Visualizzare, riorganizzare e ripulire i dati

Formez PA

Unione europea
Fondo sociale europeo

Ministero del Lavoro e delle Politiche Sociali
Direzione Generale per le Politiche Attive e Passive del Lavoro

Governo italiano
Presidenza del Consiglio dei Ministri
Dipartimento della Funzione pubblica

FSE per il tuo futuro
Programmi operativi nazionali per la formazione e l'occupazione

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

Data: Ottobre 2015

Visualizzare, riorganizzare e ripulire i dati

La visualizzazione grafica dei dati non solo è uno degli esempi più interessanti di riuso che si possono fare sui dati stessi, ma spesso può anche essere un modo essenziale per cercare di capire le informazioni che sono contenute all'interno dei dati all'interno dei dataset. Visualizzare le informazioni spesso può essere un'ottima opportunità per cogliere la complessità delle relazioni che esistono all'interno delle informazioni riportate in un dataset, oppure per comprendere l'ampiezza dei fenomeni, soprattutto nel momento in cui si ha a che fare con cifre particolarmente grandi o particolarmente piccole, difficili da cogliere assolutamente in maniera immediata.

Quindi, provare a rapportarle l'una all'altra in maniera grafica e in maniera visuale, può essere un ottimo modo per cercare di comprendere quelle informazioni al primo sguardo, in maniera semplice e in maniera diretta. Così, come ci si trova in possesso di informazioni georeferenziate, cioè che hanno un proprio tag geografico, una latitudine e una longitudine, collocare queste informazioni all'interno di mappe come visto in questa slide, può essere un ottimo modo per comprendere la localizzazione dei fenomeni, capire le relazioni che esistono rispetto agli spazi, rispetto a quello che sta avvenendo e alle informazioni che vengono detenute all'interno dei dataset.

In questa videolezione avremo modo di capire come, partendo dalle informazioni strutturate in formato tabellare cioè partendo dal dataset, con opportune modifiche e opportuni interventi sulle formazione iniziali, è possibile trarre visualizzazioni grafiche da quelle informazioni. Proviamo a partire dal grafico rappresentato alle mie spalle, si tratta dell'età media della popolazione per provincia in Emilia Romagna. Il grafico mostra in maniera immediata, al primo sguardo, come in tutte le province la popolazione sia cresciuta e quale sia il rapporto di crescita tra le singole province. Questi dati sono tratti dall'Istat e sono una visualizzazione grafica di informazioni che inizialmente vengono strutturate in formato tabellare, cioè per righe e colonne e poi rappresentate in questo modo.

Per generare il grafico che abbiamo appena visto abbiamo utilizzato un'applicazione che si chiama Datawrapper che è quella che vedete rappresentata nella slide alle mie spalle. Datawrapper è un'applicazione, come ne esistono centinaia, che permette di partire dai dati e visualizzare queste informazioni in forma grafica, spesso in forma grafica estremamente accattivante. Abbiamo deciso di raccontare la generazione di un grafico a partire dai dati attraverso questa applicazione perché è gratuita, perché è localizzata in italiano, perché

semplicemente è accessibile a chiunque ma sono decine le applicazione che ci permettono di fare esattamente la stessa operazione con pochi click. Datawrapper funziona in questo modo: si prendono dei dati si danno i dati alle applicazioni in un formato che adesso vedremo esattamente com'è strutturato e l'applicazione in automatico genera un grafico, o meglio permette a chi lo sta utilizzando di scegliere tra una serie di opportunità grafiche per la restituzione finale delle informazioni.

Il dataset che abbiamo utilizzato per questo esempio di rappresentazione grafica è la serie storica dell'età media delle province dell'Emilia Romagna. È stato scaricato dal sito dell'Istat, da dati.istat.it. Inizialmente il dataset si presentava in questo modo. Questo modo non è esattamente quello che serve a Datawrapper per generare il grafico quindi è stato necessario lavorare su quelle informazioni fino a ottenere un grafico di questo tipo, esattamente quello uguale a quello che vedete rappresentato in questa slide. Per farlo abbiamo deciso di utilizzare un'applicazione che in questo caso è Microsoft Excel, ma ne esistono decine di fogli di calcolo che permettono di riprendere le stesse informazioni, e utilizzare funzioni per rielaborarle, in particolare quella utilizzata per questo grafico è la tabella pivot, che ci ha permesso di prendere le informazioni così com'erano strutturate all'inizio e di renderle così come servivano a Datawrapper per poterle rielaborare.

Una volta ottenute le informazioni così come le volevamo all'inizio per essere rappresentate nella maniera più giusta, siamo tornati sull'applicazione Datawrapper e abbiamo fatto il primo dei 4 step che ci portano dalle informazioni in formato strutturato, in formato tabellare fino alla loro rappresentazione grafica. Il primo passaggio è quello di caricare i dati, quindi abbiamo lavorato sul dataset iniziale e adesso riversiamo i dati che abbiamo ottenuto all'interno dell'applicazione; come secondo passaggio, l'applicazione ci chiede se la comprensione dei dati è quella più corretta e quindi in questo caso bisogna dargli un ok che l'applicazione si assicuri che le informazioni caricate siano quelle corrette. Infine possiamo scegliere qual è il tipo di grafico che possiamo estrarre a seconda di come desideriamo rappresentare le informazioni iniziali e quindi pubblicare il grafico così come l'abbiamo visto all'inizio di questa parte della video lezione. Abbiamo appena visto come spesso la parte più difficile dell'elaborazione grafica di informazioni strutturate non sia tanto nell'elaborazione del grafico, quanto nel lavorare su quelle informazioni iniziali perché assumano la forma che desideriamo darle alla fine nell'esposizione grafica.

Esistono vari tool, vari strumenti che ci consentono di prendere queste informazioni in formato tabellare e rimetterci mano e ripulirle e riorganizzarle come meglio desideriamo rappresentarle alla fine. Uno di questi è Open Refine che ci consente di fare tutta una serie di

operazioni sulle informazioni, ad esempio, nella tabella che vedete rappresentata alle mie spalle, il dataset delle centraline di rilevamento della qualità dell'aria dell'acqua in Puglia. Cerchiamo di capire, ad esempio in questo caso, se il valore della colonna comune è uguale a tutti o ci sono degli errori all'interno di un dataset; in questo caso siamo alla funzione "facet" dell'applicazione Open Refine e troviamo che nella tabella comune ci sono delle occorrenze differenti o meglio, che in questo caso, il Comune di Monopoli è riportato due volte in maniera diversa quindi l'applicazione ci consente di intervenire rapidamente attraverso un paio di passaggi e attraverso un paio di funzioni ripulendo il dataset e ottenendo quindi un dataset pulito esattamente così come lo volevamo all'inizio.

L'esempio che abbiamo appena visto di ripulitura del dataset ottraverso Open Refine è estremamente banale però ci è funzionale a comprendere come strumenti di questo tipo possono essere estremamente utili qualora ci si trovi di fronte a file che hanno una grandissima quantità di informazioni. Qui è necessario probabilmente fare una riflessione più ampia cioè comprendere come spesso le informazioni che desideriamo lavorare in maniera grafica o per un'applicazione insomma per qualunque uso desideriamo farne non si trovano soltanto su una fonte ma spesso sono l'accorpamento di fonti differenti che ci portano un po' come è rappresentato nella slide che vedete alle mie spalle ad avere un dataset che contiene tutte le informazioni ma che quasi certamente sarà sporco, avrà dei problemi, non avrà esattamente la forma che desideriamo dargli alla fine. Quindi, tool come quello che abbiamo appena visto sono necessari per rimodellare le informazioni e per giungere al dataset finale, cioè quello pulito, quello che utilizzeremo per rappresentare dei grafici o per fare qualunque altro riutilizzo delle informazioni.