ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Linee Guida di Ateneo per la Gestione dei Dati della Ricerca

Linee guida a cura di:

Alma Mater Studiorum – Università di Bologna

ARIC – Area Ricerca, Settore Coordinamento Servizi Ricerca e Progetti di Area, Data Steward

Con il contributo di:

ARIN – Area Innovazione, Settore Knowledge Transfer Office

ARPAC – Area Patrimonio Culturale, Settore Gestione e sviluppo della biblioteca digitale d'Ateneo – AlmaDL; Settore Gestione e sviluppo della Biblioteca delle risorse elettroniche – AlmaRE

GLOS – Gruppo di Lavoro Open Science

Referenti Open Science dei Dipartimenti

SSRD – Staff Rettore e Direttore Staff Rettore e Direttore Generale, Unità Professionale "Protezione dei dati personali"

APPC – Area Pianificazione, Programmazione e Comunicazione - Settore Comunicazione - Ufficio Graphic design per la comunicazione

Queste linee guida accompagnano la *Policy di Ateneo per la Gestione dei Dati della Ricerca*, che stabilisce criteri e principi da seguire per trattare i dati in modo corretto e consapevole, in linea con gli standard internazionali e le peculiarità proprie dell'ambito disciplinare.

I destinatari di questo documento sono **tutti i ricercatori e le ricercatrici dell'Università di Bologna**, a qualsiasi stadio della propria carriera e in qualsiasi disciplina.

Gestire i dati della ricerca significa **curarli e organizzarli in modo consapevole durante l'intero ciclo di ricerca** con l'obiettivo di:

- rendere il processo di ricerca il più efficiente possibile;
- rendere i dati stessi interpretabili, comprensibili e rintracciabili nel tempo;
- favorire l'integrità della ricerca;
- stimolare la collaborazione con altri ricercatori.

La gestione dei dati va **pianificata attentamente** all'inizio della ricerca, accompagna tutte le fasi operative di **produzione, raccolta e analisi dati** e si conclude con la preservazione (archiviazione a lungo termine dei dati stessi) e preferibilmente la **condivisione degli stessi**.

In questa guida, vedremo a uno a uno gli aspetti più rilevanti di questo processo, fornendo indicazioni procedurali e strumenti utili ad affrontare, in modo puntuale, ciascuna delle azioni necessarie ad una corretta gestione dei dati della ricerca.

Maggiori dettagli sugli argomenti presentati in queste linee guida si possono trovare nelle schede di approfondimento indipendenti che le accompagnano. I riferimenti alle schede di approfondimento sono presentati nel testo come segue **Nome scheda**.

Link utili

Policy di Ateneo sulla Gestione dei Dati della Ricerca:
https://www.unibo.it/it/ateneo/chi-siamo/open-access-e-open-science

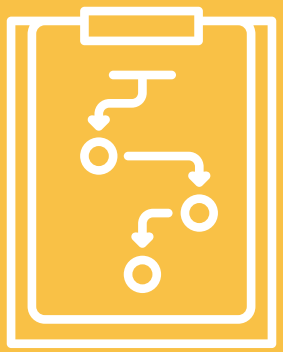
Rassegne video dell'Università di Bologna:

- "Dati: conoscerli e gestirli per valorizzare la ricerca"
  https://www.youtube.com/playlist?list=PLaUmBQ7P5K-AyDDnv1f8upAyEOtAF2gj3
- "Open Access e Open Science"
  https://www.youtube.com/playlist?list=PLaUmBQ7P5K-A83TlY96DyUl6t3rCryRK_

Video "Dati: conoscerli e gestirli per valorizzare la ricerca. Il Research Data Management"
https://www.youtube.com/watch?v=5WZF00pVSrY&list=PLaUmBQ7P5K-AyDDnv1f8upAyEOtAF2gj3&index=2

Research Data Management Decision Tree
https://doi.org/10.5281/zenodo.7190005

PIANIFICAZIONE DEL FLUSSO DATI



Principali azioni di questa fase

- **Identificare** i tipi di **dati**,
  - **decidendo se generarne di nuovi** e/o **riutilizzare** quelli già disponibili, prodotti da fonti esistenti;
  - avendo consapevolezza dei principi etici e delle norme privacy e di proprietà intellettuale da rispettare.
- **Identificare i metadati** fondamentali – per poter descrivere in modo significativo i dati generati o prestare attenzione ai metadati associati quando si riusano dati esistenti.
- **Pianificare** l'organizzazione strutturata dei dati in **dataset**.
- **Redigere un Data Management Plan** per tenere traccia delle proprie scelte.

Identificare i tipi di dati

I dati della ricerca sono **informazioni**, in qualsiasi formato, utilizzate nell'ambito di una specifica **attività di ricerca** e necessarie per **validarne i risultati**. Esiste un'enorme varietà di dati, e la loro identificazione e classificazione dipendono spesso dal dominio disciplinare di appartenenza **Dati della ricerca**.

In ogni caso, già all'inizio della tua ricerca:

- **Individua i tipi di dati** con cui lavorerai: per farlo, esamina ogni fase della tua ricerca per valutare **quali tipi di informazioni dovrai raccogliere e/o utilizzare** per rispondere alle tue domande di ricerca.
- Verifica sempre se **esistono dati già pubblicati** da altri ricercatori che possano essere utili per rispondere alle tue domande di ricerca.

Se produci dati nuovi:

- Annota sempre anche i metadati relativi ai dati stessi e al loro processo di creazione [vedi Identificare i metadati fondamentali].
- Se la ricerca coinvolge persone vulnerabili, animali, o tecnologie particolari (intelligenza artificiale, possibili usi militari, …), assicurati di gestire i dati in modo etico e di possedere le necessarie autorizzazioni per la loro raccolta, contattando gli appositi comitati etici quando opportuno.
- Nel caso i tuoi dati includano dati personali, assicurati di prendere le precauzioni necessarie a gestirli **Il rispetto della privacy**.
- Valuta se esistono altre ragioni per tenere i tuoi dati riservati, ad esempio la valorizzazione commerciale di un risultato collegato o vincoli assunti nei confronti di soggetti terzi.

Se riutilizzi dati esistenti:

- Controlla bene i metadati loro associati, che contengono informazioni fondamentali per il loro corretto riutilizzo [vedi Identificare i metadati fondamentali].
- Attraverso i metadati scopri anche se puoi riutilizzare i dati per i tuoi scopi, o se ci sono limitazioni imposte dagli autori **Diritto d'autore**.
- Assicurati sempre di riutilizzare i dati nel rispetto dei principi etici e delle norme privacy **Il rispetto della privacy** e di proprietà intellettuale.

Link utili

"Data steward all'Università di Bologna" https://youtu.be/6Ic0isyefs8?si=JqCFzHIh2Hz1U2ee

Materiali di approfondimento sui dati:

- The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-data
- Research Data Alliance https://www.rd-alliance.org/
- Digital Curation Centre (DCC) https://www.dcc.ac.uk/

Identificare i metadati fondamentali

I metadati sono **informazioni strutturate** che accompagnano i dati della ricerca [vedi Identificare i tipi di dati]. Questi "dati sui dati" servono a facilitarne la comprensione e il riutilizzo, a favorire l'identificazione e l'indicizzazione dei dati da parte di motori di ricerca e portali di aggregazione **Metadati e documentazione**.

I metadati sono in genere strutturati secondo **schemi standard**, in molti casi specifici per **ambito disciplinare**, consolidati a livello delle comunità di ricerca e implementati dalle infrastrutture di archiviazione e accesso a lungo termine **I repository**. Uno schema di metadati che adotta un vocabolario controllato **Utilizzare diversi tipi di standard** garantisce l'intellegibilità e la comprensione delle informazioni.

Già all'inizio della tua ricerca:

- Individua il prima possibile lo **schema di metadati** più adatto alla tua ricerca. Ricorda che questa scelta può essere condizionata dal repository dove archivierai a lungo termine i tuoi dataset.
- Identifica **le informazioni** richieste dallo schema di metadati che utilizzerai e assicurati di raccoglierle durante tutto il corso della ricerca.
- Dove possibile, ricordati di utilizzare **vocabolari controllati**.
- Al momento del deposito nel repository, non dimenticare di inserire tra i metadati queste informazioni fondamentali: **i nomi e le affiliazioni** (i.e., Università di Bologna) di tutti coloro che hanno collaborato alla creazione del dataset (per documentare adeguatamente l'autorialità), **il Persistent Identifier**, spesso abbreviato in PID (i.e., il Digital Object Identifier, DOI) e **il tipo di licenza** associati al dataset (vedi anche Depositare i dati secondo i principi FAIR).
- Ricorda che tutte le informazioni necessarie agli altri per comprendere i tuoi dati, se non possono essere strutturate nei metadati, devono essere associate ai dataset sotto forma di **documentazione** aggiuntiva [vedi Raccogliere la documentazione].

Link utili

Videopillola "5 Minute Metadata- What is metadata?" https://www.youtube.com/watch?v=L0vOg18ncWE

Risorse per la ricerca di schemi di metadati:

- FAIR Sharing standard registry https://fairsharing.org/search?fairsharingRegistry=Standard
- RDA metadata standards directory https://rd-alliance.github.io/metadata-directory/standards/
- DCC guidance on disciplinary metadata https://www.dcc.ac.uk/guidance/standards/metadata

Pianificare l'organizzazione dei dati in dataset

Un dataset è un **insieme strutturato di dati** (vedi Identificare i tipi di dati) in **relazione** tra loro, in qualsiasi formato, creati e/o raccolti con uno scopo comune (i.e., rispondere alla stessa domanda di ricerca) e **organizzati** per riflettere i risultati di un'attività di ricerca **Approfondimento sui dataset**. L'attenzione e la cura nella gestione dei dataset sono cruciali per garantire la qualità e l'utilità dei risultati della ricerca. Un dataset ben organizzato e gestito correttamente deve essere sempre accompagnato da metadati informativi (vedi Identificare i metadati fondamentali).

Già all'inizio della tua ricerca:

- **Organizza** ed eventualmente suddividi i tuoi dati in modo strutturato.
- Adotta una **chiara nomenclatura** per file e cartelle che comporranno il tuo dataset, così da rendere chiari i contenuti e le relazioni tra elementi.
- Inizia a r**accogliere la documentazione** necessaria per rendere il dataset il più comprensibile e riutilizzabile possibile (vedi Raccogliere la documentazione).
- Comincia a riflettere su quale potrebbe essere il repository più adatto per **archiviare il tuo dataset** a lungo termine (vedi Scegliere il repository più appropriato). Il contenuto e le dimensioni del dataset potrebbero limitarti nella tua scelta.

Redigere un Data Management Plan

Spesso abbreviato in DMP, il Data Management Plan (o piano di gestione dei dati) è lo strumento fondamentale per **documentare tutte le scelte relative alla gestione dei dati** del progetto. Solitamente consiste in un documento di testo, ed è una buona pratica che accompagni e testimoni la gestione dei dati **fin dalle fasi iniziali** del processo. Molti enti finanziatori richiedono un DMP da contratto.

- Inizia a scrivere il DMP appena inizi a ragionare su come gestirai i tuoi dati.
- Ricorda che la gestione dei dati **evolve durante il corso della ricerca** e che le valutazioni fatte inizialmente potrebbero non essere definitive.
- Ricorda che, per essere davvero utile, il DMP dovrà **essere tenuto aggiornato**.

All'interno del documento:

- Fornisci una **panoramica dettagliata dei tuoi dati di ricerca**, sia nuovi che riutilizzati (vedi Identificare i tipi di dati).
- Specifica **metodologie, strumenti e software** utilizzati per raccogliere, creare o analizzare i tuoi dati **Gestire il software**.
- Indica le strategie che utilizzerai per assicurare la **qualità dei dati** ed evitare possibili imprecisioni o incoerenze (vedi Assicurare la qualità dei dati).
- Riporta le **strategie di conservazione sicura** dei tuoi dati, ad esempio in termini di condivisione con i tuoi collaboratori o di creazione di copie di backup (vedi Conservare i dati in appositi spazi di storage).
- Descrivi le **strategie di archiviazione a lungo termine** (vedi Valutare quali dati è necessario depositare e Scegliere il repository più appropriato) e come hai applicato i **principi FAIR** ai tuoi dati (vedi Depositare i dati secondo i principi FAIR).
- Descrivi **i ruoli e le responsabilità** all'interno del team di ricerca.
- Documenta **i costi**, anche in termini di tempo, associati alla gestione dei tuoi dati.
- Tratta gli aspetti di gestione dei dati legati alla **privacy** **Il rispetto della privacy**, ai **diritti di proprietà intellettuale** **Diritto d'autore** e all'**etica**.

Link utili

Video "Dati conoscerli e gestirli per valorizzare la ricerca. Il Data Management Plan"
https://www.youtube.com/watch?v=SlOSrQdrhtQ
Video "Dati della ricerca: Il Data Management Plan" https://www.youtube.com/watch?v=QnHzMpib2Hc
Linee guida per compilare un DMP: Science Europe templates e linee guida
https://scienceeurope.org/our-priorities/research-data/research-data-management/
Strumenti per la compilazione online del DMP:

- Elixir Data Stewardship Wizard https://ds-wizard.org/
- DCC DMPonline https://dmponline.dcc.ac.uk/
- ARGOS https://argos.openaire.eu/home

Template di Data Management Plan:

- Horizon Europe Data Management Plan Template
  https://www.openaire.eu/images/Guides/HORIZON_EUROPE_Data-Management-Plan-Template.pdf
- Science Europe Data Management Plan Template
  https://scienceeurope.org/media/411km040/se-rdm-template-3-researcher-guidance-for-data-management-plans.docx

PRODUZIONE, RACCOLTA E ANALISI

Principali azioni di questa fase

- **Conservare** i dati in appositi **spazi di storage**, provvedendo ai necessari **backup**.
- **Assicurare** la **qualità** dei dati attraverso dei processi metodici e
  - **tenendo traccia** delle diverse **versioni** dei files;
  - **organizzando file e cartelle** in modo gerarchico, assegnando loro nomi coerenti;
  - **scegliendo i formati** più adatti per i dati, possibilmente standard e aperti, per favorire interoperabilità e riusabilità.
- **Raccogliere** come **documentazione** tutte le informazioni necessarie a spiegare e comprendere i dati.

Conservare i dati in appositi spazi di storage

Con **storage** intendiamo la conservazione dei dati (vedi Identificare i tipi di dati) a **breve o medio termine**, durante le fasi attive del processo di ricerca. Quando parliamo di **sistemi di storage** ci riferiamo ad esempio ad hard drive esterni, servizi cloud, oppure server.

Durante la tua ricerca:

- Pianifica precocemente in che modo conserverai i tuoi dati, decidendo **dove salvarli** e **preventivando eventuali spese** da sostenere.
- Ricorda che la scelta del sistema di storage più adatta dipende dalla **natura dei dati, dal loro volume** e dalla frequenza con cui persone diverse si trovano a **collaborare** sugli stessi file.
- Crea periodicamente delle **copie di backup** dei dati, su supporti diversi, per evitare di perderli.
- Assicurati di proteggere i dati, utilizzando delle **password efficaci** e aggiornando periodicamente gli **antivirus** del computer.
- Ricorda che la **gestione e conservazione dei dati personali** richiede un livello ulteriore di protezione, per cui puoi utilizzare una password specifica o ricorrere a programmi che criptano file e cartelle **Il rispetto della privacy**.
- Se devi condividere dati col tuo gruppo di ricerca, utilizza piattaforme di storage che permettano l'accesso da remoto, stabilendo da subito i **diritti di accesso di ognuno a file e cartelle**.
- Assicurati che l'infrastruttura di storage che hai scelto ti permetta di verificare quali **modifiche** sono state apportate ai dati, da chi sono state apportate e di poter eventualmente recuperare le **versioni precedenti**. Questi aspetti sono cruciali in un contesto di ricerca collaborativa.

Link utili

Videopillola "Dati conoscerli e gestirli per valorizzare la ricerca. Salvare e condividere i dati"
https://youtu.be/VQ0yK0tQ1N4?feature=shared.

Materiali di approfondimento sulla scelta del sistema di storage:

- The Research Data Management toolkit
  https://rdmkit.elixir-europe.org/storage#what-features-do-you-need-in-a-storage-solution-when-collecting-data
- The Turing Way Guide for Reproducible Research
  https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-storage

Strumenti per la criptazione di file e cartelle:

- Veracrypt https://www.veracrypt.fr/en/Home.html
- BitLocker https://docs.microsoft.com/it-it/windows/security/information-protection/bitlocker/bitlocker-overview

Strumenti per la valutazione dei costi del data management:

- OpenAIRE RDM cost calculator https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs
- UK data service RDM cost calculator https://ukdataservice.ac.uk/app/uploads/costingtool.pdf

Assicurare la qualità dei dati

Al di là delle peculiarità di ogni disciplina, a un livello basilare la qualità dei dati (vedi Identificare i tipi di dati) è assicurata attraverso un **insieme di processi metodici** che ne permettono la rintracciabilità, l'uso corretto e il riuso da parte di altri.

Durante la tua ricerca:

- Scegli un'**organizzazione chiara delle cartelle e dei file** e tieni traccia delle loro diverse versioni.
- Scegli uno **standard per la nomenclatura** dei file che sia chiaro e leggibile. Ad esempio, è bene riportare il nome dell'autore o la provenienza dei dati, la data di creazione e il numero di versione, evitando spazi o caratteri speciali (es. FocusGroup1_20240502_v2.rtf).
- Scegli i **formati più adatti** per i dati, possibilmente standard e aperti, per favorire interoperabilità e riusabilità **Dati della ricerca: tipologie, formati, metodi**.
- **Valida e verifica i tuoi dati** per evitare che siano imprecisi, incompleti o incoerenti. Alcune strategie comprendono la convalida dell'immissione dei dati, il controllo dell'intervallo di dati, la rimozione/registrazione di variabili imprecise o mancanti, il controllo di scale di dati coerenti.
- Definisci una **metodologia e dei processi standard** per l'analisi e l'elaborazione dei dati, specialmente se fai ricerca in contesti collaborativi. Ad esempio, puoi definire quali dati salvare in quali cartelle, in quale fase documentare i dati, o con quali sistemi condividerli con i tuoi collaboratori (vedi Conservare i dati in appositi spazi di storage).
- Se i tuoi dati sono collegati ad altri risultati della ricerca, descrivi nella documentazione come sono collegati e fornisci la **citazione completa del materiale collegato** (es. un dataset derivato da un altro dataset già esistente) (vedi Raccogliere la documentazione).

Link utili

Materiali di approfondimento sulle strategie di controllo qualità dei dati:

- The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-data-curation
- The Research Data Management toolkit https://rdmkit.elixir-europe.org/data_quality

Strumenti per la definizione della nomenclatura:

- Bulk Rename Utility (Free File Renaming Utility for Windows) https://www.bulkrenameutility.co.uk/
- File naming conventions https://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf

Strumenti per la standardizzazione e condivisione delle metodologie:

- Protocol manager https://protocols.io

Raccogliere la documentazione

I dati e i dataset (vedi Pianificare l'organizzazione dei dati in dataset), per essere **intelligibili e interpretabili da persone diverse** da coloro che li hanno generati, devono essere accompagnati da documentazione a corollario. Un esempio di documentazione è il README file, un documento testuale, leggibile e interpretabile dagli individui ("human readable") che **spiega i dati e la loro organizzazione** all'interno del dataset.

Durante la tua ricerca:

- Cura la scrittura della documentazione durante tutte le fasi attive della raccolta e dell'analisi dei dati.
- Documenta i **dati** che compongono il dataset, le **relazioni** che li legano, la loro **provenienza**.
- Inserisci informazioni esaustive circa le **metodologie** (declinate sia in termini di protocolli che di specifiche tecniche e di eventuali strumenti utilizzati) **che sono state applicate** per la raccolta e/o riutilizzo e/o generazione dei dati.
- Documenta i **processi di garanzia della qualità** nella generazione/analisi dati (vedi Assicurare la qualità dei dati).
- Inserisci informazioni su eventuali **strumentazioni o software necessari** per aprire, leggere o interpretare i dati stessi **Gestire il software**.
- **Archivia** la documentazione **insieme ai dati** nel momento in cui questi vengono depositati in un repository (vedi Valutare quali dati è necessario depositare).
- Salva il file di documentazione in un formato aperto e accessibile (es .rtf, .md).

Link utili

CESSDA Data Management Expert Guide. Documentation and metadata
https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

Utrecht University. Research data management support Guides. Metadata and documentation
https://www.uu.nl/en/research/research-data-management/guides/during-research/metadata-and-documentation

PRESERVAZIONE E CONDIVISIONE

Principali azioni di questa fase

- **Valutare quali dati è necessario depositare** per garantire la comprensione, verifica e riproducibilità della ricerca.
- **Scegliere il repository** più appropriato per il deposito dei propri dati, per garantirne la preservazione e la condivisione a lungo termine.
- **Depositare i dati**, organizzati in dataset, secondo **i principi FAIR**.
- **Associare** ai propri dati **una licenza** che ne garantisca, dove possibile, **il più ampio riutilizzo**, qualora non sussistano vincoli con terzi o strategie di valorizzazione che possano limitare questa scelta.

Valutare quali dati è necessario depositare

Il deposito indica l'archiviazione dei dati, **organizzati in dataset** (vedi Pianificare l'organizzazione dei dati in dataset), in un'infrastruttura digitale pensata per la loro preservazione a lungo termine, chiamata **repository** (vedi Scegliere il repository più appropriato). Può avvenire al termine di un'attività di ricerca, ma deve comunque precedere la pubblicazione dei risultati in un articolo scientifico.

Il deposito garantisce la **preservazione dei dati** e la loro **visibilità** ben oltre la fine del progetto di ricerca che li ha generati.

Durante la tua ricerca:

- Opera una **selezione** su quali dati depositare per permettere la validazione delle conclusioni e la riproducibilità della ricerca. Ricorda che la **responsabilità del deposito** dei dati è in capo al ricercatore che li genera.
- **Ricorda di depositare** dataset o software originali, dati grezzi ottenuti dall'analisi di campioni fisici, e dati osservativi che non possono essere riprodotti **Gestire il software**.
- **Non depositare necessariamente** dati facilmente riottenibili, oppure che siano troppo voluminosi rispetto alla loro effettiva utilità.
- **Non depositare** dati che sono già disponibili, ad esempio dati che stai riutilizzando perché sono stati già depositati da qualcun altro.
- **Documenta** accuratamente in ogni caso la **provenienza** dei dati che depositi, insieme alle **metodologie** con cui sono stati prodotti e gestiti (vedi Raccogliere la documentazione).
- Puoi depositare anche i dati che devono rimanere inaccessibili a terzi per motivi di privacy, etica o proprietà intellettuale **Diritto d'autore**, avendo cura di scegliere un repository adatto.

Link utili

Video "Dati conoscerli e gestirli per valorizzare la ricerca. Conservare i dati a lungo termine"
https://www.youtube.com/watch?v=J3VyrUzzj_E

Materiali di approfondimento sulla preservazione a lungo termine dei dati:

- The Research Data Management toolkit https://rdmkit.elixir-europe.org/data_publication
- Stanford University Library guidelines on data management and sharing https://laneguides.stanford.edu/DataManagement/
- Digital Curation Centre, How to Appraise and Select Research Data for Curation https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data

Scegliere il repository più appropriato

I repository sono **infrastrutture che garantiscono l'archiviazione a lungo termine** dei dataset (vedi Pianificare l'organizzazione dei dati in dataset). Possono essere di natura disciplinare, istituzionale o generalista. Possono essere ufficialmente certificati o meno, ma se sono affidabili assegnano un **PID** (come, ad esempio, un DOI) e permettono l'associazione di **metadati** (vedi Identificare i metadati fondamentali) e licenze **Diritto d'autore** a ciascun dataset caricato (vedi anche Depositare i dati secondo i principi FAIR).

Durante la tua ricerca:

- Usa dei **registri** (vedi link utili) per cercare uno o più repository che si adattino alle tue necessità e alle tipologie di dato che vuoi depositare.
- Scopri se c'è un repository **specifico per il tuo ambito di ricerca**. Un repository disciplinare ti permette di descrivere i tuoi dati usando uno schema di metadati specifico per la materia e li rende più visibili alla tua comunità scientifica di riferimento.
- Verifica se la tua **istituzione di appartenenza** mette uno o più repository a disposizione dei propri membri. L'Università di Bologna offre i repository istituzionali AMS Acta e AMS Historica, per i quali garantisce anche un servizio di supporto, validazione e curatela dei dati **I repository**.
- Ricorda che esistono anche repository **generalisti** (come ad esempio Zenodo) che tendono a raccogliere dati e materiali eterogenei.
- Scegli un repository che abbia un profilo di sicurezza adeguato e consenta un **accesso controllato** se i tuoi dati devono rimanere inaccessibili a terzi per motivi di privacy **Il rispetto della privacy**, etica o proprietà intellettuale.
- Ricorda che i servizi di cloud storage (vedi Conservare i dati in appositi spazi di storage), i siti web personali o di progetto e le piattaforme di social networking come ResearchGate e Academia.edu non sono repository perché non garantiscono la preservazione dei dati nel tempo.
- Presta attenzione a editori o singole riviste che, soprattutto in alcune discipline, si stanno dotando di repository su cui consigliano (o, in alcuni casi, obbligano) gli autori a depositare i propri dati. In queste situazioni, è sempre consigliabile pubblicare preventivamente i propri dati (anche) su un altro repository, disciplinare, istituzionale o generalista.

Link utili

Video "Dati conoscerli e gestirli per valorizzare la ricerca. Conservare i dati a lungo termine"
https://www.youtube.com/watch?v=J3VyrUzzj_E

Registri per l'individuazione dei repository:

- Re3data https://www.re3data.org/search?query=
- OpenAIRE explore https://www.openaire.eu/find-trustworthy-data-repository
- FAIRsharing repository database https://fairsharing.org/search?fairsharingRegistry=Database

Repository di ateneo:
AMS Acta https://amsacta.unibo.it/ | AMS Historica https://historica.unibo.it/

Per maggiori informazioni sui repository di Ateneo: "Preservare e disseminare i dati della ricerca in AMS Acta" (https://sba.unibo.it/it/almadl/servizi-almadl/preservare-disseminare-dati-della-ricerca-in-ams-acta); "Preservare e valorizzare il patrimonio culturale digitale" (https://sba.unibo.it/it/almadl/servizi-almadl/preservare-valorizzare-patrimonio-culturale-digitale).

Repository generalisti: Zenodo https://zenodo.org/ | Figshare https://figshare.com/ | Open Science Framework https://osf.io/

Depositare i dati secondo i principi FAIR

Il deposito o archiviazione a lungo termine dei dati è parte integrante di una gestione responsabile dei dati di ricerca in linea coi principi FAIR.

I principi FAIR, pubblicati per la prima volta nel 2016, sono dei suggerimenti improntati a migliorare la **riusabilità dei dati di ricerca**, da parte degli individui e dei sistemi informatici. Gestire i dati in linea coi principi FAIR significa renderli **Rintracciabili** ("**Findable**" in inglese), **Accessibili, Interoperabili e Riutilizzabili** **I principi FAIR**.

Durante la tua ricerca:

- Depositare i tuoi dati in un **repository** **I repository** è il primo passo per renderli FAIR: in questo modo, ogni dataset (vedi Pianificare l'organizzazione dei dati in dataset) è accompagnato da metadati (vedi Identificare i metadati fondamentali), tra cui un identificativo persistente (PID) e una licenza **Diritto d'autore**.
- Scegli **formati standard e aperti** **Dati della ricerca: tipologie, formati, metodi** e utilizza, dove possibile, vocabolari, ontologie e tassonomie **Utilizzare diversi tipi di standard** per rendere i tuoi dati comprensibili, interoperabili e riutilizzabili.
- Se pubblichi i dati seguendo i principi FAIR rendi **maggiormente citabili e valorizzabili** sia i dati in sé che le analisi e le pubblicazioni che ne derivano e rendi la tua ricerca più **trasparente e verificabile**, in linea con quanto richiesto da sempre più enti finanziatori, come ad esempio l'Unione Europea.
- I dati FAIR non sono sempre liberamente accessibili a chiunque mentre i metadati, solitamente, lo sono (vedi Associare ai dati una licenza).

Link utili

Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Video "Dati della ricerca: la European Open Science Cloud e i principi FAIR" https://www.youtube.com/watch?v=eNiHNaU6MrQ

Materiali di approfondimento sui principi FAIR:

- GOFAIR. FAIR Principles https://www.go-fair.org/fair-principles
- FAIRsFAIR Fostering Fair Data Practices in Europe https://www.fairsfair.eu/
- How to FAIR https://howtofair.dk/

Associare ai dati una licenza

Scegliere una licenza da associare ai dataset depositati in linea con i principi FAIR permette di specificare quali sono le modalità con cui questi dataset possono essere riutilizzati.

La pubblicazione dei dati ad accesso aperto favorisce una **ricerca aperta e collaborativa** ed è una pratica che si colloca all'interno del movimento dell'Open Science (Scienza Aperta). I dati ad accesso aperto (Open Data) sono **distribuiti con licenze che ne garantiscono la libertà di accesso, utilizzo, modifica e condivisione** da parte di chiunque, prevedendo al massimo restrizioni che riconoscano l'attribuzione di paternità e limitino alcune utilizzazioni in casi specifici, preservandone l'apertura **Diritto d'autore**.

Open Science è un movimento che ha per obiettivo **l'accesso senza barriere al sapere scientifico da parte della comunità scientifica e dei cittadini**, si basa sui principi di trasparenza, inclusione, correttezza, equità e condivisione. È obiettivo strategico dell'Unione Europea dal 2015, dell'UNESCO dal 2021 e del Ministero della Ricerca italiano dal 2022 col *Piano nazionale della Scienza Aperta*.

È sempre fondamentale ricordare che la scelta della licenza deve essere guidata dal principio "as open as possible, as closed as necessary", avendo consapevolezza che in alcuni casi è opportuno limitare l'accesso ai dati qualora ciò risulti funzionale alla valorizzazione della ricerca per finalità di natura commerciale, in linea con la missione dell'Ateneo di favorire il trasferimento dei risultati di ricerca per ottenere un impatto sull'economia e sulla società.

Durante la tua ricerca:

- Gestisci i tuoi dati secondo i **principi FAIR** **I principi FAIR** durante tutto il loro ciclo di vita e scrivi un **Data Management Plan** (vedi Redigere un Data Management Plan).
- Distribuisci i tuoi dati ad accesso aperto **se non sussistono vincoli** derivanti da diritti di terze parti o altri divieti di legge, e se questo non pregiudica le opportunità di valorizzazione commerciale dei risultati di ricerca.
- Per fare Scienza Aperta e distribuire i tuoi dati apertamente, scegli **licenze molto permissive**, ossia che ne consentano qualsiasi uso, con qualsiasi mezzo e formato e per qualsiasi fine, anche di natura commerciale. Esempi di licenze di questo tipo sono CC0 1.0, CC BY 4.0, CC BY-SA 4.0.

Link utili

Video "I Principi dell'Open Science" https://www.youtube.com/watch?v=qzktH6YlOf8

Pagina di Ateneo su Open Science https://www.unibo.it/it/ricerca/open-science

The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/open/open-data

Open Definition "Defining Open in Open Data, Open Content and Open Knowledge" https://opendefinition.org/od/2.1/en/

Per maggiori informazioni sul diritto d'autore e sulla tutela e valorizzazione del patrimonio culturale: https://sba.unibo.it/it/almadl/servizi-almadl/supporto-giuridico-per-la-gestione-del-diritto-dautore-e-la-tutela-del-patrimonio-culturale

Conclusioni

La gestione dei dati della ricerca è un insieme di buone pratiche per valorizzare i dati durante tutto il loro ciclo di vita, dalle fasi iniziali fino al deposito e alla condivisione.

Gestire correttamente i dati presenta una serie di vantaggi per il ricercatore in termini di valorizzazione, qualità e impatto della ricerca.

Il processo di gestione dei dati della ricerca si articola, come abbiamo visto, su azioni ben definite: affrontare ciascun punto con consapevolezza e al momento giusto è una condizione fondamentale per assicurare un processo corretto e di qualità.

Punto di partenza imprescindibile per percorrere queste azioni è l'identificazione consapevole dei tipi di dati con cui si lavora. Ogni ricerca si basa su dati, diversi per tipologie, origine e utilizzo. Nel rispetto dunque delle diversità disciplinari, il processo di gestione del dato consta di fasi e azioni che sono standardizzabili. Ogni area disciplinare ha le sue specificità nella definizione della tipologia di dati prodotti e nelle strategie di gestione.

Contatti

All'interno dell'Università di Bologna sono presenti diversi tipi di supporto alla gestione corretta dei dati della ricerca.

Per supporto riguardo la gestione dei dati potete rivolgervi ai data stewards che lavorano in ARIC – area della ricerca: aric.datasteward@unibo.it.

Per supporto all'utilizzo dei repository di ateneo (AMS Acta o AMS Historica) potete rivolgervi a: almadl@unibo.it.

Per supporto in materia di diritto d'autore e di altri diritti connessi al suo esercizio e di tutela e valorizzazione del patrimonio culturale, potete rivolgervi a: almadl@unibo.it.

Per supporto riguardo la valorizzazione commerciale 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.

Checklist: i punti fondamentali per la corretta gestione del dato di ricerca.

FASE DI PIANIFICAZIONE DEL FLUSSO DATI

- [ ] **Identificare i tipi di dati**.
  - [ ] Decidere se **generare dati nuovi o riutilizzare** quelli già disponibili.
  - [ ] Avere consapevolezza **della normativa vigente**, ad esempio quella relativa al trattamento dei dati personali.
- [ ] Identificare i **metadati fondamentali**.
- [ ] Pianificare l'**organizzazione dei dati in dataset**.
- [ ] Redigere un **Data Management Plan**.

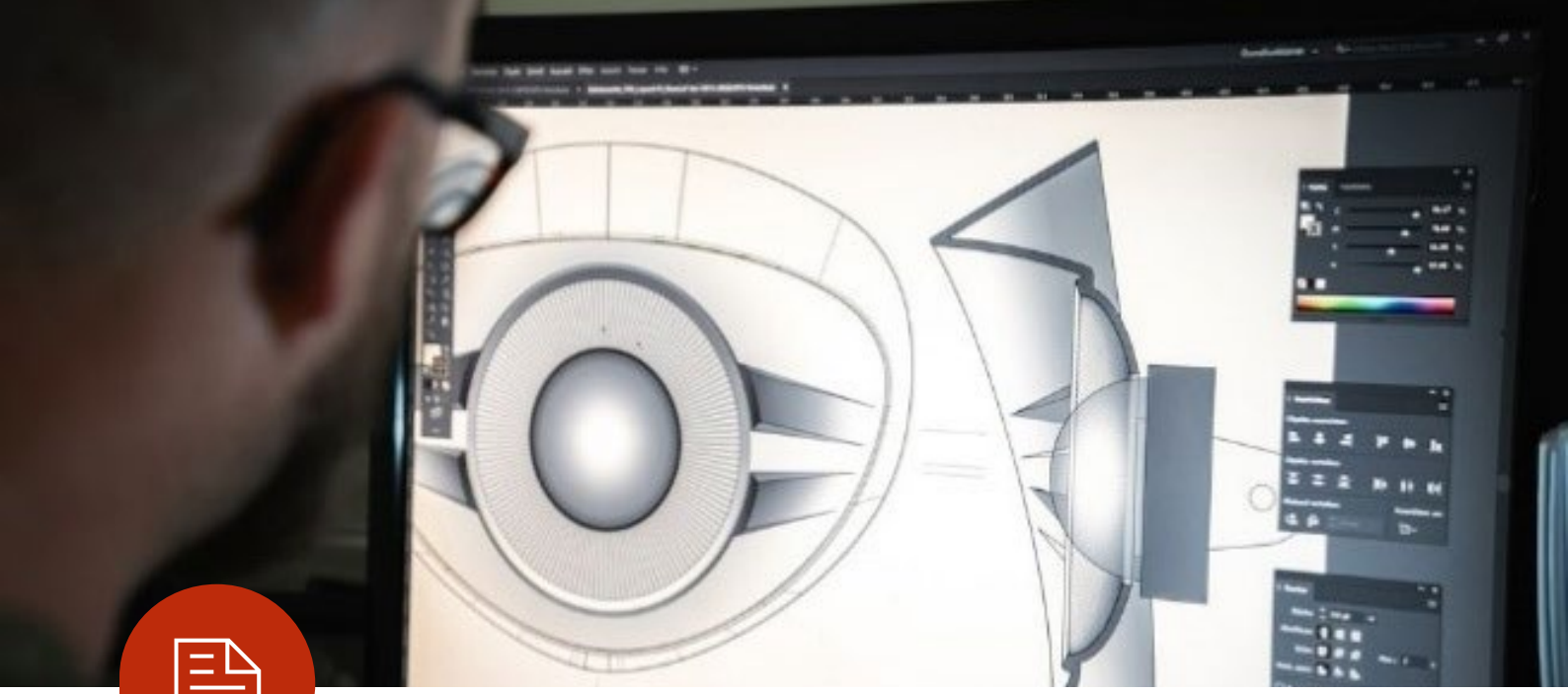
FASE DI PRODUZIONE, RACCOLTA E ANALISI DEI DATI

- [ ] Conservare i dati in **appositi spazi di storage**, con **backup**.
- [ ] Assicurare la **qualità** dei dati.
  - [ ] Tenere traccia delle diverse **versioni**.
  - [ ] **Organizzare file e cartelle** in modo coerente.
  - [ ] **Scegliere i formati** più adatti per i dati.
- [ ] **Documentare** la raccolta dei dati.

FASE DI ARCHIVIAZIONE E CONDIVISIONE DEI DATI

- [ ] Identificare i **dati da preservare a lungo termine**.
- [ ] Scegliere il **repository più appropriato** per il deposito.
- [ ] **Depositare** i dati, organizzati in dataset, secondo i **principi FAIR**.
- [ ] Associare la **licenza più appropriata** ai dati da depositare, tenendo a mente possibili vincoli esistenti.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

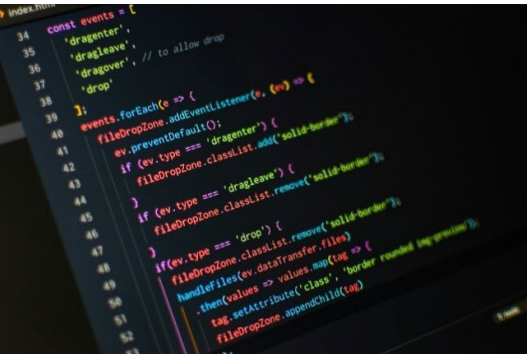


Dati della ricerca: tipologie, formati, metodi

Sono record fattuali raccolti, generati o riutilizzati nella pratica di ricerca, come base di analisi, ragionamenti, discussioni o calcoli.

Esiste un'enorme **varietà di tipi di dati**, che è possibile classificare in modi diversi. Sono esempi di dati: osservazioni, esperienze, fonti edite e inedite, riferimenti bibliografici, testi, immagini, creati e/o raccolti in formato digitale, nonché altri output digitali della ricerca come, ad esempio, modelli 3D e codice sorgente.

Le tipologie di dati variano anche in base all'**ambito** in cui si fa ricerca.



Sul campo!

***Faccio una ricerca di tipo teorico e non produco dati** – queste linee guida mi riguardano?*

Sì. Ogni tipo di ricerca produce o riusa dei dati (definiti in modo estremamente generico), anche se ogni area disciplinare ha le sue specificità. Sicuramente utilizzi delle fonti, primarie o secondarie, per rispondere alle tue domande di ricerca e produci quindi una raccolta di metadati bibliografici, organizzati in modo più o meno sistematico. In questo contesto, ricordati sempre di utilizzare i PID delle risorse che citi, e pensa a come puoi valorizzare questo risultato della tua ricerca, per esempio pubblicandolo online come open data.

Tipologie di dati: come categorizzarli

Conoscere e classificare i dati della propria ricerca permette di scegliere le strategie più adatte per gestirli consapevolmente e responsabilmente, per evitare che vengano persi o corrotti, per scegliere i metodi di raccolta, archiviazione e analisi appropriati e sicuri.

La **natura digitale, digitalizzata o non digitale** dei dati influenza la pratica di ricerca: mentre la gestione del dato digitale o digitalizzato può seguire esclusivamente dei protocolli informatizzati, le pratiche di gestione dei dati non digitali possono essere sia digitali che non.

Che siano digitali o non digitali, i dati possono essere descritti in base al **contenuto**: numerico, testuale, audiovisivo, e molti altri.

Dati con lo stesso contenuto possono avere forme diverse e quindi la loro struttura dal punto di vista digitale può cambiare. Ad esempio, dati testuali possono essere raccolti tanto nella forma di fogli di calcolo quanto nella forma di documenti di testo.

Dati con lo stesso contenuto e raccolti nella stessa forma possono avere formati (e quindi estensioni) differenti. Ad esempio, dati numerici possono essere raccolti in un foglio di calcolo che può essere scritto in formato file "comma-separated values" (CSV) con estensione del file .csv, così come in formato OpenDocument Spreadsheet (ODS), con estensione del file .ods, o ancora in formato Microsoft Excel, con estensione del file .xls o .xlsx.

Per garantire che i dati rimangano accessibili e riutilizzabili, è opportuno scegliere di raccoglierli, salvarli, condividerli e depositarli in formati aperti non proprietari, piuttosto che proprietari chiusi.

- **Formato proprietario**: di proprietà di e sviluppato da una particolare società o altra entità.
  - **Proprietario e chiuso**: chi ha sviluppato il formato detta quale software può utilizzare il formato. Ad esempio, .indd per i file del software Adobe InDesign, prodotto da Adobe e rivolto all'editoria professionale.
  - **Proprietario e aperto**: chi ha sviluppato il formato non ha ristretto i software possibili che possono utilizzarlo. Ad esempio, per i file audio esiste il formato aperto MP3 che tuttavia è soggetto a brevetto in alcuni paesi. Oppure, il formato XLS era un tempo chiuso da Microsoft, cioè eseguibile solo dal loro software proprietario Microsoft Excel, ma è stato poi aperto, come il formato `.xlsx` che, essendo basato su XML (formato aperto) può essere utilizzato anche da altri software, come LibreOffice Calc.
- **Formato non proprietario e aperto**: le specifiche del formato sono disponibili apertamente, e chiunque può creare software in grado di utilizzarlo. Ad esempio, i csv per i dati tabulari possono essere aperti da molti software diversi.

Alcuni esempi di formati aperti per le tipologie di dati più comuni sono:

- **Dati quantitativi e qualitativi tabulari**: SPSS (`.sav`), Stata (`.dta`), CSV (`.csv`);
- **Dati geospaziali, vettoriali e raster**: ESRI Shapefile (essential – `.shp`, `.shx`, `.dbf`, optional – `.prj`, `.sbx`, `.sbn`), Geo-referenced TIFF (`.tif`, `.tfw`), CAD data (`.dwg`), e Tabular GIS attribute data
- **Dati qualitativi testuali**: eXtensible Mark-up Language (XML), Rich Text Format (`.rtf`), Plain text data, ASCII (`.txt`). Accettato anche MS Word (`.doc` /`.docx`).

- **Immagini, audio e video**: TIFF (`.tif`, `.tiff`), JPEG (`.jpeg`, `.jpg`), Adobe Portable Document Format (PDF/A, PDF) (`.pdf`), PNG (`.png`), Free Lossless Audio Codec (FLAC) (`.flac`), MPEG-1 Audio Layer 3 (`.mp3`), Audio Interchange File Format (`.aif`), Waveform Audio Format (`.wav`), MPEG-4 (`.mp4`), MOV (`.mov`), Windows Media Video (WMV) (`.wmv`).

Sul campo!

Sono un ricercatore che lavora con dati organizzati in tabelle – quali sono i formati più utilizzati?

I formati più utilizzati per i dati tabulari sono

- "Comma Separated Values" (CSV, `.csv`): un formato testuale, non proprietario, in cui i dati sono separati solitamente da virgole.
- "OpenDocument Spreadsheet" (ODS, `.ods`): un formato standard aperto per fogli di calcolo, memorizza i dati in celle organizzate in righe e colonne. I file .ods possono anche essere aperti in Microsoft Excel e salvati come file XLS o XLSX.
- "Excel Workbook" (XLS/XLSX, `.xls`/`.xlsx`): il formato Excel, proprietario ma molto comune, che permette di creare, manipolare e analizzare dati tabulari in fogli di calcolo.

***Nella mia ricerca ho necessità di raccogliere dati attraverso survey** – quale strumento posso utilizzare?*

Le survey possono essere condotte tramite interviste o questionari di persona, telefonici o online.

A seconda della popolazione della quale si vuole estrarre un campione, della dimensione dello stesso, del disegno campionario, che può essere semplice o complesso, trasversale o longitudinale, può essere necessario integrare l'uso di queste tecniche e degli strumenti con servizi di supporto per la gestione del dato, la privacy, l'etica e/o il diritto d'autore.

Tra gli strumenti online per creare una survey ci sono: Microsoft Forms, Google Forms, LimeSurvey, SurveyMonkey, Qualtrics. Nel caso di raccolta di dati personali è necessario scegliere uno strumento come Microsoft Forms, fornito dall'Ateneo, o verificare eventuali licenze in uso presso il proprio Dipartimento (LimeSurvey, SurveyMonkey, Qualtrics) e non usare licenze personali.

In una survey trasversale per la quale non si ha bisogno di contattare la persona una seconda (o più volte) si possono adottare tecniche di privacy by-design in modo da avere dei dati anonimi alla fonte, quindi privi di problematiche privacy.

***Nella mia ricerca lavoro con dati di imaging biomedico** – come posso scegliere in quale formato salvarle e archiviarle?*

Il Digital Imaging and Communications in Medicine (DICOM) è lo standard per la comunicazione e la gestione delle informazioni di imaging medico e dei relativi dati. Un file DICOM, oltre all'immagine vera e propria, include anche una intestazione che contiene tutti i metadati acquisiti in associazione all'immagine (dati del paziente, luogo del tumore, durata e dose delle radiazioni ecc.).

Per conservare e condividere le immagini di imaging medico anche il TIFF è un formato appropriato: si tratta di un formato di file raster-grafico che supporta la compressione dei dati senza perdita di dati (lossless) e per questo adatto all'archiviazione e alla stampa di immagini e foto ad alta risoluzione. Tutti i metadati rilevanti possono essere salvati in un file a parte, in formato TXT.

Sulla raccolta dati e metodologie

Nella pratica di ricerca si possono distinguere i dati **riutilizzati**, e che quindi sono stati raccolti o generati da terzi, da quelli **generati per la prima volta**.

Riutilizzare dati esistenti, digitali o non digitali, permette di risparmiare tempo e risorse, se i dati riutilizzati sono di qualità. Esistono degli **archivi digitali online per l'archiviazione a lungo termine dei dati**, fruibili per consultazione e download dei dati che contengono, e che possono essere specifici per un'area disciplinare **I repository**. Prima di riutilizzare dei dati, qualunque sia la loro provenienza, è opportuno assicurarsi di avere i diritti e le eventuali autorizzazioni per poterli riutilizzare **Diritto d'autore** **Il rispetto della privacy**.

Generare o raccogliere i dati può comportare pratiche molto diverse tra loro. Ad esempio, i dati possono essere di natura **sperimentale**, quando ottenuti tramite esperimenti e dimostrazioni che seguono un metodo scientifico. Oppure possono essere di natura **osservativa**, quando vengono raccolti attraverso l'osservazione critica, con l'eventuale aiuto di strumenti. Quando la ricerca è **compilativa**, i dati vengono raccolti in forma derivata/compilata da altre fonti.

Indipendentemente dalle pratiche di generazione o raccolta dati, gli **strumenti**, **software e metodi utilizzati devono essere registrati per consentire la riproducibilità della ricerca** **Gestire il software**.

Inoltre, sempre a prescindere dai metodi di raccolta o generazione dei dati, è necessario assicurarsi di essere conformi alle normative sulla privacy e sull'etica.

Se hai intenzione di sfruttare commercialmente i tuoi dati, perché possono essere utili, per esempio, per depositare una domanda di brevetto, pianifica in anticipo delle strategie di gestione dei dati che possano garantirti adeguata protezione.

Sul campo!

Sviluppo software per l'analisi e la visualizzazione dei risultati della mia ricerca *– Devo gestirlo come se fosse un dato di ricerca?*

Sì, è bene pianificare lo sviluppo del software e utilizzare strumenti che possano documentare il suo sviluppo per poterlo valorizzare come asset e oggetto principale di output e studio di una ricerca, oltre a renderlo più facilmente riutilizzabile per attività di ricerca future.

Alcuni strumenti, come i cloud notebook, possono aiutarti a documentare lo sviluppo del codice e tutte gli step del suo algoritmo. Eseguendo il codice in cloud, è possibile visualizzare l'esecuzione di ogni sua parte con i rispettivi dati di input e di output.

Una volta che il codice ha raggiunto una versione stabile eseguibile, è bene depositarlo in repository disciplinari con adeguata documentazione e metadati specifici, per assicurarsi che sia conservato correttamente a lungo termine. Un esempio di repository disciplinare per il codice sorgente è Software Heritage, che usa CodeMetda come schema di metadati e fa harvesting automatico periodico dalle forge più comuni per lo sviluppo, come GitHub. **Gestire il software** **I repository per depositare i dati**.

Lavoro con il patrimonio culturale e i miei dati sono soprattutto testi ed immagini, spesso conservati in archivi, musei o biblioteche *– Come devo comportarmi?*

Prendi contatto con l'ente che ha in custodia le fonti che vuoi utilizzare nel tuo lavoro per capire cosa fare. Se le fonti con cui lavori non sono più coperte dal diritto d'autore potrebbero essere tutelate come bene culturale e potrebbe essere necessaria una specifica autorizzazione all'uso, ad esempio per la riproduzione. Nell'ipotesi in cui le fonti con cui lavori siano ancora tutelate dal diritto d'autore l'autorizzazione deve essere rilasciata dal titolare dei diritti **Diritto d'autore**.

Link utili

Materiali di approfondimento sul tema dei formati:
https://www.loc.gov/preservation/resources/rfs/TOC.html | https://www.dicomstandard.org/
https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

Strumenti utili per le interviste:
https://forms.office.com | https://www.qualtrics.com/it/ | https://www.limesurvey.org/it

Strumenti utili per il software:
https://datasciencenotebook.org/ | https://www.softwareheritage.org/
https://www.codemeta.github.io/codemeta-generator/ | https://github.com/

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dataset: selezionare e organizzare i dati

Un dataset è una **raccolta di dati organizzati in modo ordinato** e strutturati secondo criteri precisi. Il dataset deve raccogliere, insieme ai dati, anche i metadati che li descrivono, localizzano e li mettono in relazione.

Dati, dataset e metadati: un esempio dalla vita di tutti i giorni

DATI: una serie di fotografie disorganizzate

DATASET: una raccolta di fotografie raggruppate con uno scopo comune

METADATI: informazioni che descrivono ogni fotografia e l'intero album

→ le foto possono così essere rintracciate all'interno dell'album e comprese

Image credits (rispettivamente): Foto di Roman Kraft su Unsplash ; Foto di Laura Fuhrman su Unsplash ; Foto di Frelo Design su Unsplash

Il dataset rappresenta un pilastro fondamentale per la replicabilità delle analisi condotte. La sua corretta gestione e organizzazione sono cruciali per garantire l'affidabilità e l'utilità dei dati, consentendo agli altri ricercatori di esplorare, comprendere e approfondire le conoscenze.

Un dataset ben organizzato non è solo una raccolta di dati casuali, ma risponde a criteri precisi che ne garantiscono la qualità e la fruibilità. I **principi FAIR** (Findable, Accessible, Interoperable, Reusable) forniscono una guida essenziale per valutare la bontà di un dataset **I principi FAIR**.

Ogni dataset deve essere **riproducibile**, per consentire ad altri ricercatori di replicare l'analisi e validare i risultati. Deve essere **trasparente** e permettere così di comprendere il processo di ricerca e le metodologie adottate. Un dataset **accessibile** risulta facile da condividere e, quando associato a chiare condizioni di accesso, favorisce la collaborazione tra ricercatori. Infine, quando un dataset è **riutilizzabile** è disponibile per nuove ricerche e permette di ridurre duplicazioni e costi.

Selezionare i dati da inserire in un dataset

Pensare di depositare tutti i dati associati alla ricerca non è sostenibile: è necessario valutare quali dati sono necessari per garantire la comprensione, verifica e riproducibilità della ricerca. È una valutazione che può dipendere dall'ambito disciplinare ed è in capo al ricercatore stesso.

Esempi di dati da depositare assolutamente sono:

- Dataset e/o codice software originale.
- Dati grezzi ottenuti dall'analisi di campioni fisici.
- Dati osservativi che non possono essere rigenerati.
- Dataset non originali ma non facilmente disponibili (se si ha il permesso).

Organizzare i dataset

Una singola ricerca può produrre tanti tipi diversi di dati, che concorrono a rispondere alla stessa domanda scientifica. Strutturarli all'interno di un dataset permette di organizzarli e relazionarli tra loro, **rendendo chiaro il processo** che ha portato ad ottenere il risultato.

L'organizzazione dei dati in dataset, se **correttamente pianificata** all'inizio di una ricerca, ne semplifica la gestione durante tutto il ciclo di vita ed è un investimento fondamentale per il successo di un progetto di ricerca perché aumenta:

- **Efficienza**. L'organizzazione strutturata e ben definita dei dati ne facilita la ricerca e l'accesso nel momento in cui sono necessari per l'analisi, evitando perdite di tempo e frustrazione. Previene inoltre la duplicazione dei file in diverse posizioni, risparmiando spazio di archiviazione e semplifi-

cando la gestione, oltre a garantire che i membri del team possano collaborare più facilmente e tenere traccia delle modifiche apportate ai dati.

- **Riproducibilità**. Un'organizzazione chiara e documentata rende la ricerca più trasparente e riproducibile, consentendo ad altri ricercatori di comprendere le metodologie e i risultati. Inoltre garantire l'accesso ai dati grezzi e ai metadati facilita la verifica e la convalida dei risultati da parte di altri ricercatori.
- **Affidabilità**. Una accurata organizzazione protegge i dati da perdite, corruzioni o accessi non autorizzati. L'implementazione di misure di sicurezza adeguate protegge i dati sensibili da intrusioni e violazioni e l'organizzazione dei dati facilita la conformità alle normative e agli standard di dominio.

Sul campo!

Sono un ricercatore e devo capire come strutturare i miei dati in dataset *– da dove comincio?*

Inizia facendo una stima dei dati con cui intendi lavorare, che caratteristiche hanno e come si possono relazionare tra loro per rendere la tua ricerca più efficiente e comprensibile.

Organizzali in cartelle dalla nomenclatura chiara ed efficace.

Documenta accuratamente i dataset fornendo tutte le informazioni e i metadati necessari.

Individua e adotta soluzioni di storage per la salvaguardia dei dataset durante le fasi attive della tua ricerca.

Pianifica accuratamente, tenendo conto delle possibili criticità e scegliendo un repository adeguato alle tue esigenze, con un'ottica di lungo periodo orientata alla preservazione finale.

Sono un ricercatore di area sociale che vuole creare un dataset per analizzare delle variabili socio-demografiche *– Esempio*

Voglio studiare la ricchezza in Italia, per esempio i valori dei redditi nelle regioni, per arrivare alla produzione di un articolo in cui mostro graficamente i miei dati tramite mappe. Durante la fase di pianificazione decido di strutturare il mio dataset come segue:

- File di input, contengono i dati grezzi . Questi dati includono informazioni sulla popolazione e sulle unità territoriali.
- Il codice che ho utilizzato per analizzare i dati. Lo script carica i dati, pulisce i dati, esegue l'analisi e genera i file di output (per esempio usando R che è un software Open Source).
- I file di output contengono i risultati.Includono dei dati tabulari con i valori delle variabili elaborate con i dati di output, e delle immagini, ovvero le mappe che avevo deciso di creare in fase di pianificazione

L'insieme di questi dati costituisce il mio dataset, che è stato prodotto per rendere la mia ricerca più comprensibile e riproducibile.

Metadati e documentazione

I metadati sono informazioni sui dati, **descrittori** che ne facilitano la catalogazione e la reperibilità. Per gestire correttamente i dati della ricerca è di fondamentale importanza documentare tutte quelle informazioni necessarie per renderli intelligibili agli altri (i.e. a chi non ha lavorato alla loro raccolta, a chi vuole riutilizzarli, al "te" ricercatore del futuro che non ricorda tutti i dettagli della loro generazione). I metadati sono quindi una particolare forma di documentazione: raccogliendo le informazioni in maniera **strutturata**, descrivono, spiegano, localizzano e facilitano l'uso dei dati, rendendoli in particolare **leggibili dalle macchine** e facilitandone quindi la ricerca in archivi e motori di ricerca, consentendo a software/servizi di accedere ai dati, comprenderli e trasformarli, esportarli e/o importarli in altri software.

Tipologie di metadati

I metadati possono essere di diverso tipo, ad esempio:

- **Metadati strutturali**: descrittori che forniscono informazioni sull'oggetto (i.e. file, dataset…) in sé e consentono agli utenti di capire come questo è organizzato. Ad esempio, metadati strutturali descrivono il layout di una tabella o le relazioni tra gli elementi che la compongono.
- **Metadati descrittivi**: forniscono, attraverso una piccola stringa, una specifica categoria di informazioni sul contenuto di un oggetto (i.e. file, dataset…) e lo descrivono in modo minimo ed essenziale. Ad esempio, metadati descrittivi sono il nome del luogo e l'orario in cui è stata scattata una fotografia.
- **Metadati di qualità**: raccolgono elementi specifici per definire gli indicatori relativi alla qualità di un oggetto. Possono comprendere metriche quantitative, come la varianza o l'errore standard di una variabile, o aspetti più qualitativi (i.e. una classifica discreta).
- **Metadati amministrativi**: comprendono elementi come gli identificatori degli oggetti (ad esempio l'Identificativo Persistente associato a un dataset). Rientrano tra i metadati amministrativi anche i
  - **Metadati di provenienza**, documentano informazioni sugli autori e sui processi che hanno prodotto l'oggetto. Rientrano in questa categoria, ad esempio, i nomi e le affiliazioni degli autori.
  - **Metadati legali**, attraverso cui si documentano le condizioni di (ri)utilizzo dei dati. Rientrano in questa categoria, ad esempio, le informazioni sulle licenze.

TITOLO: Panorama di Bologna

LUOGO: Bologna

DATA: 01/01/2024

ESPOSIZIONE: 1/30s

Image credits: Foto di Bianca Ackermann su Unsplash

Schemi standard di metadati

I metadati sono in genere strutturati secondo schemi standard, i pacchetti di informazioni che vengono raccolti all'interno di un set di metadati sono cioè organizzati secondo **convenzioni standardizzate**. L'utilizzo di uno standard garantisce numerosi vantaggi in termini di interoperabilità in quanto permette di **associare un significato univoco e non ambiguo ad ogni elemento**, raccogliendo le informazioni minime essenziali sui dati e permettendo di confrontare dati eterogenei provenienti da più fonti, domini e discipline.

Gli schemi standard di metadati possono essere **sia generici che disciplinari**.

Gli schemi generici di metadati, come Dublin Core, raccolgono un set minimo di informazioni standard ad alto livello, tendono a essere facili da usare e sono ampiamente adottati. Quando è necessario coprire informazioni più specifiche però devono spesso essere ampliati.

Gli schemi specifici del dominio hanno un vocabolario e una struttura molto più ricchi e tendono a essere altamente specializzati, spesso comprensibili solo dai ricercatori di quel settore.

Un ulteriore possibile livello di standardizzazione si può avere a livello dei **vocabolari** da utilizzare per riempire i campi dei metadati **Utilizzare diversi tipi di standard**.

Sul campo!

Sono un ricercatore e devo trovare uno schema disciplinare di metadati per descrivere i miei dati *– da dove comincio?*

Indipendentemente dall'ambito in cui fai ricerca, puoi interrogare diversi portali online per individuare gli schemi di metadati più adatti:

1) FAIR Sharing, una risorsa curata manualmente che raccoglie standard di dati e metadati;
2) il catalogo degli standard di metadati curato dalla Research Data Alliance (RDA);
3) il catalogo del Digital Curation Centre (DCC) che raccoglie standard di metadati disciplina-specifici.

Sono un ricercatore di area biomedica e non riesco a trovare uno schema di metadati adatto per i miei dati – come posso fare?

Se fai fatica ad individuare uno schema specifico di metadati, è opportuno cercare quali sono le "informazioni minime raccomandate" da associare allo specifico tipo di dati nella specifica disciplina (Minimum Information about your topic, i.e. MIAME, MIAPE o MIAPPE).

Metadati e repository

La **scelta del repository** **I repository** in cui si decide di depositare i propri dataset influenza la scelta dello schema di metadati. Al momento del deposito viene richiesto un set di informazioni, organizzate secondo uno schema standard, che di fatto costituiscono i metadati dell'oggetto depositato.

Avere chiaro dalle prime fasi del progetto in quale repository si intende depositare i propri dataset permette di avere già chiaro in mente quali campi di metadati dovranno essere compilati. Un buon repository guida il ricercatore nell'aggiunta dei metadati giusti per collegare i dati ad altri software e sistemi.

Sul campo!

***Ho scelto il repository dove depositerò i miei dati** – come posso verificare quale schema di metadati adotta?*

Questa informazione può essere facilmente rintracciata nelle pagine di documentazione del repository o attraverso il registro re3data. AMS Acta, il repository di Ateneo, adotta lo schema Dublin Core e DataCite, così come Zenodo.

***Il repository che ho scelto per depositare i miei dati disciplinari adotta uno standard di metadati generico** – come posso aggiungere più informazione ai miei dati?*

Se il repository adotta uno schema generico, è comunque possibile aggiungere alla documentazione un file di testo con metadati specifici disciplinari. Per trovare uno schema adatto, fai riferimento al primo punto di questa sezione.

Metadati e documentazione

Se i metadati sono un tipo particolare di documentazione, strutturata e pensata per essere leggibile e interpretabile dalle macchine ("machine-readable"), parliamo di documentazione vera e propria quando ci riferiamo a dei documenti, ad esempio testuali come nel caso dei README file o dei codebooks, che sono **leggibili e interpretabili dagli individui** ("human readable") e che raccolgono tutte le informazioni ulteriori necessarie a capire ed interpretare i dati a cui fanno riferimento. Solitamente, se un README file contiene informazioni di alto livello (progetto o studio), come scopo e contesto della ricerca, fondi, metodologie (…), il codebook è pensato per documentare il significato dei nomi delle variabili, delle eventuali unità di misura (…) a livello di raccolta dati.

Il README file pensato per accompagnare un dataset deve includere:

- Informazioni generali sul progetto, come titolo e obiettivi del progetto e del dataset, nomi/ruoli/contatti dei ricercatori coinvolti, informazioni sui finanziamenti, PID (…).
- Struttura delle cartelle e dei file e sistema di denominazione, nomenclatura dei file e struttura delle cartelle, relazioni e dipendenze tra i file, descrizione dei contenuti di ogni file principale (…).
- Informazioni su metodi e software utilizzati per la raccolta dei dati (compresi riferimenti, documentazione, link, condizioni sperimentali, standard e calibrazione degli strumenti…), metodi e software utilizzati per l'elaborazione dei dati, formati dei file e descrizione del sistema di sistema di controllo di versione, procedure di controllo della qualità applicate (…).

- Informazioni sulle possibilità di riutilizzo e collegamento con altri materiali, quindi su licenze e restrizioni poste su (parti del) dataset, link a pubblicazioni basate sul dataset, relazione con altri dataset e con altre risorse utilizzate come fonte per la raccolta dei dati (libri, articoli, ecc.).

Il codebook, che può anche essere compreso dentro al README file, è pensato solitamente per documentare il significato dei nomi delle variabili e delle eventuali unità di misura. A livello di raccolta dati, è un tipo di documentazione organizzato di solito in un file tabulare e riporta:

- Definizione di codici, simboli e abbreviazioni utilizzati nei file.
- Elenco delle variabili con nome completo e definizione.
- Definizione delle intestazioni delle colonne e delle etichette delle righe per i dati tabellari.
- Unità di misura e formati dei dati (ad es. AAAAMMGG).
- Trattamento dei dati mancanti (codice, ecc.).
- Il file contenente la documentazione che descrive il dataset deve essere archiviato insieme ai dati nel momento in cui questi vengono depositati in un repository. È opportuno che il file di documentazione sia salvato in un formato aperto e accessibile (es .rtf).

Link utili

FAIR and the notion of metadata https://faircookbook.elixir-europe.org/content/recipes/introduction/metadata-fair.html

RDM kit, Documentation and metadata https://rdmkit.elixir-europe.org/metadata_management

The Turing Way Guide for Reproducible Research https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-metadata

FAIRify your data: data documentation and metadata, Flora D'Anna https://osf.io/wbr7t

Research Data Management: Metadata (University College Dublin Library) https://libguides.ucd.ie/data/metadata

Dublin Core Metadata Standard https://www.dublincore.org/specifications/dublin-core/dces/

Risorse per la ricerca di schemi di metadati:

- FAIR Sharing standard registry https://fairsharing.org/search?fairsharingRegistry=Standard
- RDA metadata standards directory https://rd-alliance.github.io/metadata-directory/standards/
- DCC guidance on disciplinary metadata https://www.dcc.ac.uk/guidance/standards/metadata

Minimum Information Standards:

- MIAME https://www.fged.org/projects/miame
- MIAPE https://www.psidev.info/miape
- MIAPPE https://www.miappe.org/

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

I principi FAIR

L'obiettivo alla base dei principi FAIR è quello di fornire alla comunità scientifica i dati, le metodologie e gli strumenti di ricerca necessari per convalidare (e in alcuni casi replicare) le conclusioni di una ricerca.

Il primo passo per (ri)utilizzare i dati è **trovarli**. Una volta trovati, l'utente deve sapere come **accedere** (autenticazione, autorizzazione...). I dati devono essere **integrati** con altri dati e devono **interoperare** con applicazioni o flussi di lavoro. L'obiettivo finale di FAIR è **ottimizzare il riutilizzo** dei dati.

I principi FAIR descrivono come i risultati della ricerca dovrebbero essere organizzati per essere più facilmente **accessibili, compresi, scambiati e riutilizzati**. Non tutti i dati gestiti correttamente sono FAIR e non tutti i dati FAIR sono dati Open.

Dati e principi FAIR

I repository, archivi che conservano i dati rendendoli persistenti nel tempo e rintracciabili in rete, **facilitano l'adesione ai principi FAIR** nel momento in cui il ricercatore vi deposita i dati. Per essere rintracciabili (Findable), infatti, i dati/dataset devono essere accompagnati da un identificatore persistente (PID), ovvero un riferimento unico e univoco di lunga durata. Esempi di PID sono DOI, Handle o URN. I dati/dataset devono inoltre essere accompagnati da metadati e parole chiave significative (contenenti il PID).

Quando si parla di dati accessibili non si parla di dati necessariamente aperti e tantomeno di dati aperti senza controllo, quanto di **dati associati a chiare condizioni di accesso**. I dati possono essere apertamente accessibili (default) o accessibili attraverso un sistema di autenticazione e autorizzazione, se la natura dei dati ne impedisce l'apertura. Privacy e protezione della pro-

prietà intellettuale sono alcune delle motivazioni che possono portare alla necessità di controllare l'accesso ai dati. In questi i casi, è opportuno che almeno i metadati associati a dati chiusi siano apertamente accessibili attraverso l'uso di protocolli standard.

I dati, per essere interoperabili, devono essere combinabili e utilizzabili con altri dati o strumenti. Questo si ottiene utilizzando **formati di dati aperti e interoperabili** da vari strumenti. I dati, così come i metadati, devono utilizzare un **linguaggio standardizzato e condiviso** a livello internazionale dai diversi servizi di indicizzazione. I repository **I repository** supportano **standard per l'interoperabilità** e possono raccomandare l'uso di **ontologie o vocabolari specifici** **Utilizzare diversi tipi di standard**.

Per garantirne la massima riusabilità, i dati devono essere **descritti e documentati** nel miglior modo possibile, per garantirne la qualità e per poter essere replicati e combinati in contesti diversi. L'elaborazione dei dati deve essere conforme agli standard riconosciuti dalle comunità scientifiche di riferimento.

Infine, ma non meno importante, accompagnare un dataset con una **licenza chiara ed accessibile, possibilmente aperta**, è fondamentale per stabilire e dichiarare le possibilità di riutilizzo **Diritto d'autore**.

Metadati e principi FAIR

I principi FAIR si applicano tanto ai dati quanto ai metadati.

Metadati ricchi e descrittivi hanno un profondo impatto sulla riusabilità dei dati, poiché ne migliorano la reperibilità, l'interoperabilità e la riutilizzabilità. Per rispettare i Principi FAIR, i metadati dovrebbero essere accessibili e associati ad una licenza permissiva (CC0 o equivalenti) anche se i dati non lo sono **Il rispetto della privacy**.

Sul campo!

Come posso valutare quanto sono FAIR i miei dati?

Per facilitare l'autovalutazione nel rispetto dei principi FAIR nella gestione dei dati della ricerca, proponiamo una checklist elaborata da EUDAT e tradotta da AlmaDL.

Findable / Rintracciabili

- ☐ È stato assegnato un identificatore persistente (es. DOI, Handle, URN) al dataset?
- ☐ Il dataset è stato descritto con metadati esaustivi, informativi e accurati?
- ☐ I metadati sono registrati in un catalogo online o in un data repository che sia indicizzato dai motori di ricerca?
- ☐ Fra i metadati è incluso anche l'identificatore persistente assegnato al dataset?

Accessible / Accessibili

- ☐ L'identificatore persistente associato al dataset risolve correttamente alla pagina dei matadati del dataset?
- ☐ Il protocollo di recupero dei dati e dei metadati rispetta un linguaggio standardizzato e riconosciuto come ad esempio quello delle pagine web (HTTP)?
- ☐ I metadati sono sempre pubblici, visibili e indicizzabili anche se i dati non sono in open access o non lo sono più?

Interoperable / Interoperabili

- ☐ I dati sono resi disponibili in formati aperti o almeno in formati documentati e diffusi?
- ☐ I metadati seguono schemi standard riconosciuti e condivisi?
- ☐ Sono stati utilizzati quanto più possibile vocabolari controllati tesauri o ontologie?
- ☐ Sono resi disponibili link o relazioni con altre risorse rilevanti per la comprensione dei dati come pubblicazioni o rapporti tecnici o applicazioni software?

Re-usable / Riutilizzabili

- ☐ I dati sono accurati, completi e descritti in modo che siano facilmente comprensibili e riproducibili?
- ☐ Al dataset è stata attribuita una licenza che ne specifica le possibilità di riutilizzo?
- ☐ Sono chiare dai metadati e dalla documentazione allegata le responsabilità scientifiche e finalità dei dati prodotti?
- ☐ I dati e i metadati rispettano gli standard e i protocolli di qualità del dominio di ricerca di riferimento?

Link utili

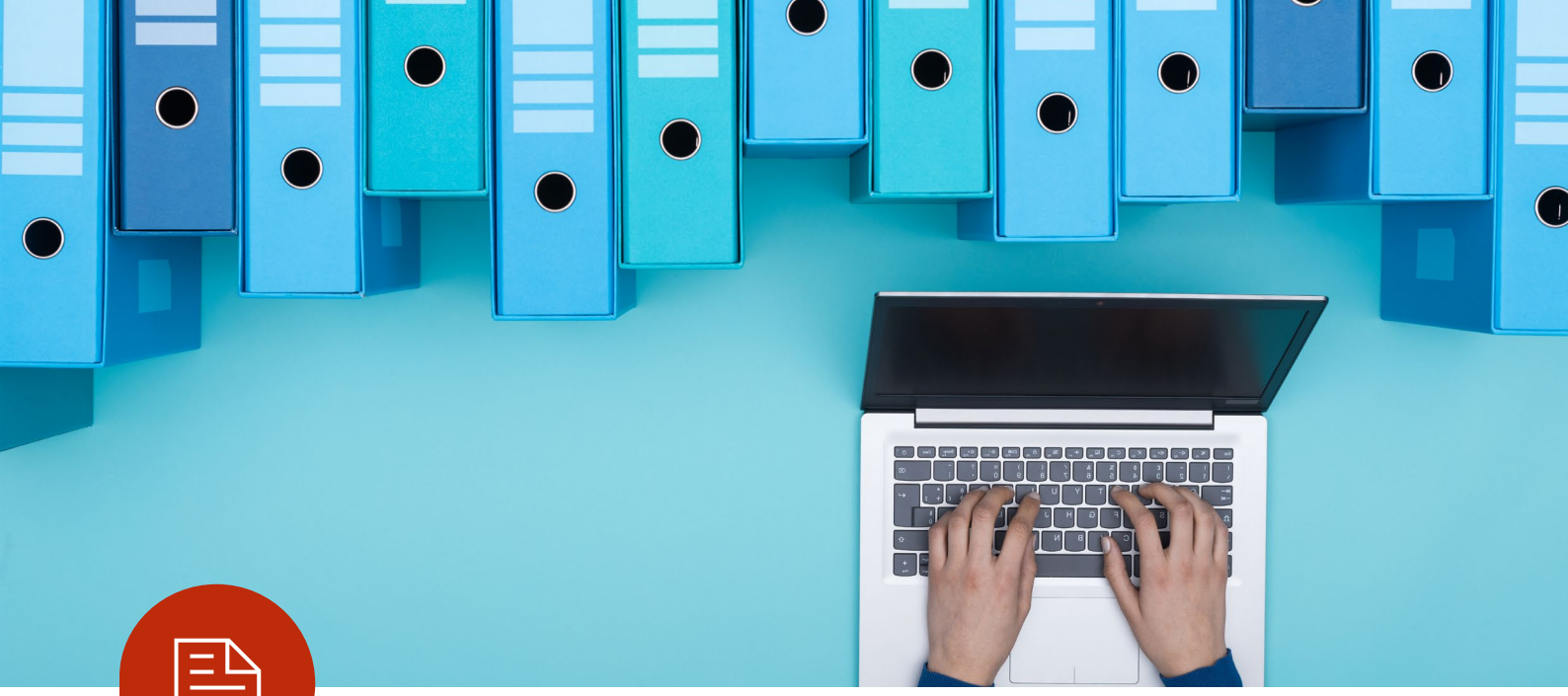
Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Video "Dati della ricerca: la European Open Science Cloud e i principi FAIR" https://www.youtube.com/watch?v=eNiHNaU6MrQ

Materiali di approfondimento sui principi FAIR:

- GOFAIR. FAIR Principles https://www.go-fair.org/fair-principles
- FAIRsFAIR Fostering Fair Data Practices in Europe https://www.fairsfair.eu/
- How to FAIR https://howtofair.dk/

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



I repository per depositare i dati

I repository sono **infrastrutture digitali per l'archiviazione e la preservazione di lungo periodo di dati, pubblicazioni, software** e più in generale dei risultati di una ricerca.

I dati archiviati all'interno di un repository vengono **associati a metadati strutturati** secondo schemi standard **Metadati e documentazione**. I repository giocano inoltre un ruolo fondamentale nella promozione della ricerca aperta e accessibile, permettendo ai ricercatori di **aderire ai principi FAIR** **I principi FAIR** in quanto consentono di associare ai dati depositati un set di metadati descrittivi, un **Persistent Identifier (PID)** e **una licenza** **Diritto d'autore**. Il repository permette anche al ricercatore di **scegliere il livello di accesso** per i propri dati e consente, a chi vuole riutilizzarli, di scaricare dati già depositati.

Volume dei dati e repository

La scelta del repository deve essere ponderata non solo in base al tipo di dati gestiti ma anche al loro volume. L'eventuale necessità di pagare per il deposito dei dati è specificata tra i **termini e condizioni** del repository ed è sempre opportuno controllare. In generale, nella maggior parte dei casi il deposito è gratuito. Alcuni repository possono permettere il deposito dietro pagamento per singoli dataset di grande ampiezza (di solito oltre le decine di GB) oppure per depositare qualsiasi dataset.

Tipi di repository

I repository possono essere divisi in **disciplinari, istituzionali o generalisti** (multiscopo). Tutte e tre le tipologie di repository presentano dei vantaggi peculiari, per cui è importante sceglierli in base alle proprie esigenze.

- I repository **disciplinari** adottano schemi di metadati appositamente progettati per la disciplina di riferimento, favoriscono una maggiore visibilità e condivisione all'interno della comunità scientifica interessata.
- I repository **istituzionali** sono messi a disposizione dei propri membri da istituzioni accademiche e di ricerca, offrendo servizi di validazione e di supporto al deposito per garantire la qualità dei dataset depositati. L'Università di Bologna mette a disposizione dei suoi ricercatori AMS Acta e AMS Historica.
- I repository **generalisti** non sono specializzati in un settore specifico e accolgono dati e materiali provenienti da diverse discipline e contesti di ricerca. Offrono una piattaforma robusta per la preservazione, la visibilità e l'accessibilità dei dati. Zenodo è il principale repository generalista in uso.

Sul campo!

Sono un ricercatore e devo scegliere il repository nel quale depositare i miei dataset *– da dove comincio?*

Effettuo un'analisi accurata dei possibili problemi di etica, privacy, IPR dei dataset che intendo produrre per stabilire il livello di accesso

Faccio una stima del volume per vedere se potrò depositare gratuitamente o meno

Decido se voglio avvalermi di un servizio di supporto o se preferisco depositare in autonomia

Controllo che il repository individuato abbia tutte le altre caratteristiche per far sì che i miei dati depositati siano FAIR (DOI, metadati, licenze)

Repository e livelli di accessibilità

Scegliere il livello di accesso è un punto essenziale nella gestione dei dati, per questo è necessario scegliere un repository che consenta di depositare i dati secondo pratiche di Scienza Aperta, ma anche di avere maggiori restrizioni dove necessario.

Il livello di accesso può essere descritto come:

- Aperto, ovvero il dataset è apertamente accessibile a chiunque lo voglia consultare, scaricare e riutilizzare.
- Ristretto, ovvero chi vuole consultare o scaricare il dataset depositato deve richiedere un'autorizzazione. Questa richiesta può essere fatta direttamente al ricercatore che ha depositato il dataset oppure, nel caso di alcuni specifici repository, ad un comitato di accesso deputato a valutare la legittimità della richiesta.

- Embargo, un tipo di restrizione dell'accesso temporanea. Permette di depositare i dataset mantenendoli privati per un tempo limitato, scaduto il quale l'embargo scade e il dataset diventa aperto.

Sul campo!

Sono un ricercatore di area sociale che vuole depositare un dataset prodotto che contiene una survey *– come posso fare?*

Cerco su re3data un repository che mi consenta una maggiore visibilità presso la comunità della mia disciplina

Faccio attenzione al tipo di survey che intendo fare, se cross-section o longitudinale, che presentano problematiche differenti

Le survey longitudinali presentano solitamente dati personali, poiché si intervistano gli stessi individui più volte nel corso del tempo, per cui dovrò scegliere un repository che mi consenta anche un accesso ristretto ai dati depositati.

Sono un ricercatore che lavora con dati sensibili, che non possono essere depositati apertamente *– come posso fare?*

Per quanto riguarda la gestione dei dati sensibili non anonimizzabili, è essenziale scegliere un repository che garantisca un accesso controllato. Questi permettono di mantenere i dati accessibili solo a utenti autorizzati e di rendere pubblici solo i metadati e la documentazione di supporto.

Un'altra opzione è quella di depositare la metodologia per l'analisi pubblicamente, ma tenere i dati chiusi dove necessario, in modo che almeno il processo di ricerca sia replicabile.

I miei dati potrebbero essere collegati ad una procedura di brevetto *- posso depositarli?*

Ricordati, se i tuoi dati sono necessari per un brevetto, di trattali con riservatezza finchè la procedura non è conclusa, per non inficiarne la "novelty". Quindi non condividerli con chiunque e non caricarli online in modo incontrollato. Per depositarli, scegli un repository che consenta di farlo con un embargo temporaneo.

Repository di Ateneo: AMS Acta e AMS Historica

AMS Acta è il repository istituzionale per la raccolta, conservazione e disseminazione dei dati della ricerca dell'Ateneo. Permette a docenti, ricercatori, assegnisti, dottorandi, borsisti e studenti afferenti all'Università di Bologna di gestire i dati della ricerca nel rispetto dei principi FAIR e dell'Open Science, in quanto:

- garantisce la conservazione e l'accesso nel tempo ai contributi depositati;
- assegna l'identificativo persistente DOI (Digital Object Identifier);
- implementa diversi livelli di accesso (aperto, chiuso, embargo);
- implementa i metadati descrittivi standard Dublin Core e Datacite, che sono sempre accessibili e associati a licenza Creative Commons (CC0 1.0 Universale);
- implementa diverse licenze, tra cui le Creative Commons;
- è conforme agli standard internazionali di interoperabilità e trasmissione dei metadati, è registrato nel catalogo dei data repository re3data ed è integra-

to e indicizzato dai principali cataloghi (OpenAIRE, BASE, WorldCat) e motori di ricerca (Google, Google Scholar, …);

- mette a disposizione i dati statistici sulla consultazione e il download di ciascun contributo.

AMS Historica è il repository istituzionale che ospita le digitalizzazioni di fonti antiche e di pregio di interesse scientifico e culturale dell'Ateneo. Permette di consultare le digitalizzazioni di documenti unici o di difficile reperibilità, quali opere d'arte, monumenti, reperti archeologici, codici manoscritti, papiri, libri, riviste, giornali, mappe, disegni, fotografie, fonti audio e video di interesse scientifico, storico e culturale conservati presso le biblioteche, i musei e gli archivi dell'Università o frutto di progetti di ricerca nazionali e internazionali.

I contenuti sono pubblicati nel rispetto delle linee guida e degli standard nazionali e internazionali che favoriscono la conservazione nel tempo e la valorizzazione delle opere digitalizzate, in quanto AMS Historica:

- associa metadati e licenze che ne promuovono la scoperta, lo studio, la condivisione e il riuso secondo i principi dell'Open Science;
- opera su una piattaforma open source DSpace-GLAM, che offre funzionalità innovative e potenti per la consultazione e lo studio di contenuti digitali eterogenei grazie ai servizi digitali basati sul formato IIIF (International Image Interoperability Framework);
- è indicizzata dai cataloghi e servizi di aggregazione nazionali e internazionali Cultura Italia, Europeana, OpenAIRE, WorldCat e BASE.

Link utili

Registri per l'individuazione dei repository:

- Re3data https://www.re3data.org/search?query=
- OpenAIRE explore https://www.openaire.eu/find-trustworthy-data-repository
- FAIRsharing repository database https://fairsharing.org/search?fairsharingRegistry=Database

Repository di ateneo:
AMS Acta https://amsacta.unibo.it/ | AMS Historica https://historica.unibo.it/

Per maggiori informazioni sui repository di Ateneo:

- "Preservare e disseminare i dati della ricerca in AMS Acta" (https://sba.unibo.it/it/almadl/servizi-almadl/preservare-disseminare-dati-della-ricerca-in-ams-acta);
- "Preservare e valorizzare il patrimonio culturale digitale" (https://sba.unibo.it/it/almadl/servizi-almadl/preservare-valorizzare-patrimonio-culturale-digitale).

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Utilizzare diversi tipi di standard

L'utilizzo di standard nel corso della ricerca è una buona pratica che può assumere diverse forme ma che in ogni caso permette di rendere i dati consistenti e interoperabili. Si possono utilizzare **standard per codificare le metodologie** con cui i dati vengono generati o riutilizzati, oppure per **descrivere i dati stessi**.

È possibile riutilizzare gli standard esistenti o crearne di nuovi specifici per l'attività di ricerca.

Per assicurarsi di seguire una **metodologia** standard per tutti i ricercatori coinvolti nell'attività di ricerca, è possibile utilizzare ad esempio gli **standard ISO** specifici per l'area disciplinare di riferimento.

Per strutturare i dati in modo univoco e renderli interoperabili con altri dati esistenti o che verranno generati in futuro, esistono standard che prendono la forma di vocabolari, tassonomie e ontologie. Quando si raccolgono informazioni, l'uso di almeno un vocabolario può aiutare a identificare correttamente i dati e a facilitarne il riutilizzo perché permette di renderli comprensibili ad altri, che condividono un vocabolario comune.

I "**vocabolari**" o "thesauri" sono risorse lessicali composte da termini controllati per descrivere un insieme di elementi, conoscenze, dati, teorie appartenenti a un determinato campo scientifico o dominio disciplinare. Essi forniscono una terminologia standard, aumentando il valore dei dati e rendendoli utilizzabili dalle macchine.

Con le "**tassonomie**", gli elementi non solo sono descritti attraverso vocabolari di termini controllati e selezionati, ma sono anche ordinati in un sistema, generalmente gerarchico.

Con le "**ontologie**", i tipi e quindi i nomi delle relazioni tra gli elementi che compongono una tassonomia o un vocabolario sono espliciti.

Per descrivere i dati invece, è bene utilizzare degli standard specifici di metadatazione **Metadati e documentazione**.

Sul campo!

Lavoro all'interno di un progetto collaborativo in cui sono coinvolte anche aziende e centri di ricerca esterni all'Università di Bologna *– come posso assicurarmi di avere risultati di qualità e interoperabili?*

Condividere all'interno del progetto collaborativo degli standard di metodologia comuni è fondamentale per raggiungere questo scopo.

Gli standard ISO sono delle norme che descrivono il modo migliore di fare qualcosa, concordato a livello internazionale da esperti del settore. Che sia la realizzazione di un prodotto, della gestione di un processo, dell'erogazione di un servizio o della fornitura di materiali: gli standard coprono un'ampia gamma di attività.

Per esempio,standard di gestione della qualità ("Quality management standards") sono pensati per aiutare a lavorare in modo più efficiente e a ridurre i difetti dei prodotti. Oppure, standard di salute e sicurezza possono ridurre gli incidenti sul posto di lavoro. Ancora, standard di sicurezza informatica iutano a proteggere le informazioni sensibili.

Voglio utilizzare per i miei dati dei vocabolari e delle tassonomie per descrivere le informazioni che raccolgo *– quali vantaggi mi porta e dove posso cercare quelli più adatti?*

Un vocabolario può contenere dei termini standard usati in uno specifico ambito disciplinare che possono essere usati per indicare le variabili di un'analisi, che poi verranno riportate in un file tabulare come titoli delle colonne ad esempio.

In una ricerca con dati qualitativi, utilizzare termini da un vocabolario standard può aiutare ad esprimere le relazioni tra le variabili anche all'interno di un testo come un report tecnico, così che chiunque sia coinvolto sul progetto sa esattamente a quale fenomeno o tipologia di dato si fa riferimento.

Esistono schemi più generici, come la Information Artifact Ontology (che descrive le entità informative intese come informazioni codificate in qualche entità digitale o fisica) o schemi estremamente specifici, come lo IUPAC Compendium of Chemical Terminology (che contiene circa 7000 concetti chimici derivati dalle Raccomandazioni IUPAC).

Link utili

Standard ISO
https://www.iso.org/standards.html

Strumento per la ricerca di schemi di metadati, vocabolari e thesauri
https://fairsharing.org/search?fairsharingRegistry=Standard

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Gestire il software

Il software può essere sviluppato in molti ambiti disciplinari diversi, ma è sempre bene avere alcune accortezze nella gestione di questo output digitale della ricerca. Il software, contrariamente ai dati, è infatti un eseguibile e cambia frequentemente forma nel tempo, richiedendo alcune azioni specifiche per gestirlo correttamente.

In primo luogo, è bene distinguere le azioni e gli strumenti per ogni fase dello sviluppo del codice.

Nella fase di **pianificazione**, è bene definire ruoli e responsabilità nella progettazione, sviluppo e mantenimento del codice. Dall'architettura e concettualizzazione allo sviluppo, fino alla revisione e debugging.

Nella fase di **sviluppo**, piattaforme online per lo sviluppo collaborativo come GitHub e GitLab permettono di monitorare lo sviluppo con il version control e la file version history, così come garantiscono una collaborazione efficace grazie al sistema Git distribuito.

Quando il software raggiunge una versione funzionante soddisfacente, è bene anche **depositarlo** su Zenodo o repository specifici per questo tipo di oggetto digitale, come Software Heritage. Entrambi permettono l'aggiornamento semi-automatico del software nel caso si volesse continuare lo sviluppo.

Quando si **deposita** il codice è fondamentale depositare sia documentazione machine-readable che documentazione human-readable **Metadati e documentazione**. Nel primo caso, parliamo semplicemente di metadati e citation file: ne esistono di specifici per il software, come CodeMeta e file CITATION.cff. Nel secondo caso, parliamo invece principalmente di file README e documentazione in-line, cioè i commenti all'interno del file di codice. In entrambi i casi, la documentazione permette di rendere maggiormente comprensibile il codice a chiunque lo trovi (ma anche ai voi stessi del futuro!)

Il software richiede anche delle licenze specifiche una volta depositato, per chiarirne gli usi consentiti. La scelta della licenza dipende dalla situazione in cui ci si trova: se si ha bisogno di lavorare in una comunità, se si vuole una licenza semplice e permissiva, o se si desidera condividere i miglioramenti.

Una volta depositato in un repository come Software Heritage, il codice ottiene un PID (identificativo persistente), una licenza e dei metadati disciplinari che lo descrivono: è quindi possibile citarlo al pari di una pubblicazione come contributo di ricerca.

A volte però depositare il software, anche con i dati di input e di output correlati, non è sufficiente per assicurarne la **riusabilità** (non solo da parte di altri, ma anche in futuro dagli stessi sviluppatori). Questo perché spesso i linguaggi di programmazione, le librerie, i plug in, cambiano frequentemente versione, e così anche il codice sviluppato solo qualche mese fa non è più aggiornato. Per questo esistono delle soluzioni apposta, come i containers o sistemi cloud-based che permettono di simulare le variabili di ambiente di sistema in cui il codice è stato sviluppato, per poterlo eseguire nuovamente.

Sul campo!

Nel contesto del mio progetto competitivo, gran parte dell'obiettivo di ricerca è sviluppare un software per analisi dati *– quale workflow devo seguire?*

Una prima riunione con tutti quelli coinvolti nello sviluppo permette di individuare ruoli e responsabilità, quindi anche tempi e scadenze di lavoro.

Lo sviluppo può iniziare direttamente su GitHub: creando un nuovo repository condiviso tra tutti quelli che parteciperanno allo sviluppo, è più facile lavorare in maniera collaborativa tracciando le modifiche al codice, dati di input e altri materiali a supporto.

Alcuni tool cloud-based, come i notebook, possono aiutare nella documentazione del codice in ogni step del suo algoritmo, tracciando e visualizzando i dati di input e di output di ogni funzione.

Prima ancora che il codice arrivi alla sua prima versione operativa, collegare il repository a Software Heritage ne permette l'harvesting periodico automatico assicurandosi che il codice sia conservato correttamente (cosa che GitHub non garantisce!) con uno schema di metadatazione appropriato, in questo caso: CodeMeta.

Per scegliere una licenza appropriata alle esigenze di progetto da inserire fin dal repository di GitHub, esistono vari tool online interattivi che riasumono l'intera documentazione delle principali licenze in uso per il software, tra cui Choose a Licence.

Per assicurarsi che il lavoro venga riconosciuto appropriatamente, anche i file CITATION.cff possono essere generati in modo semi-automatico, risparmiando molto tempo, e possono poi esser caricati direttamente sul repository di GitHub e di conseguenza salvati in automatico su Software Heritage.

Link utili

Strumento di selezione della licenza https://choosealicense.com/

Generatore di Citation Files https://citation-file-format.github.io/cff-initializer-javascript/#/

Esempi di cloud notebook per lo sviluppo del codice https://datasciencenotebook.org/

- Software Heritage https://www.softwareheritage.org/
- CodeMeta https://www.codemeta.github.io/codemeta-generator/
- GitHub https://github.com/

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Diritto d'autore e dati della ricerca

I **dati della ricerca possono essere oggetto di tutela** autoriale quando sono opere dell'ingegno di carattere creativo; a titolo meramente esemplificativo testi, immagini creative, tabelle elaborate, database, software **Gestire il software**.

Il diritto d'autore, infatti, riconosce una tutela alle **opere dell'ingegno di carattere creativo** che appartengono alla letteratura, alla musica, alle arti figurative, all'architettura, al teatro, alla cinematografia, alle scienze, qualunque ne sia il modo o la forma di espressione. Il diritto d'autore tutela la forma espressiva di un'opera dell'ingegno e non i dati in quanto tali, infatti la tutela autoriale non copre le idee, i procedimenti, i metodi di funzionamento o i concetti matematici in quanto tali.

Riferimenti normativi

Legge 22 Aprile 1941, n. 633
"Protezione del diritto d'autore
e di altri diritti connessi
al suo esercizio."

Sul campo!

Sono un ricercatore e devo usare opere tutelate dal diritto d'autore
– come posso sapere se sussiste ancora la tutela autoriale?

Verifica se l'opera è in pubblico dominio, cioè sia scaduta la durata di protezione del diritto d'autore, calcolando 70 anni dalla morte dell'autore o 70 anni dalla prima pubblicazione per le opere collettive.

In questi casi verifica che non ci siano vincoli di tutela del patrimonio culturale.

Sono un ricercatore e devo usare opere tutelate dal diritto d'autore nella mia ricerca – cosa posso fare per non violare il copyright?

Per prima cosa verifica se esiste una licenza d'uso e controlla che gli usi che intendi fare dell'opera siamo coerenti con i termini della licenza. Ad esempio, una licenza Creative Common Attribution (CC BY) ne consente il libero uso.

In caso contrario, valuta se ai fini della tua ricerca è necessario l'uso integrale dell'opera o solo di brani o parti di essa. In questo secondo caso, verifica che l'uso che ne farai corrisponde alla previsione dell'art. 70 l.d.a per cui non è libera la riproduzione e la comunicazione al pubblico del riassunto, della citazione edi brani o di parti di se effettuati per uso di critica o di discussione, nei limiti giustificati da tali fini e purché non costituiscano concorrenza all'utilizzazione economica dell'opera e se effettuati a fini di insegnamento o di ricerca scientifica l'utilizzo deve inoltre avvenire per finalità illustrative e per fini non commerciali.

In tutti gli altri casi sarà necessario chiedere l'autorizzazione al titolare dei diritti (spesso l'editore) per utilizzare l'opera.

I diritti riconosciuti all'autore

All'autore spettano alcuni **diritti c.d. morali**, tra cui il diritto di rivendicare la paternità dell'opera, il diritto di inedito e il diritto all'integrità dell'opera. Tali diritti sono irrinunciabili, inalienabili, imprescrittibili e possono essere fatti valere senza limiti di tempo dopo la morte dell'autore. All'autore spettano anche **diritti c.d. patrimoniali**, cioè diritti di sfruttamento economico dell'opera in esclusiva. Pubblicare, digitalizzare, comunicare (anche online), modificare e tradurre, tra altri, sono annoverati tra i diritti patrimoniali cui esercizio spetta in via esclusiva all'autore. Tali diritti possono essere ceduti a titolo oneroso o gratuito, in via esclusiva o non esclusiva e sono esercitabili fino a 70 anni dopo la morte dell'autore.

Banche di dati: tra diritto d'autore e diritto connesso del costitutore

Le banche di dati sono definite come le "*raccolte di opere, dati o altri elementi indipendenti sistematicamente o metodicamente disposti ed individualmente accessibili mediante mezzi elettronici o in altro modo*". Sono protette dal diritto d'autore le banche di dati creative che per la scelta o la disposizione del materiale costituiscono una creazione intellettuale dell'autore. La tutela autoriale delle banche di dati non si estende al contenuto, rispetto al quale restano impregiudicati i diritti di terzi.

Gli investimenti per la costituzione della banca di dati o per la sua verifica o la sua presentazione, impegnando mezzi finanziari, tempo o lavoro, sono tutelati indipendentemente dalla tutela autoriale, attraverso il riconoscimento di un diritto connesso (c.d. ***diritto sui generis***) in capo al costitutore di una banca di dati. Tale diritto si esplica nel vietare le operazioni di estrazione ovvero reimpiego della totalità o di una parte sostanziale della stessa.

Il diritto sui generis ha una durata inferiore ai diritti d'autore e decorre, per 15 anni, dal primo gennaio dell'anno successivo alla data del completamento della banca di dati o alla data della prima messa a disposizione del pubblico.

Le condizioni di utilizzo di una banca di dati sono disciplinate dal titolare dei diritti attraverso specifiche licenze d'uso. Ricorda di consultare i termini d'uso.

Le licenze d'uso delle opere dell'ingegno

Per utilizzare opere e materiale protetto dal diritto d'autore o da diritti connessi è necessario avere il **preventivo consenso del titolare dei diritti**.

I diritti patrimoniali hanno ad oggetto l'opera nel suo insieme ed in ciascuna delle sue parti, pertanto, anche l'uso parziale dell'opera rientra nell'esclusiva dell'autore e deve essere autorizzata.

Attraverso contratti di concessione d'uso (licenze) l'autore dispone dei propri diritti patrimoniali e consente a terzi di utilizzare l'opera, alle condizioni concordate, restando titolare dei relativi diritti che alla scadenza del contratto tornano nella sua piena disponibilità.

I diritti patrimoniali sono tra loro indipendenti, per cui ciascun diritto può essere trasferito separatamente dagli altri. La trasmissione dei diritti deve essere provata per iscritto.

L'utilizzazione di un'opera può avvenire in assenza di un'autorizzazione da parte dei titolari dei diritti solo nell'ipotesi di eccezioni e limitazioni espressamente previste dalla legge. È sempre necessario verificare i termini della licenza d'uso associata alla pubblicazione e/o messa a disposizione di un'opera.

Le licenze Creative Commons

Le licenze Creative Commons (CC) sono le licenze maggiormente utilizzate per le opere digitali, costituiscono veri e propri contratti di licenza d'uso con i quali l'autore concede ad una generalità di soggetti indefiniti l'autorizzazione all'uso dell'opera a determinate condizioni, decidendo quali diritti riservare e quali concedere in uso.

I sei schemi di licenza disponibili si articolano sulla combinazione di quattro clausole base che l'autore può scegliere e combinare, esplicando così le modalità d'uso della propria opera da parte degli utilizzatori finali.

A ciascuna clausola base è associato un simbolo grafico allo scopo di renderne più facile il riconoscimento:


BY – attribuzione: è sempre presente


NC – non è consentito l'uso commerciale


SA – condividi allo stesso modo


ND – non sono consentite opere derivate

Le licenze CC sono disponibili in tre forme:

- il Commons Deed (i simboli user friendly riassuntivi dei termini delle licenze);
- il Legal Code (il vero e proprio contratto di licenza per esteso);
- il CC REL – Creative Commons Rights Expression Language (l'insieme di informazioni leggibili dal computer).

Le licenze Creative Commons in linea con i principi dell'Open Science, frequentemente associate ai dataset sono:

- CC BY, "Attribution": indica la possibilità di riusare e modificare liberamente l'opera, attribuendo sempre la citazione dell'opera originaria.
- CC BY-SA, "ShareAlike": indica la possibilità di riusare e modificare liberamente l'opera, attribuendo sempre la citazione dell'opera originaria e distribuendo l'opera così modificata con la stessa licenza dell'opera originale.
- CC0, "No Rights Reserved": indica pubblico dominio e/o rinuncia a tutti i diritti sull'opera.

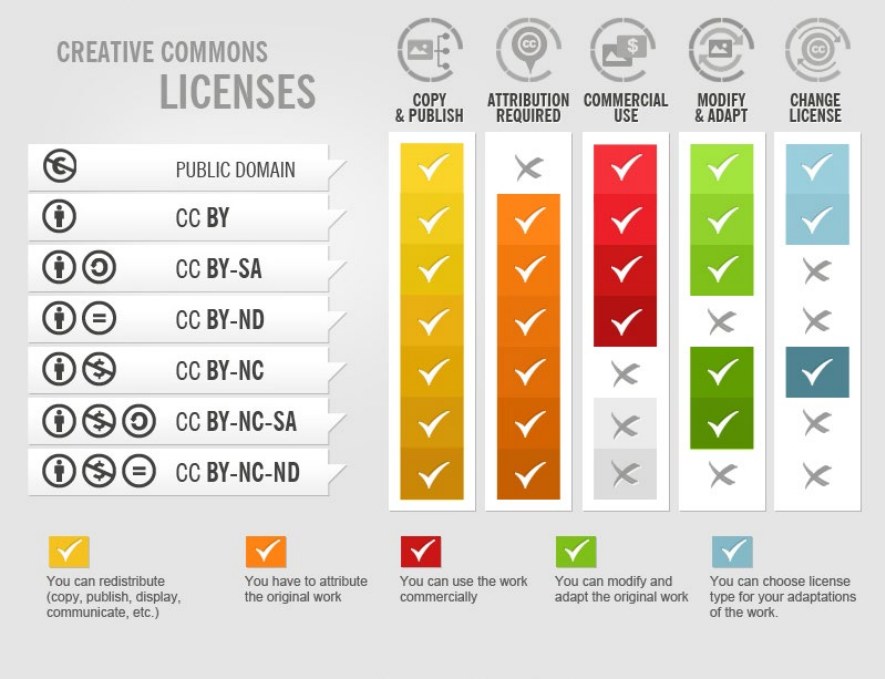


Image credits:
JoKalliauer; foter, CC BY-SA 3.0
https://foter.com/blog/how-to-attribute-creative-commons-photos/
via Wikimedia Commons

Sul campo!

Sono un ricercatore e voglio utilizzare nella mia ricerca materiale che ho trovato online in vari siti web *– come posso capire se sono legittimato all'uso?*

Ricorda la regola generale per cui il web non è esente dall'obbligo di rispettare regole e limitazioni. Verifica sempre in cascun sito web i termini d'uso del sito e del materiale messo a disposizione: l'utilizzo di opere liberamente e gratuitamente disponibili in rete non è in automatico libero.

Ricorda che se estrai dati da un database online, leggi un articolo scientifico da una rivista elettronica assicurati di rispettare i termini di licenza a questi associati.

Se non sono associate licenze Creative Commons, cerca nella pagina web i termini d'uso spesso indicati come "Terms of use". In assenza di espresse indicazioni sui termini d'uso tutti i diritti devono intendersi riservati e gli usi devono essere autorizzati contattando il titolare dei diritti.

Sono un ricercatore e voglio utilizzare nella mia ricerca immagini che ho scaricato da una digital library *- come posso capire se sono legittimato all'uso?*

Verifica la licenza d'uso associata, spesso sono usate licenze Creative Commons.

Sono un ricercatore e voglio associare una licenza CC0 ai dati che ho generato – cosa implica in pratica?

Verifica che la stessa licenza possa essere applicata ai dati in esso contenuti senza pregiudizio ai diritti di terzi e nel rispetto di vincoli di legge o di altri accordi.

Ricorda che la rinuncia si estende a tutti i diritti sull'opera inclusi tutti i diritti connessi al diritto d'autore o affini, è irrevocabile ed è valida in tutto il mondo, nella misura consentita dalla legge.

Ricorda che l'applicazione di una licenza CC0 a un dataset consente a qualsiasi utilizzatore di copiare, modificare, distribuire e utilizzare il dataset e i dati in esso contenuti, anche per fini commerciali, senza chiedere alcun permesso.

Link utili

Normativa di riferimento: Legge 22 aprile 1941, n. 633 "Protezione del diritto d'autore e di altri diritti connessi al suo esercizio" https://www.gazzettaufficiale.it/eli/id/1941/07/16/041U0633/sg

Risorse utili:

- Schemi di licenze Creative Commons https://creativecommons.org/
- Versione estesa della licenza CC0 https://creativecommons.org/publicdomain/zero/1.0/legalcode.it

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Il rispetto della privacy nella gestione dei dati di ricerca

La dimensione etica rientra sotto il cappello più ampio del principio dell'integrità della ricerca, fondamentale per garantirne l'affidabilità, la qualità e la trasparenza.

Coinvolgere persone è fondamentale in diversi tipi di ricerca, dagli studi clinici su pazienti passando alla raccolta di dati demografici, dagli studi antropologici a quelli linguistici. Una gestione responsabile dei dati personali rende necessario prevedere e affrontare possibili problematiche legate alla privacy, per predisporne il trattamento a norma di legge (vedi box a fianco) e secondo il **principio di "privacy by design"**, che richiede attenzione su questi temi sin dalla progettazione di un'attività di trattamento di dati personali. Un altro principio fondamentale da tenere a mente è il **principio di minimizzazione dei dati**, che invece impone di raccogliere e trattare solamente quei dati personali necessari per il raggiungimento dello scopo (in questo caso, scientifico).

Si definisce "trattamento" qualsiasi operazione o insieme di operazioni applicate a dati personali o insiemi di dati personali. Esempi di trattamento sono la raccolta, uso, organizzazione, conservazione, e distruzione dei dati personali.

Riferimenti normativi

"Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica" e "Prescrizioni relative al trattamento dei dati personali effettuato per scopi di ricerca scientifica (aut. Gen. N. 9/2016)" emanate dal garante privacy; "General Data Protection Regulation", regolamento (EU) 2016/679

I dati personali

Sono definiti dati personali le **informazioni che identificano una persona fisica** o che la rendono identificabile direttamente (ad esempio i dati anagrafici, come il nome e il cognome, o le immagini personali) o indirettamente (ad esempio dati relativi a abitudini, stile di vita, stato di salute, situazione economica, o un codice attribuito nell'ambito di una ricerca scientifica).

I dati che rivelano l'origine razziale od etnica, le convinzioni religiose, filosofiche, le opinioni politiche, l'appartenenza sindacale, relativi alla salute o alla vita sessuale, i dati genetici, i dati biometrici, i dati relativi all'orientamento sessuale e quelli giudiziari, relativi a condanne penali e reati dei soggetti coinvolti nella ricerca sono considerati **dati personali "particolari"**, ed è necessario riservare loro maggiore attenzione perché possono comportare discriminazioni.

Pianificare la gestione della privacy

Durante la fase di pianificazione, è bene innanzitutto domandarsi se sia davvero necessario raccogliere e trattare dati personali per la specifica attività di ricerca e decidere se raccogliere dei nuovi dati o riutilizzarne di già raccolti in passato (eventualmente da soggetti terzi).

Tutti i potenziali partecipanti alla ricerca devono essere informati in modo chiaro, trasparente e appropriato in merito al progetto di ricerca e al relativo trattamento dei dati personali attraverso lo strumento dell'**informativa**, che deve contenere informazioni circa chi tratterà i dati in oggetto, per quali finalità e secondo quali modalità. Nell'informativa devono essere indicate anche le basi giuridiche del trattamento di dati personali (e.g. consenso, che può essere raccolto con diverse metodologie – cartaceo, online, videoregistrazione, ecc.) e i diritti che i soggetti possono esercitare sui loro dati.

Per la redazione dell'informativa è quindi necessario **definire in partenza** per quanto tempo i dati personali raccolti saranno conservati in forma identificativa (quindi non anonimizzati) e con quali soggetti dovranno essere condivisi e per quali scopi.

È importante ricordare che la normativa italiana prevede obbligatoriamente che le ricerche cliniche, che prevedono il reclutamento di pazienti per il loro svolgimento, ricevano l'approvazione del comitato etico preposto. Nel caso specifico dell'Università di Bologna, il comitato competente per l'approvazione delle ricerche cliniche è il comitato etico indipendente di Area Vasta Emilia Centro (CE-AVEC).

Per tutte le altre ricerche non cliniche che raccolgono però dati personali, il passaggio al comitato di bioetica è opportuno ma non necessario a meno di una richiesta esplicita, che può provenire ad esempio dall'ente finanziatore o dall'editore della pubblicazione a cui i dati sottendono.

Sul campo!

Sono un ricercatore nel campo delle scienze umane e sociali e raccolgo dati osservativi e da questionari – Come gestisco gli aspetti trasversali come la privacy?

Se raccogli dati personali, già dalla fase di pianificazione prepara un'informativa che dev'essere firmata dai soggetti a cui verranno fatte le domande o da cui dovrai prendere informazioni. Segui la normativa vigente, come il GDPR, e chiedi aiuto agli uffici preposti per redigere l'informativa per il consenso informato.

Misure per garantire la sicurezza dei dati personali

Le misure tecniche e organizzative da mettere in atto durante le fasi attive della ricerca (raccolta, analisi, conservazione) per garantire la sicurezza dei dati personali devono essere valutate in relazione alla specifica attività di ricerca.

Il primo aspetto da considerare è quello della **scelta di un opportuno sistema di storage** che garantisca la protezione del dato. Se è necessario utilizzare un sistema cloud per facilitare la collaborazione con partner terzi per il trattamento dei dati, è fondamentale, per rispettare quanto previsto dalla General Data Protection Regulation (GDPR), assicurarsi di scegliere una soluzione i cui server siano localizzati in uno dei paesi che, come definito della decisione di adeguatezza della CE, garantiscano un appropriato livello di protezione dei dati personali. Sempre nell'ottica di **bilanciare collaborazione e protezione**, è necessario chiarire quanti e quali ricercatori devono necessariamente avere accesso ai dati nella loro forma identificativa per portare avanti le attività di ricerca e gestire di conseguenza gli accessi alle cartelle in cui i dati personali sono conservati. Può essere opportuno, sulla base dei dati trattati, prendere in considerazione anche l'opzione di criptare le cartelle utilizzando strumenti specifici.

Altro aspetto fondamentale da prendere in considerazione nelle fasi di analisi e conservazione nel breve periodo è la possibilità di **anonimizzare o pseudonimizzare i dati**. Anonimizzazione e pseudonimizzazione sono tecniche utilizzate per modificare i dati personali al fine di, rispettivamente, eliminare o ridurre la possibilità che questi siano identificabili. Entrambi i processi comportano la rimozione o la modifica degli identificatori personali diretti e indiretti e possono ridurre la qualità e l'utilità dei dati di ricerca, perché comportano una perdita di informazioni.

Si definisce anonimizzazione l'elaborazione dei dati in modo tale che nessun individuo e/o nessuno strumento possano più identificare i soggetti coinvolti a partire dai dati stessi. Dati anonimi lo sono per tutti, anche per i ricercatori che li hanno raccolti in prima istanza. Per anonimizzare i dati è necessario individuare tutti i possibili identificatori, sia diretti che indiretti, e modificarli con la strategia più appropriata. È di fondamentale importanza prestare attenzione alle combinazioni di attributi che possono individuare singole persone e ai campioni di piccole popolazioni. Esempi di misure per evitare l'identificazione a ritroso, o *inference disclosure*, sono la generalizzazione, l'aggregazione, la codifica dei limiti superiori e inferiori per nascondere gli outlier identificabili, la perturbazione dei dati per spostare i valori (…). I dati anonimizzati non sono più da considerare dati personali ai sensi del GDPR.

Si parla di pseudonimizzazione quando i dati identificativi vengono elaborati in modo da non essere più attribuibili a una persona specifica senza l'uso di informazioni aggiuntive come una chiave di codifica. I dati pseudonimizzati sono ancora dati personali ai sensi del GDPR. Per pseudonimizzare i dati è necessario eliminare o codificare tutte le informazioni direttamente identificabili. È consigliato utilizzare per la codifica codici utili e casuali per ogni persona e conservare la chiave di codifica, possibilmente criptata, separatamente dal file dei dati codificati.

Sul campo!

Devo pseudonimizzare due set di dati, uno di tipo quantitativo e uno di tipo qualitativo *– come posso fare?*

Per pseudonimizzare dei dati di tipo quantitativo:

- Rimuovi o codifica tutte le informazioni che rendono possibile un'identificazione diretta (e.g. nome, cognome, indirizzo, numero di telefono, indirizzo e-mail, indirizzo IP…) utilizzando un codice randomico per ogni persona.
- Cripta la chiave di codifica e conservala separatamente dal file codificato.
- Generalizza o rimuovi gli identificatori indiretti (e.g. età, genere, impiego…).

Per pseudonimizzare dei dati di tipo qualitativo:

- Testi, es. trascrizioni di interviste: utilizza pseudonimi e descrizioni generiche e indica le sostituzioni con [parentesi quadre]. Esempio: [Persona 1] lavora per [un'organizzazione finanziaria] in Belgio.
- Audio e/o video: sfoca i volti e distorci le voci con strumenti appositi.

Sono un ricercatore di area medica e devo anonimizzare i miei dati quantitativi
– quali strategie posso mettere in atto e con quali risultati?

- La generalizzazione o la rimozione di identificatori indiretti riduce il dettaglio dei dati, ad esempio cambiando "età 27" in "gruppo di età 21-30"; cambiando "Disturbo schizoide di personalità" in "Disturbo mentale e comportamentale".
- La codifica dei limiti superiori e inferiori nasconde i valori anomali (outlier) nei dati, ad esempio gruppo di età superiore a 70 anni, stipendio inferiore a 1.658 euro/mese (...).
- La perturbazione modifica il valore dei dati numerici aggiungendo "rumore", sostituendo i valori con valori simulati o valori medi.

Conservazione a lungo termine dei dati personali

La durata della conservazione dei dati personali nella loro forma identificabile deve essere stabilita già nella fase di pianificazione della ricerca ed essere presentata attraverso l'informativa ai soggetti che partecipano alla ricerca. Non è quindi legittimo conservare i dati oltre il tempo necessario per raggiungere le finalità per le quali sono stati raccolti. Se è necessario per garantire la trasparenza e la riproducibilità della ricerca depositare ad accesso aperto dei dati personali nella loro forma identificabile, questo è possibile solo se sussiste un'idonea base giuridica (e.g. il consenso dell'interessato) che lo permetta. Nel caso in cui questo non fosse possibile, diventa obbligatorio valutare un eventuale deposito con modalità protetta, ad esempio su repository che prevedono un accesso controllato ai dati anche attraverso un apposito comitato di valutazione delle richieste di consultazione e riutilizzo dei dati.

Come abbiamo visto, i dati anonimizzati non sono più da considerarsi dati personali ai sensi del GDPR. Per questo motivo, una attenta anonimizzazione è la strategia da prediligere alla fine della propria ricerca per poter rendere disponibili i dati della propria ricerca in un repository di dati.

Link utili

Pagina intranet sul Trattamento di dati personali per finalità di ricerca scientifica
https://intranet.unibo.it/Ateneo/Web1/Pagine/PrivacyRicerca.aspx

Leggi e regolamenti:

- Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9069637
- Prescrizioni relative al trattamento dei dati personali effettuato per scopi di ricerca scientifica (aut. gen. n. 9/2016) https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9124510#5
- General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) https://eur-lex.europa.eu/eli/reg/2016/679/oj.
- Decisione di adeguatezza https://ec.europa.eu/newsroom/article29/items/614108 https://www.garanteprivacy.it/temi/trasferimento-di-dati-all-estero.

Strumenti utili:

- Criptare i dati https://www.veracrypt.fr/en/Home.html https://docs.microsoft.com/it-it/windows/security/information-protection/bitlocker/bitlocker-overview.
- Alterare i volti dei soggetti in video https://coehelp.uoregon.edu/using-openshot-to-blur-a-face-in-a-video/.
- Alterare le voci dei soggetti registrati https://www.qualitative-research.net/index.php/fqs/article/view/512/1106.
- Anonimizzare i dati https://amnesia.openaire.eu/ | https://arx.deidentifier.org/ | https://github.com/sdcTools/sdcMicro.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA