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