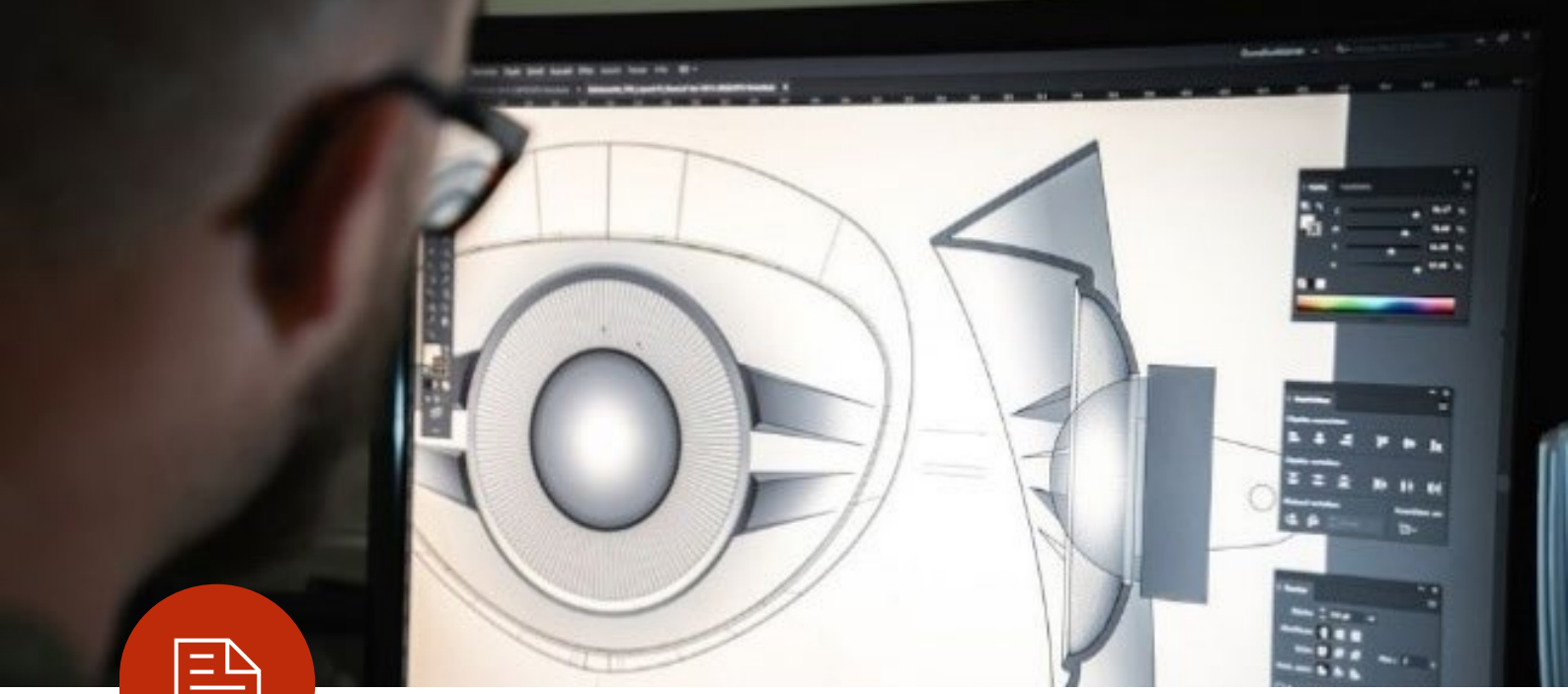
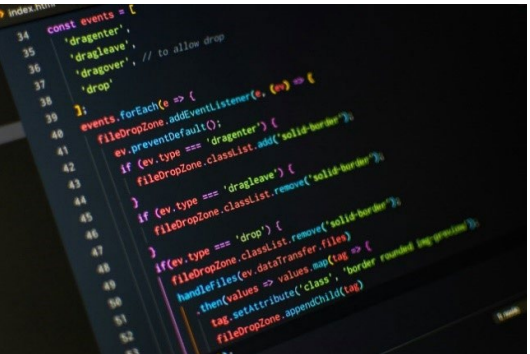


Dati della ricerca: tipologie, formati, metodi

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