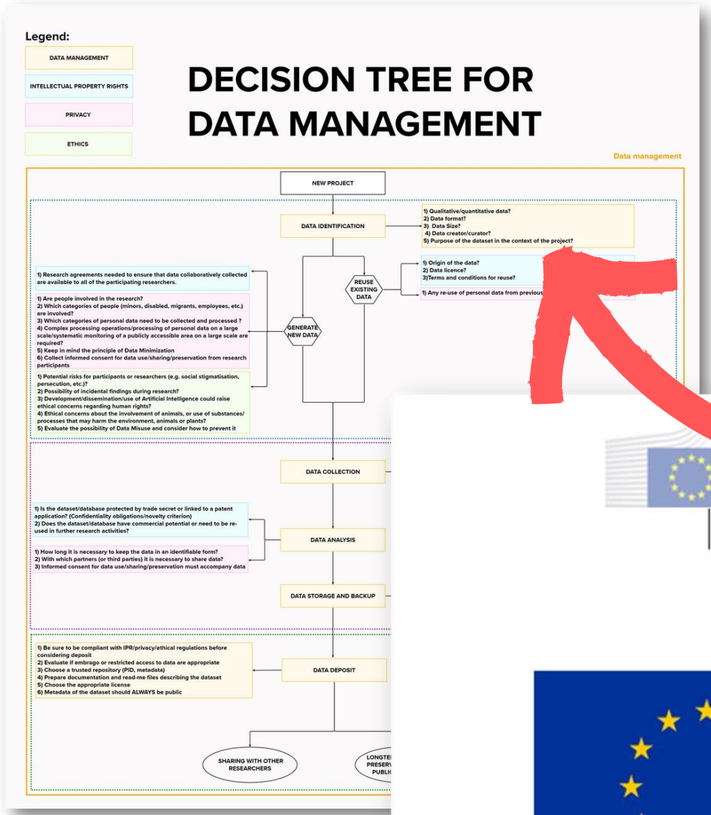
- requisiti di privacy/etica
- legislazione sui diritti di proprietà intellettuale
- principi FAIR

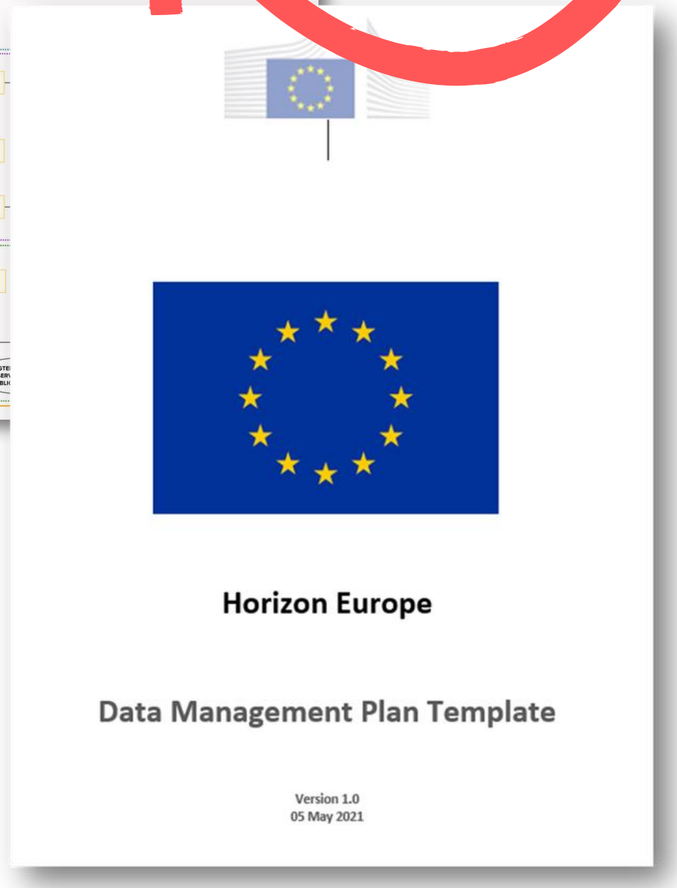
DATA MANAGEMENT PLAN: DOCUMENTA LE RISPOSTE!

Il DMP viene redatto all'inizio del progetto e aggiornato nel corso del tempo.

Un DMP è una tabella di marcia, con una serie di domande a cui rispondere per:

- Gestire i dati evitando problemi come duplicazione, perdita di dati, violazioni della sicurezza.
- Sviluppare procedure efficienti e coerenti per la gestione dei dati, risparmiando tempo in seguito.
- Essere consapevoli di come state adempiendo a tutti i vostri obblighi legali ed etici.





RIUTILIZZARE DATI

Durante la pianificazione, controllate se esistono dati già pubblicati da altri ricercatori utili per la vostra ricerca.

L'elemento più importante da considerare è la possibilità di riutilizzare legalmente i dati trovati.

- Sono coperti da diritti di proprietà intellettuale (copyright, diritti sulle banche dati)?
- Se il set di dati è accompagnato da una licenza, questa vi dirà cosa potete o non potete fare con esso.
- Se il set di dati non è accompagnato da una licenza, dovrete contattare l'autore originale e chiederlo.

ESEMPI DI REPOSITORY DOVE CERCARE DATI DA RIUTILIZZARE:

DISCIPLINARI

- CESSDA, Data Catalogue: https://datacatalogue.cessda.eu/
- NASA, EarthData: https://www.earthdata.nasa.gov/
- British Film Institute: https://www.bfi.org.uk/industry-data-insights

GENERALISTI

- Google Dataset Search: https://datasetsearch.research.google.com/
- Harvard University Dataverse: https://dataverse.harvard.edu/

SIA SE RIUTILIZZATE CHE SE GENERATE DATI, NELLA RICERCA POTRESTE DOVER AFFRONTARE I SEGUENTI ASPETTI:

DIRITTI IPR

- Cosa è consentito fare legalmente con i dati creati da altri?
- Come volete che gli altri riutilizzino i vostri dati?
- Il vostro lavoro ha un potenziale di valorizzazione commerciale (ad esempio, trasferimento tecnologico)?
- Gli accordi esistenti con terzi limitano lo sfruttamento o la diffusione dei dati che usate?

PRIVACY

- Raccoglierete dati da soggetti umani?
- Come potete proteggere la privacy dei vostri soggetti di ricerca?
- Come potete assicurarvi che la vostra ricerca rispetti la legislazione sul trattamento dei dati personali (ad esempio, il GDPR)?

ETICA

- Svolgete ricerche su animali?
- La vostra ricerca ha applicazioni militari?
- La vostra ricerca può influenzare la vita dei vostri soggetti di ricerca (ad esempio, scoperte accidentali, discriminazioni)?
- Nella vostra ricerca sviluppate/utilizzate Intelligenza Artificiale?

L'UNIVERISTÀ DI BOLOGNA PUÒ SUPPORTARVI PER AFFRONTARE QUESTI ASPETTI: A CHI POTETE RIVOLGERVI?

DIRITTI IPR

Per supporto riguardo la proprietà intellettuale dei risultati di ricerca UniBo potete rivolgervi al Knowledge Transfer Office: **kto@unibo.it**

PRIVACY

Per supporto sulle questioni relative alla privacy potete scrivere all'indirizzo: **privacy@unibo.it**

ETICA

Per supporto sulle questioni relative alle questioni etiche potete consultare la pagina: **https://www.unibo.it/it/ricerca/strutture-di-ricerca/comitati-etici-1**

COSTI E RESPONSABILITÀ

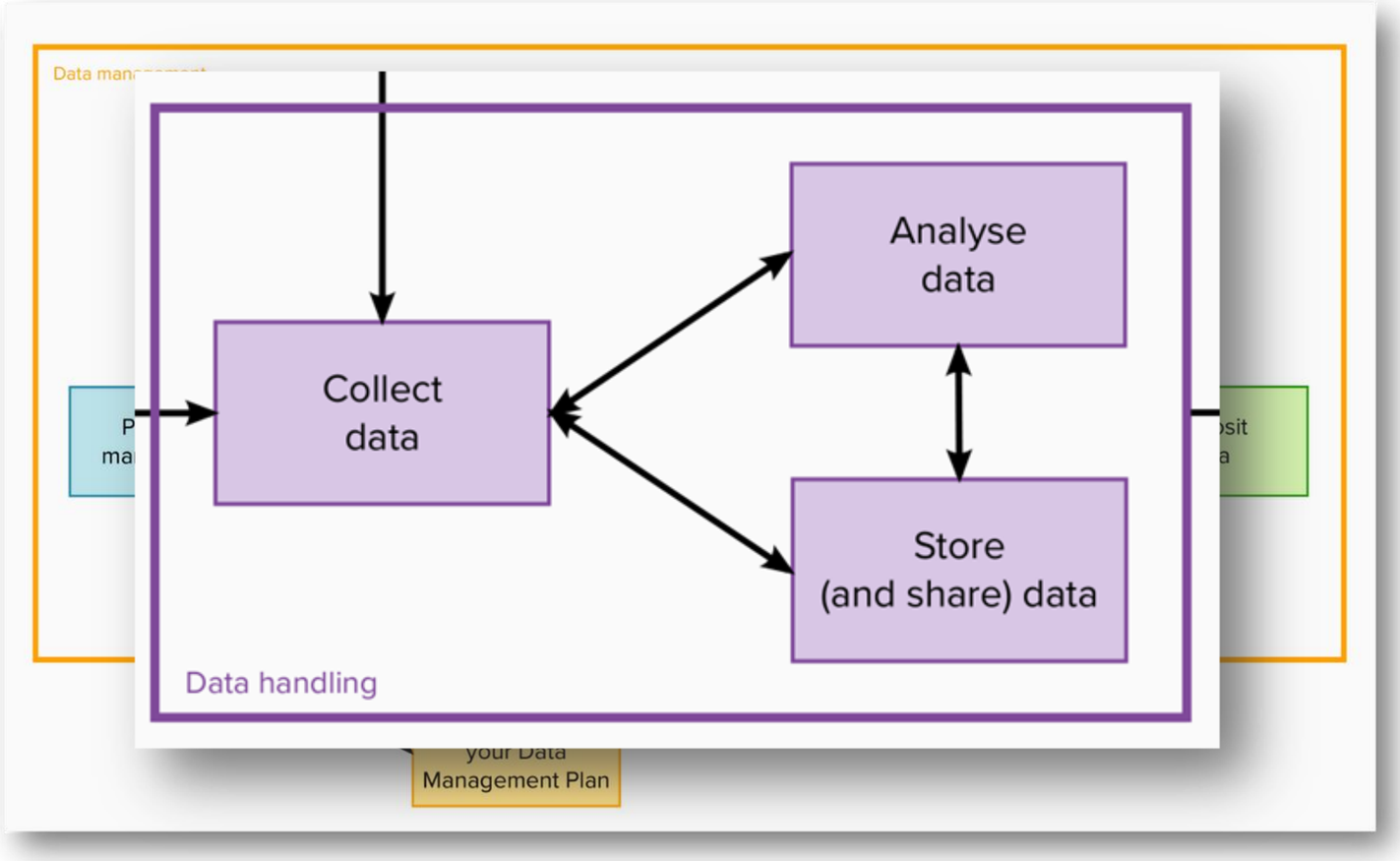
La corretta RDM è estremamente utile ai ricercatori, ma può essere dispendiosa sia in termini di tempo che di denaro. I costi possono essere collegati a:

- Raccolta dei dati: e.g. acquisizione dataset esterni, formattazione, organizzazione, trascrizione.
- Descrizione dei dati, i metadati e la documentazione: richiedono molto tempo, soprattutto se effettuati in una fase successiva del progetto.
- Archiviazione: e.g. raccogliendo dati di grandi dimensioni.
- Accesso ai dati e la sicurezza: e.g. accesso remoto tramite VPN.
- Conservazione: e.g. convertire i dati/file in formati aperti.
- Riutilizzo dei dati: e.g. anonimizzazione, diritti d'autore, condivisione dei dati.

COSTI E RESPONSABILITÀ

Nel contesto della gestione dei dati di ricerca, il termine "responsabilità" indica l'individuo che è incaricato di rispondere di tutti gli aspetti dell'esecuzione di una DMP.

RDM NELLE FASI ATTIVE DELLA RICERCA



RDM NELLE FASI ATTIVE DELLA RICERCA

Le fasi attive della ricerca sono quelle più concitate!

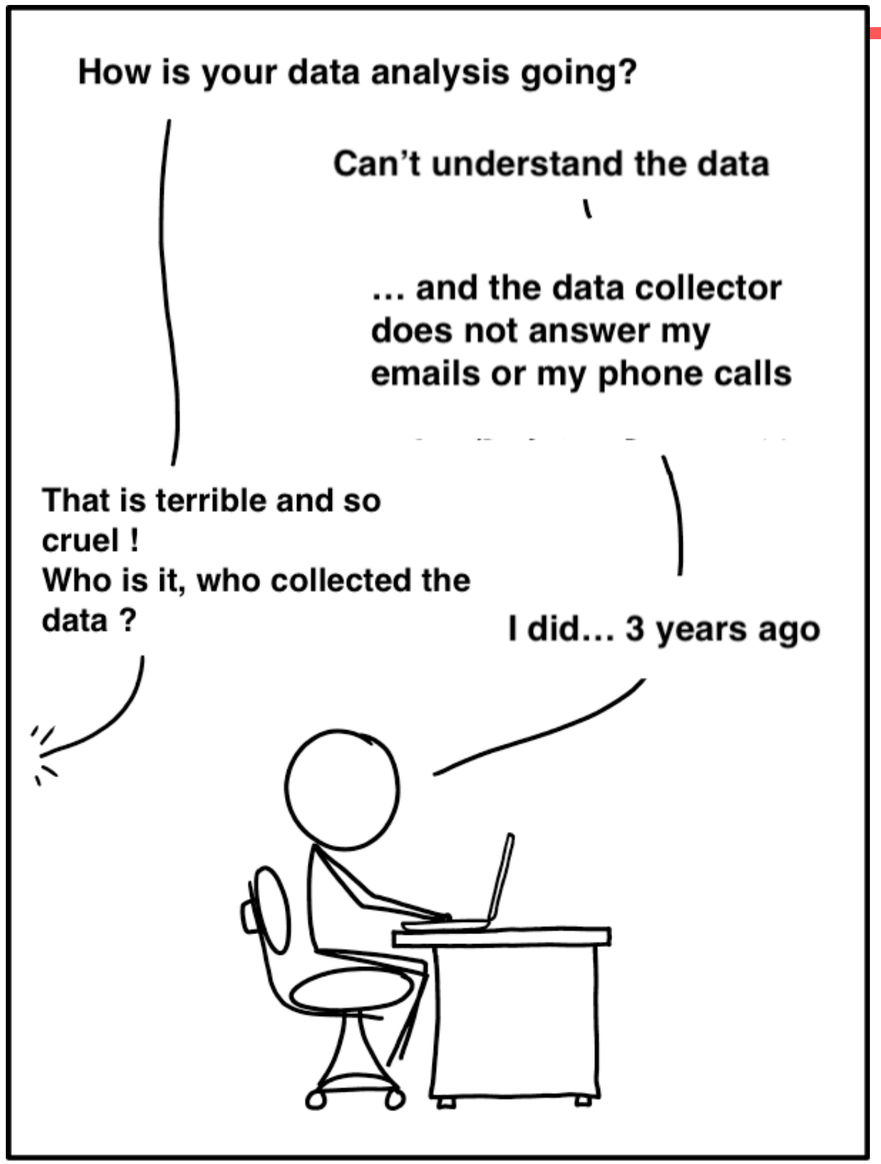
- Aver preso le giuste decisioni prima e averle scritte in un DMP aiuta a portare avanti la ricerca in modo coerente.
- Tutte le buone pratiche adottate in queste fasi sono funzionali per il prosieguo della ricerca e per risparmiare tempo/fatica.
- Raccogliere durante queste fasi quante più informazioni relative ai dati utilizzati/generate aiuta a garantirne qualità, trasparenza e riproducibilità.

METADATI

- Informazioni che aggiungono struttura ai dati per renderli machine-readable.
- Strutturali (su un oggetto in sé) o descrittivi (sul contenuto di un oggetto).
- È possibile utilizzare vocabolari controllati.
- Esistono schemi standard di metadati, sia generici che specifici per disciplina.

DOCUMENTAZIONE

- Informazioni che rendono i dati più comprensibili agli altri, permette di capirli e interpretarli anche molto dopo la raccolta.
- Human-readable.
- Molti ricercatori accompagnano i dati con un README file, che spiega come sono stati raccolti i dati, cosa significano i metadati ecc.
- Può anche essere "in-file", come i commenti nel codice.



Your first collaborators are your future selves, be nice to them !

your future self, by Julien Colomb, CC-BY-NC, derived from .NORM Normal File Format, CC-BY-NC, by Randall Munroe

METADATI...

...DOCUMENTAZIONE...

...MA ANCHE METODOLOGIE!

- Metodi aggiornati e versionabili
- Associati a un DOI al momento della pubblicazione
- Metodi persistentemente archiviati e recuperabili

https://www.protocols.io/

protocols.io

CONTROLLARE E GARANTIRE LA QUALITÀ DEI DATI

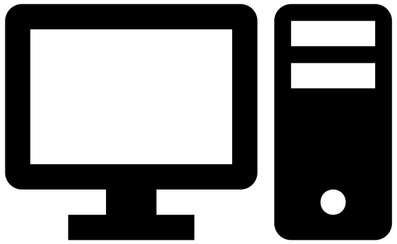
Descrizione: comprende la registrazione delle definizioni (ciò che è stato effettivamente misurato) e delle unità di misura. Le descrizioni delle variabili si traducono in metadati.

Organizzazione: comprende principalmente l'identificazione dei dati. Quali dati sono disponibili e come possono essere utilizzati? Questo include anche l'identificazione di dataset esterni che possono essere riutilizzati.

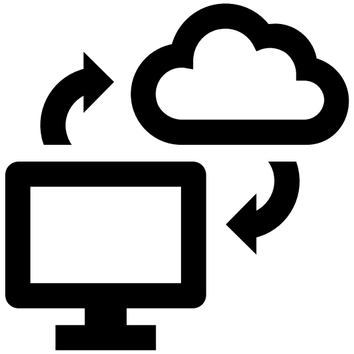
Conservazione: consiste nel valutare come e dove archiviare i dati per assicurarsi che siano disponibili per un riutilizzo futuro.

Pulizia: identificazione dei valori mancanti e non plausibili e il controllo della distribuzione delle variabili.

STORAGE DEI DATI: PERCHÈ PREOCCUPARSENE?



Per gestire i dati in sicurezza ed evitare di incorrere in problemi come la perdita parziale o totale dei dati!



- Che dimensione avranno i vostri dati?
- Dove immagazzinerete i vostri dati?
- Come condividerete i dati con i collaboratori?
- Utilizzerete soluzioni cloud?
- Farete il backup dei dati? Dove e come?

BACKUP: PERCHÈ È FONDAMENTALE?

BUONE PRATICHE :

- Non archiviare i dati solo sul portatile;
- Utilizzare una soluzione di archiviazione locale con backup automatici;
- Backup automatici di tutto il contenuto del laptop/PC (Windows: Cronologia file; Mac: Time Machine);
- Tenere conto delle diverse versioni dei file.

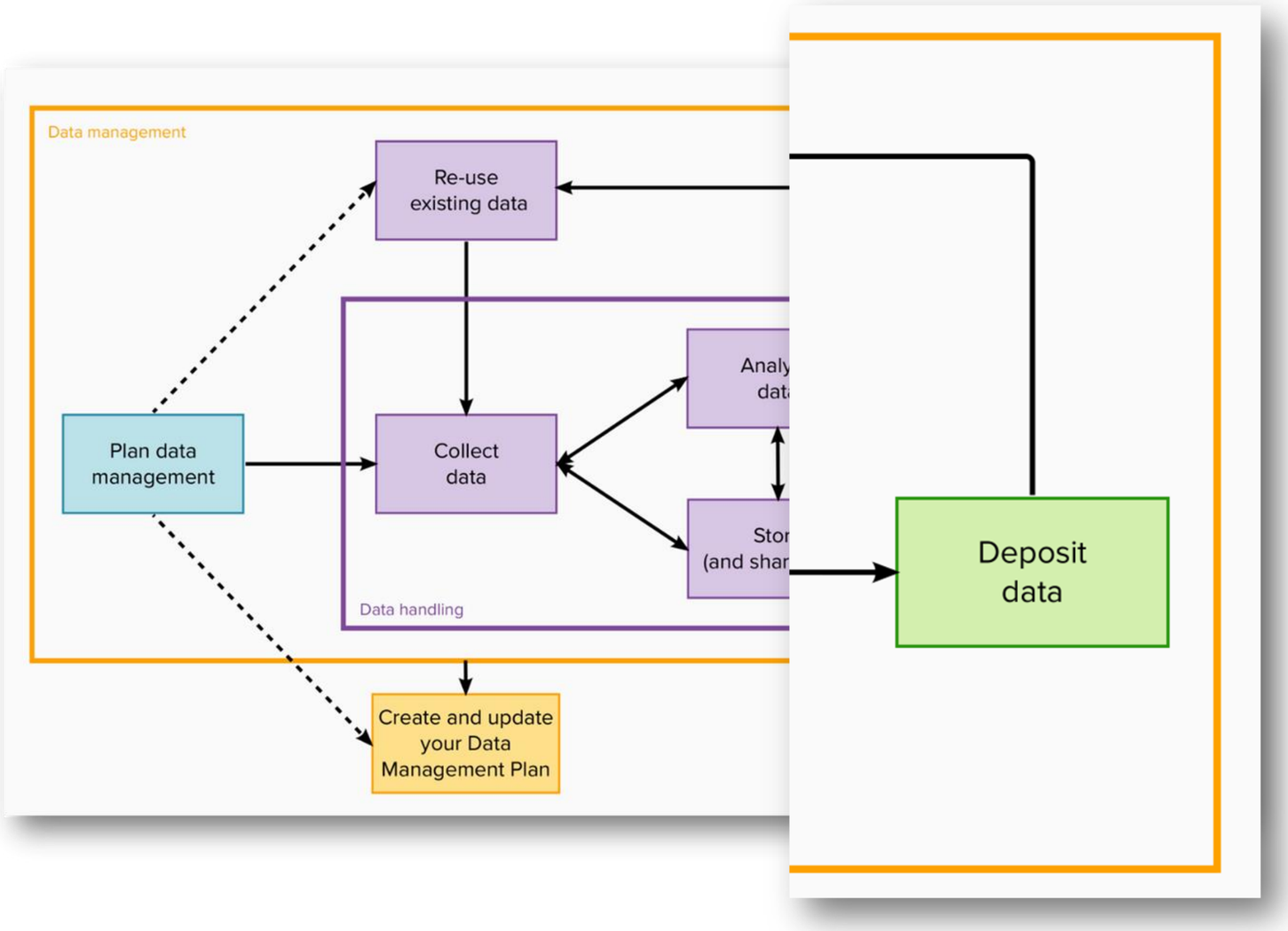
STORAGE DEI DATI: STRATEGIE AGGIUNTIVE PER DATI SICURI

Per assicurarti che i tuoi dati siano protetti contro modifiche e accessi non autorizzati.

Per proteggere i dati personali e le informazioni riservate.

- Aggiornare l’antivirus
- I file spostati nel cestino non sempre sono cancellati correttamente: usare un software certificato per la cancellazione
- Proteggere con password i software di lavoro
- Prendere in considerazione la cifratura dei dati (processo di codifica delle informazioni digitali in modo tale che solo utenti autorizzati possano visualizzarle)

DEPOSITO DEI DATI: LA FASE FINALE DEL CICLO DI VITA DEL DATO



Gualandi, Bianca, Caldoni, Giulia, & Marino, Mario. (2022). Research Data Management: Data Lifecycle. Zenodo. https://doi.org/10.5281/zenodo.7249051

STORAGE E DEPOSITO: DUE CONCETTI DIFFERENTI

Storage: immediatamente dopo la raccolta dei dati, può comportare la condivisione con il team/partner.

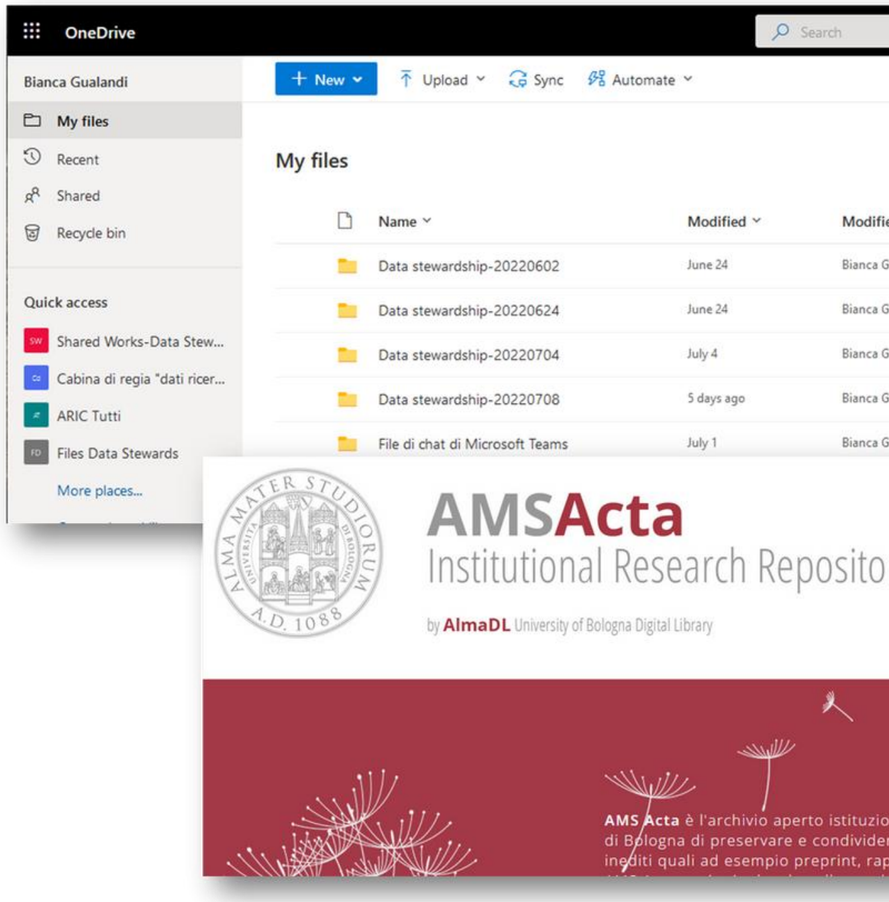
Comporta:

- Backup e sicurezza dei dati
- Organizzazione e denominazione di file e cartelle

Deposito: viene effettuato una volta o (preferibilmente) in modo iterativo.

Comporta:

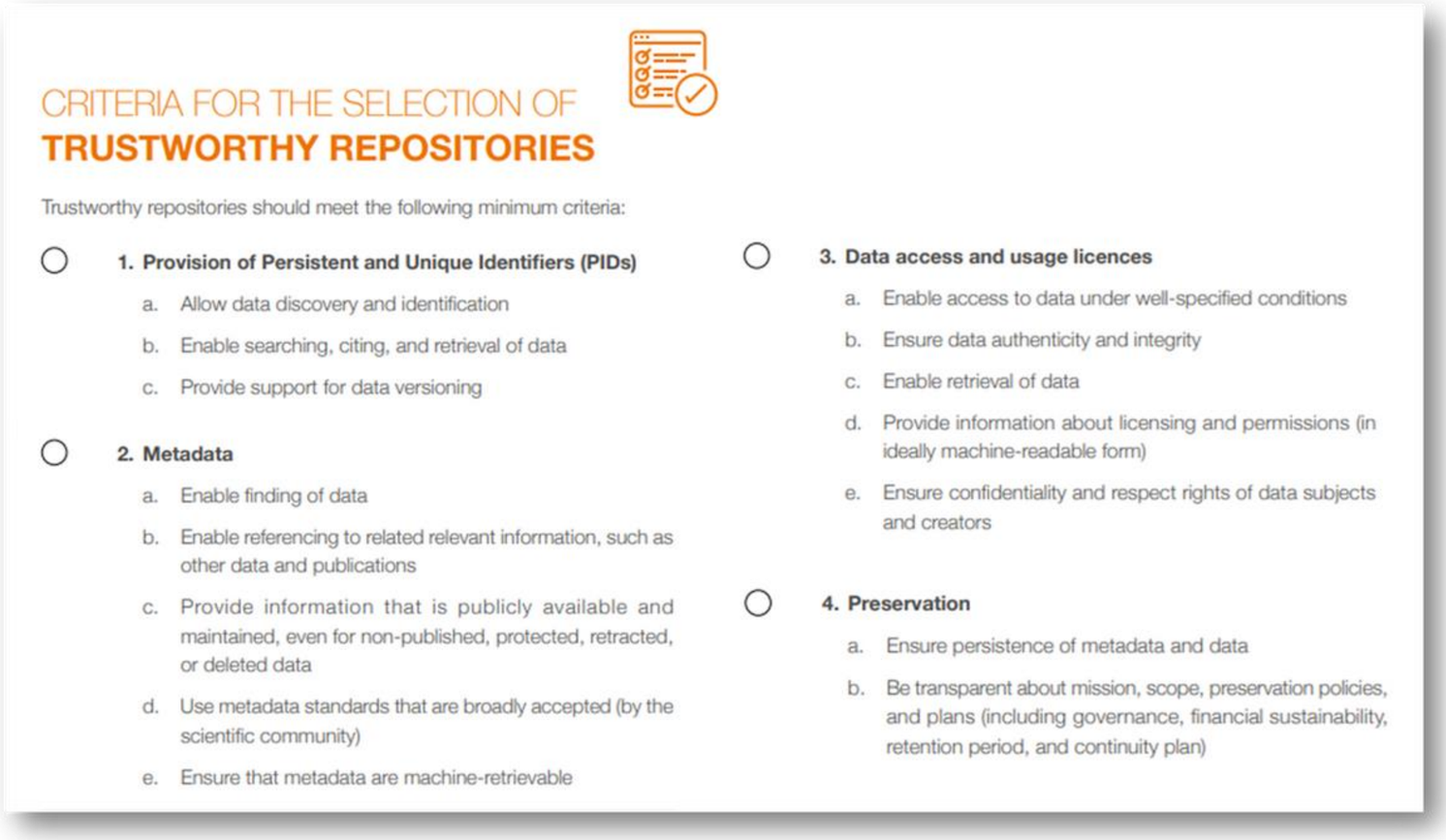
- Passaggio di responsabilità per la preservazione dal ricercatore al repository
- Scelta della licenza e del livello di accesso



LE CARATTERISTICHE DI UN (BUON) REPOSITORY

Il repository è un'infrastruttura digitale in cui i dati vengono depositati dai ricercatori per garantirne la conservazione a lungo termine.

E' uno degli elementi chiave per l'implementazione pratica dei principi FAIR e delle pratiche di Open Science:



CRITERIA FOR THE SELECTION OF
TRUSTWORTHY REPOSITORIES

Trustworthy repositories should meet the following minimum criteria:

1. Provision of Persistent and Unique Identifiers (PIDs)

a. Allow data discovery and identification
b. Enable searching, citing, and retrieval of data
c. Provide support for data versioning

2. Metadata

a. Enable finding of data
b. Enable referencing to related relevant information, such as other data and publications
c. Provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data
d. Use metadata standards that are broadly accepted (by the scientific community)
e. Ensure that metadata are machine-retrievable

3. Data access and usage licences

a. Enable access to data under well-specified conditions
b. Ensure data authenticity and integrity
c. Enable retrieval of data
d. Provide information about licensing and permissions (in ideally machine-readable form)
e. Ensure confidentiality and respect rights of data subjects and creators

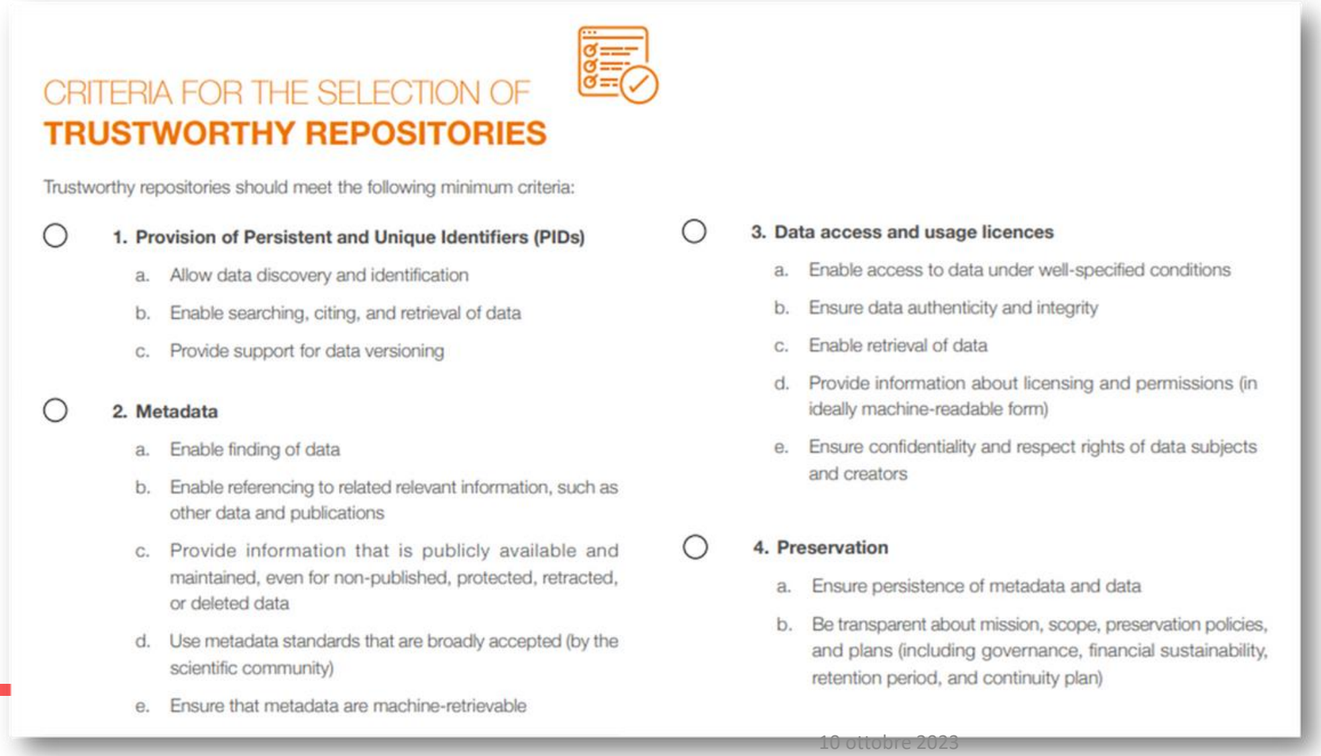
4. Preservation

a. Ensure persistence of metadata and data
b. Be transparent about mission, scope, preservation policies, and plans (including governance, financial sustainability, retention period, and continuity plan)

https://scienceeurope.org/our-resources/practical-guide-to-the-international-alignment-of-research-data-management/

LE CARATTERISTICHE DI UN (BUON) REPOSITORY

- Consente di assegnare identificatori persistenti (PID);
- Supporta schemi di metadati standard e strutturati (es: dublincore, datacite, ecc);
- Consente di attribuire una licenza ai dati;
- Supporta standard di interoperabilità;
- Consente ai dati di essere preservati nel lungo periodo.



CRITERIA FOR THE SELECTION OF
TRUSTWORTHY REPOSITORIES

Trustworthy repositories should meet the following minimum criteria:

1. **Provision of Persistent and Unique Identifiers (PIDs)**
   a. Allow data discovery and identification
   b. Enable searching, citing, and retrieval of data
   c. Provide support for data versioning
2. **Metadata**
   a. Enable finding of data
   b. Enable referencing to related relevant information, such as other data and publications
   c. Provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data
   d. Use metadata standards that are broadly accepted (by the scientific community)
   e. Ensure that metadata are machine-retrievable
3. **Data access and usage licences**
   a. Enable access to data under well-specified conditions
   b. Ensure data authenticity and integrity
   c. Enable retrieval of data
   d. Provide information about licensing and permissions (in ideally machine-readable form)
   e. Ensure confidentiality and respect rights of data subjects and creators
4. **Preservation**
   a. Ensure persistence of metadata and data
   b. Be transparent about mission, scope, preservation policies, and plans (including governance, financial sustainability, retention period, and continuity plan)

https://scienceeurope.org/our-resources/practical-guide-to-the-international-alignment-of-research-data-management/

COME SCEGLIERE COSA DEPOSITARE?

Non tutti i dati della ricerca sono necessari per la sua comprensione, verifica e riproducibilità.

Alcune idee su cosa non conservare:

- dati facilmente replicabili,
- dati di prova, pilota o intermedi,
- dati che non possono essere utilizzati da altri per una serie di motivi.

Image credit: http://aukeherrema.nl CC-BY

COME SCEGLIERE DOVE DEPOSITARE I DATI?

1. Utilizzare re3data per trovare un repository disciplinare OPPURE

2. Utilizzare il repository della tua istituzione OPPURE

3. Utilizzare un repository generalis come Zenodo

AMS ACTA: IL NOSTRO REPOSITORY DI ATENEO

- Permette agli autori di auto-archiviare i propri contributi con una procedura semplice
- Assegna il DOI a ciascun contributo depositato
- Garantisce la conservazione e l'accesso nel tempo ai contributi depositati
- Implementa licenza d’uso tra cui Creative Commons
- Implementa le Linee Guida OpenAire
- Ciascun contributo è indicizzato e reso accessibile dai principali motori di ricerca (Google, Google Scholar...)

https://amsacta.unibo.it/

Per supporto sulle questioni relative al deposito potete contattare: almadl@unibo.it

https://sba.unibo.it/it/almadl/servizi-almadl/disseminare-pubblicazioni-e-dati-della-ricerca-in-open-access

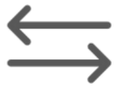
QUALCHE PUNTO CHIAVE

- Gestire attentamente i dati aiuta a rendere il processo della ricerca più efficiente.
- La gestione dei dati della ricerca è un processo che deve seguire tutte le fasi del ciclo di vita del dato.
- Le decisioni prese durante la fase di pianificazione devono essere riportate nel Data Management Plan.
- Le fasi attive della ricerca sono le più concitate: aver pianificato può aiutarti a portare avanti la ricerca in modo coerente.
- Non tutti i dati della ricerca sono necessari per la sua comprensione, verifica e riproducibilità: fai una scelta su cosa condividere, depositandoli in un repository dei dati alla fine della ricerca.

QUALCHE PUNTO CHIAVE

Per supporto sulla gestione dei dati potete rivolgervi ai data stewards: aric.datasteward@unibo.it

Per supporto sulle questioni relative al deposito nel repository di ateneo potete contattare: almadl@unibo.it



Per supporto riguardo la proprietà intellettuale dei risultati di ricerca UniBo potete rivolgervi al Knowledge Transfer Office: kto@unibo.it

Per supporto sulle questioni relative alla privacy potete scrivere all'indirizzo: privacy@unibo.it

GRAZIE