Introduzione agli Open Data
Stumenti per dati a 2 e 3 stelle

Cristiano Longo
longo@dmi.unict.it

Università di Catania

Riepilogo - 5 stelle

Classificazione a 5 stelle:[1]

1. dati disponibili rilasciati con una licenza aperta;
2. dati in un formato leggibile da un agente automatico;
3. dati in un formato aperto;
4. dati resi disponibili con le tecnologie del web semantico;
5. dati collegati ad altri dataset.

Vedremo alcuni strumenti per trattare dati a 2 e 3 stele.

[1] vedi anche `http://5stardata.info/`

Riepilogo - 5 stelle

Classificazione a 5 stelle:[1]

1. dati disponibili rilasciati con una licenza aperta;
2. dati in un formato leggibile da un agente automatico;
3. dati in un formato aperto;
4. dati resi disponibili con le tecnologie del web semantico;
5. dati collegati ad altri dataset.

Vedremo alcuni strumenti per trattare dati a 2 e 3 stele.

[1] vedi anche `http://5stardata.info/`

Dati ad una e due stelle

Esempi di dati ad una e due stelle: files pdf.

Le scansioni pdf non sono trattabili. Vedi ad esempio

`http://www.comune.messina.it/informazioni/trasparenza-valutazione-e-merito/allegato/pubblicazione-art-1-c-735-lf-2007.pdf` .

Le tabelle dentro i PDF *selezionabili* (PDFa) possono invece essere estratte. Vedi ad esempio

`http://www.comune.messina.it/informazioni/trasparenza-valutazione-e-merito/dati-relativi-a-incarichi-e-consulenze/allegati/report-dipendenti-2010.pdf` .

Dati ad una e due stelle

Esempi di dati ad una e due stelle: files pdf.

Le scansioni pdf non sono trattabili. Vedi ad esempio

`http://www.comune.messina.it/informazioni/trasparenza-valutazione-e-merito/allegato/pubblicazione-art-1-c-735-lf-2007.pdf` .

Le tabelle dentro i PDF *selezionabili* (PDFa) possono invece essere estratte. Vedi ad esempio

`http://www.comune.messina.it/informazioni/trasparenza-valutazione-e-merito/dati-relativi-a-incarichi-e-consulenze/allegati/report-dipendenti-2010.pdf` .

Dati ad una e due stelle

Esempi di dati ad una e due stelle: files pdf.

Le scansioni pdf non sono trattabili. Vedi ad esempio

`http://www.comune.messina.it/informazioni/trasparenza-valutazione-e-merito/allegato/pubblicazione-art-1-c-735-lf-2007.pdf` .

Le tabelle dentro i PDF *selezionabili* (PDFa) possono invece essere estratte. Vedi ad esempio

`http://www.comune.messina.it/informazioni/trasparenza-valutazione-e-merito/dati-relativi-a-incarichi-e-consulenze/allegati/report-dipendenti-2010.pdf` .

Da due a tre stelle: Tabula

Tabula[2] è uno strumento open source che permette di estrarre le tabelle nei files pdf e convertirle in formato CSV.

Sul sito principale sono presenti i pacchetti per mac e windows. Per gli altri sistemi operativi è necessario scaricare `tabula-jar.zip`.
Per avviare l'applicazione, scompattare il pacchetto ed eseguire il seguente comando

java -Dfile.encoding=utf-8 -Xms256M -Xmx1024M -jar tabula.jar

Nella form disponibile al seguente indirizzo, importare il file PDF da convertire:

`http://127.0.0.1:8080/` .

Nell'anteprima, selezionare le aree contenenti le tabelle da convertire. Se l'anteprima dell'import non mostra problemi, scaricare il CSV. Nel caso in cui la tabella sia su più pagine, è possibile selezionare più di un'area. Quando tutte le aree di interesse sono state selezionate, premere su *Download all data*.

[2] `http://tabula.technology/`

Da due a tre stelle: Tabula

Tabula[2] è uno strumento open source che permette di estrarre le tabelle nei files pdf e convertirle in formato CSV.

Sul sito principale sono presenti i pacchetti per mac e windows. Per gli altri sistemi operativi è necessario scaricare `tabula-jar.zip`.
Per avviare l'applicazione, scompattare il pacchetto ed eseguire il seguente comando

java -Dfile.encoding=utf-8 -Xms256M -Xmx1024M -jar tabula.jar

Nella form disponibile al seguente indirizzo, importare il file PDF da convertire:

`http://127.0.0.1:8080/` .

Nell'anteprima, selezionare le aree contenenti le tabelle da convertire. Se l'anteprima dell'import non mostra problemi, scaricare il CSV. Nel caso in cui la tabella sia su più pagine, è possibile selezionare più di un'area. Quando tutte le aree di interesse sono state selezionate, premere su *Download all data*.

[2] `http://tabula.technology/`

Da due a tre stelle: Tabula

Tabula[2] è uno strumento open source che permette di estrarre le tabelle nei files pdf e convertirle in formato CSV.

Sul sito principale sono presenti i pacchetti per mac e windows. Per gli altri sistemi operativi è necessario scaricare `tabula-jar.zip`.
Per avviare l'applicazione, scompattare il pacchetto ed eseguire il seguente comando

java -Dfile.encoding=utf-8 -Xms256M -Xmx1024M -jar tabula.jar

Nella form disponibile al seguente indirizzo, importare il file PDF da convertire:

`http://127.0.0.1:8080/` .

Nell'anteprima, selezionare le aree contenenti le tabelle da convertire. Se l'anteprima dell'import non mostra problemi, scaricare il CSV. Nel caso in cui la tabella sia su più pagine, è possibile selezionare più di un'area. Quando tutte le aree di interesse sono state selezionate, premere su *Download all data*.

[2] `http://tabula.technology/`

Da due a tre stelle: Tabula

Tabula[2] è uno strumento open source che permette di estrarre le tabelle nei files pdf e convertirle in formato CSV.

Sul sito principale sono presenti i pacchetti per mac e windows. Per gli altri sistemi operativi è necessario scaricare `tabula-jar.zip`.
Per avviare l'applicazione, scompattare il pacchetto ed eseguire il seguente comando

java -Dfile.encoding=utf-8 -Xms256M -Xmx1024M -jar tabula.jar

Nella form disponibile al seguente indirizzo, importare il file PDF da convertire:

`http://127.0.0.1:8080/` .

Nell'anteprima, selezionare le aree contenenti le tabelle da convertire. Se l'anteprima dell'import non mostra problemi, scaricare il CSV. Nel caso in cui la tabella sia su più pagine, è possibile selezionare più di un'area. Quando tutte le aree di interesse sono state selezionate, premere su *Download all data*.

[2] `http://tabula.technology/`

Trattamento di Dati Tabellari - Datawrapper

Datawrapper[3] è uno strumento open-source per la creazione di grafici a partire da dati tabellari in formato CSV. I grafici creati con il servizio online possono essere visualizzati nelle proprie pagine web.

Per creare un nuovo grafico è necessario innanzitutto caricare un file CSV che contiene i dati da mostrare. È necessario specificare la fonte, che poi verrà riportata nel grafico finale come meta-dato.

Il secondo passo è speficare i tipi delle colonne, nel caso in cui il detect automatico non abbia funzionato. È necessario che sia presente almeno un campo con tipo number.

Infine si sceglie il tipo di grafico e si esporta in PDF o altro formato.

[3] `https://datawrapper.de/`

Trattamento di Dati Tabellari - Datawrapper

Datawrapper[3] è uno strumento open-source per la creazione di grafici a partire da dati tabellari in formato CSV. I grafici creati con il servizio online possono essere visualizzati nelle proprie pagine web.

Per creare un nuovo grafico è necessario innanzitutto caricare un file CSV che contiene i dati da mostrare. È necessario specificare la fonte, che poi verrà riportata nel grafico finale come meta-dato.

Il secondo passo è speficare i tipi delle colonne, nel caso in cui il detect automatico non abbia funzionato. È necessario che sia presente almeno un campo con tipo number.

Infine si sceglie il tipo di grafico e si esporta in PDF o altro formato.

[3] `https://datawrapper.de/`

Trattamento di Dati Tabellari - Datawrapper

Datawrapper[3] è uno strumento open-source per la creazione di grafici a partire da dati tabellari in formato CSV. I grafici creati con il servizio online possono essere visualizzati nelle proprie pagine web.

Per creare un nuovo grafico è necessario innanzitutto caricare un file CSV che contiene i dati da mostrare. È necessario specificare la fonte, che poi verrà riportata nel grafico finale come meta-dato.

Il secondo passo è speficare i tipi delle colonne, nel caso in cui il detect automatico non abbia funzionato. È necessario che sia presente almeno un campo con tipo number.

Infine si sceglie il tipo di grafico e si esporta in PDF o altro formato.

[3] `https://datawrapper.de/`

Trattamento di Dati Tabellari - Datawrapper

Datawrapper[3] è uno strumento open-source per la creazione di grafici a partire da dati tabellari in formato CSV. I grafici creati con il servizio online possono essere visualizzati nelle proprie pagine web.

Per creare un nuovo grafico è necessario innanzitutto caricare un file CSV che contiene i dati da mostrare. È necessario specificare la fonte, che poi verrà riportata nel grafico finale come meta-dato.

Il secondo passo è speficare i tipi delle colonne, nel caso in cui il detect automatico non abbia funzionato. È necessario che sia presente almeno un campo con tipo number.

Infine si sceglie il tipo di grafico e si esporta in PDF o altro formato.

[3] `https://datawrapper.de/`

Trattamento di Dati Tabellari - Datawrapper - incidenti stradali

Vediamo come usare Datawrapper per visualizzare i dati sugli incidenti stradali nel comune di catania. In particolare evidenziamo il numero di incidenti per mese, tralasciando invece il numero di morti e feriti.

Passo 1 - scaricare i dati in CSV dal seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/incidenti-2012` .

Passo 2 - creare un nuovo grafico su Datawrapper, caricando il dataset e indicando la fonte.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - incidenti stradali

Vediamo come usare Datawrapper per visualizzare i dati sugli incidenti stradali nel comune di catania. In particolare evidenziamo il numero di incidenti per mese, tralasciando invece il numero di morti e feriti.

Passo 1 - scaricare i dati in CSV dal seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/incidenti-2012` .

Passo 2 - creare un nuovo grafico su Datawrapper, caricando il dataset e indicando la fonte.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - incidenti stradali

Vediamo come usare Datawrapper per visualizzare i dati sugli incidenti stradali nel comune di catania. In particolare evidenziamo il numero di incidenti per mese, tralasciando invece il numero di morti e feriti.

Passo 1 - scaricare i dati in CSV dal seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/incidenti-2012` .

Passo 2 - creare un nuovo grafico su Datawrapper, caricando il dataset e indicando la fonte.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - incidenti stradali

Vediamo come usare Datawrapper per visualizzare i dati sugli incidenti stradali nel comune di catania. In particolare evidenziamo il numero di incidenti per mese, tralasciando invece il numero di morti e feriti.

Passo 1 - scaricare i dati in CSV dal seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/incidenti-2012` .

Passo 2 - creare un nuovo grafico su Datawrapper, caricando il dataset e indicando la fonte.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - incidenti stradali

Vediamo come usare Datawrapper per visualizzare i dati sugli incidenti stradali nel comune di catania. In particolare evidenziamo il numero di incidenti per mese, tralasciando invece il numero di morti e feriti.

Passo 1 - scaricare i dati in CSV dal seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/incidenti-2012` .

Passo 2 - creare un nuovo grafico su Datawrapper, caricando il dataset e indicando la fonte.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - incidenti stradali

Vediamo come usare Datawrapper per visualizzare i dati sugli incidenti stradali nel comune di catania. In particolare evidenziamo il numero di incidenti per mese, tralasciando invece il numero di morti e feriti.

Passo 1 - scaricare i dati in CSV dal seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/incidenti-2012` .

Passo 2 - creare un nuovo grafico su Datawrapper, caricando il dataset e indicando la fonte.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - Risultati Elettorali

Vediamo come usare Datawrapper per visualizzare i risultati elettorali delle elezioni amministrative del 2013 del comune di Catania.

I dati sono disponibili in CSV al seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/elezioni-amministrative-2013-voti-liste-consiglio-comunale` .

Si crea il grafico con Data Wrapper come visto prima. La prima colonna deve essere nascosta e di dati trasposti.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - Risultati Elettorali

Vediamo come usare Datawrapper per visualizzare i risultati elettorali delle elezioni amministrative del 2013 del comune di Catania.

I dati sono disponibili in CSV al seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/elezioni-amministrative-2013-voti-liste-consiglio-comunale` .

Si crea il grafico con Data Wrapper come visto prima. La prima colonna deve essere nascosta e di dati trasposti.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - Risultati Elettorali

Vediamo come usare Datawrapper per visualizzare i risultati elettorali delle elezioni amministrative del 2013 del comune di Catania.

I dati sono disponibili in CSV al seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/elezioni-amministrative-2013-voti-liste-consiglio-comunale` .

Si crea il grafico con Data Wrapper come visto prima. La prima colonna deve essere nascosta e di dati trasposti.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - Risultati Elettorali

Vediamo come usare Datawrapper per visualizzare i risultati elettorali delle elezioni amministrative del 2013 del comune di Catania.

I dati sono disponibili in CSV al seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/elezioni-amministrative-2013-voti-liste-consiglio-comunale` .

Si crea il grafico con Data Wrapper come visto prima. La prima colonna deve essere nascosta e di dati trasposti.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - Risultati Elettorali

Vediamo come usare Datawrapper per visualizzare i risultati elettorali delle elezioni amministrative del 2013 del comune di Catania.

I dati sono disponibili in CSV al seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/elezioni-amministrative-2013-voti-liste-consiglio-comunale` .

Si crea il grafico con Data Wrapper come visto prima. La prima colonna deve essere nascosta e di dati trasposti.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Datawrapper - Risultati Elettorali

Vediamo come usare Datawrapper per visualizzare i risultati elettorali delle elezioni amministrative del 2013 del comune di Catania.

I dati sono disponibili in CSV al seguente indirizzo

`http://opendata.comune.catania.gov.it/dataset/elezioni-amministrative-2013-voti-liste-consiglio-comunale` .

Si crea il grafico con Data Wrapper come visto prima. La prima colonna deve essere nascosta e di dati trasposti.

Passo 3 - accertarsi che le colonne *INCIDENTI CON FERITI* e *INCIDENTI MORTALI* siano rilevate come colonne numeriche. Eliminare le due colonne *FERITI* e *MORTI* dalla visualizzazione.

Passo 4 - Inserire il titolo (scheda *Annotate*) e selezionare il tipo di grafico (scelta consigliata: istogramma raggruppato). Per confrontare il numero di incidenti mortali e con feriti per mese è conveniente trasporre il grafico.

Passo 5 - Infine pubblicare o esportare il grafico.

Trattamento di Dati Tabellari - Pulizia

A volte, anche se forniti in formati tabellari aperti, i dati non sono adatti ad essere processati automaticamente. Ad esempio, lo stesso nome può essere indicato tutto in maiuscolo e, nella stessa tabella, solo con la prima lettera maiuscola.

Nell'esempio visto prima riguardante gli incarichi del comune di messina, il simbolo dell'euro e l'utilizzo della virgola come separatore della parte decimale per i compensi rende impossibile il trattamento di questi dati con Datawrapper.

Trattamento di Dati Tabellari - Pulizia - Open Refine (1/3)

Open Refine[4] è un progetto open-source che, assieme ad altre funzionalità, fornisce alcuni strumenti per la pulizia e la trasformazione dei dati.

Per utilizzare open-refine è sufficiente scompattare il pacchetto fornito sul sito e avviare l'applicatovo `refine`. All'avvio viene comunicato l'indirizzo attraverso il quale accedere all'applicativo.[5]

[4] `http://openrefine.org`

[5] Solitamente `http://127.0.0.1:3333/` .

Trattamento di Dati Tabellari - Pulizia - Open Refine (1/3)

Open Refine[4] è un progetto open-source che, assieme ad altre funzionalità, fornisce alcuni strumenti per la pulizia e la trasformazione dei dati.

Per utilizzare open-refine è sufficiente scompattare il pacchetto fornito sul sito e avviare l'applicatovo `refine`. All'avvio viene comunicato l'indirizzo attraverso il quale accedere all'applicativo.[5]

[4] `http://openrefine.org`

[5] Solitamente `http://127.0.0.1:3333/` .

Trattamento di Dati Tabellari - Pulizia - Open Refine (2/3)

Su Open Refine è possibile caricare i dati da diverse tipologie di fonti (web, file locale, ...). Carichiamo il file sugli incarichi al comune di Messina precedentemente creato con Tabula. In questa fase è necessario specificare che la prima riga contiene l'header e che il file è in formato CSV con i campi separati da virgole.

Quando l'anteprima non rileva errori o imprecisioni, scegliere un nome e creare il progetto.

Effettuiamo le operazioni di pulizia. Nel nostro caso dobbiamo rimuovere il simbolo dell'euro e sostituire la virgola con il punto nella colonna `Importo Erogato`. Per fare questo selezionare `Edit Cells - Transform` nel menù a tendina relativo alla colonna. Si apre a questo punto una form nella quale specificare le modifiche da effettuare in linguaggio GREL.[6] Nel nostro caso inserire la seguente espressione:

```
value.replace("€","").replace(".","").replace(",",".").trim()
```

Infine, si esporta il progetto in formato CSV.

[6] Per il linguaggio gRel vedi
https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Expressions e
https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions .

Trattamento di Dati Tabellari - Pulizia - Open Refine (2/3)

Su Open Refine è possibile caricare i dati da diverse tipologie di fonti (web, file locale, ...). Carichiamo il file sugli incarichi al comune di Messina precedentemente creato con Tabula. In questa fase è necessario specificare che la prima riga contiene l'header e che il file è in formato CSV con i campi separati da virgole.

Quando l'anteprima non rileva errori o imprecisioni, scegliere un nome e creare il progetto.

Effettuiamo le operazioni di pulizia. Nel nostro caso dobbiamo rimuovere il simbolo dell'euro e sostituire la virgola con il punto nella colonna `Importo Erogato`. Per fare questo selezionare `Edit Cells - Transform` nel menù a tendina relativo alla colonna. Si apre a questo punto una form nella quale specificare le modifiche da effettuare in linguaggio GREL.[6] Nel nostro caso inserire la seguente espressione:

```
value.replace("€","").replace(".","").replace(",",".").trim()
```

Infine, si esporta il progetto in formato CSV.

[6] Per il linguaggio gRel vedi
https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Expressions e
https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions .

Trattamento di Dati Tabellari - Pulizia - Open Refine (2/3)

Su Open Refine è possibile caricare i dati da diverse tipologie di fonti (web, file locale, . . .). Carichiamo il file sugli incarichi al comune di Messina precedentemente creato con Tabula. In questa fase è necessario specificare che la prima riga contiene l'header e che il file è in formato CSV con i campi separati da virgole.

Quando l'anteprima non rileva errori o imprecisioni, scegliere un nome e creare il progetto.

Effettuiamo le operazioni di pulizia. Nel nostro caso dobbiamo rimuovere il simbolo dell'euro e sostituire la virgola con il punto nella colonna `Importo Erogato`. Per fare questo selezionare `Edit Cells - Transform` nel menù a tendina relativo alla colonna. Si apre a questo punto una form nella quale specificare le modifiche da effettuare in linguaggio GREL.[6] Nel nostro caso inserire la seguente espressione:

value.replace("€","").replace(".","").replace(",",".").trim()

Infine, si esporta il progetto in formato CSV.

[6] Per il linguaggio gRel vedi
https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Expressions e
https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions .

Trattamento di Dati Tabellari - Pulizia - Open Refine (2/3)

Su Open Refine è possibile caricare i dati da diverse tipologie di fonti (web, file locale, ...). Carichiamo il file sugli incarichi al comune di Messina precedentemente creato con Tabula. In questa fase è necessario specificare che la prima riga contiene l'header e che il file è in formato CSV con i campi separati da virgole.

Quando l'anteprima non rileva errori o imprecisioni, scegliere un nome e creare il progetto.

Effettuiamo le operazioni di pulizia. Nel nostro caso dobbiamo rimuovere il simbolo dell'euro e sostituire la virgola con il punto nella colonna `Importo Erogato`. Per fare questo selezionare `Edit Cells - Transform` nel menù a tendina relativo alla colonna. Si apre a questo punto una form nella quale specificare le modifiche da effettuare in linguaggio GREL.[6] Nel nostro caso inserire la seguente espressione:

value.replace("€","").replace(".","").replace(",",".").trim()

Infine, si esporta il progetto in formato CSV.

[6] Per il linguaggio gRel vedi
https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Expressions e
https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions .

Trattamento di Dati Tabellari - Pulizia - Open Refine (3/3)

Altre caratteristiche interessanti di Open Refine sono:

- *esportazione dei template* che permette di riapplicare le stesse trasformazioni ad un file di formato simile in termini di colonne;
- *suddivisione di celle* ad esempio per separare nome e cognome, o via e numero civico;
- *conversione in maiuscolo*.

Trattamento di Dati Tabellari - Pulizia - Open Refine (3/3)

Altre caratteristiche interessanti di Open Refine sono:

- *esportazione dei template* che permette di riapplicare le stesse trasformazioni ad un file di formato simile in termini di colonne;
- *suddivisione di celle* ad esempio per separare nome e cognome, o via e numero civico;
- *conversione in maiuscolo.*

Trattamento di Dati Tabellari - Pulizia - Open Refine (3/3)

Altre caratteristiche interessanti di Open Refine sono:

- *esportazione dei template* che permette di riapplicare le stesse trasformazioni ad un file di formato simile in termini di colonne;
- *suddivisione di celle* ad esempio per separare nome e cognome, o via e numero civico;
- *conversione in maiuscolo*.

Una delle funzionalità più utilizzate di Open Refine è il clustering.

Il *clustering* permette di accorpare, usando delle euristiche predefinite, campi con valori differenti ma che, secondo le informazioni ottenute utilizzando le suddette euristiche, rappresentano lo stesso oggetto.

Ad esempio, le stringhe "`Cristiano Longo`", "`Longo, Cristiano`", "`cristiano longo`" sarano identificate come appartententi allo stesso cluster, e sarà possibile sostituirle con un'unica stringa (ad esempio, "`Cristiano Longo`").

Trattamento di Dati Tabellari - Pulizia - Esempi di Clustering: Basi Dati delle PA

La tecnica del clustering è stata utilizzata per raffinare le basi dati della Pubblica Amministrazione comunicati all'Agenzia dell'Italia Digitale.

`http://basidati.agid.gov.it/catalogo/download.html`

Tra tutti i file disponibili, consideriamo quello relativo alle *Basi di Dati*. Cliccando sulla colonna amministrazione e selezionando Edit cells - Cluster and edit si apre la piattaforma di clustering.

Tra i vari cluster riconosciuti, il primo contiene i seguenti valori: `ISTITUTO COMPRENSIVO`, `ISTITUTO COMPRENSIVO - ISTITUTO COMPRENSIVO`, `Istituto Comprensivo`. È possibile decidere di sostituire tutti questi valori con un'unica stringa, specificata nell'ultima colonna, selezionando il flag `Merge`.

Il processo di clustering può essere applicato iterativamente.

Trattamento di Dati Tabellari - Pulizia - Esempi di Clustering: Basi Dati delle PA

La tecnica del clustering è stata utilizzata per raffinare le basi dati della Pubblica Amministrazione comunicati all'Agenzia dell'Italia Digitale.

`http://basidati.agid.gov.it/catalogo/download.html`

Tra tutti i file disponibili, consideriamo quello relativo alle *Basi di Dati*. Cliccando sulla colonna amministrazione e selezionando Edit cells - Cluster and edit si apre la piattaforma di clustering.

Tra i vari cluster riconosciuti, il primo contiene i seguenti valori: `ISTITUTO COMPRENSIVO`, `ISTITUTO COMPRENSIVO - ISTITUTO COMPRENSIVO`, `Istituto Comprensivo`. È possibile decidere di sostituire tutti questi valori con un'unica stringa, specificata nell'ultima colonna, selezionando il flag `Merge`.

Il processo di clustering può essere applicato iterativamente.

Trattamento di Dati Tabellari - Pulizia - Esempi di Clustering: Basi Dati delle PA

La tecnica del clustering è stata utilizzata per raffinare le basi dati della Pubblica Amministrazione comunicati all'Agenzia dell'Italia Digitale.

`http://basidati.agid.gov.it/catalogo/download.html`

Tra tutti i file disponibili, consideriamo quello relativo alle *Basi di Dati*. Cliccando sulla colonna amministrazione e selezionando Edit cells - Cluster and edit si apre la piattaforma di clustering.

Tra i vari cluster riconosciuti, il primo contiene i seguenti valori: `ISTITUTO COMPRENSIVO`, `ISTITUTO COMPRENSIVO - ISTITUTO COMPRENSIVO`, `Istituto Comprensivo`. È possibile decidere di sostituire tutti questi valori con un'unica stringa, specificata nell'ultima colonna, selezionando il flag `Merge`.

Il processo di clustering può essere applicato iterativamente.

Trattamento di Dati Tabellari - Pulizia - Esempi di Clustering: Basi Dati delle PA

La tecnica del clustering è stata utilizzata per raffinare le basi dati della Pubblica Amministrazione comunicati all'Agenzia dell'Italia Digitale.

`http://basidati.agid.gov.it/catalogo/download.html`

Tra tutti i file disponibili, consideriamo quello relativo alle *Basi di Dati*. Cliccando sulla colonna amministrazione e selezionando Edit cells - Cluster and edit si apre la piattaforma di clustering.

Tra i vari cluster riconosciuti, il primo contiene i seguenti valori: `ISTITUTO COMPRENSIVO`, `ISTITUTO COMPRENSIVO - ISTITUTO COMPRENSIVO`, `Istituto Comprensivo`. È possibile decidere di sostituire tutti questi valori con un'unica stringa, specificata nell'ultima colonna, selezionando il flag `Merge`.

Il processo di clustering può essere applicato iterativamente.

Dati Geografici

I dati geospaziali e geolocalizzati rappresentano una grossa fetta degli open data disponibili. Vengono forniti in diversi formati e modalità. Vedi come esempio l'elenco delle farmacie pubblicato dal comune di catania.[7]

Esamineremo alcuni tool per l'utilizzo di dati geo-referenziati.

Spesso la posizione di alcuni oggetti di interesse (scuole, asili, discariche, ...) viene fornita senza le coordinate geospaziali. Vedremo alcuni servizi di *geocoding* per ottenere le coordinate dagli indirizzi.

[7] `http://opendata.comune.catania.gov.it/dataset/test-geo`

Dati Geografici - CartoDB - Importare un Dataset

Sono disponibili alcuni strumenti online per visualizzare dati geografici. *CartoDB*[8] è un servizio web che permette di realizare mappe da esporre sul proprio sito realizzate a partire da dataset contenenti informazioni georeferenziate.

Una volta entrati nel proprio account, CartoDB fornisce due viste: una per le mappe ed una per i dataset. Un dataset può essere caricato da file o importato da diverse sorgenti (Dropbox, Google sheet, ...), recuperato online oppure caricato da un file locale.

I dataset così ottenuti vengono salvati su un database relazionale interno a CartoDB. È possibile selezionare gli oggetti da mostrare attraverso dei filtri in linguaggio SQL.

[8] `https://cartodb.com/`

Dati Geografici - CartoDB - Importare un Dataset

Sono disponibili alcuni strumenti online per visualizzare dati geografici. *CartoDB*[8] è un servizio web che permette di realizare mappe da esporre sul proprio sito realizzate a partire da dataset contenenti informazioni georeferenziate.

Una volta entrati nel proprio account, CartoDB fornisce due viste: una per le mappe ed una per i dataset. Un dataset può essere caricato da file o importato da diverse sorgenti (Dropbox, Google sheet, ...), recuperato online oppure caricato da un file locale.

I dataset così ottenuti vengono salvati su un database relazionale interno a CartoDB. È possibile selezionare gli oggetti da mostrare attraverso dei filtri in linguaggio SQL.

[8] `https://cartodb.com/`

Dati Geografici - CartoDB - Importare un Dataset

Sono disponibili alcuni strumenti online per visualizzare dati geografici. *CartoDB*[8] è un servizio web che permette di realizare mappe da esporre sul proprio sito realizzate a partire da dataset contenenti informazioni georeferenziate.

Una volta entrati nel proprio account, CartoDB fornisce due viste: una per le mappe ed una per i dataset. Un dataset può essere caricato da file o importato da diverse sorgenti (Dropbox, Google sheet, ...), recuperato online oppure caricato da un file locale.

I dataset così ottenuti vengono salvati su un database relazionale interno a CartoDB. È possibile selezionare gli oggetti da mostrare attraverso dei filtri in linguaggio SQL.

[8] `https://cartodb.com/`

La vista *Maps* permette invece di realizzare delle mappe. È possibile collegare una mappa ad un dataset quando la mappa viene creata, oppure specificare il dataset in un momento successivo.

Per creare una mappa, selezionare innanzitutto la *cartografia* da utilizzare (*change base-map*). Inoltre, specificare i contenuti degli info-box.

Infine, si può pubblicare la mappa. Fatto questo sarà possibile includerla nel proprio sito.

La vista *Maps* permette invece di realizzare delle mappe. È possibile collegare una mappa ad un dataset quando la mappa viene creata, oppure specificare il dataset in un momento successivo.

Per creare una mappa, selezionare innanzitutto la *cartografia* da utilizzare (*change base-map*). Inoltre, specificare i contenuti degli info-box.

Infine, si può pubblicare la mappa. Fatto questo sarà possibile includerla nel proprio sito.

La vista *Maps* permette invece di realizzare delle mappe. È possibile collegare una mappa ad un dataset quando la mappa viene creata, oppure specificare il dataset in un momento successivo.

Per creare una mappa, selezionare innanzitutto la *cartografia* da utilizzare (*change base-map*). Inoltre, specificare i contenuti degli info-box.

Infine, si può pubblicare la mappa. Fatto questo sarà possibile includerla nel proprio sito.

Trattamento di Dati Geografici - uMap

uMap[9] è un servizio libero e open source per la creazione e fruizione di mappe basate su *Open Street Map*.[10]

Open Street Map è un database collaborativo realizzato sul modello di wikipedia (tutti gli utenti possono aggiungere contenuti). Tutti i dati presenti su Open Street Map sono rilasciati con licenza aperta.

[9] `https://umap.openstreetmap.fr`

[10] `http://www.openstreetmap.org/`

Con uMap è possibile creare mappe personalizzate da conservare nel proprio account.

Oltre ad aggiungere punti di interesse (POI) manualmente, è possibile importare punti caricando file in diversi formati o importandone alcuni disponibili sul web.

I marker relativi ai punti di interesse sono personalizzabili in termini di forma, colori ed eventuali simboli. ATTENZIONE: è possibile inserire una immagine solo per marker di tipo *derivato*.

Infine, anche la mappa utilizzata come sfondo può essere selezionata tra quelle disponibili.

Trattamento di Dati Geografici - uMap - Creare Mappe

Con uMap è possibile creare mappe personalizzate da conservare nel proprio account.

Oltre ad aggiungere punti di interesse (POI) manualmente, è possibile importare punti caricando file in diversi formati o importandone alcuni disponibili sul web.

I marker relativi ai punti di interesse sono personalizzabili in termini di forma, colori ed eventuali simboli. ATTENZIONE: è possibile inserire una immagine solo per marker di tipo *derivato*.

Infine, anche la mappa utilizzata come sfondo può essere selezionata tra quelle disponibili.

Trattamento di Dati Geografici - uMap - Creare Mappe

Con uMap è possibile creare mappe personalizzate da conservare nel proprio account.

Oltre ad aggiungere punti di interesse (POI) manualmente, è possibile importare punti caricando file in diversi formati o importandone alcuni disponibili sul web.

I marker relativi ai punti di interesse sono personalizzabili in termini di forma, colori ed eventuali simboli. ATTENZIONE: è possibile inserire una immagine solo per marker di tipo *derivato*.

Infine, anche la mappa utilizzata come sfondo può essere selezionata tra quelle disponibili.

Con uMap è possibile creare mappe personalizzate da conservare nel proprio account.

Oltre ad aggiungere punti di interesse (POI) manualmente, è possibile importare punti caricando file in diversi formati o importandone alcuni disponibili sul web.

I marker relativi ai punti di interesse sono personalizzabili in termini di forma, colori ed eventuali simboli. ATTENZIONE: è possibile inserire una immagine solo per marker di tipo *derivato*.

Infine, anche la mappa utilizzata come sfondo può essere selezionata tra quelle disponibili.

Nei dataset geografici a volte non sono indicate le coordinate, ma solo gli indirizzi. Un esempio è il dataset *ANAGRAFE DEGLI EDIFICI PUBBLICI*[11] del comune di Palermo.

Col termine *geocoding* si intende un processo che permetta di ottenere da un indirizzo (stato, città, via, civico) le corrispondenti coordinate.

[11] http://www.comune.palermo.it/opendata_dld.php?id=319

Trattamento di Dati Geografici - Geocoding - Esempio: Beni Confiscati Palermo

I beni confiscati in gestione al comune di Palermo sono indicati nel dataset *ANAGRAFE DEGLI EDIFICI PUBBLICI*[12] del comune di Palermo. Vediamo come visualizzarli usando CartoDB.

Innanzitutto si importi su CartoDB il dataset visto prima. Successivamente selezioniamo tra quelli importati solo i beni confiscati. Questi sono quelli che nella tabella rilasciata dal comune contengono riferimenti alla legge 575/65. La query da effettuare è la seguente:

```
SELECT * FROM _4 WHERE destinazione LIKE '%575/65%'
```

Alla creazione della mappa viene richiesto su quali colonne effettuare il geocoding. Selezionare l'opzione `Street Addresses` e indicare `indirizzo` come Street Address e `no.civico` come componente aggiuntiva dell'indirizzo (tasto "+"). L'applicazione del geocoding aggiungerà dei campi con le indicazioni geografiche nel dataset.

[12] `http://www.comune.palermo.it/opendata_dld.php?id=319`