

Cataloghi per i dati aperti

Formez_{PA}



Questo materiale didattico è stato realizzato da Formez PA nel *Progetto PerformancePA*, Ambito A Linea 1, in convenzione con il Dipartimento della Funzione Pubblica, organismo intermedio del Programma Operativo Nazionale Governance e Azioni di Sistema (PON GAS), Asse E Capacità istituzionale. Il PON GAS è cofinanziato dal Fondo Sociale Europeo ed è a titolarità del Ministero del Lavoro e delle Politiche Sociali.

L'opera è distribuita con Licenza [Creative Commons Attribuzione - Condividi allo stesso modo 4.0 Internazionale](https://creativecommons.org/licenses/by-sa/4.0/).



Autore: Vincenzo Patruno

Creatore: Formez PA

Diritti: Dipartimento della Funzione Pubblica

Data: Ottobre 2015

Cataloghi per i dati aperti

Come sono fatti i dataset

Un dataset è una serie di dati, solitamente rappresentati con una tabella, in cui possiamo trovare il dato, più una serie di attributi che servono a descrivere il dato stesso.

Per fare un esempio un foglio elettronico con l'elenco di tutte le farmacie d'Italia può essere considerato un dataset.

In questo caso il dato può essere il nome della farmacia mentre la partita iva, l'indirizzo, il codice Istat del Comune sono elementi identificativi del dato, servono a dettagliare il dato, possono essere identificati come "dimensioni" e definiscono i metadati strutturali, cioè definiscono la struttura stessa del dataset.

Publicare i dataset online

Pensiamo ad un Ente che si trova nella situazione di dover pubblicare svariati dataset di vario genere: semplici elenchi come quelli delle farmacie, le fermate delle stazioni meteo, gli orari dei musei, delle biblioteche, la posizione degli hotel; i dati relativi ad esempio alla quantità di raccolta differenziata, i dati sulla qualità dell'aria, i dati sulle gare e appalti, i dati sulla spesa sanitaria e così via.

Una buona idea è organizzare i dataset all'interno di un catalogo di dati su Internet. Questa soluzione ha un duplice scopo:

- consentire all'Ente di avere un sistema web di riferimento per la diffusione dei dataset;
- permettere a possibili utilizzatori di avere un sistema abbastanza agevole per la consultazione del catalogo e l'identificazione dei dataset di interesse

Publicare i dataset in un catalogo

Il caso più semplice di pubblicazione di un dataset in un catalogo online è quando abbiamo un elenco sotto forma di foglio elettronico, con informazioni ben strutturate e facilmente gestibile come un file.

Publicare un dataset su un catalogo vuol dire trasferire fisicamente il file sulla piattaforma. E' un'operazione che viene fatta da utenti autorizzati che attraverso delle credenziali possono accedere all'interfaccia di gestione del catalogo.

Identificare e descrivere i dataset pubblicati sul catalogo

Una volta che i dataset sono stati caricati sul catalogo online vanno descritti con ulteriori metadati quali: il titolo, la descrizione, il tipo di licenza d'uso, la data di pubblicazione e la categoria.

Questi metadati sono di tipo diverso da quelli visti precedentemente e identificano l'intero dataset.

Se un dataset è descritto correttamente può essere trovato più facilmente dagli utenti non solo sul web grazie ad una migliore indicizzazione nei motori di ricerca, ma soprattutto all'interno del catalogo sul quale è stato pubblicato ad esempio selezionando una categoria.

Per quanto riguarda la categoria, il dataset delle farmacie lo andremo ad esempio a collocare nella categoria "Sanità".

Un esempio di catalogo di dataset

A titolo di esempio vediamo come funziona il catalogo Open Data della Regione Lombardia.

Sulla home page sono messe bene in evidenza le singole categorie dei dataset. Cliccando sulla categoria Sanità si va a visualizzare una lista di dataset. Leggendo titoli e descrizione dei dataset pubblicati sarà possibile identificare il dataset a cui si è interessati. Che nel caso del catalogo della Regione Lombardia verrà visualizzato sullo schermo, in questo caso sotto forma di tabella, prima di poter essere scaricato.

Dare valore ai dataset con i dati georeferenziati

Abbiamo visto che il dataset relativo all'elenco delle farmacie ha una dimensione chiamata "indirizzo" dove è presente la via della farmacia stessa. Con l'operazione di geocoding, è possibile associare ad un indirizzo le sue coordinate geografiche, ossia aggiungere al dataset due colonne, quelle relative alla latitudine e alla longitudine. Il risultato sarà un dataset geografico. Il dato di partenza, in questo caso la singola farmacia, verrà quindi georiferita sul territorio, permettendo di visualizzare i dati su una mappa (direttamente all'interno del catalogo).

I cataloghi di dati danno la possibilità di pubblicare e visualizzare anche dati geografici, cioè dati associati a elementi geometrici vettoriali come punti, linee o poligoni.

L'informazione relativa alla posizione geografica del dato è espressa in un definito sistema geodetico di riferimento.

dati.gov.it, il catalogo nazionale dei dati aperti

Pubblicato nel 2011 dati.gov.it è il catalogo nazionale dei dati aperti.

Questo “metacatalogo” raccoglie in modo automatico i metadati dei dataset contenuti nei cataloghi di diverse pubbliche amministrazioni, e consente di ricercare agevolmente informazioni sugli oltre 10.000 datasets disponibili.

E’ da notare in questo caso quanto sia importante la descrizione di un dataset. Questo catalogo infatti non contiene fisicamente i dataset ma solo la loro descrizione (metadati). I dati sono disponibili sul catalogo della PA che li ha pubblicati.

Il metacatalogo

Per fare sì che tutto questo possa avvenire in modo automatico servono sostanzialmente due cose.

- che il contenuto del singolo catalogo sia accessibile in formato “*machine to machine*”¹;
- che si seguano alcune specifiche tecniche, cioè una serie di standard che consentono di riportare ad uno schema comune tutti i metadati di tutti i dataset pubblicati sui diversi cataloghi.

Questa è una operazione che viene fatta generalmente attraverso l’utilizzo di API², ossia di speciali interfacce che vengono utilizzate da applicazioni software e che consentono al metacatalogo di connettersi direttamente ai dati pubblicati, in questo caso al database del singolo catalogo che contiene i metadati di ciascun dataset.

Per far sì che tutti i cataloghi dei Paesi europei “parlino la stessa lingua” è stato definito uno standard comune chiamato DCAT³, dove vengono definiti quali sono i metadati obbligatori come ad esempio il titolo del dataset, la descrizione, l’area tematica e quali quelli facoltativi.

I cataloghi internazionali

Tra i cataloghi internazionali, vanno sicuramente citati data.gov.uk del Regno Unito e data.gov, il catalogo Open Data degli Stati Uniti d’America, che ad oggi catalogano rispettivamente circa 25.000 e 190.000 datasets. Questi due cataloghi sono stati i primi meta cataloghi rilasciati da governi centrali e hanno fatto da apripista, assieme all’Australia e alla Nuova Zelanda, per tutto quanto sta accadendo ora a livello mondiale in termini di governo aperto, trasparenza, accountability, agenda digitale e opportunità per il mercato. La

¹ Un formato cioè fatto apposta per essere letto da applicazioni software.

² Acronimo di Application Programming Interface (intraducibile in italiano con “Interfaccia di Programmazione di un’Applicazione”), le API permettono ad applicazioni esterne di interagire con altre piattaforme informatiche sulla base di un set specifico di regole.

³ DCAT è uno standard europeo progettato per facilitare l’interoperabilità tra i cataloghi di dati pubblicati sul Web.

pubblicazione di questi due cataloghi di dati aperti nel 2010 ha dato un impulso determinante per la diffusione su scala mondiale della cultura dell'Open Data e Open Government, ridefinendo il rapporto tra cittadini e amministrazioni, aziende e territorio, innovazione e pubblica amministrazione.

Alcune piattaforme per esporre dati in formato aperto

L'esposizione dei dati aperti rilasciati dalla pubblica amministrazione avviene su piattaforme tecnologiche che hanno specifiche funzioni dedicate alla catalogazione dei dataset. Questo tipo di piattaforme tecnologiche prende il nome di data store, traducibile come "negoziato di dati". Con l'affermarsi su scala internazionale della pratica dell'open data le tipologie di data store offerte dal mercato sono moltiplicate, tra le più rappresentative: CKAN⁴, DKAN⁵ che ha il vantaggio di integrarsi direttamente con Drupal⁶ e Socrata⁷.

⁴ Piattaforma Open source sviluppata da Open Knowledge, organizzazione no profit con sede in UK che promuove la conoscenza aperta come elemento di innovazione del sistema sociale ed economico. Grazie alla esaustiva documentazione e alla sua versatilità in fase di sviluppo, CKAN è adottata da molte pubbliche amministrazioni di differenti livelli e nazionalità che vanno dal Governo centrale degli Stati Uniti d'America al Comune di Matera.

⁵ Piattaforma Open source sviluppata da NuCivic, azienda americana con base a New York che promuove l'innovazione all'interno della pubblica amministrazione e delle strutture governative. DKAN ha il vantaggio di integrarsi nativamente con il CMS Drupal. Il Governo italiano si è orientato su DKAN come scelta tecnologica per il catalogo del portale nazionale dei dati aperti dati.gov.it.

⁶ Drupal è una piattaforma software di content Management System (CMS), modulare, aperta, ovvero un programma che permette la creazione e distribuzione di complessi siti web dinamici.

⁷ Piattaforma sviluppata dalla società omonima che fornisce soluzioni per digital government, si tratta di una soluzione proprietaria che punta non solo sulla catalogazione dei dati, ma anche sulla loro visualizzazione dinamica online. Ad esempio è la soluzione scelta dalla Regione Lombardia per il proprio datastore.