Corso di Laurea Magistrale in Economia

Data Science

A.A. 2018/2019

Lez. 7 – Open Data

Introduzione

- Conosci esattamente quanta parte delle tue tasse è destinata all'illuminazione stradale o alla ricerca contro il cancro?
- Qual è l'itinerario più breve, sicuro e panoramico per raggiungere in bici il tuo ufficio da casa tua?
- E cosa c'è nell'aria che respiri durante il tragitto?
- Dove troverai le migliori opportunità di lavoro nella tua regione, e dove il maggior numero di alberi da frutta pro-capite?
- Quand'è che puoi influenzare attivamente le decisioni sui temi che ti stanno più a cuore e con chi dovresti parlarne?
- Le nuove tecnologie permettono di creare servizi per rispondere automaticamente a queste domande.
- Molti dei dati necessari a rispondere a queste questioni sono in effetti prodotti da organismi pubblici. Tuttavia spesso tali dati non sono disponibili in formati che li rendano facili da manipolare.

Definizione

- La nozione di dati aperti – *open data*, e più specificatamente dati aperti del settore pubblico - *open government data*, è intesa come informazione, pubblica o no, accessibile e riutilizzabile da chiunque e per qualunque fine
- L'uso comune del concetto inizia nel 2009, quando diversi governi (come gli Stati Uniti d'America, il Regno Unito, il Canada e la Nuova Zelanda) hanno annunciato nuove iniziative per l'apertura della loro informazione pubblica.

Perché gli Open Data

- Molte persone e molte organizzazioni raccolgono, per svolgere i loro compiti, una vasta gamma di dati diversi.
- Quello che fa il Governo è particolarmente importante in questo senso, non solo per la quantità e centralità dei dati raccolti, ma anche perché la maggior parte dei dati governativi sono pubblici per legge, e quindi dovrebbero essere resi aperti e disponibili all'uso per chiunque.
- Ci sono molte circostanze in cui possiamo attenderci che i dati aperti abbiano un valore rilevante e molti esempi in cui questo già accade.
- Ci sono anche numerose categorie di soggetti e organizzazioni che possono trarre beneficio dalla disponibilità di dati aperti, inclusa la pubblica amministrazione.

Perché gli Open Data

- È già possibile indicare un vasto numero di aree dove i dati pubblici aperti stanno creando valore. Tra di esse citiamo:
  - la trasparenza e il controllo democratico;
  - la partecipazione;
  - l'auto-responsabilizzazione;
  - il miglioramento o la creazione di prodotti e servizi privati
  - l'innovazione
  - il miglioramento dell'efficienza dei servizi governativi
  - la misura dell'impatto di determinate politiche
  - l'estrazione di nuova conoscenza dalla combinazione di diverse fonti di dati e dall'identificazione di regolarità che emergono dall'analisi di grandi masse di dati

Perché gli Open Data

- I dati aperti governativi possono inoltre aiutare a prendere decisioni migliori nella nostra vita privata, o renderci più attivi nell'ambito della società civile. In Danimarca, una sviluppatrice ha creato findtoilet.dk che permette di accedere alla lista di tutti i bagni pubblici del paese, così anche chi soffre di problemi di incontinenza ora si sente più rassicurato dovendo uscire di casa.
- In Olanda il servizio vervuilingsalarm.nl ti avvisa quando la qualità dell'aria del tuo quartiere raggiunge una soglia critica da te definita.
- A New York puoi facilmente capire dove puoi portare a spasso il tuo cane, così come trovare altre persone che usano il tuo stesso parco.
- Tutti questi esempi utilizzano dati aperti rilasciati dai governi.

Perché gli Open Data

- Anche dal punto di vista economico i dati aperti hanno un'enorme importanza. Svariati studi hanno stimato il valore economico dei dati aperti in diverse decine di miliardi di euro ogni anno, nella sola Europa.
- Nuovi prodotti e nuove aziende stanno ri-usando dati aperti. Il sito danese husetsweb.dk aiuta a trovare i modi migliori di risparmiare energia elettrica in casa, inclusa la pianificazione finanziaria e la possibilità di contattare gli artigiani che potranno eseguire il lavoro. Funziona grazie al riutilizzo di dati catastali, a informazioni sugli incentivi governativi e al registro delle imprese locali.
- Il Ministero olandese dell'Istruzione ha pubblicato on-line tutti i dati relativi al sistema educativo consentendone il ri-uso. Da allora il numero di domande ricevute è sceso, riducendo il carico di lavoro e i costi, e anche per i dipendenti pubblici è ora più facile rispondere alle domande residue, perché ora è chiaro dove possono essere trovati i dati che servono per rispondere. I dati aperti rendono anche il governo più efficace, il che in ultima analisi riduce anche i costi.

Caratteristiche degli Open Data

- *Disponibilità e accesso*: i dati devono essere disponibili nel loro complesso, per un prezzo non superiore ad un ragionevole costo di riproduzione, preferibilmente mediante scaricamento da Internet. I dati devono essere disponibili in un formato utile e modificabile.
- *Riutilizzo e ridistribuzione*: i dati devono essere forniti a condizioni tali da permetterne il riutilizzo e la ridistribuzione. Ciò comprende la possibilità di combinarli con altre basi di dati.
- *Partecipazione universale*: tutti devono essere in grado di usare, riutilizzare e ridistribuire i dati. Non ci devono essere discriminazioni né di ambito di iniziativa né contro soggetti o gruppi. Ad esempio, la clausola 'non commerciale', che vieta l'uso a fini commerciali o restringe l'utilizzo solo per determinati scopi (es. quello educativo) non è ammessa.

Interoperabilità

- L'interoperabilità è importante perché permette a componenti diverse di lavorare insieme. L'abilità di rendere ciascun dato un componente e di combinare insieme vari componenti è essenziale per la costruzione di sistemi sofisticati. In assenza di interoperabilità ciò diventa quasi impossibile – come nel mito della Torre di Babele, in cui l'impossibilità di comunicare (e quindi di Inter-operare) dà luogo a un fallimento sistemico della costruzione della torre.
- Nel caso dei dati ci troviamo in una situazione simile. Il punto cruciale di un bacino di dati (o linee di codice) accessibili e utilizzabili in modo condiviso è il fatto che potenzialmente possono essere liberamente "mescolati" con dati provenienti da altre fonti anch'esse aperte.
- L'interoperabilità è la chiave per realizzare il principale vantaggio pratico dell'apertura: aumenta in modo esponenziale la possibilità di combinare diverse basi di dati, e quindi sviluppare nuovi e migliori prodotti e servizi.

Come «aprire» i dati

- *Scegliere la semplicità*. Cominciare con un progetto piccolo, semplice e veloce. Non è necessario aprire tutti i dati in una sola volta. Inizialmente va bene aprire anche un solo dataset, o anche una sua parte – naturalmente, più dati si aprono, meglio è.
  - Da ricordare che è innovazione. Muoversi il più in fretta possibile è bene, perché significa prendere slancio e imparare dall'esperienza – innovare comporta successi ed errori, e non tutte le banche dati saranno utili.
- *Coinvolgere gli utenti fin dall'inizio e coinvolgerli spesso*. Cercare presto e spesso il confronto con i potenziali utilizzatori dei dati fra cittadini, imprese o sviluppatori Ciò aumenterà la rilevanza dell'iniziativa durante tutto il suo percorso.
  - È essenziale tenere presente che gran parte dei dati non raggiungeranno gli utenti finali direttamente, ma tramite 'info-intermediari'. Queste sono le persone che prendono i dati e li trasformano o li remixano per la presentazione. Ad esempio, la maggior parte di noi non vuole o non ha bisogno di un grande database di coordinate GPS, preferiamo decisamente una mappa.
  - Così coinvolgete da subito gli info-intermediari, in modo che essi possano riutilizzare e riadattare i vostri dati.
- Affrontare i timori e le incomprensioni diffuse. Questo è importante soprattutto se lavori in o con grandi organizzazioni come le istituzioni governative. Nell'aprire i dati sorgeranno molte domande e timori. È importante (a) identificare le più rilevanti, e (b) darvi una risposta il più presto possibile.

Passi

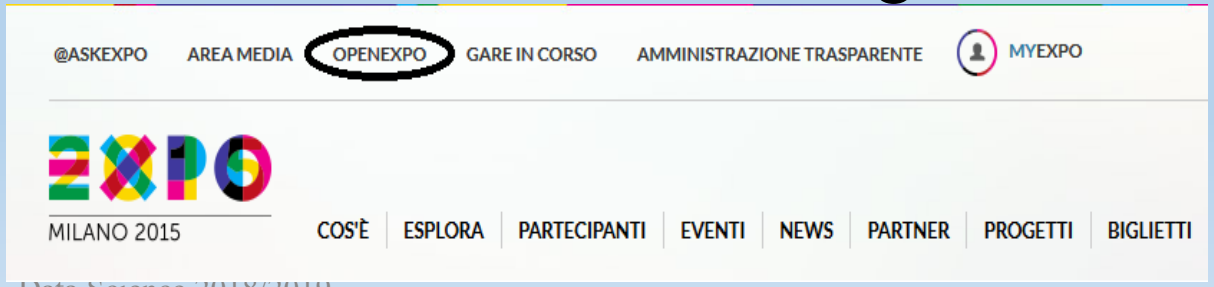
- *Scegliere i dataset.* Scegliere ciò che si intende rendere aperto, ricordando che si può (ovvero potrebbe essere necessario), rivedere questo passaggio se si incontrano problemi nelle fasi successive.
- *Utilizzare una licenza open.*
  - Determinare quali sono i diritti di proprietà intellettuale che insistono sui dati.
  - Applicare una adeguata licenza 'open' che copra tutti i diritti identificati, compatibile con la definizione di apertura discussa in precedenza
  - Se ciò non è possibile, si ritorni al punto 1 e si riprovi con una banca dati diversa.
- *Rendere i dati disponibili* - in gran quantità e in un formato utile. Si possono prendere in considerazione anche metodi alternativi come la distribuzione attraverso API.
- *Renderlo disponibile* - pubblicare sul Web e forse organizzando un catalogo centrale dove elencare l'insieme dei dati aperti.

Dove trovare i dati

- Esistono una serie di strumenti già presenti sul web che sono specificamente progettati per rendere i dati facilmente trovabili.
- Uno di quelli di maggior successo è DataHub ed è un catalogo e deposito di dati di dataset provenienti da ogni parte del mondo. Il sito rende facile, per singole persone ed organizzazioni il modo di pubblicare il materiale e agli utenti di trovare i dati che a loro servono.
- In aggiunta, ci sono decine di cataloghi specializzati per settori e luoghi differenti.
- Molte comunità scientifiche hanno creato un sistema di catalogo per i loro campi, visto che spesso è obbligatorio pubblicare i dati delle loro ricerche.

Open Data EXPO 2015

- Open Expo è un'iniziativa di Expo 2015 volta ad assicurare la totale trasparenza all'Esposizione Universale del 2015 attraverso la pubblicazione in formato aperto di tutte le informazioni riguardanti la gestione, la progettazione, l'organizzazione e lo svolgimento dell'evento.
- In questa logica OpenExpo2015.it intende arricchire le informazioni già rilasciate da Expo 2015 S.p.A. nella sezione «Amministrazione trasparente» rendendo disponibili i dati relativi alla gestione economica dell'evento (entrate e uscite, acquisti, pagamenti e relativi beneficiari), alle opere realizzate (cantieri, descrizione delle opere, importi previsti per la loro realizzazione) e alle eventuali varianti nello svolgimento della manifestazione.



Open Data EXPO 2015

- Nella sezione *Open Data* sono presenti 5 dataset.
  - *Sintesi Gare Lavori*: anagrafica delle gare per i lavori dei cantieri di Expo2015'. In questo dataset, oltre agli estremi delle gare di appalto, sono inclusi i dettagli degli inviti, la composizione delle commissioni giudicatrici, le esclusioni e gli aggiudicatari.
  - *Gare Beni e Servizi*: dettaglio delle Gare per Beni e Servizi di Expo 2015.
  - *Fornitori:* anagrafica completa dei fornitori di EXPO 2015.
  - *Dettaglio Gare Lavori:* lavori da realizzare per l'Expo 2015, distinti in lotti. Il dataset contiene i dettagli di ciascun lotto funzionale.
  - *Dati Cruscotti*: i dati riepilogativi con i relativi indicatori utilizzati nel cruscotto dei lavori di OpenExpo 2015. Il calcolo degli indicatori dipende dallo stato dei lavori.

  All'interno di ogni dataset sono presenti diversi file; ogni file è disponibile in diversi formati (CSV, XML, JSON)

Open Data EXPO 2015 - Esempio

- Questo file contiene l'elenco dei commissari delle gare per i lavori dei cantieri di Expo 2015. Vengono riportati il numero e l'oggetto della gara, il nome del componente della commissione e il ruolo ricoperto (presidente, commissario, etc).
- Estratto dalla sorgente, il file presentava il seguente aspetto:

	A	B	C	D	E	F
1	Numero ga	Oggetto de	Membro	Ruolo		
2	9	Progetto vi	Antonio Ac	PRESIDENTE		
3	9	Progetto vi	Benedetta	COMMISSARIO SUPPLENTE		
4	9	Progetto vi	Gianluca L	COMMISSARIO		
5	9	Progetto vi	Davide Am	COMMISSARIO SUPPLENTE		
6	9	Progetto vi	Donatello [	COMMISSARIO		
7	9	Progetto vi	Anna Font	SEGRETARIO		
8	10	Procedura	Ciro Mariai	-		
9	10	Procedura	Simona Mi	-		
10	10	Procedura	Donatello [	-		
11	10	Procedura	Susanna M	-		
12	10	Procedura	Angelo Pa	-		
13	11	Progetto vi	Davide Am	COMMISSARIO SUPPLENTE		
14	11	Progetto vi	Donatello [	COMMISSARIO		
15	11	Progetto vi	Anna Font	SEGRETARIO		
16	11	Progetto vi	Antonio Ac	PRESIDENTE		
17	11	Progetto vi	Benedetta	COMMISSARIO SUPPLENTE		
18	11	Progetto vi	Gianluca L	COMMISSARIO		
19	13	INTERVEN	Antonio Ac	COMMISSARIO SUPPLENTE		
20	13	INTERVEN	Gianluca L	PRESIDENTE		
21	13	INTERVEN	gGacomo	COMMISSARIO		
22	13	INTERVEN	Davide Am	COMMISSARIO SUPPLENTE		
23	13	INTERVEN	Susanna M	COMMISSARIO		
24	14	realizzazio	Gianluca L	COMMISSARIO SUPPLENTE		
25	14	realizzazio	Anna Font	SEGRETARIO		
26	14	realizzazio	Antonio Ac	PRESIDENTE		
27	14	realizzazio	Donatello [	COMMISSARIO		
28	14	realizzazio	Cristina Mi	COMMISSARIO		

Open Data EXPO 2015 - Esempio

- Il file è composto da 49 tuple (48 relative alle gare, più una tupla di intestazione). Le voci:
  - Numero gara (colonna A);
  - Oggetto della gara (colonna B);
  - Membro (colonna C);

  sono risultate già «pronte», e non hanno necessitato di alcuna pulizia e/o modifica.
- L'unica operazione ad esse apportate è stata l'allargamento della colonna di appartenenza (per avere una visuale più ampia e buona dei dati).
- L'unica voce alla quale è stata apportata una modifica è *Ruolo* (colonna D). Tale modifica è consistita nel sostituire il simbolo "-" (indicante che non si era a conoscenza del ruolo ricoperto dal membro in questione), con la stringa "non disponibile".
- E' stato scelto di fare ciò per una questione di stile e di migliore presentabilità.
- Le stesse modifiche sono state apportate anche ai file "Registro delle imprese invitate alle gare" e "Registro delle imprese che hanno partecipato alle gare".
- Il file "Registro delle imprese escluse dalle gare" non è stato preso in considerazione perché, per nessuna delle aziende presenti in elenco era riportato il motivo dell'esclusione