

# Formati aperti

Formez<sup>PA</sup>



Questo materiale didattico è stato realizzato da Formez Pa nell'Ambito 2 Linea 2 del Programma Operativo di Assistenza Tecnica 2012-2015 (POAT), Ob.II.4 PON GAT (FESR) 2007-2013 a titolarità del Ministero dell'Economia e delle Finanze. Il Dipartimento della Funzione Pubblica opera come Organismo intermedio per l'Obiettivo operativo II.4.

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:** Sergio Agostinelli, Piero Zilio

**Creatore:** Formez PA

**Diritti:** Dipartimento della Funzione Pubblica

**Data:** Ottobre 2015

## Dati, dataset, banca dati e documenti di tipo aperto

Il patrimonio informativo pubblico è composto da tutti i contenuti prodotti da una pubblica amministrazione che tradizionalmente è stato reso disponibile per la lettura e visualizzazione, in forma scritta o altra forma (immagine, audio, video).

Da studi della Commissione europea emerge ancora una forte difficoltà a reperire e riutilizzare queste informazioni e dati, in particolare per il riuso con sistemi informatici. Per sormontare gli ostacoli al riutilizzo e ridurre la frammentazione dei mercati dei dati, la Commissione ha quindi adottato un pacchetto di misure, il cui elemento cardine è la direttiva relativa al riutilizzo dell'informazione del settore pubblico.

Esistono diversi livelli di definizioni e scelte da considerare per una corretta gestione dell'apertura dei dati:

- **Dato:** è una descrizione di fatti priva di contesto (es. un numero, una parola, una immagine)
- **Dataset:** è una serie di dati (es. un elenco di numeri, una tabella di indirizzi, un insieme di "log" ossia eventi caratteristici di un'applicazione informatica)
- **Banca dati:** è un insieme organizzato di dati opportunamente strutturato per lo svolgimento di determinate attività (Es. Banca dati delle spese, Banca dati dei progetti)
- **Informazione:** è un elemento che consente la conoscenza di fatti (una notizia, un progetto ma anche informazioni descrittive delle categorie di dati come i metadati)
- **Documento:** rappresenta una collezione di informazioni che consentono la conoscenza di fatti (atti, pubblicazioni editoriali, rapporti, grafici, materiali didattici, contenuti multimediali)

Le raccomandazioni all'apertura per le serie di dati (dataset), richiedono un'attenzione particolare sui vincoli tecnici all'acquisizione ed all'utilizzabilità da parte dei sistemi informativi automatici, come per esempio la pubblicazione in formati leggibili meccanicamente da un programma software e la rappresentazione formale e standardizzata dei dati.

La normativa italiana, in linea con le direttive europee, definisce nell'art. 68, comma 3, del CAD (Codice dell'amministrazione digitale) un formato dei dati di tipo aperto, "*un formato di dati reso pubblico, documentato esaustivamente e neutro rispetto agli strumenti tecnologici necessari per la fruizione dei dati stessi*".

## Cos'è il formato di un file

Un formato di file, in informatica, indica la convenzione che viene usata per leggere, scrivere e interpretare i contenuti di un file.

Il formato di un file è comunemente indicato attraverso l'estensione, che è una serie di lettere (in genere tre, per motivi storici) unita al nome del file attraverso un punto. Ad esempio, "contenuto.txt" è un file di testo (o meglio, il suo contenuto va interpretato come testo), mentre "contenuto.jpg" è un'immagine.

Sebbene i file siano composti dagli stessi elementi (i byte), ogni file può archiviare contenuti molto diversi fra loro: testi, video, audio, immagini, ecc.

Grazie al formato è possibile interpretare correttamente i byte di un file a seconda del tipo di contenuto, e indicare al programma software come interpretare questi dati.

## **Formati proprietari e aperti**

I formati possono essere costruiti in modo da limitare la lettura del contenuto soltanto ad alcuni programmi software. In questo caso parliamo di formato proprietario, ossia coperto da diritti dei creatori (copyright) e le loro specifiche tecniche non vengono distribuite pubblicamente. Questo comporta la necessità di usare particolari programmi per aprire e modificare i file e vincola la circolazione dei dati al possesso di un software specifico, che potrebbe non essere disponibile a tutti gli utenti interessati. Non c'è reale indipendenza tra contenuto sistema di lettura.

Al contrario, un formato aperto è un formato non proprietario che si basa su una specifica di costruzione pubblica e che garantisce la lettura da parte di qualsiasi programma. Il formato aperto è realizzato rispettando gli standard dichiarati e non presenta alcuna restrizione legale per il suo utilizzo.

I formati aperti incoraggiano la concorrenza tra sviluppatori di funzionalità software invece di garantire il controllo di un solo produttore su tutti i contenuti realizzati dagli utenti attraverso il formato proprietario. In questi casi si ha il fenomeno del lock-in ossia quando, individualmente o collettivamente, si è "catturati" da una scelta tecnologica rispetto ad altre disponibili.

## **L'uso dei formati tecnici aperti negli Open Data**

La libera circolazione dei dati sul web si basa sulla accessibilità totale, e può essere assicurata con l'adozione di un formato tecnico aperto in grado di eliminare i problemi legati ai diritti legali e alle specifiche tecniche del software.

Il formato aperto favorisce quindi la circolazione dei dati e ne garantisce un uso più completo, dalla semplice consultazione, all'aggiornamento, al mash-up (l'incrocio con altri dati).

Il valore dell'apertura è maggiore quando i dati sono strutturati o semi-strutturati, ossia organizzati completamente o parzialmente secondo schemi standard che ne facilitano la comprensione per il riutilizzo.

La valorizzazione massima dei dati si ha quando sono utilizzati strutturalmente, oltre ai formati tecnici, gli standard di rappresentazione dei dati e le ontologie per implementare i modelli di dati e metadati.

Per la pubblica amministrazione italiana il documento di riferimento su questi elementi è rappresentato dalle *Linee guida per la valorizzazione del patrimonio informativo pubblico* elaborate dall'AGID.

## **Formati per la pubblicazione di dati aperti**

I formati aperti più comuni per la pubblicazione dei dati sono l'XML molto diffuso e usato per la definizione dei dati sul web il CSV usato per rappresentare informazioni con struttura tabellare ed il JSON un formato aperto per la rappresentazione e lo scambio di dati

I dati geografici possono essere rappresentati con diversi formati aperti. Uno dei più diffusi è lo shapefile. Creato dalla ESRI è oggi uno standard de-facto per la rappresentazione di dati dei sistemi informativi geografici.

Google invece ha contribuito a diffondere il formato KML oggi standard OGC. Un formato molto diffuso nello sviluppo di applicazioni web è il geoJSON, una combinazione tra il formato JSON coordinate e geometrie.

In fine c'è il GML una grammatica XML per lo scambio di dati territoriali.

**XML (eXtensible Markup Language)**, un formato standardizzato dal W3C e molto diffuso per la definizione dei dati sul web attraverso tag o annotazioni personalizzati (e.g., sistema XBRL per la rappresentazione dei bilanci, o sistema Normattiva per la rappresentazione di documenti informatici in ambito giuridico, ecc).

**CSV (Comma Separated Values)**. È un formato di file testuale usato per rappresentare informazioni con struttura tabellare. Esso è spesso usato per importare ed esportare il contenuto di tabelle di database relazionali e fogli elettronici. Le righe delle tabelle corrispondono a righe nel file di testo CSV e i valori delle celle sono divisi da un carattere separatore, tipicamente, come indica il nome stesso, la virgola. Il W3C ha recentemente istituito un gruppo "CSV on the web" al quale partecipano, tra gli altri, esperti di Google e dell'Open Data Institute (ODI). L'iniziativa mira a produrre uno standard attraverso cui formalizzare un'opportuna sintassi per l'introduzione dei metadati all'interno dei CSV, e fornire suggerimenti per la conversione verso vari formati.

**JSON (JavaScript Object Notation)**. E' un formato aperto per la rappresentazione e lo scambio di dati semi-strutturati che risulta leggibile anche dagli umani e che mantiene, rispetto a formati simili come l'XML, una sintassi poco prolissa. Questo aspetto ne fa un formato flessibile e compatto. Esso nasce dalla rappresentazione di strutture dati semplici nel linguaggio di programmazione JavaScript, ma mantiene indipendenza rispetto ai linguaggi di programmazione. Ha avuto ampia diffusione per il fatto di essere flessibile e compatto.

**(Shapefile)**. E' il formato standard de-facto per la rappresentazione dei dati dei sistemi informativi geografici (GIS). Esso è creato dalla società ESRI che rende pubbliche le sue specifiche. Questo ha consentito lo sviluppo di strumenti in grado di gestire e creare tale formato. I dati sono di tipo vettoriale. Seppur impropriamente ci si riferisca a uno shapefile, nella pratica si devono considerare almeno tre file: un.shp contenente le forme geometriche, un .dbf contenente il database degli attributi delle forme geometriche e un file .shx come indice delle forme geometriche. A questi tre si deve anche accompagnare un file .prj che contiene le impostazioni del sistema di riferimento.

**KML**. È un formato basato su XML per rappresentare dati geografici. Nato con Google, è diventato poi uno standard OGC. Le specifiche della versione 2.2 presentano una serie di entità XML attraverso cui archiviare le coordinate geografiche che rappresentano punti, linee e poligoni espressi in coordinate WGS84 e altre utili a definire gli stili attraverso cui i dati andranno visualizzati. Eventuali attributi delle geometrie vanno espressi invece attraverso la

personalizzazione di alcune entità. Molti strumenti di conversione non si occupano tuttavia di creare questa struttura dati e delegano gli attributi delle geometrie allo stile di visualizzazione. Si consiglia pertanto di distribuire questo dato prestando attenzione o, eventualmente, accompagnando il dataset assieme ad un altro formato aperto per i dati geografici (es., .shp, .geojson). Spesso viene distribuito in modalità compressa attraverso file con estensione .kmz.

**geoJSON.** E' un formato aperto per la rappresentazione e l'interscambio dei dati territoriali in forma vettoriale, basato su JSON (JavaScript Object Notation). Ogni dato è codificato come oggetto che può rappresentare una geometria, una feature o una collezione di feature. Ad ogni oggetto è associato un insieme di coppie nome/valore (membri). I principali nomi di membri che rappresentano le caratteristiche dei dati geografici sono: "type" che serve ad indicare il tipo di geometria (punto, linea, poligono o insieme multi-parte di questi tipi); "coordinates" attraverso cui sono indicate le coordinate dell'oggetto in un dato sistema di riferimento; "bbox" attraverso cui sono indicate le coordinate di un riquadro di delimitazione geografica; "crs" (opzionale) per l'indicazione del sistema di riferimento. È possibile inoltre associare all'oggetto specifici attributi attraverso il membro con nome "properties". Si tratta di un formato molto diffuso e supportato da diversi software, ampiamente utilizzato in ambito di sviluppo web.

**GML (Geography Markup Language).** E' una grammatica XML che rappresenta un formato di scambio aperto per i dati territoriali. Essa, definita originariamente da OGC e diventata lo Standard ISO 19136:2008, fornisce la codifica XML (schemi XSD) delle classi concettuali definite in diversi Standard ISO della serie 19100 e di classi aggiuntive appositamente definite: geometrie, oggetti topologici, unità di misura, tipi di base, riferimenti temporali, feature, sistemi di riferimento, copertura

## Formati per la pubblicazione di documenti aperti

Per quanto riguarda i documenti come delibere e atti amministrativi da pubblicare on line in Italia i formati aperti di riferimento sono:

l'ODT per documenti testuali, l'ODS per fogli di calcolo e l'ODP per documenti di presentazione. Questi tre formati sono basati sullo standard XML e sono stati adottati come formato principale in alcune suite per l'automazione d'ufficio come OpenOffice.org e LibreOffice inoltre sono supportata da altre suite come Microsoft Office, Google Drive e IBM Lotus.

Il più popolare formato per la rappresentazione di documenti contenenti testo e immagini che sia indipendente dalla piattaforma di lettura è Il formato PDF creato da Adobe ed oggi standardizzato dall'ISO.

**ODT (Open Document Text).** E' uno standard aperto per documenti testuali basato su XML. Fa parte dello standard OASIS Open Document Format for Office Applications. E' stato adottato come formato principale per i testi in alcune suite per l'automazione d'ufficio come OpenOffice.org e LibreOffice ed è supportato da altre come Microsoft Office, Google Drive e IBM Lotus.

**ODS (Open Document Spreadsheet).** E' uno standard aperto per fogli di calcolo basato su XML. Fa parte dello standard OASIS Open Document Format for Office Applications. Come nel caso precedente, è stato adottato come formato principale per i fogli di calcolo in alcune suite per l'automazione d'ufficio come OpenOffice.org e LibreOffice ed è supportato da altre come Microsoft Office, Google Drive e IBM Lotus.

**ODP (Open Document Presentation).** E' uno standard aperto per documenti di presentazione basato su XML. Fa parte dello standard OASIS Open Document Format for Office Applications. E' stato adottato come formato principale per i documenti di presentazione in alcune suite per l'automazione d'ufficio come OpenOffice.org e LibreOffice ed è supportato da altre come Microsoft Office, Google Drive e IBM Lotus.

**PDF (Portable Document Format).** E' un formato aperto creato da Adobe per la rappresentazione di documenti contenenti testo e immagini che sia indipendente dalla piattaforma di lettura (applicativo, sistema operativo e hardware). E' stato successivamente standardizzato dall'ISO (ISO/IEC 32000-1:2008) con una serie di formati differenti, ognuno avente una propria prerogativa (e.g., PDF/UA per l'accessibilità, PDF/H per documenti sanitari, PDF/A per l'archiviazione, ecc.).

Spesso c'è l'esigenza di pubblicare documenti contenenti dati solo in forma tabellare, come elenchi (come ad esempio nel caso dei tassi di assenza del personale delle pubbliche amministrazioni), senza elementi aggiuntivi (e.g., grafici, formule). In questi casi si può evitare di utilizzare formati come PDF o ODT ma si consiglia di utilizzare il formato CSV. Negli altri casi di dati in forma tabellare con elementi aggiuntivi si raccomanda l'uso di ODS. Nel caso di documenti in generale, si raccomanda, ove esistano, standard XML documentali internazionali o nazionali.

Una regola di qualità nella pubblicazione resta il multiformato, ossia rendere un documento disponibile in più formati.



Esempio: Open Expo  
I multiformati aperti e proprietari del sito  
<http://dati.openexpo2015.it/catalog/dataset>



## Dal web semantico al modello a 5 stelle

Publicare online i dati non rende il web “semantico”, popolato di contenuti contestualizzati e interpretabili in maniera automatica dai sistemi informativi come per esempio i motori di ricerca. Per raggiungere questo obiettivo è necessario migliorare il collegamento fra i dati e renderli disponibili all'uomo e alle sue applicazioni.

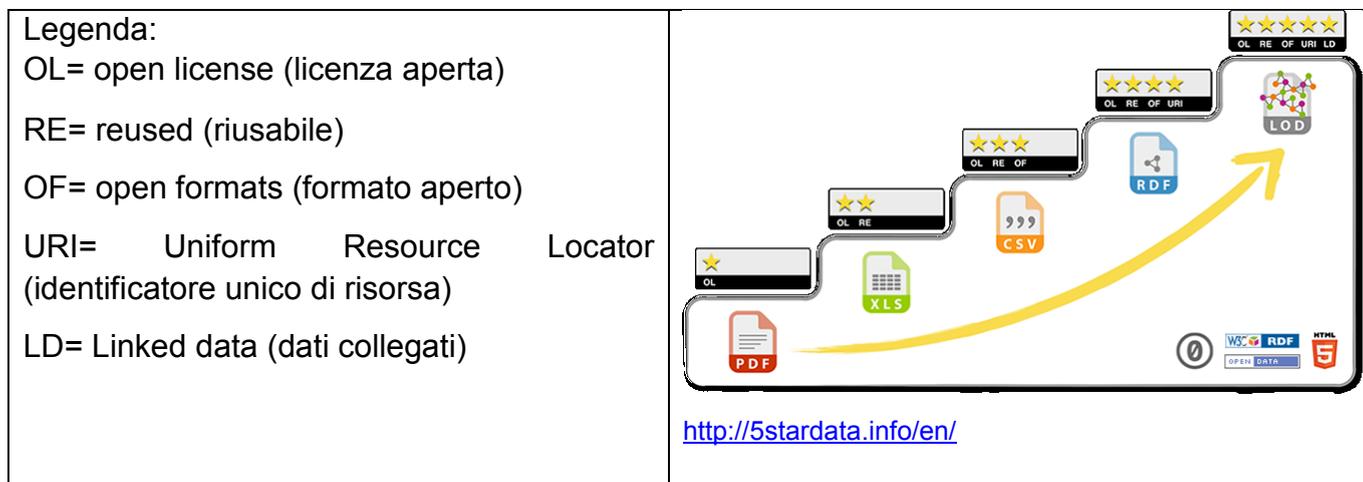
In quest'ottica Tim Berners-Lee ha sviluppato un sistema di classificazione dei dati basato su cinque livelli, chiamato “5 ★ Open Data”.

## Il modello a 5 stelle

Secondo il modello di Tim Berners-Lee

- ★ Il file è disponibile sul web (in un qualsiasi formato) con licenza aperta, ma non strutturato
- ★★ Il file è disponibile sul web in un formato comprensibile da un software e quindi riusabile (per esempio un file Excel e non la semplice scansione di una tabella stampata)
- ★★★ Il file è disponibile sul web in un formato aperto (per esempio CSV e non Excel)
- ★★★★ Il file è disponibile sul web e usa gli standard aperti del W3C e quindi si utilizzano URI (Uniform Resource Identifier) che, oltre a identificare una risorsa digitale, fornisce mezzi per agire su di essa (per esempio RDF)
- ★★★★★ Il file è contestualizzato e collegato ad altri dati

Ecco la scala del sistema di classificazione dei dati aperti:



## Costi e benefici dei dati aperti con ★

Spesso troviamo dati disponibili sul web ad una stella ossia file in qualsiasi formato, ad esempio documenti contenenti testi o tabelle in formato Adobe Pdf, non processabili automaticamente dalla macchina (anche se rilasciati con licenza legale di utilizzo dei contenuti aperta).

In questi casi i dati sono incorporati all'interno di documenti senza struttura, sono per esempio scansioni di tabelle stampate e quindi leggibili e interpretabili solo grazie a un intervento umano. Per utilizzare, riusare o abilitare qualche servizio a partire dai dati contenuti nei documenti, è necessario attivare significativi interventi umani di estrazione ed elaborazione dei

possibili dati.

I cittadini possono guardare, stampare, scaricare e condividere questi dati. Dal punto di vista della PA il dato risulta semplice da pubblicare

<p><b>Esempio 1</b></p> <p>Una tabella chiusa in File PDF non strutturato per essere processabile automaticamente da un sistema informatico, ma pensato per la semplice visualizzazione umana.</p>	<table border="1"><thead><tr><th colspan="2"><b>Temperature forecast for Galway, Ireland</b></th></tr><tr><th><b>Day</b></th><th><b>Lowest Temperature (°C)</b></th></tr></thead><tbody><tr><td>Saturday, 13 November 2010</td><td>2</td></tr><tr><td>Sunday, 14 November 2010</td><td>4</td></tr><tr><td>Monday, 15 November 2010</td><td>7</td></tr></tbody></table>	<b>Temperature forecast for Galway, Ireland</b>		<b>Day</b>	<b>Lowest Temperature (°C)</b>	Saturday, 13 November 2010	2	Sunday, 14 November 2010	4	Monday, 15 November 2010	7
<b>Temperature forecast for Galway, Ireland</b>											
<b>Day</b>	<b>Lowest Temperature (°C)</b>										
Saturday, 13 November 2010	2										
Sunday, 14 November 2010	4										
Monday, 15 November 2010	7										
<p><b>Esempio 2</b></p> <p>Schede ed elenchi scaricabili in formato PDF non strutturato per essere processabile automaticamente da un sistema informatico, ma pensato per la semplice visualizzazione umana.</p>	<p>Contratti Pubblici - Anno 2012</p> <p>Schede dei singoli lavori ed elenco sintetico di aggiudicazione lavori delle stazioni appaltanti di interesse regionale per importi superiori a 150.000 Euro nell'anno 2012.</p> <p>Data e Risorse</p> <table><tr><td></td><td><b>Schede per singoli lavori 2012</b> Schede per singoli lavori 2012</td><td><a href="#">Esplora</a></td></tr><tr><td></td><td><b>Elenco sintetico delle aggiudicazioni - lavori ...</b> Elenco sintetico delle aggiudicazioni - lavori 2012</td><td><a href="#">Esplora</a></td></tr></table>		<b>Schede per singoli lavori 2012</b> Schede per singoli lavori 2012	<a href="#">Esplora</a>		<b>Elenco sintetico delle aggiudicazioni - lavori ...</b> Elenco sintetico delle aggiudicazioni - lavori 2012	<a href="#">Esplora</a>				
	<b>Schede per singoli lavori 2012</b> Schede per singoli lavori 2012	<a href="#">Esplora</a>									
	<b>Elenco sintetico delle aggiudicazioni - lavori ...</b> Elenco sintetico delle aggiudicazioni - lavori 2012	<a href="#">Esplora</a>									

### Costi e benefici dei dati aperti con ★★

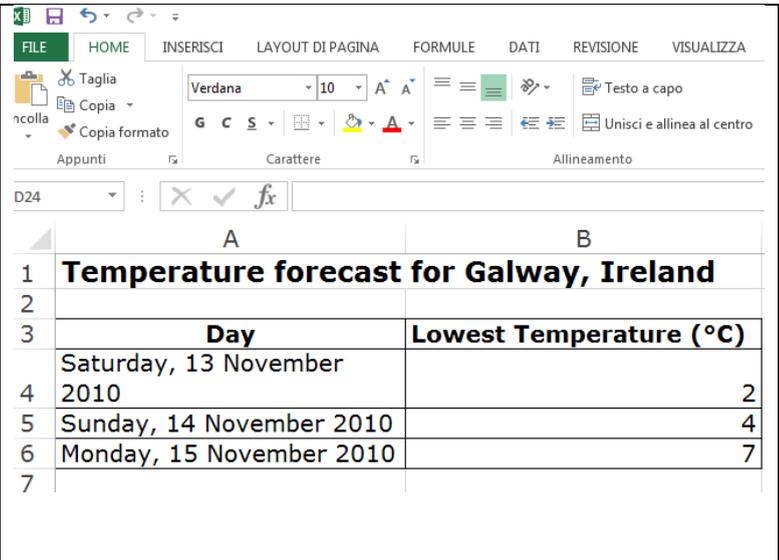
Molti dei dati a disposizione li troviamo sul web in un formato comprensibile da un software, ossia processabile automaticamente ma proprietario (per esempio un file Excel e non la semplice scansione di una tabella stampata).

Con questo tipo di file il cittadino può lavorare sui dati con software proprietario ed esportare i dati in un altro formato.

La PA ha il vantaggio che questo formato di file è semplice da pubblicare

### Esempio

Una tabella scaricabile in formato Excel (.xls), quindi strutturato per essere processabile automaticamente da un sistema informatico, ma proprietario e quindi vincolato all'uso prescritto dalle specifiche del produttore, in questo caso Microsoft.



The screenshot shows the Microsoft Excel interface with a table titled "Temperature forecast for Galway, Ireland". The table has two columns: "Day" and "Lowest Temperature (°C)". The data rows are:

Day	Lowest Temperature (°C)
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

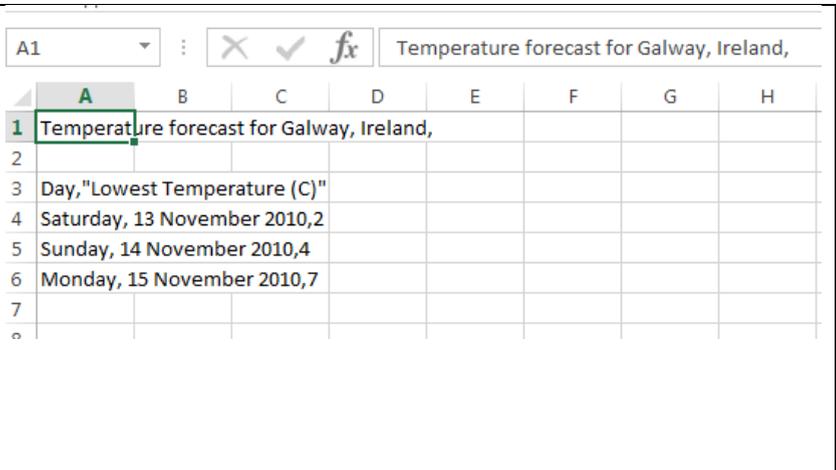
### Costi e benefici dei dati aperti con ★★★

Possiamo cominciare a parlare di dati realmente aperti quando i dati sono disponibili sul web in un file in un formato comprensibile da un software, ossia processabile automaticamente e non proprietario (per esempio un file CSV e non XLS). Quando il cittadino trova un dato pubblicato a tre stelle può sicuramente manipolare il dato in ogni modo, senza le restrizioni legali e tecniche imposte dal software proprietario.

Questo tipo di dati sono sempre semplici da pubblicare per la PA ma potrebbe essere necessaria una conversione dal formato proprietario a quello aperto.

### Esempio

Una tabella scaricabile in formato CSV, quindi strutturato per essere processabile automaticamente da un sistema informatico, e non proprietario e quindi non vincolato nell'uso dalle specifiche del produttore.



The screenshot shows a CSV file with the following data:

Day,"Lowest Temperature (C)"
Saturday, 13 November 2010,2
Sunday, 14 November 2010,4
Monday, 15 November 2010,7

### Costi e benefici dei dati aperti con ★★★★★

Siamo di fronte a progetti avanzati di apertura dei dati quando i dati sono disponibili con l'adozione di standard aperti definiti dal W3C per quanto riguarda il web semantico e quindi si utilizzano URI (Uniform Resource Locator) che servono ad identificare una risorsa digitale in modo univoco, si forniscono i mezzi per agire su di essa (per esempio RDF) e per completare viene fornita una modalità di accesso (endpoint SPARQL) per effettuare delle interrogazioni dei dati (query) personalizzate.

Con i dati a quattro stelle il cittadino puntare al dato da qualsiasi luogo (con un collegamento dal web o locale), riusare parte dei dati, degli strumenti e delle librerie, capire la struttura di un grafico di dati RDF, anche se richiede competenze maggiori rispetto ai dati con struttura

tabellare (per esempio Excel/CSV) o ad albero (per esempio XML/JSON) e realizzare mash-up.

Per le PA invece aumenta il controllo sui singoli dati e si possono ottimizzare gli accessi e le richieste, altre PA possono collegarsi ai dati pubblicati.

Si richiede però un maggior investimento nel trattamento dei dati, è necessario assegnare URIs ai singoli dati e pensare a come rappresentarli, infine è necessario adattare modelli esistenti o crearne ex novo

**Esempio**

Una tabella on line semi- strutturata per essere processabile automaticamente da un sistema informatico con un indirizzo identificativo unico sul web (URI) e descritto in RDF.

**Temperature forecast for Galway, Ireland**

Day	Lowest Temperature (C)
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

```

<table border="1">
  <thead>
    <tr>
      <th style="text-align: left;">Day
      <th style="text-align: left;">Lowest Temperature (C)
    </tr>
  </thead>
  <tbody>
    <tr>
      <td>Saturday, 13 November 2010
      <td>2
    </tr>
    <tr>
      <td>Sunday, 14 November 2010
      <td>4
    </tr>
    <tr>
      <td>Monday, 15 November 2010
      <td>7
    </tr>
  </tbody>
</table>

```

### Costi e benefici dei dati aperti con ★★★★★

Il livello massimo della scala a 5 stelle si raggiunge quando i dati sono contestualizzati e collegati ad altri dati. Solo in questi casi si parla di LOD (Linked Open Data) .

I cittadini possono adesso scoprire nuovi dati connessi a quelli visualizzati e imparare direttamente dallo schema dei dati. Il rischio è quello di incappare in link interrotti o rischiare di fraintendere il significato dei dati, arrivando a conclusioni arbitrarie

La PA ha il vantaggio di rendere i dati rilevabili. I dati così hanno un valore più elevato e la PA potrà godere dei benefici derivanti dal collegamento tra i dati.

Dal punto di vista dei costi la PA deve prevedere delle risorse per collegare i dati e riparare eventuali link interrotti.

**Esempio**

Una tabella on line strutturata con un indirizzo identificativo unico sul web (URI) e descritto in RDF e con collegamenti semantici a risorse del web.

<http://5stardata.info/en/>

Day	Lowest Temperature (C)
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

```

Lowest
<a class="highlight" rel="rdfs:seeAlso" href="http://en.wikipedia.org/wiki/Temperature" resource="http://dbpedia.org/resource/Temperature">Tem
(<span class="highlight" rel="owl:sameAs" resource="http://dbpedia.org/resource/Celsius">°C/</span>)

```