



PUBBLICARE UN DATASET COME OPEN DATA IN UNIMI

L'ESEMPIO DI AIR



Sommario

0. INTRODUZIONE	3
1. COSA SONO GLI OPEN DATA?	3
2. Avviare un progetto di pubblicazione di open data	4
3. Quali dati?	4
4. Chi si occupa del progetto?	5
5. Dove pubblicare in open data	5
6. Selezione, trattamento e estrazione dei dati	5
7. L'invio dei dati al portale regionale	6
8. La pubblicazione sul portale regionale	7
9. Conclusione: si può fare!	7
10. Riferimenti	7
11. Appendici	8
Open Data Lombardia per immagini	8
I metadati descrittivi di Open Data Lombardia	10



0. INTRODUZIONE

Queste linee guida nascono da una esperienza specifica, ovvero la creazione e esposizione di un set di dati (dataset) delle pubblicazioni dei ricercatori UNIMI degli ultimi 5 anni.

L'obiettivo strategico era mettere a disposizione i propri dati ai portatori di interesse, secondo modalità facilmente fruibili. In concreto, i due principali obiettivi e bisogni che il progetto ha soddisfatto sono:

1. Esporre le informazioni sulle pubblicazioni di Ateneo depositate dai ricercatori dell'Ateneo nell'archivio istituzionale AIR in un formato che ne consente il riuso informatico (machine-to-machine) a utenti interni e esterni all'Ateneo
2. Pubblicare il dataset di cui sopra sul portale Open Data Lombardia

L'attività si è inserita in un contesto nazionale, che sarà brevemente delineato nel primo capitolo, in cui questo tipo di attività sta diventando non solo una prassi consolidata delle Pubbliche amministrazioni di qualsiasi genere (dai comuni agli enti di istruzione e ricerca), ma anche un obiettivo specifico legato al contesto più generale della digitalizzazione, obiettivo per il quale vengono indicate alle amministrazioni tempistiche e risultati anche quantitativi da raggiungere.

L'attività, che ha seguito l'iter descritto nei successivi capitoli, è stata promossa e condotta dall'Ufficio Qualità dei dati della Direzione Performance, Assicurazione Qualità, Valutazione e Politiche di Open Science, in collaborazione per la parte informatica con la Direzione ICT.

Queste linee guida intendono dare indicazioni operative che consentano ad altre strutture dell'Ateneo di intraprendere progetti analoghi; per questo motivo si è cercato e voluto un testo breve e privo di tecnicismi di tipo informatico.

1. COSA SONO GLI OPEN DATA?

Gli open data o dati aperti sono dati la cui principale caratteristica è quella di essere liberamente accessibili, e quindi riutilizzabili, da chiunque.

In particolare queste linee guida si occupano di open data della Pubblica amministrazione: in questo ambito gli open data sono considerati un bene comune e corrispondono al patrimonio informativo dell'amministrazione stessa messo a disposizione del più largo pubblico tramite tecnologie digitali e gratuitamente.

Le tre principali caratteristiche che devono avere gli open data della Pubblica amministrazione sono:

- 1) una licenza che ne permetta il più ampio riuso, anche commerciale;
- 2) un formato aperto, ovvero leggibile senza bisogno di strumenti informatici commerciali, e metadati descrittivi anch'essi in formato aperto;
- 3) la gratuità

Gli open data sono uno degli ambiti dell'AGID, l'Agenzia per l'Italia Digitale, l'ente incaricato di definire strategie e fornire supporto tecnico per sostenere la trasformazione digitale delle Pubbliche amministrazioni, delle imprese e dei cittadini.



In ambito di open data AGID si occupa di monitorare il raggiungimento degli obiettivi previsti dal Piano Triennale per l'informatica nella Pubblica Amministrazione. Nel caso delle Università, la roadmap prevede per il 2024 la pubblicazione di 10 insiemi di dati (dataset).

I dataset devono essere pubblicati nel catalogo nazionale [dati.gov](https://dati.gov.it), e questo può avvenire direttamente oppure attraverso un portale regionale che per il nostro Ateneo è [Open Data Lombardia](https://open.data.lombardy.it).

2. Avviare un progetto di pubblicazione di open data

Un progetto di pubblicazione di open data richiede una struttura interessata che abbia sia dati da rilasciare in formato aperto e che possano essere di interesse internamente ed esternamente all'Ateneo, sia personale da dedicarvi.

Serve inoltre decidere dove pubblicare il dataset, scelta che determina il tipo di operazioni tecnico informatiche da effettuare; i dati devono essere corredati da una serie di metadati che ne garantiranno il ritrovamento e l'uso corretto.

La struttura avrà inoltre bisogno di supporto informatico: per ottenerlo è possibile rivolgersi al Responsabile della Transizione al Digitale di Ateneo (Direzione ICT). L'Ufficio Qualità dei Dati sulla Ricerca è a disposizione per chiarimenti su queste linee guida.

3. Quali dati?

La prima cosa da fare è individuare i dati che possono, opportunamente trattati, costituire un dataset di interesse per portatori interni e esterni all'Ateneo.

L'ufficio o struttura UNIMI deve quindi analizzare le proprie attività individuando quelle che producono dati che possono sia essere diffusi ad accesso aperto che essere di interesse in termini di riuso.

Per quanto riguarda la diffusione ad accesso aperto, vanno considerati tutti quei dati privi di aspetti di riservatezza: nel caso del progetto relativo alle informazioni sulle pubblicazioni sono stati esclusi tutti i dati personali relativi ai ricercatori dell'Ateneo; in particolare le informazioni sulle pubblicazioni non presentano problemi per un progetto di pubblicazione come open data perché già pubbliche in altra forma (ad esempio sulle pubblicazioni stesse, ma anche in una quantità di banche dati). Questo fa capire che non è necessario scegliere dati mai pubblicati o esposti ad accesso aperto, ma anzi può essere utile partire da dati che sono già pubblici in altri formati o sedi e adattarli alle specifiche open data.

Per quanto riguarda l'interesse al riuso, nel caso delle informazioni sulle pubblicazioni erano arrivate richieste di consentire il recupero automatico di tali informazioni attraverso procedure automatiche, a cui non si era potuto rispondere perché non esisteva una modalità di accesso ai dati che non richiedesse o una forma di autorizzazione, o complesse procedure di selezione dei dati. I dataset open data, al contrario, sono esposti senza alcuna intermediazione umana necessaria, e sono direttamente accessibili da un computer.

Il riuso quindi è inteso principalmente come interazione diretta computer-computer.

Nell'attività di selezione dei dati, infine, può essere utile fare riferimento a esperienze già concluse da strutture analoghe di altre istituzioni: nel caso del nostro progetto è stato particolarmente utile il contatto con l'Università di Trento, che ha condiviso le specifiche del proprio progetto già concluso.

4. Chi si occupa del progetto?

Internamente alla struttura che intende rilasciare dataset open data è richiesto il coinvolgimento di figure esperte di dominio, possibilmente anche con conoscenze tecnico informatiche che consentano la gestione sia degli aspetti legati alla produzione del dataset (selezione dei dati e loro estrazione da banca dati/applicativo) oltre che per la gestione di progetto.

E' altresì utile coinvolgere una figura della direzione ICT per la verifica di aspetti tecnici e operativi sul progetto (documentazione tecnica e logiche implementative), oltre che per la condivisione degli aspetti strategici che coinvolgono anche la direzione ICT.

5. Dove pubblicare in open data

Una volta selezionato il set di dati da pubblicare in open data, va definito dove verranno pubblicati: in questo momento a seguito del progetto delle pubblicazioni l'Ateneo è formalmente un contributor del portale regionale Open Data Lombardia, pertanto si consiglia di utilizzare questo canale.

Dato che in questo momento nel portale regionale è presente il solo progetto delle informazioni sulle pubblicazioni, andranno definite le modalità di partecipazione di eventuali altre strutture che decidano di esporre propri dataset. Questa operazione andrà condotta in modo da ottimizzare l'esposizione dei dati relativi all'Ateneo e alle strutture che pubblicano i dataset e sono quindi responsabili della loro manutenzione.

L'utilizzo del portale lombardo ha una ricaduta sui passaggi di trattamento dei dati che vedremo nei prossimi capitoli, perché il portale utilizza una specifica piattaforma opensource, SOCRATA, che ha precise modalità di raccolta dei dati in parte diverse da quelle di altri portali regionali. Inoltre il portale deposita i dati nel portale nazionale dati.gov, secondo modalità tecniche anche in questo caso specifiche; infine, il dataset viene pubblicato anche sul portale europeo [European data](#).

6. Selezione, trattamento e estrazione dei dati

Per selezione dei dati in questa parte del progetto si intende la redazione di un preciso elenco di quali dati saranno esposti nel portale regionale.

Nel caso delle pubblicazioni, ad esempio, si è scelto di includere l'anno di pubblicazione escludendo il mese e il giorno anche se presenti, la tipologia di pubblicazione (articolo, libro...) ma non la sottotipologia (recensione, poster...).

In questa fase va fatto riferimento alla banca dati da cui i dati vanno estratti, e quindi l'elenco sarà un elenco di campi che verranno indicati con l'etichetta assegnata all'interno del sistema, meglio ancora se nel database o vista che si userà per l'estrazione e non nell'interfaccia utilizzata per inserire e/o consultare i dati.

Nell'elenco, che si consiglia di redigere come una tabella o foglio di calcolo, andranno indicate eventuali modifiche da effettuare sui valori che risulterebbero poco leggibili (ad es. per le pubblicazioni si è trasformato il valore 'openAccess' in 'Open').

E' anche necessario definire eventuali filtri di selezione, ad esempio nel caso delle pubblicazioni si è deciso di esporre quelle degli ultimi 5 anni, e di escludere determinate tipologie (ad esempio le tesi di dottorato).



Una volta ottenuto l'elenco dei campi e definite le modifiche necessarie, si deve procedere all'estrazione dei dati dall'applicativo in cui risiedono.

L'estrazione non deve risultare in un file statico, ma deve essere condotta a livello del database, realizzando una vista dei dati ad hoc. In sintesi, una vista su un database è una interrogazione che restituisce un sottoinsieme dell'intero database sulla base di specifici parametri di ricerca: il sottoinsieme sono i campi dell'elenco (con le modifiche eventualmente definite), i parametri di selezione sono i filtri di ricerca.

7. L'invio dei dati al portale regionale

Il portale Open Data Lombardia offre diverse modalità di caricamento dei dataset.

La modalità di caricamento manuale prevede la creazione di un nuovo dataset (click su "Crea", e poi su "Seleziona" sotto "Set di dati", indicazione del nome del dataset e click su "Crea Dataset") e poi l'utilizzo della funzione "Aggiungi dati": al click sul pulsante omonimo si potrà caricare, anche per trascinamento, i dati prodotti al punto precedente in un formato a scelta tra quelli supportati. Si può utilizzare la procedura di caricamento manuale anche per l'aggiornamento dei dati, e in questo caso non sarà più necessario creare un nuovo dataset ma basterà individuare quello già creato e usare la funzione "Aggiorna dati".

Questa procedura di caricamento dei dati mediante un file è particolarmente indicata per il primo caricamento dei dati prodotti, perché è veloce, non richiede particolari competenze e consente di concentrarsi sulla verifica dei dati che vengono caricati.

In particolare, una volta effettuato il primo caricamento con questa modalità va controllato quanto il sistema scriverà nella sezione "Info sulla categoria" del dataset, in particolare la sezione "Colonne" in cui ci dovranno essere tutti i campi definiti al capitolo precedente, e "Dati", in cui invece si dovrà verificare che tutte le righe del file siano presenti; è infine consigliato effettuare verifiche a campione su qualche riga per verificare che i dati siano stati tutti caricati senza errori e visualizzati correttamente.

Una volta che nel portale sia presente una versione corretta dei dati, si potrà definire una modalità di caricamento e aggiornamento automatica dei dati. Nel portale Open data Lombardia la via privilegiata e ottimale è lo sviluppo di API ad hoc secondo quanto descritto al link <https://dev.socrata.com/publishers/>

Le API (Application Programming Interface) sono, in sintesi e con riferimento agli open data, procedure informatiche che vanno scritte secondo specifiche particolari in modo da permettere l'invio e l'aggiornamento di dati tra due applicazioni informatiche.

Lo sviluppo di API per la pubblicazione di dati in formati "aperti" verso il pubblico è particolarmente indicato per soddisfare sempre più crescenti richieste di utenti interni e esterni all'Ateneo per un riuso dei dati machine-to-machine.

Una valida alternativa, che è stata quella utilizzata per il progetto alla base di queste linee guida, è lo sviluppo di una procedura di estrazione, trasformazione e caricamento (ETL) che consente il trasferimento dei dati (come ottenuti al punto precedente) a SOCRATA a intervalli periodici in modo da garantirne l'aggiornamento; va precisato che la frequenza di aggiornamento non è solo un parametro tecnico ma un indicatore di qualità dei dati open data.



8. La pubblicazione sul portale regionale

I dataset open data vengono esposti con una serie di metadati descrittivi (ad es. titolo, frequenza di aggiornamento, formato dei dati, licenza ecc.). Questi metadati vengono definiti sul portale regionale stesso, e non solo esposti insieme al dataset, ma anche inviati al portale nazionale dati.gov; è importante prestare attenzione alla loro compilazione, a cui queste linee guida dedicano un'appendice.

La definizione dei metadati sul portale regionale, in concreto, viene effettuata con la funzione "Edita meta dati del set di dati": al click sul pulsante "Modifica Metadati" viene prodotta una maschera di inserimento dei metadati descritti uno per uno in appendice.

Una volta caricati i dati e inseriti i metadati si può procedere alla pubblicazione vera e propria: basta un click sul pulsante "Pubblica dataset" e definire che vogliamo che i nostri dati diventino pubblici: una volta confermata la pubblicazione sarà effettiva, e nella pagina del dataset sarà indicato "Pubblico" a fianco del nome.

Per modificare un dataset pubblicato bisogna utilizzare il pulsante Modifica che il sistema mostra al posto di "Pubblica dataset" per i dataset che sono stati pubblicati: questo non "spubblica" il dataset, ma crea una copia detta "bozza" che può essere modificata e che sostituirà il dataset pubblicato solo una volta che sarà stata pubblicata a sua volta.

Sul portale è presente una interessante funzionalità di realizzazione di viste sul dataset: si possono cioè suddividere i dati del dataset in parti che si ha interesse a esporre in modo separato e in aggiunta all'intero dataset (ad esempio per il progetto delle pubblicazioni si è realizzata una vista per ognuno dei 5 anni compresi nel dataset complessivo); al momento in cui queste linee guida sono state scritte non è possibile ordinare le viste di un dataset.

Un'altra funzionalità che può essere di interesse è quella di generazione di visualizzazioni (grafici, mappe, ecc) che possono essere incorporati in altri portali web sia tramite codice HTML (IFrame) che Javascript.

9. Conclusione: si può fare!

La pubblicazione di dataset open data è ormai un preciso adempimento delle pubbliche amministrazioni; AGID ha fissato anche un determinato numero di dataset da esporre in un dato periodo di tempo a seconda della tipologia di pubblica amministrazione, e non è quindi più una attività eludibile.

Tuttavia, è importante considerare che quanto sopra risponde a criteri di trasparenza e qualità della risposta alle esigenze dei portatori di interesse che il nostro Ateneo condivide e sostiene.

Il progetto delle informazioni sulle pubblicazioni, inoltre, ha dimostrato che pubblicare dataset open data è fattibile e non presenta criticità particolari per UNIMI.

Invitiamo quindi tutte le strutture a partecipare alla realizzazione di propri dataset open data con convinzione e senza particolari timori.

10. Riferimenti

Codice dell'Amministrazione Digitale <https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07:82>

AGID <https://www.agid.gov.it/>



AGID – Open Data <https://www.agid.gov.it/it/ambiti-intervento/open-data>

AGID – Linee guida in ambito Dati e open data <https://www.agid.gov.it/it/linee-guida#index-1>

Portale dati.gov.it <https://www.dati.gov.it/>

Portale regionale Open Data Lombardia <https://www.dati.lombardia.it/>

Software SOCRATA alla base del portale Open data Lombardia <https://dev.socrata.com/>

Licenze Creative Commons <https://creativecommons.it/chapterIT/index.php/license-your-work/>

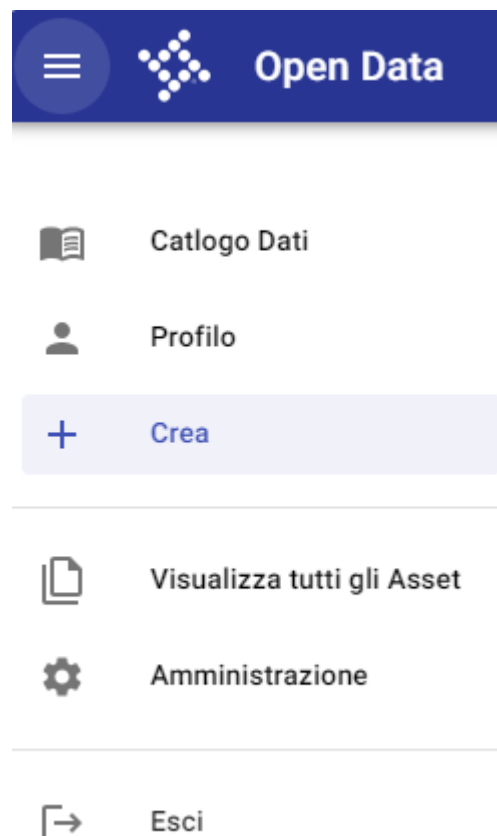
API Application programming interface https://it.wikipedia.org/wiki/Application_programming_interface

ETL Extract, transform, load https://it.wikipedia.org/wiki/Extract_transform_load

11. Appendici

Open Data Lombardia per immagini

Di seguito le schermate che descrivono la creazione di un dataset, il caricamento manuale dei dati e l'inserimento dei metadati





Crea un Nuovo Dataset

Per iniziare, scegli un nome per il tuo dataset

Crea Dataset

☰

Open Data

Cerca

+

?

🔔

Prova Bozza Pubblica dataset

Prova

Azioni di dati

+

Aggiungi dati

Aggiungi tabellare, geospaziale, non analizzabile o dati con link esterno

Aggiungi dati

🗄️

Rivedi e configura i dati

Revisiona i dati

✎

Edita meta dati del set di dati

Aggiungi contesto per aiutare gli utenti a capire i tuoi dati

Modifica Metadata

✎

Edita metadati della colonna

Modifica Metadata

Dati

Carica un file dati

Importa dati da URL

Connessioni

Collegamento a una Fonte Dati Esterna

Trascina Qui un Nuovo File di Dati

...oppure clicca Sfoglia per selezionare un file dal tuo computer

Sfoglia

I tipi di file supportati sono: Valori separati da virgola (CSV), Valori separati da tabulazione (TSV), Microsoft Excel (XLS), Microsoft Excel (OpenXML), Archivio ZIP (shapefile), Formato JSON (GeoJSON), Formato GeoJSON, Keyhole Markup Language (KML), Zipped Keyhole Markup Language (KMZ)

Sono accettati anche altri tipi di file, ma saranno disponibili solo per il download. I file di dimensioni inferiori a 1 GB saranno sottoposti a una scansione di sicurezza per individuare eventuali problemi. È inoltre possibile caricare il file solo per il download.

Il caricamento delle cartelle non è supportato

Università degli Studi di Milano – DIREZIONE Performance, Assicurazione Qualità, Valutazione e Politiche di Open Science
Via Festa del Perdono, n°7 - 20122 Milano, Italy



Modifica metadati

X

TITOLO E DESCRIZIONE

Fai in modo che il titolo e la descrizione siano il più semplici e chiari possibile.

* CAMPO OBBLIGATORIO

* Titolo dell'attività

Prova

Breve descrizione

Inserisci una descrizione

Etichetta della riga

Descrivere cosa rappresenta ciascuna riga della risorsa
(se del caso).

Etichetta della riga

ad es: crimine, persona, ospedale

CATEGORIE E TAG

Classificate le vostre attività per facilitarne la ricerca.

Categoria

-- Selezione un valore --

Tag e parole chiave

LICENZE E ATTRIBUZIONE

Tipo di licenza pubblica (se applicabile)

-- Nessuna Licenza --

Dati forniti da

Annullamento

Risparmiare

I metadati descrittivi di Open Data Lombardia

Di seguito vengono presentati e discussi i metadati richiesti dal portale Open Data Lombardia per ogni set di dati da pubblicare.

I metadati vengono inseriti e/o modificati con la funzione “Edita meta dati del set di dati” presente una volta effettuato il login nel portale (vedi anche appendice precedente).

Titolo del set di dati [obbligatorio]: è il nome che si vuole dare al dataset; dovrebbe essere specifico e consentire sia di identificare i dati che di facilitarne il ritrovamento

Breve descrizione [obbligatorio]: descrizione breve ma esaustiva del contenuto del dataset; anche in questo caso è utile pensare a una descrizione che faciliti il ritrovamento del dataset

Etichetta della riga: da non compilare

Categoria [obbligatorio]: va scelta una tra le categorie proposte; nella scelta va però considerato che quanto indicato sarà poi esportato nel portale nazionale dati.gov che presenta una propria categorizzazione (<https://docs.italia.it/italia/daf/linee-guida-cataloghi-dati-dcat-ap-it/it/stabile/temi.html>) in linea con il vocabolario controllato europeo Data theme (<https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/data-theme>). In concreto per la scelta della categoria è opportuno o confrontarsi con il portale Open Data Lombardia oppure verificare dopo la pubblicazione quale categoria risulta in dati.gov

Tag e parole chiave: si tratta di parole chiave libere che descrivono il dataset. Mentre vengono digitate il sistema le autocompleta se già presenti e in questo caso vanno selezionate col mouse, altrimenti va



premuto invio da tastiera per inserire una nuova parola o frase. Il sistema di tag di Open Data Lombardia è in corso di revisione, quindi conviene contattare il portale Open Data Lombardia per capire quali inserire, anche in questo caso tenendo conto di come saranno gestite da dati.gov

Tipo di licenza pubblica: si consiglia di indicare dalla tendina "[Creative Commons 1.0 Universal \(Public Domain Dedication\)](#)", indicata per set di dati che possono essere riutilizzati a qualunque scopo; in alternativa la scelta "[Creative Commons Attribution 4.0 International \(CC BY 4.0\)](#)" è indicata per set di dati per i quali si vuole che venga esplicitato chi li ha prodotti

Dati forniti da: si consiglia di indicare l'ufficio titolare dei dati preceduto dall'Ateneo, ad es. "Università degli Studi di Milano - Ufficio Qualità dei Dati sulla Ricerca"

Link alla fonte: indicare a che indirizzo web possono essere trovati i dati nella loro forma originale, ad es. "https://air.unimi.it"

Indirizzo e-mail: non utilizzare

Nome della risorsa: non utilizzare

Allegati: è possibile caricare file (ad es. pdf) che si ritiene possano essere utili a contestualizzare il dataset

Dataset Page: non utilizzare

Referente: ripetere l'ufficio indicato in "Dati forniti da" senza l'Ateneo, ad es. "Ufficio Qualità dei Dati sulla Ricerca"

Email referente: indicare la mail personale o di ufficio di riferimento per i dati originali (ad es. air@unimi.it)

Metodo di pubblicazione: non utilizzare

Frequenza (di aggiornamento): indicare la frequenza con cui i dati vengono aggiornati, ad es. "giornaliera"

Data ultima modifica: se l'aggiornamento è automatico, si consiglia di indicare "Vedere la data di Ultimo aggiornamento Dati"

Nome Serie: non utilizzare

Link alla Serie: non utilizzare

Microsite: non utilizzare

Ulteriori informazioni (Direzione, Fonte, Viewer geografico, Dettaglio metadati, Link_allegati): non utilizzare se non pertinenti al dataset da pubblicare