ALLEGATO B

AGENDA DIGITALE LOMBARDA

Processi e Standard
del portale Open Data di Regione Lombardia

Novembre 2018

Scopo del documento

Il presente documento definisce i processi e gli standard a cui **devono** attenersi tutti i soggetti autorizzati a pubblicare sul portale Open Data di Regione Lombardia, esposto su internet all'indirizzo www.dati.lombardia.it.

Lo scopo della definizione degli standard è quelli di creare le condizioni per una pubblicazione ordinata e di qualità dei dati e favorire, in sostanza, la ricercabilità e la fruibilità dei dati.

Nel capitolo **Errore. L'origine riferimento non è stata trovata.** "Processi***Errore. L'origine riferimento non è stata trovata.***" vengono presentati i processi e le differenti modalità di pubblicazione che il portale Open Data di Regione Lombardia mette a disposizione dei soggetti che pubblicano.

Nel capitolo 2 "*Convenzioni sui nomi*" sono illustrate le convenzioni sui nomi del pubblicatore, dei dataset e dei "token" di pubblicazione.

Nel capitolo 3 "*Tipologia di dataset pubblicabili*" sono illustrate le diverse tipologie di dataset pubblicabili e le viste che è possibile derivare dagli stessi.

Nel capitolo 4 "*Tipologia di dati pubblicabili*" sono illustrate le tipologie di dati che possono essere pubblicate all'interno di dataset tabellari e le indicazioni per un utilizzo ottimale.

Nel capitolo 5 "*Metadatazione*" è trattato il tema della metadatazione, illustrato lo standard di metatadazione del portale Open Data della Regione Lombardia e come questo viene mappato nel profilo dello **standard DCAT-AP_IT**.

Nel capitolo 6 "*Vocabolari controllati*" sono illustrati i vocabolari standardizzati e come sono mappati con vocabolari standard europei o nazionali.

Nel Capitolo 7 "Definizione delle serie di dati" è illustrato come gestire e descrivere le serie di dati.

Per chiarimenti o suggerimenti in merito al presente documento, scrivere a: admin@dati.lombardia.it

Sommario

1 Processi

Aprire i dati a disposizione dell'Amministrazione comporta una serie di passaggi che è opportuno strutturare in maniera esplicita attraverso un processo organizzato che prenda in considerazione le diverse variabili esistenti.

L'approccio per processi qui descritto, declinato nell'attività di apertura dei dati dell'Amministrazione, sottolinea l'importanza di:

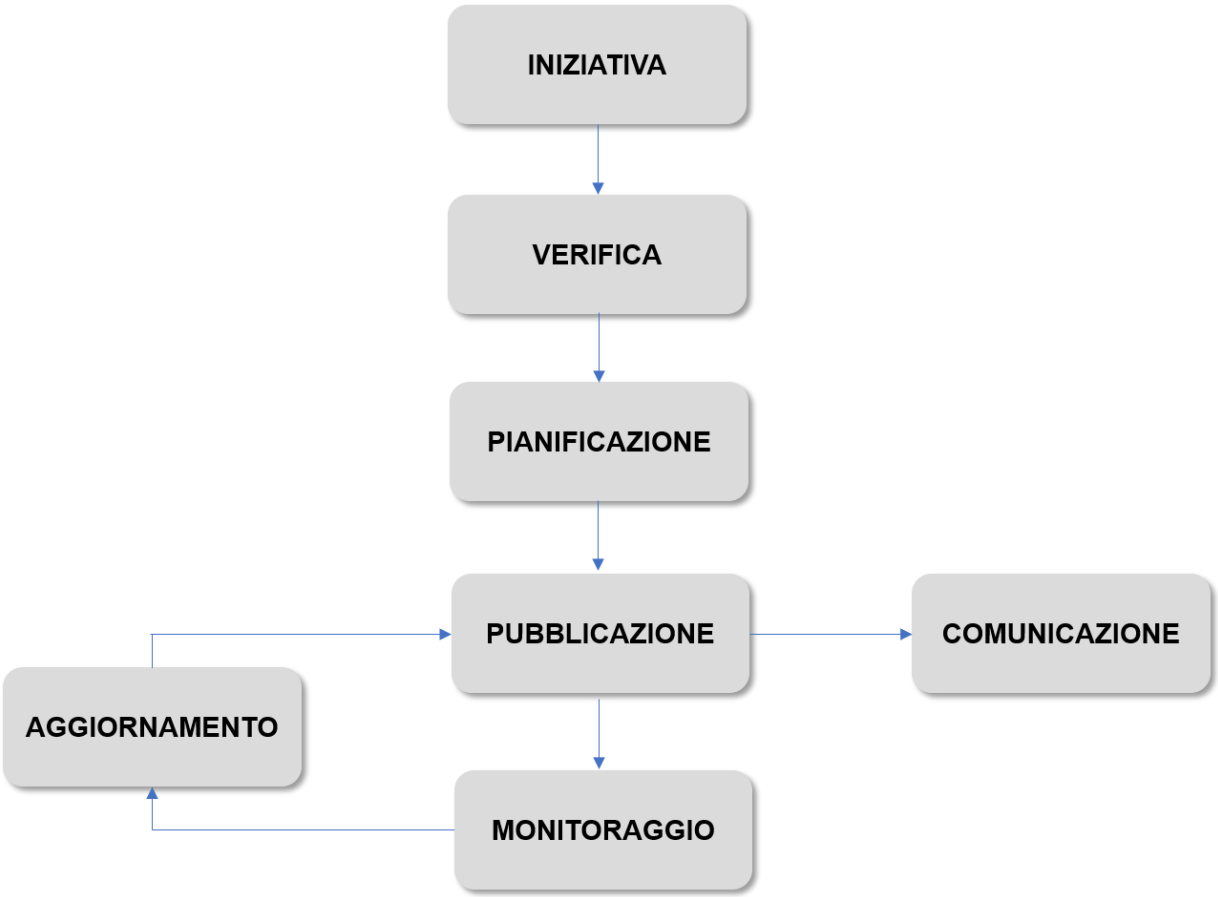
- aver chiaro i requisiti di partenza, nelle diverse fasi connesse all'apertura dei dati;
- valutare le attività previste dai processi;
- conseguire risultati misurabili in termini di efficacia complessiva dell'azione di apertura dei dati;
- ottenere un miglioramento continuo dei processi, basato su misurazioni oggettive definite dal Team Open Data.

Descrivere i processi connessi all'apertura dei dati a disposizione di un'Amministrazione Pubblica vuol dire identificare e descrivere la sequenza strutturata di azioni che sottendono tale attività, partendo dall'identificazione del dato da pubblicare e passando per la sua elaborazione, per arrivare alla pubblicazione e diffusione attraverso i canali più idonei.

Spostandosi da una visione complessiva e sistemica a una maggiormente operativa e analitica, a valle della decisione di aprire un insieme di dati in possesso dell'Amministrazione per metterli a disposizione del cittadino e della comunità, è possibile identificare un processo strutturato nelle seguenti fasi:

1. Iniziativa di pubblicazione dei dataset
2. Verifica dei dataset da pubblicare
3. Pianificazione dei dataset da pubblicare
4. Pubblicazione dei dataset
5. Comunicazione e promozione dei dataset pubblicati
6. Monitoraggio dell'efficacia
7. Aggiornamento dei dataset pubblicati

Descrizione del “Ciclo di vita” del dataset e delle attività di gestione del dato:



Di seguito sono dettagliate le singole fasi del processo.

1.1 Iniziativa di pubblicazione dei dataset

Questa fase prevede l'identificazione dei dati da pubblicare in base a un’analisi che risponda alle seguenti domande:

1. quali sono i dati a disposizione dell’Amministrazione?
2. quali sono i dati d’interesse per la Comunità?

1.1.1 Quali sono i dati a disposizione dell’Amministrazione?

L’Amministrazione Pubblica, come qualsiasi organizzazione complessa, gestisce quotidianamente una grande quantità di dati che servono per garantire l’operatività e l’erogazione dei servizi al cittadino. Con lo sviluppo dell’Information & Communication Technology (ICT) e la diffusione dei sistemi di e-Government, tali dati sono ormai disponibili in formato digitale e possono quindi essere condivisi e diffusi. Per farlo, però, è necessario identificarli nell’ambito della struttura amministrativa dell’Ente.

Definire dove sono i dati a disposizione, vuol dire:

- identificare quali sono le strutture funzionali dell'Amministrazione che detengono dati potenzialmente utili;
- comprendere, in ciascun ambito, quali sono le fonti di dati di rilievo; in altri termini, è necessario **identificare archivi strutturati, elenchi, basi dati** prodotte da software in uso presso l'amministrazione.

Le organizzazioni complesse non sempre sono a completa conoscenza dell'intero corpus di dati che producono o che gestiscono. In molti casi, infatti, il dato è il risultato di un processo o il sottoprodotto di un'elaborazione, funzionale ad altri scopi.

In generale, è possibile identificare queste tipologie di dati in possesso dell'Amministrazione:

- **dati necessari alla gestione delle attività:** sono tutti quei dati che servono all'Amministrazione per il corretto funzionamento dei suoi processi. Sono i dati in ingresso (o come abbiamo già citato, in input) che sorreggono e riforniscono i processi gestiti dall'Ente;
- **dati prodotti come risultato dell'attività**: sono tutti quei dati che l'Amministrazione produce, spesso alla conclusione di un iter procedurale "monolitico", nell'esercizio delle sue funzioni. Sono i dati in uscita (output) ottenuti come risultato dei processi e dei procedimenti gestiti dall'Ente;
- **dati prodotti nella gestione delle attività:** sono quei dati che l'Amministrazione ottiene o produce come sottoprodotto di un processo, che derivano da lavorazioni intermedie o che sono prodotti con la finalità di essere dati in pasto ad altri procedimenti.

1.1.2 Quali sono i dati d'interesse per la Comunità?

Nella La DGR XI/723 del 05/11/2018 che ha ri-definito i "Criteri per l'Open Data in Regione Lombardia" è suggerito di "*adottare criteri che tengano conto della potenziale utilità dei dati nella creazione di valore sociale e di mettere in atto modalità che possano permettere ai potenziali riutilizzatori di esprimere il proprio interesse, al fine di attuare una politica di apertura sempre più guidata dalla domanda. Allo scopo, sul portale Open Data, è presente in prima pagina la funzione "Suggerisci un dataset", che permette a chiunque di richiedere la pubblicazione di un dato."*.

I possessori dei dati sono invitati ad individuare iniziative che possano essere utili a coinvolgere sia gli stakeholder interni all'amministrazione che quelli esterni (es. (e.g., studenti universitari, soggetti preposti a indagini e analisi statistiche e/o economiche, datajournalist, startup e aziende).

È tuttavia necessario riconoscere che non essendo semplice valutare l'interesse che susciterà un determinato dataset, si suggerisce, quando un dato è pubblicabile senza costi significativi, di provvedere alla sua 'apertura' anche nel momento in cui non se ne ravveda un'utilità immediata.

1.2 Verifica dei dataset da pubblicare

Una volta identificate le possibili fonti di dataset, è necessario procedere ad analisi più approfondite per valutare l'effettiva "pubblicabilità" dei dati come Open Data.

Questa attività deve essere svolta da chi sa come e dove i dati sono gestiti fisicamente (referente informatico dell'Ente), in cooperazione con chi conosce il dato e le sue funzionalità, il responsabile del procedimento.

Per ogni singolo dataset potenziale occorre:

1. verificare se i dati sono distribuibili sotto il profilo giuridico;
2. analizzare la qualità dei dati;
3. definire i processi di produzione del dataset;
4. produrre documentazione di supporto.

1.2.1 Verificare se i dati sono distribuibili sotto il profilo giuridico

Nel caso in cui un Ente terzo voglia pubblicare, sul portale Open Data di Regione Lombardia, dataset contenenti dati personali di cui l'Ente è Titolare del trattamento, Regione Lombardia deve essere nominata dall'Ente – con specifico atto di nomina – quale Responsabile del trattamento.

La valutazione sulla liceità e necessità del trattamento dei dati personali resta in capo all'Ente Titolare del trattamento.

In qualità di Titolare del trattamento dei dati, l'Ente determina le finalità e i mezzi del trattamento (art. 4, punto 7 GDPR) nel rispetto del combinato disposto tra art. 5, 6 e considerando n. 50 del GDPR. Ciò significa che i responsabili dei procedimenti, con il supporto del Privacy Officer e/o del Responsabile della Protezione dei Dati (RPD), decidono se è opportuno pubblicare dati personali previa valutazione sulla liceità e necessità del trattamento. A questa valutazione sono legate anche la scelta della **Licenza** di utilizzo dei dati e della «**persistenza**» dei dati (ovvero l'intervallo di tempo in cui è lecito che ciascun dataset resti pubblicato), che devono essere indicate nella scheda metadati di ciascun dataset.

In questa fase il Titolare del dato deve anche confermare che il dato è pubblicabile con licenza aperta (IODL2.0 o CC-BY 4.0) o scegliere una licenza più restrittiva.

1.2.2 Analizzare la qualità dei dati

Dopo diversi anni di esperienze in ambito Open Data, anche in Italia, si è compreso che la qualità dei dati è un aspetto fondamentale da curare e che può incidere significativamente sulla probabilità di riuso da parte degli stakeholder.

È quindi importante eseguire verifiche qualitative sui dati, prima della loro pubblicazione e mettere in atto eventuali azione correttive laddove necessario.

Gli aspetti fondamentali di cui tenere conto sono:

- Accuratezza sintattica - Il formato dei dati dovrebbe essere controllato perché sia coerente con la realtà (es. una data dovrebbe essere riportata sempre nel formato corretto);
- Accuratezza semantica - Il valore dei dati deve essere il più possibile affidabile: ciò si ottiene utilizzando soprattutto delle fonti interne o esterne certificate e/o controllate.
- Tempestività – È importante garantire che l'aggiornamento del **dataset pubblicato in Open Data avvenga in tempi congrui** con quelli di aggiornamento dell'informazione reale. **Il livello adeguato di aggiornamento** dipende dal tipo di dato: vi sono alcuni dati che hanno bisogno di essere aggiornati molto frequentemente e altri per i quali è sufficiente una frequenza di aggiornamento periodica. La frequenza di aggiornamento viene calcolata come il rapporto tra il periodo che passa tra due aggiornamenti del dataset e la periodicità stimata di variazione delle informazioni reali.

- Completezza – Occorre valutare **la completezza del dato e comunicarla:** viene calcolata come la percentuale di rappresentatività del dataset rispetto all'universo complessivo reale (ad esempio: il dataset biblioteche pubbliche della Lombardia' quante biblioteche contiene rispetto al totale?)

La pretesa di ottenere la perfezione assoluta non è realistica, anche perché richiederebbe dei costi eccessivi rispetto ai benefici, ma occorre verificare **se sono necessarie azioni correttive** per migliorare la qualità del dato, con la consapevolezza che questo intervento può essere realizzato in modalità graduale e progressiva.

È importante comunque giudicare se si è raggiunto un **livello sufficiente di correttezza, aggiornamento e completezza del dato** da consentire a enti terzi elaborazioni utili; tale livello dipende dal tipo di dato e dall'universo trattato.

È inoltre ampliamente dimostrato dall'esperienza come la scelta di aprire il dataset possa essere funzionale a migliorarne la qualità, anche attraverso processi di coinvolgimento della Comunità di riferimento. Sono numerosi i casi di segnalazioni di incompletezza o non accuratezza del dato che hanno favorito un aumento della qualità dei dati.

1.2.3 Definire i processi di produzione del dataset

Dopo aver eseguito l'analisi della qualità del dataset, si passa all'analisi da parte dei referenti tecnici che valutano prima di tutto la complessità ed il tempo necessario per gli interventi utili ad estrarre i dati ed eventualmente creare un "job" di pubblicazione/aggiornamento automatico.

Quest'attività tecnica necessita della stretta collaborazione tra il responsabile del procedimento (che conosce i processi di produzione del dato) e le figure tecniche del Team Open Data, che conoscono le funzionalità del portale Open Data e sono in grado di mettere a disposizione strumenti e buone pratiche.

In generale, la stima dei costi e dei tempi per la produzione del dataset, è influenzata da una serie di fattori:

1) **La fonte dei dati è identificata?** Occorre sapere dove sono fisicamente collocate le basi dati su cui si va a lavorare, chi ne ha la gestione, quali tecnologie sono in uso.
2) **La struttura dei dati è nota?** I moderni data base hanno una struttura che può essere anche molto complessa; occorre conoscerla per estrapolare i dati che si intende pubblicare.
3) **Sono necessarie trasformazioni dei dati?** Si è già accennato al fatto che talvolta si può decidere che il dataset sia composto di campi che sono a loro volta risultato di alcune elaborazioni, la cui complessità può essere variabile e incide sui tempi di produzione del dataset. Per fare alcuni esempi di possibili interventi che possono rendersi necessari, ecco alcune domande che chi vuole esportare il dataset dovrà porsi:
    a. **esistono già strumenti con funzionalità di esportazione dei dati?** In molti casi sono già state previste tali funzioni, in altre occorreranno interventi per realizzarle;
    b. **il dato è geo-localizzato o deve esserlo?** Se una base dati contiene già le coordinate per geo-localizzare l'informazione è cosa utile poterla esportare. In alcuni casi si hanno solo gli indirizzi che possono però anche essere utilizzati per geo-localizzare il dato;

questa però è un'attività che va preventivata e può richiedere l'impegno anche del gestore del dato che deve intervenire per i casi in cui l'indirizzo incompleto non permette una corretta geo-localizzazione;

c. **ci sono dati che devono essere resi anonimi?** Se sì, è necessario, prima che il dato sia trasmesso per la sua pubblicazione, prevedere l'intervento con opportune funzioni che trasformano in dati anonimi le informazioni;
d. **alcuni dati devono essere di sintesi**? Se non esiste già una funzione che restituisce dati aggregati o di sintesi dei dati originari, questa dovrà essere realizzata.

4) **Ogni quanto tempo il dataset deve essere aggiornato?** Questo dipende dalla frequenza di aggiornamento della base dati di origine. Considerando che i dataset sono immagini statiche del patrimonio informativo devono essere previsti congrui tempi di aggiornamento; per basi dati che hanno aggiornamenti periodici (es.: rilevazioni annuali), potrà essere previsto l'aggiornamento del dataset di conseguenza; nel caso invece di basi dati che subiscono aggiornamenti in modo costante, si dovrà prevedere una congrua tempistica di aggiornamento del dataset. Tutte le tempistiche vanno opportunamente indicate nei metadati (vedi par. successivo). È importante considerare che, una volta messo in linea, il dataset dovrà essere tenuto aggiornato e va deciso anche se aggiornare sostituendo il dataset pubblicato in precedenza o mantenendo la serie storica dei dataset.

1.2.4 Produrre documentazione di supporto

Nel corso dell'analisi dovrà essere predisposta un'opportuna documentazione che sarà allegata ai dataset al momento della pubblicazione.

1.2.4.1 *Metadata*

La descrizione dei dati tramite "metadata" è fondamentale per la comprensione dei dati da parte dei fruitori. A questo argomento è dedicato il ***Capitolo 5 Metadatazione.***

1.2.4.2 *Scheda descrizione dataset*

Se i metadati comunicano all'utente informazioni di base sul dataset, per favorirne il riutilizzo può essere certamente utile descriverne i contenuti qualora il nome degli attributi non sia auto-esplicativo o siano presenti delle codifiche.

La scheda di descrizione del dataset dà all'utente dettagli su com'è organizzato; possibili contenuti della scheda sono:

- descrizione dettagliata del dataset, da dove è originato, per quali scopi è realizzato;
- legenda dei nomi delle colonne (es. IDEXT = Identificativo Esterno);
- legenda di sigle, acronimi o altre abbreviazioni utilizzate nei contenuti (es. 1=Sì, 0=No).

Queste informazioni – in particolare la legenda sul significato dei nomi delle colonne – sono fondamentali per un corretto utilizzo dei dataset in seno ad applicazioni sviluppate da soggetti terzi, e devono essere prodotte da chi gestisce la base dati in collaborazione con la parte tecnica.

1.3 Pianificazione dei dataset da pubblicare

I dataset individuati e verificati possono essere predisposti e pubblicati in autonomia dall'Ente oppure con il supporto del Team Open Data di Regione Lombardia.

In caso di pubblicazione autonoma da parte dell'Ente (ad esempio gli Enti Locali che hanno aderito all'iniziativa di co-finanziamento), il Team Open Data di Regione Lombardia deve essere informato della pianificazione della pubblicazione effettuata dall'Ente stesso.

Nel caso in cui l'Ente si affidi al Team Open Data di Regione Lombardia per la pubblicazione (come accade ad esempio nel caso degli Enti del Sistema Regionale), è necessaria una pianificazione condivisa. In questo caso, sulla base delle informazioni raccolte in fase di verifica, il Team Open Data farà una stima dei giorni uomo previsti per le attività di pubblicazione del dataset (incluso la creazione di eventuali "job" automatici) ed una previsione di quando la stessa potrà essere realizzata in base alle attività in corso in quel periodo.

Eventuali urgenze devono essere concordate con il Team Open Data.

NOTA BENE

Nel caso il dataset contenga dati personali, il titolare del trattamento (Data Owner interno a Regione Lombardia o Ente terzo) deve, prima di pubblicare il dato o di chiederne la pubblicazione al team Open Data, inoltrare al Team Open Data una comunicazione nella quale dichiari di aver valutato i presupposti di liceità e necessità e deve assumersi la responsabilità della pubblicazione e delle scelte in merito alla licenza d'uso ed alla persistenza dei dati.

1.4 Pubblicazione dei dataset

Solo dopo le fasi di iniziativa, verifica e pianificazione è dato il via all'effettiva estrazione e pubblicazione del dataset sul portale Open Data anche se alcune delle fasi di estrazione possono essere anticipate, per dare supporto alle fasi di iniziativa e verifica.

Queste attività sono svolte dalla parte tecnica che deve, comunque, restare in contatto con chi gestisce le basi dati coinvolte nell'intervento e con il Team Open Data, nel caso l'Ente si affidi a Regione Lombardia per il processo di pubblicazione.

Le principali attività per la pubblicazione sono riassunte nei paragrafi seguenti.

1.4.1 Definizione del dataset

Partendo da quanto stabilito in analisi, si deve a questo punto stabilire con precisione quali informazioni saranno esportate, individuando quali dati e in che formato è possibile esporli (ad es.: formato tabellare, shape file, etc).

1.4.2 Estrazione

Nel caso in cui l'Ente pubblichi in autonomia le attività di estrazione sono interne all'Ente.
Nel caso l'Ente si affidi a Regione Lombardia per il processo di estrazione e pubblicazione, l'Ente realizzerà gli strumenti che permettono di mettere i dati a disposizione del Team Open Data (es. viste

sul DB, ETL, file in apposita area di scambio dati, etc) concordando modalità e tempi con il Team Open Data.

1.4.3 Geolocalizzazione

La geolocalizzazione delle informazioni ne facilita il riutilizzo all'interno delle applicazioni che fanno uso di mappe; se però non è realizzata alla fonte è richiesto un lavoro d'integrazione che può essere fatto anche in un momento successivo alla prima pubblicazione del dataset, ma del quale occorre tenere conto in termini di tempo e dell'impegno organizzativo.

Perché sia possibile procedere con la geolocalizzazione è necessario che nei dati siano presenti "indirizzo e numero civico" il più possibile "normalizzati".

Nel caso in cui l'Ente voglia chiedere il supporto al Team Open Data di Regione Lombardia per la geolocalizzazione è necessario considerare che l'attività richiede una pianificazione e che, siccome è probabile che il processo di geolocalizzazione produca una serie di errori, sarà necessario un suo intervento per la gestione dei casi non gestibili in automatico.

1.4.4 Pubblicazione sul portale Open Data

Il portale Open Data di Regione Lombardia offre due modalità di pubblicazione:

- pubblicazione tramite interfaccia Web amministrativa
- pubblicazione tramite API

La prima modalità risponde all'esigenza dei processi di pubblicazione "manuale" ed è **indicata laddove la frequenza con cui si pubblica o aggiorna il dato sia una tantum o al più annuale**. Essa è assimilabile all'utilizzo di un CMS per pubblicare contenuti redazionali su un portale informativo ed allo stesso modo ne condivide alcune criticità: la necessità di un intervento umano (costoso, poco efficiente e non sempre disponibile) e la possibilità di introduzione di errori umani.

La seconda modalità risponde all'esigenza dei processi di pubblicazione "automatizzati" ed **è fortemente consigliata laddove la frequenza di pubblicazione e/o aggiornamento è inferiore all'anno** e rende la pubblicazione manuale poco sostenibile. Tramite le API è possibile sia creare automaticamente nuovi dataset, che aggiornare gli esistenti intervenendo sui metadati e sui contenuti; oltre ad una maggior efficienza, si ha anche un maggior controllo sul formato dei dati, sulla presentazione dei dati (larghezza colonne) e la possibilità di verificare la qualità dei contenuti, attraverso controlli automatici.

Per i motivi di cui sopra, laddove possibile, si consiglia sempre e comunque la pubblicazione automatizzata.

1.4.4.1 *Processi di pubblicazione automatizzata*

Per poter procedere autonomamente alla pubblicazione, i soggetti autorizzati **devono**:

1) registrarsi autonomamente sul sito www.dati.lombardia.it
2) comunicare l'indirizzo email di registrazione dell'utente scrivendo ad admin@dati.lombardia.it

3) ottenuta conferma della profilazione come “editor”, **devono** registrare un “token” per la pubblicazione dei dati tramite le API

I meccanismi, le API ed i tool per la pubblicazione automatica sono documentati qui:
https://dev.socrata.com/publishers/

Particolarmente utile il tool DataSync (http://socrata.github.io/datasync/) e le API per gestire i metadati (https://socratametadataapi.docs.apiary.io/#).

1.4.5 Comunicazione, promozione dei dataset

La pubblicazione dei dataset è importante che sia accompagnata da un’attività di comunicazione e promozione. Regione Lombardia promuove in generale il tema Open Data, il suo programma ed il portale, ma è a cura dell’Ente che pubblica i dati informare i propri interlocutori più diretti delle attività di pubblicazione in chiave Open Data.

1.4.6 Monitoraggio

Il Team Open Data di Regione Lombardia garantisce un costante monitoraggio al quale tuttavia sono chiamati a concorrere tutte le strutture coinvolte a vario titolo nel programma Open Data.

Attraverso strumenti automatici il Team Open Data monitorerà costantemente la qualità dei dati e segnalerà ai titolari dei dati eventuali anomalie quali ad esempio la non completezza dei metadata, il non rispetto dei vocabolari controllati o il mancato aggiornamento dei dati nei tempi stabiliti.

Inoltre, in considerazione del fatto che i dati pubblicati creano valore solo se qualcuno li usa, sarà analizzato l’interesse che suscitano i diversi dataset e tracciate le applicazioni che ne fanno uso.

Il Team Open Data estrarrà dal portale informazioni quali il conteggio dei download e del numero di visualizzazioni del dataset, in grado di fornire un primo indicatore dell’interesse per un determinato set di dati.

L’identificazione dei casi di riuso, considerato che l’accesso ai dati è anonimo e non tracciabile, è particolarmente difficile. Tuttavia, dall’analisi dei “referral” e dalle interazioni con gli utenti che scrivono all’amministratore del portale, è possibile intercettare alcuni casi d’uso che saranno censiti in un apposito dataset al fine di darne evidenza pubblica.

un apposito dataset al fine di darne evidenza pubblica.

2 Convenzioni sui nomi

2.1 Data Owner

Nel definire il NOME i pubblicatori **devono** rispettare le seguenti regole:

- se il soggetto è presente sull'Indice PA (IPA), indicare l'esatta denominazione dell'ente come è registrata su IPA
- se il soggetto non è presente su IPA, indicare la ragione sociale

2.2 Token

Nel denominare il "token", che nella sezione "il mio profilo" è chiamato "Applicazione", i pubblicatori **devono** rispettare le seguenti regole:

- per il "token" utilizzato per pubblicare i dati denominarlo "Nome Ente - pubblicazione"
- nel caso che l'ente sviluppi una o più App che utilizzano le API per accedere ai dati, denominarlo "Nome Ente - Nome applicazione"

2.3 Dataset

Nel denominare i dataset che vengono pubblicati è **mandatorio** adottare, per tutti i soggetti al di fuori di Regione Lombardia ed Enti del Sistema Regionale, la seguente regola:

"Nome Ente Titolo del dataset"

Il nome dell'Ente si intende che sia "contratto", per esempio:

- SI →Provincia Monza Brianza Titolo dataset
- NO →Provincia di Monza e della Brianza – Titolo del dataset

3 Tipologia di dataset pubblicabili

Il portale Open Data Lombardia consente di pubblicare essenzialmente due tipologie di dataset:

- dataset tabellari
- dataset geografici

I dataset **tabellari** possono essere creati manualmente a partire da file CSV o Excel, avendo cura di definire in modo appropriato alcune tipologie di dati quali numeri e date.

A partire da dataset tabellari il portale permette di creare diverse rappresentazioni, quali viste filtrate, grafici e calendari; se il dataset tabellare comprende anche le coordinate geografiche sarà possibile creare anche delle mappe (vedi dettagli in capitolo *4 Tipologia di dati pubblicabili*).

I dataset di tipo geografico possono essere pubblicati a partire da shapefile (vedi: https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf). Se si utilizza un sistema di coordinate diverso dal WGS-84, occorre specificare questa informazione e includere il sistema utilizzato nello shapefile.

4 Tipologia di dati pubblicabili

Il portale Open Data Lombardia consente di inserire in un dataset i seguenti formati di dati:

- Numero
- Testo
- Web URL
- Booleano
- Data
- Valuta (sconsigliato e di difficile gestione in aggiornamento)
- Percentuale
- Coppia di coordinate geografiche (location)

La piattaforma ingloba una serie di funzionalità avanzate sui dataset di tipo tabellare, che permette di aggiungere funzionalità fruibili direttamente on-line, quali ad esempio l'ordinamento crescente/decrescente, link ad un elemento esterno e altro.

Per i formati «ordinabili» (in particolare Numero e Data), è fortemente consigliato che siano rappresentati con il corretto formato e non come generico Testo, in quanto ciò migliora la qualità del dato e permette all'utente di fruire della funzionalità di ordinamento dei dati.

Per una rappresentazione "percentuale", occorre scegliere un numero con il punto come separatore dei decimali.

Per creare una rappresentazione geografica puntuale, derivata da dati tabellari, occorre inserire nella tabella una coppia di dati, in formato numerico e con sistema di coordinate **WGS-84 EPSG 4326** (ad es. 46.4039704, 9.3668236), in cui i numeri devono avere un punto come separatore. Un campo contiene la latitudine e l'altro la longitudine: questi due campi servono a compilare una colonna di tipo "location", usata per creare una vista derivata di tipo "mappa". I due campi dovranno essere denominati WGS84_X e WGS84_Y. Il terzo campo dovrà essere denominato "location" e viene compilato automaticamente dal sistema indicando, nel momento della creazione della colonna, di usare latitudine e longitudine esistente. Questa procedura (manuale) prevede che le colonne WGS84_X e WGS84_Y siano state indicate al sistema come data type "number".

5 Metadatazione

I dati aperti pubblicati sul portale Open Data di Regione Lombardia utilizzano uno schema di metadati comune descritto in questo capitolo.

In accordo con le linee guida nazionali per la valorizzazione del patrimonio informativo pubblico sono state recepite le indicazioni relative all'utilizzo del profilo nazionale DCAT-AP_IT.

Il profilo di metadatazione adottato dal portale Open Data di Regione Lombardia, tuttavia, contiene un insieme di metadati aggiuntivi ritenuti rilevanti per migliorare il riuso dei dati pubblicati, come già previsto dalle linee guida nazionali: “Le pubbliche amministrazioni possono integrare i metadati previsti dal modello DCAT-AP_IT con metadati aggiuntivi, secondo le proprie necessità seppur nel pieno rispetto delle regole di conformità come definite nella specifica DCAT-AP_IT”.

Non si intende in questo capitolo approfondire i dettagli tecnici della specifica DCAT-AP_IT, già ampiamente discussi nei documenti ufficiali, ma focalizzarsi su due aspetti specifici che riguardano l’introduzione della specifica DCAT-AP_IT nel contesto della pubblicazione dei dati aperti sul portale Open Data di Regione Lombardia.

In questo capitolo verranno presentati:

A. il profilo di metadatazione standard del portale Open Data di Regione Lombardia;
B. le relazioni tra i metadati della specifica DCAT-AP_IT e lo schema di metadati adottato dal portale Open Data di Regione Lombardia.

5.1 Profilo di metadatazione del portale Open Data di Regione Lombardia

In questo paragrafo è illustrato il profilo standard di metadati per i dataset pubblicati sul portale Open Data di Regione Lombardia a cui i soggetti che pubblicano devono conformarsi.

Di seguito i metadati che è possibile definire con la funzione di modifica metadati messa a disposizione degli utenti con profilo "publisher" o "editor" dal portale www.dati.lombardia.it.

Alcuni metadati sono definiti a solo uso interno, per processi di gestione, e non sono resi pubblici.

NOTE:

- i metadati contrassegnati con * **sono obbligatori**
- i metadati contrassegnati con A **sono attribuiti automaticamente** dal portale
- i metadati contrassegnati con V sono quelli per cui esiste un **vocabolario controllato**
- nella colonna Pubblico è definito se il metadato sarà visibile agli utenti del portale o solamente agli amministratori del portale a scopo di gestione del catalogo

Metadato	Descrizione	Pubblico
Ente possessore del dato A	Ente titolare del dato (derivato dal nome dell'utente che pubblica)	**SI**
Data creazioneA	Data in cui il dataset è reso disponibile on-line	**SI**
Ultimo aggiornamento datiA	Data di ultima modifica del dataset	**SI**
Ultimo aggiornamento metadatiA	Data di ultima modifica dei metadati del dataset	**SI**
Titolo del dataset*	Nome del dataset che viene mostrato all'utente quando consulta il catalogo, ad es. "Elenco dei siti turistici visitabili" NB: (vedi paragrafo 2.3)	**SI**
Breve descrizione*	Descrizione breve ma esaustiva del contenuto	**SI**
Categoria *V	Categoria a cui associare il dataset (vedi paragrafo 6.1)	**SI**
Tag/Parole chiave*	Parole chiave che riguardano i contenuti del dataset	**SI**
Etichetta della riga	DA NON COMPILARE!	
Licenza*V	Tipo di licenza applicata (vedi capitolo relativo)	**SI**
Dati forniti da*	Nome dell'ente "titolare" dei dati	**SI**
Link della sorgente*	Eventuale URL ad una pagina che descrive il dato o in mancanza di questa alla home page del sito istituzionale	**SI**
Dataset Page	DA NON COMPILARE!	
Frequenza*V	Frequenza di aggiornamento (vedi paragrafo 6.2)	**SI**
Persistenza*	Data di scadenza	**SI**
Data ultima modifica*	Data di ultima modifica dei dati contenuti nel dataset (vedi capitolo relativo) nella forma DD/MM/YYYY	**SI**
Nome serie	Eventuale nome della serie a cui appartiene il dataset (vedi capitolo 7)	**SI**
Link alla serie	Eventuale link alla serie a cui appartiene il dataset (vedi capitolo 7)	**SI**
Direzione	Eventuale nome della struttura interna qualora lo si intenda rendere visibile	**SI**

Metadato	Descrizione	Pubblico
Fonte	DA NON COMPILARE !	
Viewer geografico	Eventuale URL puntuale del viewer geografico sul proprio portale cartografico (per dati geografici)	**SI**
Dettaglio metadati	Eventuale descrizione aggiuntiva o link a pagine web contenenti dettagli esplicativi del contenuto del dataset	**SI**
Link allegati	Eventuali link a file pubblicati sul web contenenti dettagli esplicativi del contenuto del dataset, o del processo di formazione dello stesso (es. moduli per il censimento / raccolta del dato)	**SI**
Referente	Nome del referente del dataset	**NO**
Email referente	Email del referente del dataset	**NO**
Metodo di pubblicazione [v]	Metodo di pubblicazione (vedi Vocabolario paragrafo 6.37)	**NO**
Direzione	Eventuale Direzione (ad uso interno)	**NO**
Microsite	Solo per enti che hanno un "microsite"	**SI**
Semantica e RDF	DA NON COMPILARE!	
API – nome della sorgente	DA NON COMPILARE!	
Immagine di anteprima	Eventuale immagine da associare al contenuto del dataset	**SI**
Allegati	Eventuale file contenente una descrizione campi dettagliata (significato di una sigla, range dei valori, e altro)	**SI**
Email di contatto	Eventuale email del referente del dataset specifico	**NO**

[1] Solo nel caso che il dataset contenga dati personali.

5.2 Mapping con standard DCAT-AP_IT

In accordo con le linee guida nazionali per la valorizzazione del patrimonio informativo pubblico è necessario recepire le indicazioni relative all'utilizzo del profilo nazionale DCAT-AP_IT, le cui specifiche sono disponibili qui:

- http://www.dati.gov.it/content/dcat-ap-it-v10-profilo-italiano-dcat-ap-0
- http://linee-guida-cataloghi-dati-profilo-dcat-ap-it.readthedocs.io/it/latest/dcat-ap_it.html

Questo paragrafo non ha lo scopo di approfondire i dettagli tecnici della specifica DCAT-AP_IT, già ampiamente discussi nei documenti ufficiali, ma si focalizza sulla relazione tra i metadati della specifica DCAT-AP_IT e lo schema di metadati adottato dal portale di Regione Lombardia.

La specifica DCAT-AP_IT propone una struttura di metadati, basata sui concetti principali di **Catalogo**, **Dataset** e **Distribuzione**. Il Catalogo rappresenta un insieme di dataset, e pertanto i metadati relativi ad esso riguardano le proprietà dell'intero insieme di dataset (es. Organizzazione che pubblica i dati). Al Catalogo sono associati i Dataset che lo compongono. A sua volta ogni Dataset, può avere a sé associate diverse Distribuzioni, che si differenziano per il formato usato per la pubblicazione dei dati, la licenza utilizzata, e così via. Ogni Distribuzione prevede quindi metadati specifici per descrivere queste proprietà.

Uno dei principali vantaggi della produzione del catalogo secondo lo standard DCAT_AP-IT è la possibilità di attribuire la titolarità dei dati all'ente che li produce ed utilizza il portale Open Data della Regione Lombardia: l'Ente sarà definito come **dct:rightsHolder** mentre Regione Lombardia sarà definita come **dct:publisher**.

L'implementazione del DCAT_AP-IT sarà arricchita progressivamente nel tempo con le proprietà opzionali al fine di rendere sempre più esaustiva la descrizione dei dataset.

Nella tabella che segue è illustrata la mappatura tra il profilo standard di metadati per i dataset pubblicati sul portale Open Data di Regione Lombardia e la **sezione dataset** del profilo DCAT_AP-IT:

Ente possessore del dato [A]	**dct:rightsHolder**
Data creazione[A]	L'attributo non è attualmente mappato nel DCAT
Ultimo aggiornamento dati[A]	**dct:modified**
Ultimo aggiornamento metadati[A]	L'attributo non è attualmente mappato nel DCAT
Titolo del dataset*	**dct:title**
Breve descrizione*	**dct:description**
Categoria *[V]	**dcat:theme** (tabella di conversione in paragrafo 6.1)
Tag/Parole chiave*	L'attributo non è attualmente mappato nel DCAT
Etichetta della riga	L'attributo non è attualmente mappato nel DCAT
Licenza*[V]	**dct:license**
Dati forniti da*	**foaf:name**
Link della sorgente*	L'attributo non è attualmente mappato nel DCAT
Dataset Page	L'attributo non è attualmente mappato nel DCAT
Frequenza*[V]	**dct:accrualPeriodicity** (tabella di conversione in paragrafo 6.2)
Persistenza	L'attributo non è attualmente mappato nel DCAT
Data ultima modifica*	
Nome serie	L'attributo non è attualmente mappato nel DCAT
Link alla serie	L'attributo non è attualmente mappato nel DCAT
Direzione	L'attributo non è attualmente mappato nel DCAT
Fonte	L'attributo non è attualmente mappato nel DCAT
Viewer geografico	L'attributo non è attualmente mappato nel DCAT
Dettaglio metadati	L'attributo non è attualmente mappato nel DCAT
Link allegati	L'attributo non è attualmente mappato nel DCAT
Referente	L'attributo non è attualmente mappato nel DCAT
Email referente	L'attributo non è attualmente mappato nel DCAT
Metodo di pubblicazione [V]	L'attributo non è attualmente mappato nel DCAT
Direzione	L'attributo non è attualmente mappato nel DCAT
Microsite	L'attributo non è attualmente mappato nel DCAT
Semantica e RDF	L'attributo non è attualmente mappato nel DCAT
API – nome della sorgente	L'attributo non è attualmente mappato nel DCAT
Immagine di anteprima	L'attributo non è attualmente mappato nel DCAT
Allegati	L'attributo non è attualmente mappato nel DCAT
Email di contatto	L'attributo non è attualmente mappato nel DCAT

6 Vocabolari controllati

6.1 Vocabolario delle Categorie

Il portale Open Data della Regione Lombardia raggruppa i dataset i 22 categorie:

1. Agricoltura
2. Ambiente
3. Attività Produttive
4. Commercio
5. Cultura
6. Energia
7. Famiglia
8. Government
9. Istruzione
10. Mobilità e trasporti
11. Paesaggio
12. Protezione Civile
13. Sanità
14. Sicurezza
15. Solidarietà
16. Sport
17. Statistica
18. Territorio
19. Trasparenza
20. Tributi
21. Turismo
22. Università e ricerca

Si ricorda che la categoria è un attributo **mandatorio** dei metadati.

Mappatura delle categorie per temi del DCAT-AP-IT secondo il vocabolario EU Data Theme:

Agricoltura	http://publications.europa.eu/resource/authority/data-theme/AGRI
Attività Produttive	http://publications.europa.eu/resource/authority/data-theme/ECON
Bilancio	http://publications.europa.eu/resource/authority/data-theme/ECON
Commercio	http://publications.europa.eu/resource/authority/data-theme/ECON
Tributi	http://publications.europa.eu/resource/authority/data-theme/ECON
Istruzione	http://publications.europa.eu/resource/authority/data-theme/EDUC
Sport	http://publications.europa.eu/resource/authority/data-theme/EDUC
Turismo	http://publications.europa.eu/resource/authority/data-theme/EDUC
Cultura	http://publications.europa.eu/resource/authority/data-theme/EDUC
Energia	http://publications.europa.eu/resource/authority/data-theme/ENER
Ambiente	http://publications.europa.eu/resource/authority/data-theme/ENVI
Government	http://publications.europa.eu/resource/authority/data-theme/GOVE
Trasparenza	http://publications.europa.eu/resource/authority/data-theme/GOVE
Sanità	http://publications.europa.eu/resource/authority/data-theme/HEAL
Protezione Civile	http://publications.europa.eu/resource/authority/data-theme/JUST
Sicurezza	http://publications.europa.eu/resource/authority/data-theme/JUST
Paesaggio	http://publications.europa.eu/resource/authority/data-theme/REGI
Territorio	http://publications.europa.eu/resource/authority/data-theme/REGI
Famiglia	http://publications.europa.eu/resource/authority/data-theme/SOCI
Solidarietà	http://publications.europa.eu/resource/authority/data-theme/SOCI
Statistica	http://publications.europa.eu/resource/authority/data-theme/SOCI
Università e ricerca	http://publications.europa.eu/resource/authority/data-theme/TECH
Mobilità e trasporti	http://publications.europa.eu/resource/authority/data-theme/TRAN
Trasporti	http://publications.europa.eu/resource/authority/data-theme/TRAN

6.2 Vocabolario delle frequenze di aggiornamento

Una informazione particolarmente significativa e che quindi deve essere sempre presente nei metadati è quella relativa alla frequenza di aggiornamento prevista per il dato.

Essa rappresenta un impegno del Data Owner nei confronti della comunità dei riutilizzatori.

Allo scopo di normalizzare le definizioni delle frequenze di aggiornamento e di poterle rappresentare secondo il vocabolario europeo come previsto nel profilo **standard DCAT-AP_IT**, è stato definito un vocabolario che deve essere adottato dai soggetti che pubblicano i dati sul portale di Regione Lombardia:

- **Tempestiva** (Dati la cui frequenza non è prevedibile, ma al verificarsi del cambiamento vengono pubblicati tempestivamente)
- **Tempo reale** (Dati che vengono aggiornati in tempo reale o ogni N ore)
- **Giornaliera**
- **Settimanale**
- **Quindicinale**
- **Mensile**
- **Bimestrale**
- **Trimestrale**
- **Quadrimestrale**
- **Semestrale**
- **Annuale**
- **Biennale**
- **Triennale**
- **Quinquennale**
- **Decennale**
- **Mai** (Dati “storici”, che per loro natura non cambiano)
- **Non definita** (Dati soggetti a cambiamento, ma i cui processi di aggiornamento non sono ancora definiti)

Di seguito la tabella di conversione con le frequenze standard per il DCAT-AP_IT:

Vocabolario www.dati.lombardia.it	Vocabolario DCAT-AP_IT
Tempestiva	Irreg
Tempo reale	Cont
Giornaliera	Daily
Settimanale	Weekly
Quindicinale	Monthly_2
Mensile	Monthly
Bimestrale	Bimonthly
Trimestrale	Quarterly
Quadrimestrale	Annual_3
Semestrale	Annual_2
Annuale	Annual
Biennale	Biennal
Triennale	Triennal
Quinquennale	Other
Decennale	Other
Mai	Never
Non definita	Unknown

6.3 Vocabolario delle modalità di pubblicazione

Per definire nei metadati il metodo di pubblicazione gli Enti devono utilizzare le seguenti diciture:

- Manuale
- Automatico
- Semi Automatico

NB: attenzione a rispettare maiuscole, minuscole e spazi.

7 Definizione delle serie di dati

Nel caso di dati che sono soggetti ad aggiornamento nel tempo, è possibile mettere in atto politiche di aggiornamento molto diverse tra loro.

Sarebbe opportuno, in generale, evitare la proliferazione di dataset nel caso di dati per i quali esistono differenti valori nel tempo (es. serie statistiche annuali), introducendo nel dataset un attributo che rappresenti l'unità temporale a cui si riferisce. Oltre a non "gonfiare" inutilmente il catalogo rendendolo più complesso da consultare, è indubbio che un unico dataset che cresce nel tempo è più semplice da riutilizzare che una serie numerosa di singoli dataset.

Ciò premesso, vi sono casi in cui è necessario che il dato venga pubblicato attraverso una "serie di dataset", ovvero diversi dataset che rappresentano diversi aggiornamenti del dato in diversi momenti.

Potrebbe essere utile, ad esempio, quando sia richiesto che il dato sia pubblicato per un periodo massimo di N anni, come nel caso dei dati previsti dal D.lgs 33/2013 (Decreto Trasparenza). In tal caso può essere più semplice pubblicare differenti dataset per anno e cancellare i dataset una volta scaduto il termine della loro pubblicazione. Altro caso potrebbe essere quello in cui si intende mettere a disposizione un dataset che rappresenti sempre la fotografia attuale del dato, ad esempio quando l'uso prevalente sia quello di applicazioni che necessitano di conoscere lo stato aggiornato del dato, ma si voglia pubblicare nel catalogo anche versioni precedenti al fine di permettere una analisi storica.

La corretta rappresentazione della "frequenza di aggiornamento", in presenza di serie di dati, potrebbe indurre ad una interpretazione errata da parte del riutilizzatore.

L'approccio che si consiglia, nel caso in cui si crei una serie di dati, è quello di definire la frequenza di aggiornamento riferendosi al dato e non al singolo al singolo dataset, il cui contenuto non cambia nel tempo. Quindi, ad esempio, se la politica di aggiornamento prevede di creare un nuovo dataset che compone la serie di dati ogni anno, la frequenza di aggiornamento per tutti i dataset della serie dovrà essere definita "annuale"; anche se in realtà i singoli dataset "storici" non cambieranno mai.

Al fine di distinguere opportunamente dataset singoli o serie di dati è necessario inserire nei metadati l'informazione sulla serie di dati.

Nome Serie, se presente, indica che il dataset appartiene ad una serie, composta da tutti i dataset con lo stesso "Nome Serie".

Al fine di permettere, anche all'utente che consulta il catalogo "a vista", di identificare tutti i dataset appartenenti alla serie, è fortemente consigliato introdurre un tag univoco che indichi la serie in ogni dataset, così da poter valorizzare il **link alla serie** come segue:

https://www.dati.lombardia.it/browse?tags=tagdellaserie

Il tag può essere una singola parola o più di una, nel qual caso l'URL sarà composta inserendo "+" al posto degli spazi, es:

https://www.dati.lombardia.it/browse?tags=tag+della+serie