Come si producono 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.

Autore: Gianfranco Andriola

Creatore: Formez PA

Diritti: Dipartimento della Funzione Pubblica

Data: Ottobre 2015

Come si producono i dati aperti

Parlando di open data della Pubblica Amministrazione siamo certamente parlando di dati leggibili da applicazioni informatiche; questa caratteristica viene normalmente definita come machine-readable e infatti si parla di machine-readable data, cioè di dati leggibili in maniera comoda e in maniera semplice, in maniera diretta da applicazioni. In realtà la ragione principale per cui si ragiona in termini di machine-readable data è più o meno riconducibile al fatto che la maggior parte delle persone osservano questi dati attraverso applicazioni informatiche e producono questi dati attraverso applicazioni informatiche; in questo senso l'open data cerca di preservare questa sorta di catena tra scrittura e lettura, in maniera tale che le applicazioni possano semplicemente comunicare tra di loro e le persone comunicare attraverso le applicazioni.

Proviamo adesso a vedere un po' più da vicino come è fatto un file strutturato. Quello che vedete alle mie spalle, ad esempio, è un classico dataset, strutturato per righe e colonne. Partiamo dall'elemento distintivo, quello più caratteristico, quello che meglio descrive tutto il resto cioè la cella: la cella è uno spazio che contiene un'informazione univoca messe le une accanto alle altre, le celle compongono i "record" cioè una serie di righe, una dopo l'altra, come potete vedere in questo file, che fanno riferimento ad uno stesso oggetto. Guardando invece il file in verticale è possibile guardare quelli che comunemente vengono detti attributi, cioè le colonne dove ogni colonna contiene valori omogenei dello stesso attributo.

Quella che vedete proiettata nella slide alle mie spalle, è quella che comunemente viene chiamata Scala di Tim Berners Lee, cioè, è una serie di classificazioni che ci aiutano a capire quanto un file può essere considerato aperto e soprattutto ci permettono di fare un ragionamento sulla qualità dei file pubblicati in formato aperto. La scala di Tim Berners Lee prevede una serie di step successivi e propedeutici l'uno all'altro che partono dal file peggio redatto che, in questo caso, è rappresentato dal file PDF, fino ad arrivare ai dataset più evoluti, redatti con tecnologie molto avanzate dove ad esempio (si usano) i linked opendata piuttosto che l'RDF.

Proviamo a vedere in questa video lezione quello che è il file di partenza, quello più interessante dal punto di vista dell'open data, cioè il primo step che possiamo considerare davvero in dataset aperto e cioè il CSV. CSV è l'acronimo di Comma-Separated Values ed è un file formato da valori separati l'uno dall'altro sulla base di virgole. Un file CSV che normalmente viene rappresentato in questo modo dall'applicazione informatica, guardandolo

su file di testo è strutturato in questo modo: vedete le informazioni vengono staccate l'una dall'altra sulla base di virgole, mentre le informazioni di tipo alfanumerico vengono differenziate dai numeri sulla base delle virgolette.

Proviamo adesso a vedere come può essere generato un file CSV partendo da quattro tipologie diverse da (ad esempio) da un database, da un classico foglio di calcolo, vedremo in particolare formato Excel, da un documento PDF e infine dai documenti cartacei. Partiamo dal database nel caso in cui ci dovessimo trovare a estrarre delle informazioni che si trovano all'interno dei database, lo strumento con cui è possibile estrarre queste informazioni e poi tradurle il file CSV è la così detta "query" c'è un'interrogazione puntuale fatta sul database che ci permette di estrarre informazioni strutturate in righe e colonne che poi vengono salvate il formato CSV.

Partendo invece da un file Excel il procedimento è piuttosto semplice, almeno prima vista: Excel permette di salvare in formato CSV il file su questo che lavorando quindi un file strutturato in righe e colonne può essere salvato da Excel a CSV attraverso un comando. Va fatta però una distinzione: non tutte le tabelle possono automaticamente diventare dataset; in alcuni casi ci si può trovare di fronte dei problemi di formattazione una tabella strutturata in questo modo ad esempio (vedete nelle celle più in alto c'è plastica, carta e vetro) certamente quel genere di celle, sono celle unite, cioè due celle che insieme ne compongono una cioè una questione di formattazione. Il file CSV non ammette questo genere di funzioni quindi è necessario, prima di salvare il file Excel in CSV tornare sulla formattazione del file in maniera tale che sia più pulito in questo caso, ad esempio, potremmo ristrutturare questo file esattamente così come vedete strutturato in questa slide dove, nella parte più alta del record più alto troviamo le definizioni e poi più in basso i valori. Le informazioni sono le stesse di quelle viste in partenza solo rielaborate in funzione del salvataggio in CSV.

Discorso diverso certamente più complesso è quello di estrarre un CSV partendo da un file PDF, è utile può essere comodo appoggiarsi ad applicazioni esterne: una delle più utilizzate è "Tabula" che consente di partire dal PDF ed estrarre il CSV restando in righe e colonne come abbiamo visto all'inizio di questa video lezione mantenendo però la formattazione iniziale. Cosa significa? Il PDF è un formato che è come se imbrigliasse le informazioni all'interno di una formattazione standard; applicazioni tipo Tabula riescono a smontare questa formattazione e rimontare le informazioni in una formattazione più agevole che è funzionale al salvataggio in CSV. Qualora ci dovessimo trovare di fronte a PDF statici, cioè PDF che al loro interno contengono delle immagini che non hanno una formattazione di partenza da cui

partire, è sostanzialmente impossibile utilizzare applicazioni tipo Tabula, quindi diventa quasi impossibile riuscire a estrarre formazioni in un modo così agevole.

Infine vediamo come è possibile estrarre file CSV partendo dalla carta. È un discorso più difficile, più raro che accada una situazione simile, dal momento che ormai sono molti anni che le informazioni vengono prodotte attraverso applicazioni informatiche, però qualora ci si dovesse presentare il caso l'unica soluzione è appoggiarsi a programmi di riconoscimento ottico dei caratteri (OCR), che in sostanza prendono la scansione digitale del file, individuano dove riconoscono, dove è possibile riconoscere lettere e numeri ed in generale formati caratteri alfanumerici, le traducono attraverso una formattazione. In questo caso, però, è molto difficile che la formattazione di arrivo sia simile alla formattazione di partenza. Quindi è bene ricordarsi, qualora si dovesse ricorrere ad estrazioni di file di tipo OCR, provare a ritornare su queste informazioni e riadattare quella che è la resa finale del file rispetto alla situazione di partenza.

Tornando alla scala di Tim Berners Lee che abbiamo visto all'inizio, proviamo a ragionare invece sui passi successivi che possono essere fatti oltre il CSV: parliamo di linked opendata. Il CSV è certamente un ottimo file di partenza, ed è certamente opendata un file redatto in formato CSV ma per aggiungere valore all'informazione è bene partire da questo file e provare a ragionare su file diversi, tipi di file diversi, che contengono ancora più informazioni in unico file in un'unica descrizione. In questo caso l'esempio che abbiamo deciso di raccontare in questa video lezione è l'RDF acronimo di Recource Description Framework ed è uno strumento base proposto dal W3C per la codifica di informazioni in maniera linked, dove le informazioni sono linkate l'una all'altra.

Provando a ragionare con un minimo di astrazione, i file RDF potrebbero essere simili a quello che vedete proiettato in questa slide e cioè una serie di database che riescono a dialogare tra di loro sulla base di un elemento comune; questo elemento comune è definito ontologia. Un'ontologia è una rappresentazione formale di un determinato dominio di interesse, costruito in maniera tale che un'applicazione informatica possa comprenderla. Nel momento in cui parliamo di file linked, file che possono dialogare tra di loro in formato RDF, sostanzialmente stiamo facendo in modo che i dati parlino la stessa lingua: i dati diventano interoperabili tra di loro.

Sia che si parli di file CSV o che si parli di file RDF o di qualunque genere di formato utilizzato dalle Pubbliche Amministrazioni per esporre i propri dati in formato aperto, è necessario e indispensabile, in questo momento, ragionare sulla qualità della redazione dei file: avere file

redatti, ben descritti, ben meta-datati e ben fatti anche da un punto di vista formale è certamente l'elemento essenziale per aggiungere, per permettere, anche a soggetti esterni di aggiungere valore a quelle informazioni.