

## OPEN DATA EXPLORER: UN'APPLICAZIONE PER ORIENTARSI NEL MONDO DEI DATI APERTI

Salvatore IIRITANO<sup>1</sup>, Sara LAURITA<sup>2</sup>, Mariagrazia ZOTTOLI<sup>3</sup>

### SOMMARIO

Il fenomeno del *data deluge*, generato dall'incremento massivo dei dispositivi connessi alla rete, ha condotto ad un aumento della quantità dei dati prodotti, della velocità del proprio aggiornamento e della varietà di fonti che li generano. In questo contesto è andata sviluppandosi la corrente che prende il nome di *data revolution* il cui punto cardine è rappresentato dalla condivisione della conoscenza (*shared data*) attraverso il paradigma degli *open data*.

In tale movimento si collocano le iniziative della Pubblica Amministrazione che hanno portato negli ultimi anni a una rilevante produzione di dati pubblici "aperti" (*open government data*). Questo processo di "apertura" si accompagna a una serie di criticità che fanno capo alla qualità del dato, alla possibilità di riuso e all'*information retrieval*. In Italia l'iniziativa più importante di apertura dei dati pubblici è rappresentata dal portale *dati.gov.it*.

*Open Data Explorer - ODE*, interfacciandosi con tale piattaforma, fornisce uno strumento a coloro che si avvicinano al complesso mondo dei dati pubblici aperti, per selezionare dataset di qualità con standard ben definiti che ne permettono un riuso statistico.

---

<sup>1</sup> Exeura s.r.l., via Pedro Alvares Cabrai, C.da Lecco, 87036, Rende (CS), e-mail: salvatore.iiritano@exeura.eu

<sup>2</sup> Università della Calabria, via Pietro Bucci, 87036, Campus di Arcavacata di Rende (CS), e-mail: slaurita@unical.it.

<sup>3</sup> Contesti s.r.l., via Della Resistenza, 23, 87036, Rende (CS), e-mail: mgr.zottoli@gmail.com.

## 1. Introduzione<sup>4</sup>

Il fenomeno del *data deluge*, generato dall'incremento massivo dei dispositivi connessi alla rete, ha condotto ad un aumento della quantità dei dati prodotti, della velocità del proprio aggiornamento e della varietà di fonti che li generano. In questo contesto è andata sviluppandosi la corrente che prende il nome di *data revolution* il cui punto cardine è rappresentato dalla condivisione della conoscenza, si parla, a questo proposito, di *shared data*. I dati condivisi possono essere di diversa natura: *big data*, se in numero tale da non poter essere gestiti dai tradizionali strumenti di analisi e archiviazione, *linked data*, se collegati tra di essi, ed *open data*. Con tale locuzione si intendono i dati pubblicati da fonti eterogenee (tra le quali trovano spazio anche quelle della statistica ufficiale) in un formato che ne consente il riuso per scopi di interesse da parte dell'utente.

La centralità del concetto di apertura del dato, trova la propria realizzazione nella *data liberation*. In questo movimento si collocano anche le iniziative di trasparenza della Pubblica Amministrazione che hanno portato negli ultimi anni ad una rilevante produzione di dati pubblici "aperti" (*open government data*). Tale produzione, soprattutto in Italia, soffre dei seguenti problemi:

- scarsa visione sistemica (Agenzia per l'Italia Digitale, 2014);
- elevata autonomia delle PPAA con riferimento alle modalità di produzione dei dati e di gestione dei processi amministrativi, con conseguente produzione di vere e proprie "isole" di informazioni;
- carenza di qualità nei dati pubblicati: tipicamente, le Pubbliche Amministrazioni che rendono disponibili le informazioni non seguono linee guida per certificare le tecniche e/o i metodi per il reperimento e la valutazione dei dati;
- disponibilità di uno strumento per la ricerca ed il reperimento di *open data* (dal portale *dati.gov.it*) che non supporta adeguatamente gli utenti ai fini dell'individuazione di tutti e soli i dataset utili per le analisi.

La proliferazione "incontrollata" degli *open government data* ha generato inoltre una sorta di giungla dei dati, tanto da portarci a parlare di "*open data jungle*", nella quale è difficile districarsi. Ci troviamo di fronte ad un'immensa risorsa non ancora sfruttata appieno che solo in parte è riuscita a generare i benefici economici e sociali connessi a questo nuovo paradigma.

Il lavoro di ricerca, partendo dall'analisi del portale nazionale dei dati aperti della PA - *dati.gov.it*, affronta operativamente la questione del riuso dei dati concentrandosi sullo sviluppo di una piattaforma, denominata *Open Data Explorer - ODE*, che mette a disposizione funzionalità di ricerca e condivisione di *open government data*, supportando gli utenti esperti nel processo di selezione e classificazione dei dataset in base ai loro contenuti e alle principali dimensioni della qualità statistica.

Il sistema *ODE*, basato su una tecnologia web, consente, interfacciandosi con la piattaforma *dati.gov.it*, il reperimento rapido e semplice tutti i dataset in formato aperto da (ri)utilizzare per scopi statistici. Le principali funzionalità di *ODE* permettono, per i dataset individuati:

- a) la visualizzazione delle correlazioni;
- b) la visualizzazione di statistiche sul livello di qualità dei dati;
- c) la possibilità di condividere la ricerca effettuata e/o un sottoinsieme di essi.

---

<sup>4</sup> Il paper è frutto del lavoro in corso di realizzazione nell'ambito progetto di ricerca "Sinse+. Sistema di reasoning e tutoring su grandi moli di dati" finanziato dalla Regione Calabria nell'ambito dell'Agenda Strategica del Polo di Innovazione ICT Calabria - POR Calabria FESR 2007/2013, Asse I, Linea di Intervento 1.1.1.2

## 2. La rivoluzione dei dati

L'avvento di Internet ha rivoluzionato a livello mondiale il concetto di produzione dei dati ed ha avuto un notevole impatto anche sulla qualità degli stessi. Si pensi che nel 2003 il numero di dispositivi pro capite connessi alla rete era meno di uno, nel 2010 si è passati a 1,84 dispositivi per persona, per arrivare a più di tre nel 2015 (Evans, 2011).

Partendo dal presupposto che ogni apparecchio collegato ad Internet genera dati, negli ultimi anni si è assistito ad una crescita esponenziale della produzione di informazione grezza. In un sistema in cui ogni fruitore diventa potenziale produttore di informazioni, le stime in merito alla quantità di dati prodotti ogni giorno da fonti di natura differente quali, ad esempio, sensori, social network e rilevatori GPS, denotano delle cifre davvero impressionanti (Scannapieco, et al., 2013). Tali numeri sono destinati a crescere ulteriormente tenendo conto dell'avanzamento di fenomeni come il *crowdsourcing* e l'*attivismo digitale*, che rappresentano l'ultima frontiera dello scambio di dati tra i soggetti in rete. A ciò si deve aggiungere che Internet, partendo da una prima evoluzione ad *agorà* di discussione e condivisione, è oggi diventata un ambiente fatto di oggetti in grado di generare, raccogliere e distribuire dati; si parla in tal senso di *Internet Of Things* (IoT), la cui evoluzione è rappresentata dall'*Internet of Everything* (IoE), da intendersi come un *ecosistema* in grado di mettere in relazione i dati raccolti dai vari oggetti.

Il passaggio da un deserto a un diluvio di dati di varia natura, puntuali e relazionati tra di essi, disponibili ad un costo relativamente basso e in formato sempre più aperto ed accessibile, dà possibilità senza precedenti di informare e trasformare la società in un'ottica di sostenibilità ambientale (Independent Expert Advisory Group, 2014). I governi, le aziende, i ricercatori e i gruppi di cittadini organizzati in movimenti o in libere associazioni, sono impegnati con fervore nella sperimentazione, nell'innovazione e nella ricerca di nuove forme di adattamento a questo nuovo mondo di dati in cui la quantità di dati disponibili è sempre più smisurata, più veloce e più dettagliata, richiedendo avidamente, e mai appagati, ancora più dati.

L'enorme quantità di dati disponibili può determinare, però, il cosiddetto fenomeno dell'*information overloading*, ovvero della sovrabbondanza di informazioni, spesso non utili ai fini dell'analisi da eseguire, che di fatto è un elemento ostativo per chi deve prendere decisioni. Avere un "eccesso" di dati implica tempi di elaborazione esorbitanti e necessità di analizzare e scartare molta informazione ridondante o inutile. Tale fenomeno può essere attenuato solo tramite opportuni strumenti di *Information Retrieval* che consentano di selezionare facilmente tutta e sola l'informazione di valore.

Nella *open data jungle* diventa essenziale giungere alla definizione di una misura della qualità dei dati che consenta nei fatti di massimizzare il valore intrinseco del patrimonio informativo. È del tutto evidente che un'analisi basata su indicatori non calcolati correttamente avrà il rischio di essere imprecisa se non addirittura deleteria per l'organizzazione che ne subisce l'impatto.

Valutare il livello qualitativo dell'informazione grezza disponibile, non è però un'operazione banale. Le tecniche ed i metodi utilizzati tradizionalmente dalla statistica ufficiale nell'ambito del processo di rilevazione, elaborazione e diffusione dei dati, sebbene costituiscano una garanzia in questo senso, non sono sempre applicabili alle "nuove" tipologie di dati; si rivela dunque indispensabile un'evoluzione del paradigma alla base della produzione statistica ufficiale, al fine di estendere il marchio di qualità anche a queste fonti.

Si aggiunga, inoltre, che quando si parla di qualità dei dati si fa riferimento anche all'accuratezza o alla presenza della documentazione inerente la generazione e le fonti informative. Senza tali informazioni diventa difficile riutilizzare le informazioni per finalità decisionali.

Nel loro rapporto "*A world that counts*" il gruppo di esperti dell'ONU, costituito per valutare le opportunità derivanti dall'innovazione, dai progressi tecnologici e dall'esplosione del numero di produttori di dati pubblici e privati, per dare indicazioni sulle possibili forme di evoluzione dei sistemi convenzionali di produzione statistica e per proporre soluzioni per rafforzare i processi di responsabilizzazione dei governi ai vari livelli, indica con chiarezza che la vera essenza della *data revolution* è data dall'essere in grado di capire

il valore che risiede in tali enormi quantità di dati che vengono generati quotidianamente e di trovare nuove modalità per impiegare tali dati nel migliorare effettivamente la qualità della vita delle persone (ONU, 2014).

In questo contesto, si apre un ampio spazio allo sviluppo e utilizzo di soluzioni tecniche e metodologiche in grado di migliorare i sistemi di raccolta dati negli ambiti che presentano attualmente criticità (ad esempio, nel settore dell'ambiente, della condizione femminile, della povertà e ineguaglianza), di introdurre sistemi affidabili per la misurazione e valutazione delle performance e di aumentare la diffusione di dati di qualità in formato aperto.

### 2.1. *Tecniche di sopravvivenza nella data jungle*

Per affrontare e cercare di risolvere ognuna delle problematiche inerenti l'esplorazione della *data jungle* sono state individuate alcune tecniche e metodologie statistiche che rappresentano un primo set di strumenti di orientamento a disposizione degli addetti ai lavori.

Un primo strumento è l'utilizzo delle tecniche di *information retrieval* (IR) che permettono di gestire la rappresentazione, la memorizzazione, l'organizzazione e l'accesso ad oggetti contenenti informazioni quali documenti, pagine web, cataloghi online e oggetti multimediali. L'IR è un campo interdisciplinare che nasce dall'incrocio di discipline diverse e coinvolge la psicologia cognitiva, l'architettura informativa, la filosofia, il design, il comportamento umano sull'informazione, la linguistica, la semiotica, la scienza dell'informazione e l'informatica.

Nell'ambito della ricerca di dataset *opengov* gli strumenti di IR sono in grado di mettere a disposizione un'*indicizzazione multidimensionale* che consente il reperimento delle informazioni sulla base delle caratteristiche dei dataset, ma anche dei metadati associati e dei *tag* descrittivi.

Per affrontare il tema della categorizzazione dei dataset si è fatto ricorso nel tempo a due diversi approcci.

Il primo fa riferimento alla standardizzazione del formato degli output delle informazioni prodotte. Le esigenze di distribuzione e condivisione degli *open data* richiedono, infatti, che i dati siano disponibili per consentire a chiunque di effettuare delle proprie rielaborazioni. Per lungo tempo si è affrontato il problema etico sul rendere fruibili i dataset frutto delle rilevazioni ufficiali, domandandosi, ad esempio, quanto fosse legittimo che i database di un istituto di statistica fossero scaricati da un'organizzazione privata che li riorganizzasse in maniera più accattivante e li cedesse a pagamento (Speroni, 2011). Sull'accesso alle statistiche ufficiali sono stati compiuti passi in avanti fondamentali: la maggior parte dei dati ISTAT è gratuita e liberamente scaricabile, così come accade per le collezioni EUROSTAT e OECD. L'aumento della quantità dei dati resi disponibili e delle relazioni tra gli stessi, nonché degli standard per uniformarli, ha condotto alla definizione di un Web dei dati, noto anche come Web semantico, un'evoluzione del tradizionale web dei documenti in una rete di informazioni collegate tra di esse e corredate di ulteriori informazioni che consentono di collocarle in un determinato contesto.

L'ulteriore approccio si basa sulla definizione di un sorta di "rank" per misurare la qualità dei dati trattati. La qualità dei dati si traduce, di fatto, nella possibilità di aggregarli per gli scopi di interesse per l'utente: si parla in questi casi di *fitness of use*. Tim Berners-Lee, ideatore del web e pioniere nella campagna per la sensibilizzazione sul tema dei dati aperti, ha provveduto a definire un sistema di ranking basato sul formato di pubblicazione dei dati che potesse guidare gli utenti nel processo di riutilizzo. Il metodo consiste nell'attribuzione di un punteggio (da una a cinque stelle) a partire dai dataset nel formato caratterizzato dalla minore possibilità di riuso, quali, ad esempio i file in formato PDF, fino ad arrivare a quelli di tipo LOD, acronimo che sta per *Linked Open Data*, e descrive l'evoluzione del tradizionale concetto di dato, a seguito dello sviluppo dell'*Internet of Things* e parallelamente delle tecnologie del *Web semantico*. Tale formato di pubblicazione consente di collegare dati appartenenti a fonti diverse, espressi secondo standard *machine readable*: attualmente si contano circa 188 milioni di fatti rappresentati sottoforma di triple RDF (*Resource Description Framework*). Molti autori si sono posti nella condizione di definire delle metodologie che consentissero la valutazione della qualità dei dati *open* anche in formato *linked*, trovandosi però ad affrontare

delle sfide piuttosto ardue. La valutazione della qualità si basa, essenzialmente sulla definizione di euristiche per la misurazione di dimensioni che siano di interesse per l'utilizzatore.

Gli studi condotti durante le attività di ricerca hanno fatto emergere che ciò che effettivamente rende un dato riusabile è il fatto di disporre di *metadati* che consentano di valutare la qualità dell'informazione.

Secondo la definizione fornita dalla *National Information Standard Organization*, per *metadato* si intende "un'informazione strutturata che consente di descrivere, spiegare e, in alcuni casi, rendere più semplice il recupero dei dati". I metadati sono "ciò che rende i dati utili", come afferma Francis Bretherton, in quanto consentono all'utente di definire l'adattabilità dei dati al proprio caso d'uso e contribuiscono alla definizione della qualità, indubbiamente scarsa se vi è carenza di dati esplicativi a corredo (Goodchild, et al., 2002). Nel 1996 in una riunione di bibliotecari e archivisti riunitisi nella città di Dublin, in Ohio, si è giunti alla definizione di un'architettura di metadati che potesse adattarsi alle esigenze di venditori e produttori di informazioni. Il set di elementi di base da rendere disponibili nella pubblicazione dei dataset è pari a 15: a partire dall'autore del dato e dal titolo del dataset, fino ad arrivare alla data di pubblicazione/aggiornamento e alla licenza a corredo.

Questo set di elementi non basta, però, per determinare la bontà dei dati aperti ma è necessario definire metriche attraverso le quali misurare la qualità a priori evitando ulteriori dispendiosi approfondimenti e valutazioni tecniche. A parere degli autori, bisognerebbe introdurre una classificazione basata su un ranking costruito a partire dalle seguenti dimensioni: a) accuratezza, b) tempestività e puntualità; c) regolarità; d) chiarezza; e) comparabilità.

### 3. *L'open government data*

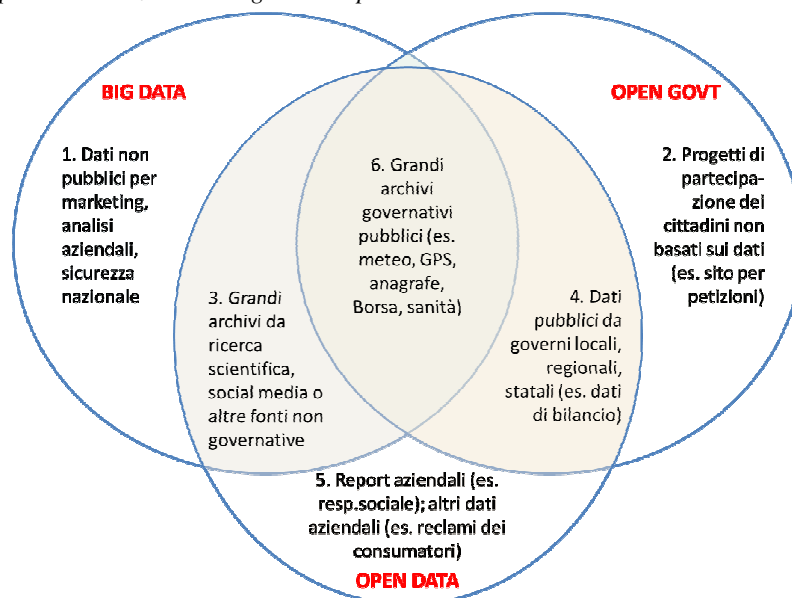
Una delle principali sfide che i governi odierni si trovano ad affrontare è quella sulla trasparenza che, associata alla risonanza avuta dal fenomeno della *data liberation* e conseguentemente degli *open data*, ha posto le pubbliche amministrazioni nella condizione non solo di imparare a gestire i dati, ma anche a presentarli in modo efficace e convincente, mantenendo un particolare riguardo alla tutela della sicurezza.

Il fenomeno dell'*Open government* (*Opengov*) è esploso a seguito della campagna elettorale dell'attuale presidente degli Stati Uniti d'America, Barack Obama. Tale corrente promuove l'apertura e l'accessibilità di tutte le attività riguardanti la *res pubblica*, al fine di favorire azioni efficaci volte a garantire il controllo sull'operato di governi e Pubbliche Amministrazioni. Si tratta di una nuova idea di democrazia associata ad una più attiva partecipazione dei cittadini che è venuta progressivamente a integrarsi con quella di creare un governo elettronico, in cui l'aggettivo *open* diventa indicatore di un mutamento nella complessa relazione tra cittadini, amministrazioni pubbliche e sistema politico, nella prospettiva di una completa integrazione (Di Donato, 2010).

I vantaggi derivanti dall'*Open Government* sono diversi e ricoprono differenti ambiti di valutazione. Da un punto di vista economico, svariati studi hanno portato alla stima del valore dei dati aperti in diverse decine di miliardi di euro ogni anno. Per quanto riguarda l'Europa, basti pensare al solo guadagno in termini di efficienza del lavoro. Nel sociale consentono di migliorare la qualità della vita dei cittadini che li utilizzano, supportandoli nel prendere decisioni sul privato e rendendoli più attivi nell'ambito della società civile: sostanzialmente un cittadino più informato è un cittadino più competente (Daprà, 2013).

Joel Gurin, utilizzando i diagrammi di Venn (Figura 1) ha mappato la relazione tra *big data* e *open data* e il modo in cui essi sono relazionati al più ampio concetto di *open government*. In particolare, gli *open government data* sono quei dati pubblici derivanti dai grandi archivi governativi e quelli prodotti dai governi locali, regionali e statali nell'ambito delle proprie funzioni (es. dati di bilancio).

Figura 1 - La mappa delle relazioni tra big data e open data



Fonte: Gurin, 2014.

Da una certa prospettiva la normativa del nostro Paese in tema di *open government data* è tra le più avanzate a livello europeo. Nel dicembre del 2012 con la pubblicazione del *decreto Crescita 2.0* è stato introdotto, facendo riferimento agli articoli 52 e 68 del C.A.D., il principio di "*open by default*" secondo il quale dati e documenti pubblicati dalle pubbliche amministrazioni senza la specifica indicazione di una licenza proprietaria si intendono automaticamente rilasciati con un taglio *open*. A gennaio 2013, l'art. 9 del D.L. 18 ottobre 2012, n. 179 (modificato dalla legge di conversione 17 dicembre 2012, n. 221) ha sostituito l'art. 52 del D.Lgs. 7 marzo 2005 n. 82 (C.A.D.) facendo degli *open data* un "obbligo di legge" nel rispetto della privacy, del segreto statistico, del diritto di autore, etc. Da ultimo, la Presidenza del Consiglio dei Ministri ha presentato il 3 marzo 2015 la *Strategia per la Crescita Digitale* conferendo all'Agenzia per l'Italia Digitale il compito di promuovere le politiche nazionali per la valorizzazione del patrimonio informativo pubblico e di indirizzare le amministrazioni verso un processo di produzione e rilascio dei dati standardizzato e interoperabile su scala nazionale e internazionale.

L'effettiva applicazione delle norme in termini di qualità e riutilizzo risulta però ancora piuttosto frammentata, con poche realtà virtuose e tante ancora molto lontane dall'obiettivo, quasi ad uno stadio volontaristico e artigianale (Iacono, 2014). Ad esempio, ad oggi sul portale dei dati aperti della PA italiana (*dati.gov.it*), sono online più di diecimila dataset, ma la maggior parte è pubblicata con un corredo di informazione che non ne consente la piena comprensione e quindi un immediato riuso.

L'impasse attuale è legata, da un lato, ad una serie di barriere che le pubbliche amministrazioni si trovano ad affrontare, a partire dalla riluttanza dei dipendenti pubblici (Stentella, 2015) fino ad arrivare alla presenza di sistemi informativi non sempre adeguati. Dall'altro, manca una attenzione al rilascio di informazioni dettagliate sul processo di generazione dei dati con la conseguenza di mancanza di qualità statistica.

#### 4. Open Data Explorer - ODE

L'idea di sviluppare un *search engine*, in grado di trovare semplicemente ed efficacemente tutti i dataset *open* che possono essere utilizzati a supporto di un'analisi statistica per un dato dominio applicativo, nasce dalla constatazione che il Portale dei dati aperti italiani *dati.gov.it* non consente:

- la ricerca multidimensionale dei dataset (es. in base ai metadati e agli attributi);

- la correlazione tra i dataset, secondo logiche definite dall'utente (es. due dataset sono correlati se hanno almeno un tag in comune), finalizzata ad individuare dati che ragionevolmente potrebbero essere analizzati insieme;
- la valutazione della qualità dei dataset in termini di riusabilità statistica, tenendo conto di un sistema di ranking;
- la condivisione di dataset e ricerche tra utenti, secondo una logica mutuata dai social network.

Questi deficit hanno come prima conseguenza quella di rallentare il processo di individuazione dei dataset utili e di non consentire un efficace riutilizzo dei dati.

L'attività di ricerca è stata diretta, da un lato, allo sviluppo di metriche di valutazione della qualità e dall'altro a migliorare gli algoritmi di selezione dei data set attraverso lo sviluppo di un'architettura informatica.

#### 4.1. *La valutazione del rank di qualità*

Il meccanismo di valutazione del rank di qualità utilizzato nella piattaforma *ODE* tiene conto di due aspetti: l'autorevolezza della fonte e la metadattazione dei dataset.

Con riferimento alla fonte, si differenziano le sorgenti di dati in base alla propria appartenenza o meno al mondo della statistica ufficiale. Le rilevazioni degli enti del SISTAN (Sistema Statistico Nazionale) che entrano a far parte del PSN (Programma Statistico Nazionale), sono effettuate con rigore metodologico e corredate di adeguata documentazione, ragion per cui in *ODE* rappresentano le fonti di maggior affidabilità.

Per quanto concerne i dataset che non rientrano nelle indagini del SISTAT, l'analisi è più complessa e deve essere svolta a più livelli di profondità. La sola valutazione della fonte non è sufficiente per definire l'affidabilità del dato pubblicato; diventa dunque necessario affiancarla ad altri indicatori, quali le misure legate alla valutazione delle caratteristiche del dataset.

In questo caso, il sistema permette con riferimento alla metadattazione di verificare:

- quali e quanti metadati del set "Dublin Core" sono stati utilizzati per descrivere il dataset;
- la presenza delle "macrodimensioni" Linked Open Data - LOD (Auer, et al., 2012):
  - accessibilità, che racchiude tutti gli aspetti legati all'accesso, all'attestazione di autenticità e alla possibilità di recuperare i dataset;
  - caratteristiche intrinseche, completamente indipendenti rispetto al contesto nel quale opera l'utente, quali validità sintattica, accuratezza semantica, consistenza, precisione e completezza;
  - caratteristiche relative al contesto del task da svolgere, quali pertinenza, attendibilità, comprensibilità, aggiornamento costante;
  - dimensioni rappresentative, che consentono di catturare degli aspetti relativi alla definizione dello schema dei dati, vale a dire precisione nella rappresentazione, interoperabilità, interpretabilità e versatilità;
- le caratteristiche peculiari del dataset, quali, ad esempio, il numero di colonne e di righe che risultano essere valorizzate, il numero di valori mancanti, una serie di misure statistiche (media, mediana, primo quartile e terzo quartile, se calcolabili) e un indice sintetico che consenta di valutare la sparsità nel dataset.

#### 4.2. *L'architettura logico-funzionale*

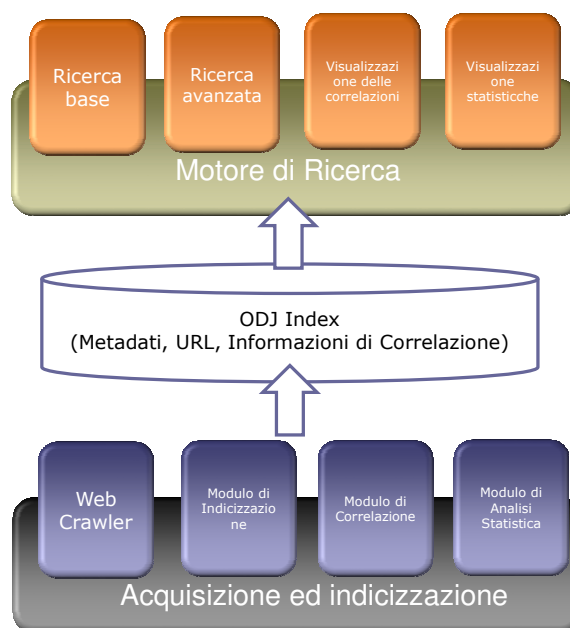
L'architettura logico-funzionale di *ODE*, illustrata in

Figura 2, è strutturata su tre livelli funzionali:

- I. il livello di acquisizione ed indicizzazione, nell'ambito del quale trovano collocazione:

- il *Web Crawler*, un agente software in grado di «simulare» il comportamento di un utente che apre alcune pagine web. Tramite opportune «API» tale agente sarà in grado di:
    - leggere il contenuto delle pagine web (formato HTML);
    - estrarre le informazioni di interesse;
  - il modulo di *indicizzazione*, che provvede a salvare le informazioni estratte nell'ambito di un database relazionale;
  - il modulo di *correlazione*, che consente di stabilire se due dataset sono collegati. Nell'ambito del progetto verranno individuati i criteri per stabilire la correlazione tra due dataset, ma, al minimo, due dataset saranno considerati correlati se condivideranno: almeno un tag; almeno un metadato; almeno un attributo;
  - il modulo di *analisi statistiche*, che consentirà di individuare, per ogni dataset, le seguenti informazioni:
    - dimensione;
    - tempo necessario per il download;
    - numero di righe;
    - numero di attributi;
    - per ogni attributo: massimo, minimo, media, mediana (se calcolabili); primo e terzo quartile (se calcolabili); moda (se calcolabile); numero di righe con valori NULL o MISSING;
- II. il livello dell'indice delle informazioni, costituito da un database relazionale che conterrà tutte le informazioni estratte;
- III. il livello della ricerca, in cui troveranno collocazione le interfacce per la ricerca e la visualizzazione delle informazioni inerenti i differenti dataset.

Figura 2 - Architettura piattaforma ODE

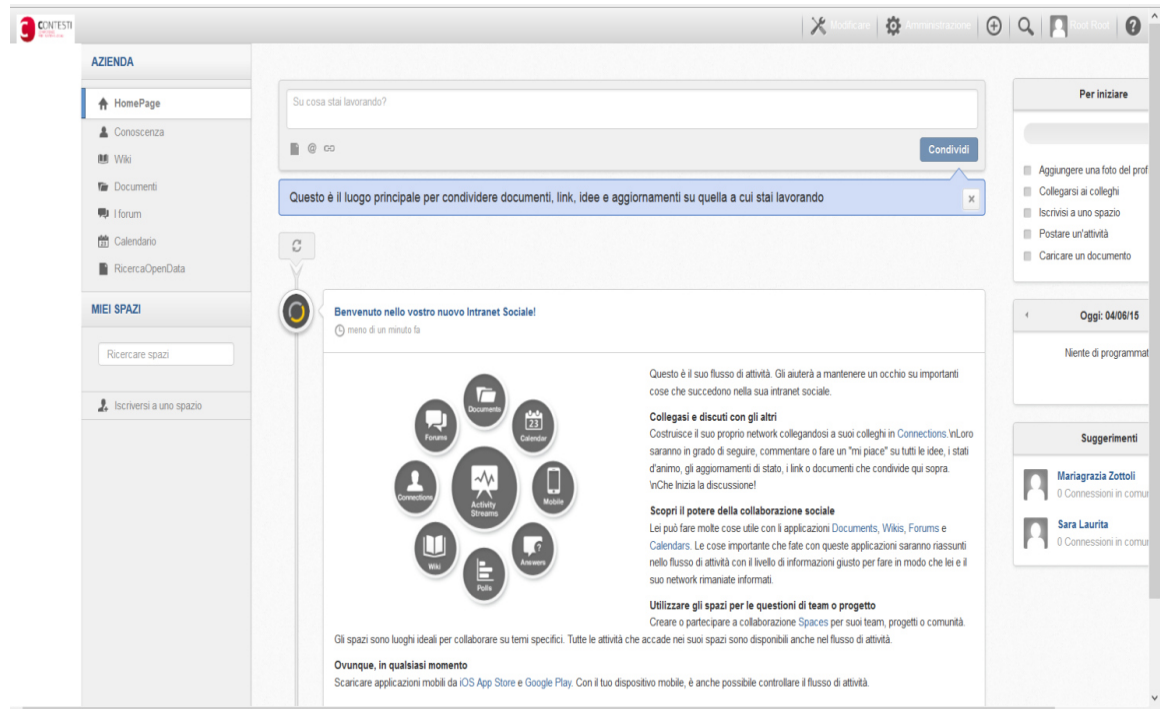




#### 4.3. Il sistema in funzione

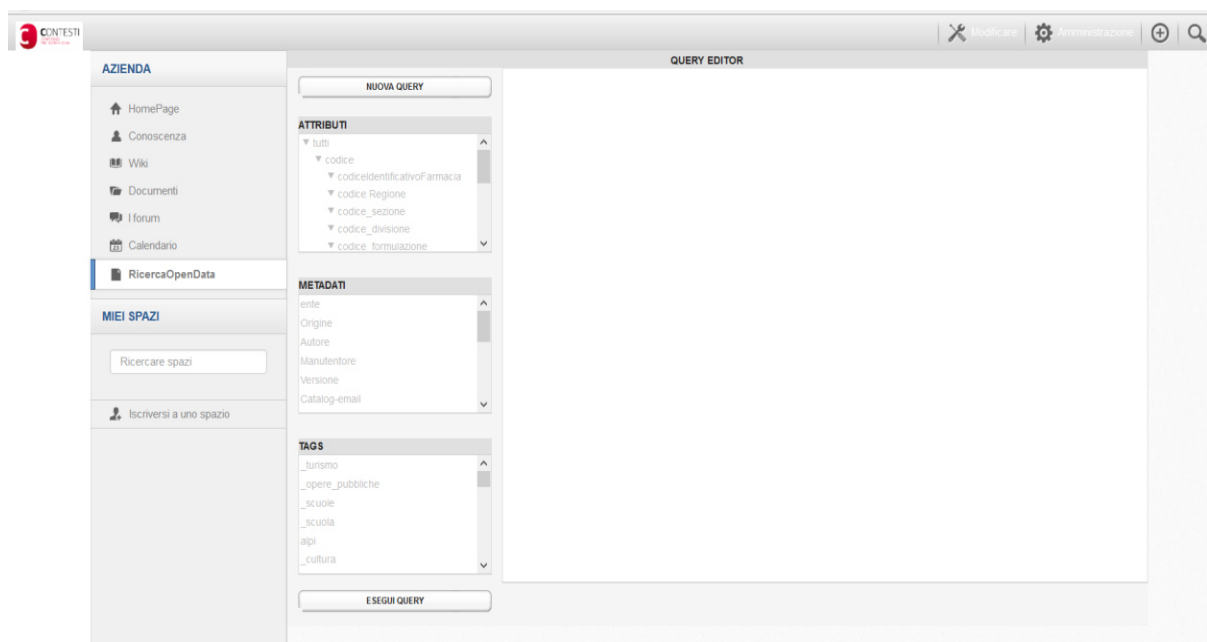
L'accesso dell'utente al portale *ODE* consente di utilizzare una serie di funzionalità tipiche dei *social network*, quali la bacheca personale, uno spazio forum, uno spazio wiki ed un sistema di archiviazione documentale (Figura 3).

Figura 3 - Schermata iniziale della piattaforma Open Data Explorer



La funzionalità “Ricerca Open Data” consente di accedere al motore di ricerca (Figura 4).

Figura 4 - Schermata del motore di ricerca



L'utente ha la possibilità di effettuare ricerche utilizzando, quali fattori di filtro, gli *attributi* del dataset, i *metadati* ed i *tag*. Le tre dimensioni possono essere utilizzate alternativamente, oppure combinando variamente gli elementi in AND o in OR; è inoltre possibile specificare meccanismi di ricerca esatti o tramite l'operatore LIKE.

Nella figura seguente si mostra la ricerca di tutti i dataset che contengono l'attributo "codice\_formulazione", sono emessi dall'ente "Comune di Milano" ed hanno il tag "turismo". Alla pressione del bottone "esegui query" il sistema mostra l'elenco dei dataset che soddisfano i criteri di ricerca (Figura 5 e 6).

Figura 5 - Esempio di formulazione di una query

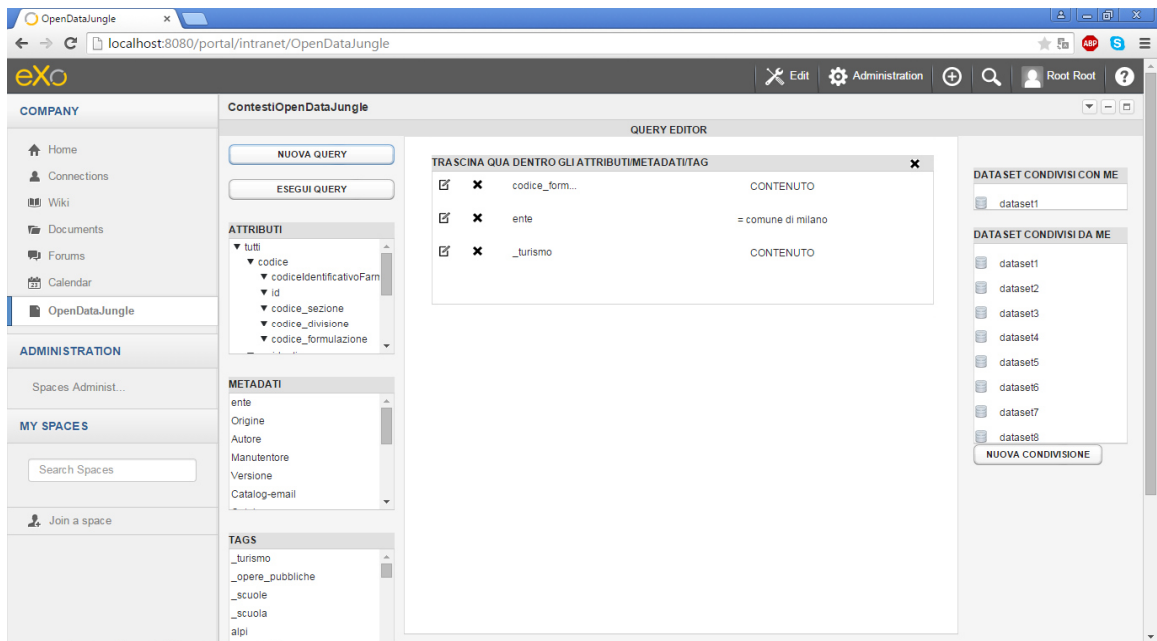
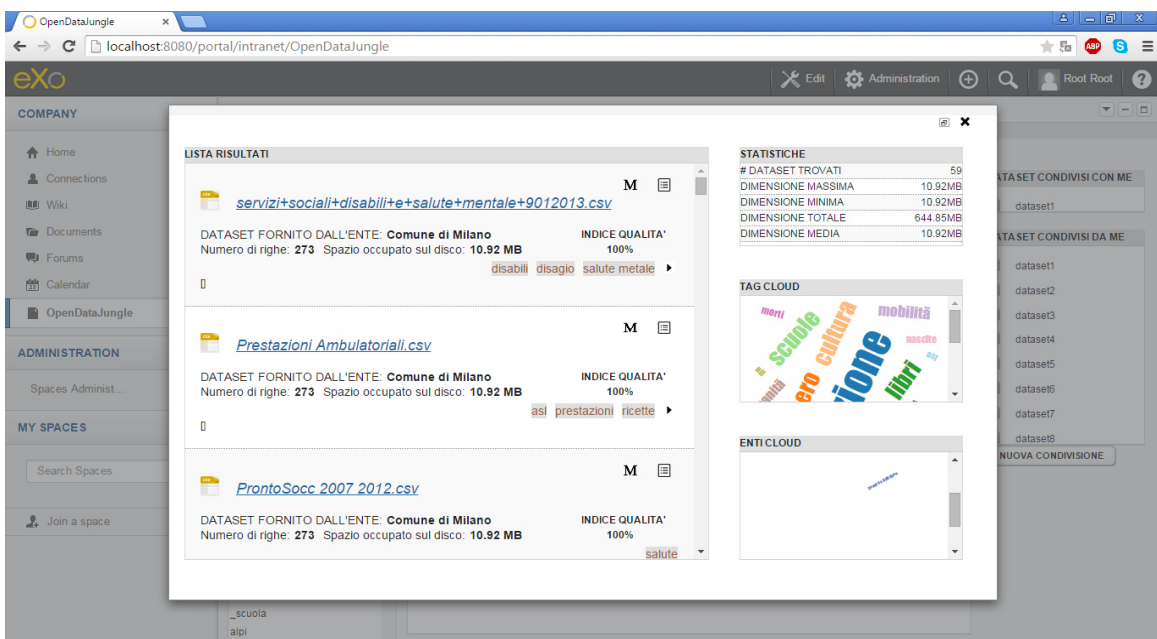


Figura 6 - Esecuzione della query e lista dei risultati



Per ogni dataset sono mostrate le caratteristiche principali (nome, dimensione, tipologia, descrizione) nonché la lista dei tag; il sistema, inoltre, rende disponibile una statistica dell'intero insieme di dataset selezionati (numero di dataset, dimensione totale, dimensione in termini di numero di tuple).

Cliccando sul pulsante “statistiche” è inoltre possibile visualizzare statistiche di dettaglio per il dataset (numero di tuple, dettaglio degli attributi). L'utente può accedere alla pagina in cui il dataset è pubblicato, ed inoltre può decidere di condividere una lista di dataset ad altri utenti della rete sociale (Figura 7).

L'interfaccia di ricerca fornisce infine un tag cloud dei tag inerenti i dataset selezionati (Figura 8). Cliccando su uno dei tag è possibile raffinare la ricerca precedentemente impostata (Figura 9).

Figura 7 - Misure statistiche di sintesi dei dataset

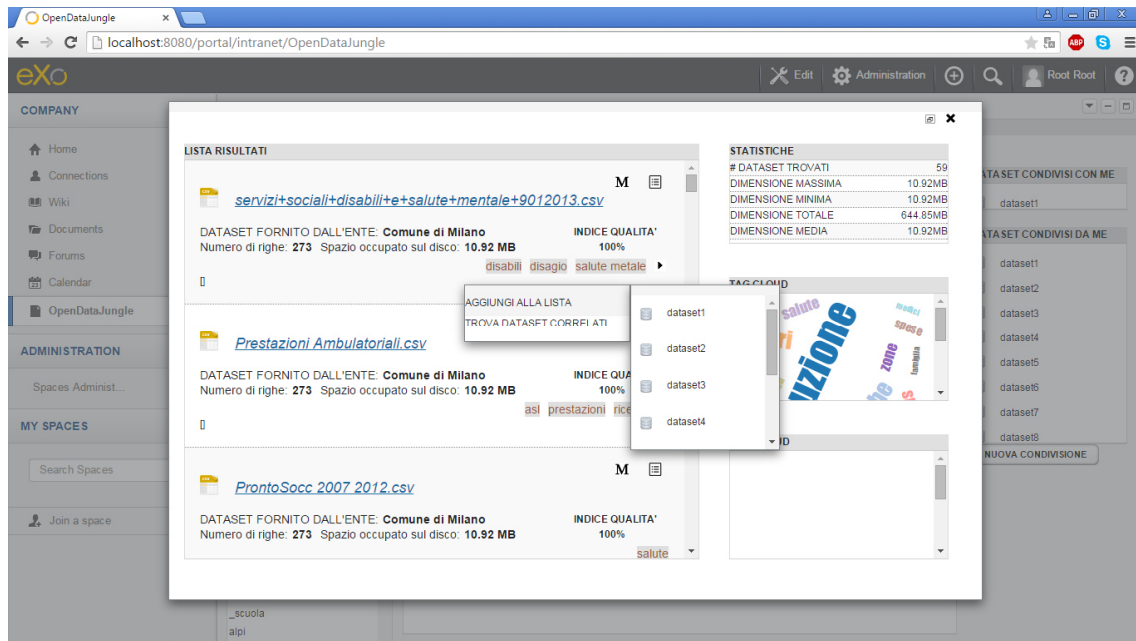


Figura 8 - Condivisione dei dataset di interesse

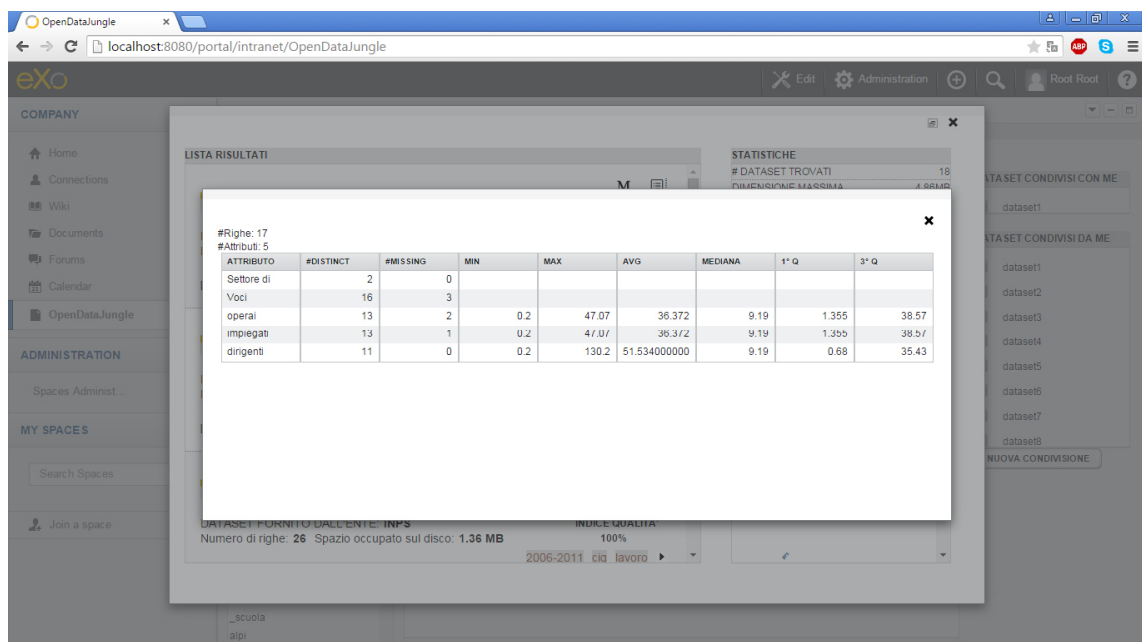
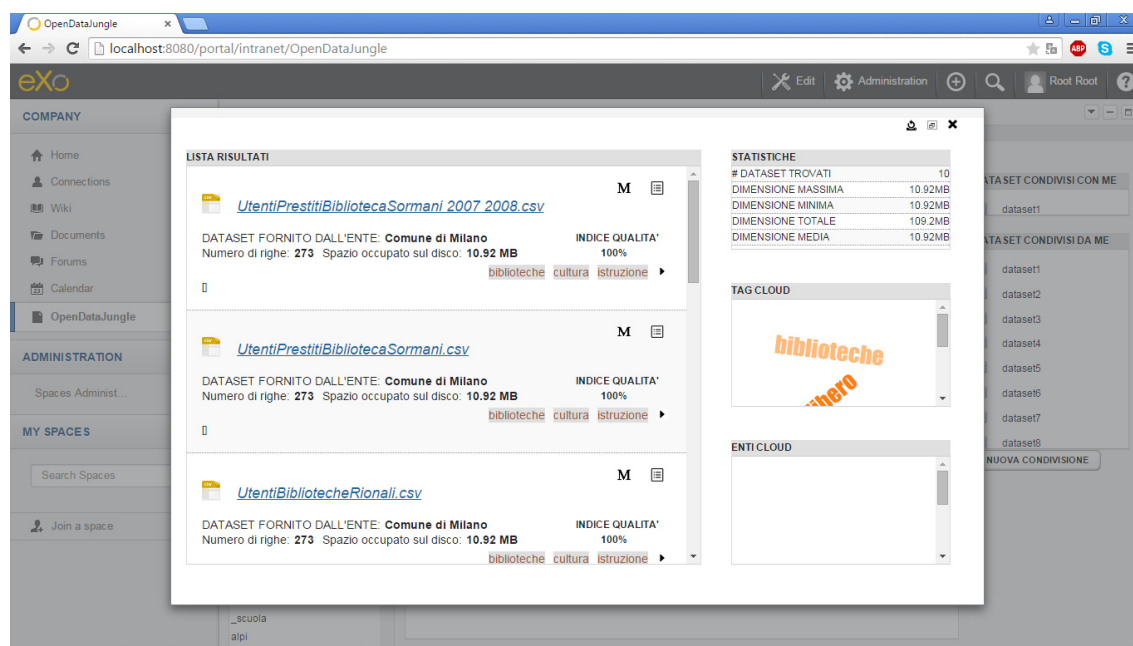


Figura 9 - Raffinamento della ricerca mediante tagcloud



## 5. Conclusioni

Alla mole di dati in continua crescita, non è corrisposta una altrettanto rapida evoluzione degli strumenti tecnologici a disposizione per l'individuazione e l'analisi di tali dati, così che si è venuto a creare un gap tra l'informazione che abbiamo a disposizione e la capacità di riuscire a sfruttarla pienamente.

Il sistema *Open Data Explorer* mette a disposizione una serie di funzionalità innovative per supportare una ricerca efficace e veloce dei dataset pubblicati in *dati.gov.it*. Il sistema è stato rilasciato in versione *beta*, al fine della valutazione dell'efficacia da parte degli utenti.

La *roadmap* prevede l'implementazione di tutte le misure per la valutazione della qualità degli *open data*, con particolare riguardo alle caratteristiche dei *Linked Open Data* - LOD, e l'applicazione ad altri siti che pubblicano dataset in formato open.

Nell'alveo della valutazione della qualità, gli autori sono direzionati verso lo sviluppo di un sistema di rank basato sull'attribuzione di un diamante (💎), in modo da consentire agli utilizzatori di valutare in maniera immediata l'usabilità dei dati di interesse.

Il lavoro di ricerca sarà direzionato, inoltre, verso lo sviluppo di funzionalità "semplificate" per un'utenza non esperta nella convinzione che i benefici della *data revolution* debbano essere alla portata di tutti i cittadini.

Nell'ottica dell'apertura dei dati, le comunità devono essere considerate il punto di partenza per utilizzare un approccio che sia non solo tecnologico, ma che, mirando prima di tutto alla *statistical literacy*, prenda in considerazione elementi di formazione, comunicazione e creazione di competenze in modo di fare del coinvolgimento, dell'ascolto e della partecipazione civica, l'elemento centrale per la creazione di valore (Open Data Knowledge Foundation, 2015).

## 6. Bibliografia

- Agenzia per l'Italia Digitale (2014) Agenda nazionale per la valorizzazione del patrimonio informativo pubblico. s.l..
- Auer S. et al. (2012) Quality Assessment for Linked Data: a survey.
- Daprà A. (2013) Open Data o Good Data? *Statistica & Società*. Statistiche ufficiali, 2013, Vol. 1, 2.
- Del Longo T. (2015) Open data, l'Italia si dia una mossa: ora coinvolgere gli utenti. *Corcom.it*. [Online] 2015. <http://www.corrierecomunicazioni.it>.
- Di Donato F. (2010) Lo stato trasparente. Linked open data e cittadinanza attiva. Pisa : Edizioni ETS, 2010.
- Evans D. (2011) The Internet of Things. How the evolution of the Internet is changing everything. s.l. : Cisco Internet Business Solution Group (IBSG), 2011.
- Goodchild M. F. e Clarke K. C. (2002) Data quality in massive data sets. *Handbook of Massive Data Sets*. s.l. : Kluwer Academic Publishers, 2002.
- Grossenbacher A. (2013) Big Data, Open Data and Official Statistics. *Blog about Stats*. [Online] 2013. [blogstats.wordpress.com/2013/04/21/big-data-open-data-and-official-statistics/](http://blogstats.wordpress.com/2013/04/21/big-data-open-data-and-official-statistics/).
- Gurin J. (2014) Open data Now: the Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation – January 7, 2014.
- Iacono N. (2014) Aprire i dati per il riuso: i ritardi italiani. *Agenda Digitale eu*. [Online] 2014. <http://www.agendadigitale.eu/>.
- Independent Expert Advisory Group (2014) *A world that Count: Mobilising the Data Revolution for Sustainable Development*. s.l. : UN, 2014.
- ONU (2014) A World that Counts. Mobilising the Data Revolution for Sustainable Development, *produced by Independent Expert Advisory Group Secretariat* - <http://www.undatarevolution.org/report/>
- Open Data Knowledge Foundation (2015) Opendata Handbook, Why Open Data? s.l. : ODKF.
- Scannapieco M., Virgillito A. e Zardetto, D. (2013). *Placing Big Data in Official Statistics: A Big Challenge?* Brussel, Belgium: March : Paper presented at the New Techniques and Technologies for Statistics conference, 2013. p. 1-2.
- Speroni D. (2011). Dati aperti, la statistica cambia a vantaggio di tutti. *Corriere della sera/blog*. [Online] 2011. <http://numerus.corriere.it/>.
- Stentella M. (2015) Data revolution: la rivoluzione possibile. Ne parliamo con Enrico Giovannini. [Online] 2015. [saperi.forumpa.it](http://saperi.forumpa.it).

## ABSTRACT

The *data deluge* phenomenon, generated by the increase of the massive network-connected devices, has led to an increase in the amount of data produced, the speed of its update of the variety of sources that generate them. In this context it has been developing the current named *data revolution* whose foundation is the sharing of knowledge (*shared data*) through the paradigm of *open data*.

This movement has led to the growth in recent years of production initiatives of public data "open" (open government data). This process of "opening" is accompanied by a number of critical issues related to data quality, ability to reuse and information retrieval.

In Italy the initiative of opening public data more important is *dati.gov.it*.

*Open Data Explorer - ODE*, interfacing with this platform provides a tool to those who approach the complex world of public open data, to select quality datasets with well-defined standards that permit reuse for statistical purposes.