ALMA MATER STUDIORUM – UNIVERSITÁ DI BOLOGNA

CAMPUS DI CESENA

SCUOLA DI SCIENZE

CORSO DI LAUREA IN SCIENZE E TECNOLOGIE INFORMATICHE

ANALISI COMPARATIVA DI SOFTWARE PER LA PUBBLICAZIONE DI OPEN DATA

Relazione finale in

Algoritmi e Strutture Dati

Relatore:

Prof.

Luciano Margara

Presentata da:

Stefano Valentini

Sessione III

Anno accademico 2012/2013

Indice

1. Introduzione

1.1 Tema della tesi:

Lo scopo di questa tesi è quella di analizzare in dettaglio i principali software usati a livello mondiale per la pubblicazione degli open data, per fornire una guida a sviluppatori che non conoscono i programmi adatti a questa fase. L'idea è nata durante il tirocinio, svolto presso un'azienda che promuove servizi per la Pubblica Amministrazione. Avendo contribuito allo sviluppo di un portale di pubblicazione di open data ho preso contatto con alcune delle soluzioni software presenti sul mercato.

Gli Open data, e in particolare gli open government data, sono una immensa risorsa ancora in gran parte inutilizzata. 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, 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 di chiunque fosse interessato.

La prima parte della tesi sarà concentrata ad introdurre il mondo degli open data, con definizioni, concetti e leggi sull'argomento. La seconda parte sarà invece il fulcro dell'analisi tra diversi software già largamente utilizzati per pubblicare open data.

1.2 La situazione in Italia

Negli ultimi anni In Italia sono state lanciate diverse iniziative d'apertura del patrimonio informativo da parte di pubbliche amministrazioni comunali e regionali. I primi Data Store italiani pubblicati online sono stati quello del Piemonte ed Emilia Romagna, seguiti dal portale dati.gov.it. Online dal 18 Ottobre 2011, il portale nazionale dei dati aperti cataloga tutti i dataset pubblicati dai portali italiani di open data. Al momento in Italia sono stati pubblicati 8980 dataset da parte di Comuni, Regioni o Enti pubblici.

Il primo grafico posto nella figura sotto, traccia l'andamento del catalogo da marzo 2012 a gennaio 2014. Nel grafico inferiore invece è riportata la ripartizione dei dataset per livello di riusabilità, da una a cinque stelle (vedi capitolo 2.4).

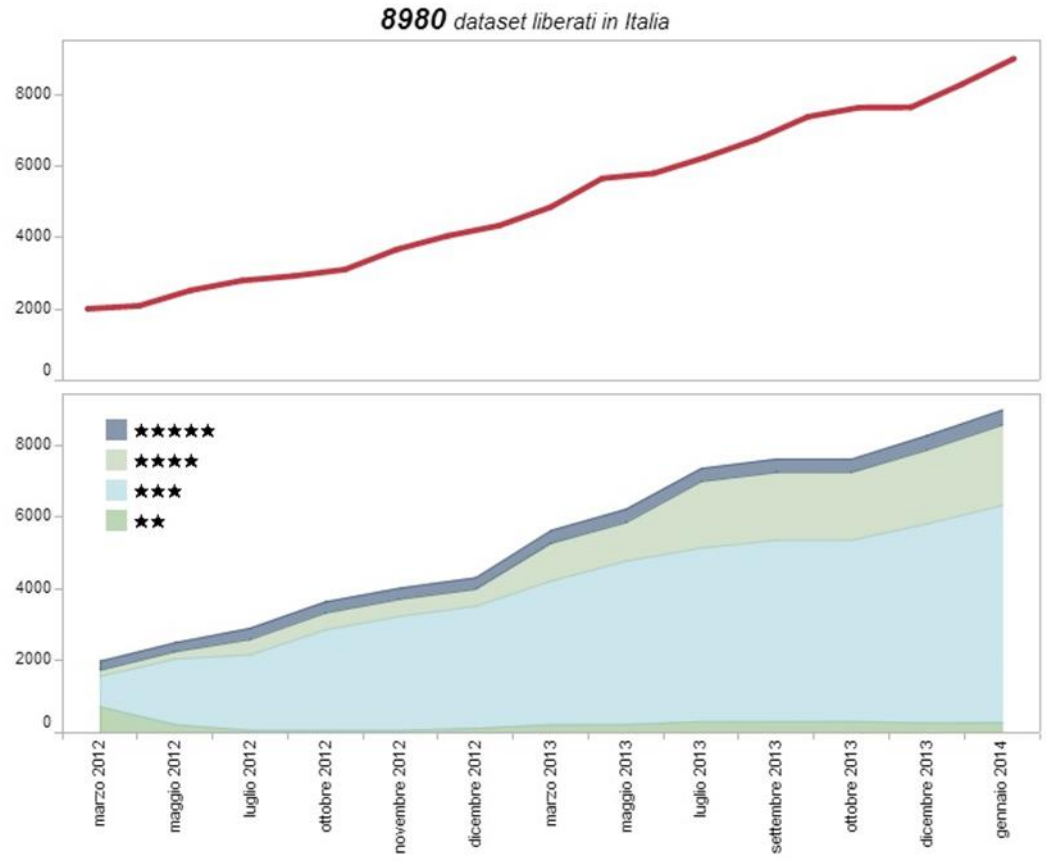


Figura 1: Grafici informativi sui dataset liberati in Italia (fonte: dati.gov.it)

Esistono due tipi di portali open data: quelli che fungono da "collettori" e motori di ricerca dei dataset pubblicati, e i siti web focalizzati sugli open data delle amministrazioni che hanno esposto i propri dataset.

Riguardo ai primi, quattro sono quelli di primaria importanza a livello nazionale:

- dati.gov.it: è l'esperienza italiana di portale nazionale dei dati aperti nato dopo una serie di data store governativi lanciati negli ultimi anni (il più celebre dai quali è il data.gov americano, lanciato nel dicembre 2009); in pochi anni la pratica degli open data e dei data store governativi si è estesa in tutto il mondo, mentre nel dicembre 2012 è stato pubblicato anche il portale europeo in versione beta (open-data.europa.eu);
- opendatahub.it: si tratta di una piattaforma di indicizzazione e ricerca dei dataset aperti disponibili in Italia. L'indice viene elaborato da un motore di ricerca che va ad individuare le fonti su web delle Pubbliche Amministrazioni e Organizzazioni/Aziende pubbliche e private che hanno pubblicizzato il rilascio di dati aperti. Il portale è gestito da Sciamlab, un'organizzazione che si occupa di open data e di forme di cooperazione e federazione delle informazioni al fine di potenziare la capacità di analisi e poter giungere a decisioni migliori;
- datiopen.it: l'idea di base di DatiOpen.it è quella di dare una spinta decisa al fenomeno OpenData Italiano mediante la raccolta e documentazione della maggior quantità possibile di dati open italiani, monitorando quotidianamente il panorama italiano alla ricerca di nuovi dati da catalogare ed inserire nel sistema, e l'opportunità di visualizzare direttamente dal sito dati, tabelle, grafici, mappe; l' iniziativa parte da un gruppo di professionisti che credono sia negli Open Data che nell'Open government e che mettono a disposizione le loro competenze e la loro passione per la diffusione nel contesto italiano;
- linkedopendata.it: il portale pubblica dati aperti e facilmente accessibili da persone e applicazioni. I dataset a disposizione, con licenze aperte e pubblicati in modalità LinkedData, possono essere direttamente interrogati da qualsiasi applicazione indipendentemente da linguaggi di programmazione e tecnologie. Il portale è gestito da un'associazione senza fini di lucro, appassionata di tecnologie Web e Semantic Web e che ritiene che la

pubblicazione di dati grezzi, istituzionali e non, sia un importante passo verso la trasparenza ma soprattutto verso l'offerta di servizi innovativi ai cittadini ed alle imprese.

Rispetto invece alle esperienze di singole amministrazioni, ISTAT è – e non potrebbe essere altrimenti data la sua natura di ente statistico – il soggetto pubblico che ha pubblicato il maggior numero di dataset sul proprio portale.

Al secondo posto c'è la Regione Lombardia, al terzo la Provincia Autonoma di Trento mentre ai due posti successivi troviamo due amministrazioni comunali: Firenze e Bologna.

Figura 2: Illustrazione enti che possiedono un portale open per i dati in Italia. Maggiore è il raggio del cerchio, maggiore è il numero di dataset pubblicati. (fonte: dati.gov.it)

2. Open Data

2.1 Cosa sono gli Open Data?

Gli open data, detti anche dati aperti, sono dati prodotti dalla pubblica amministrazione che possono essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque, soggetti eventualmente alla necessità di citarne la fonte e di condividerli con lo stesso tipo di licenza con cui sono stati originariamente rilasciati. Essi si richiamano alla più ampia disciplina dell'open government, cioè una dottrina in base alla quale la pubblica amministrazione dovrebbe essere aperta ai cittadini, tanto in termini di trasparenza quanto di partecipazione diretta al processo decisionale.

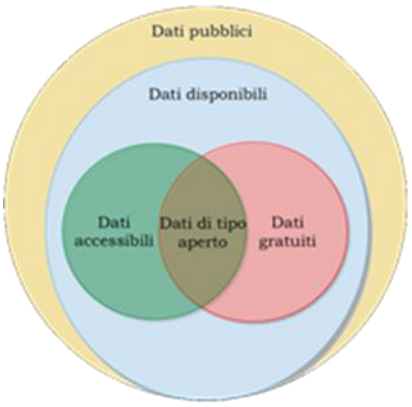


Figura 3: Illustrazione dei dati pubblicati dagli enti (fonte: Documento Interoperabilità)

Gli Open Data condividono gli aspetti della Open Definition, che espone i principi di "openness" riguardo a dati e contenuti:

- Disponibilità ed accesso: l'opera deve essere disponibile nella sua interezza e preferibilmente tramite il download gratuito via Internet.
- Riutilizzo e ridistribuzione: la licenza deve consentire la realizzazione di modifiche e di opere derivate, deve consentire la loro distribuzione negli stessi termini dell'opera originaria, non deve imporre alcuna limitazione alla vendita o all'offerta gratuita dell'opera singolarmente considerata o come parte di un pacchetto composto da opere provenienti da fonti diverse, ed infine non deve richiedere alcuna "royalty" o altra forma di pagamento per tale vendita o distribuzione.

La ragione fondamentale per cui è importante chiarire il significato di "aperto" e del perché utilizzare proprio questa definizione, può essere identificata in un termine: interoperabilità.

2.2 Interoperabilità

L'interoperabilità è la capacità di diversi sistemi e organizzazioni di lavorare insieme, e quindi interoperare. E' la chiave per rendere ciascun dato un componente, e combinare insieme vari componenti è essenziale per costruire sistemi sofisticati. Il punto cruciale di un bacino di dati accessibili e utilizzabili in modo condiviso è il fatto che, potenzialmente, possono essere liberamente "mescolati" con dati provenienti da fonti anch'esse aperte.
Il problema dell'integrazione fra dati di sistemi diversi è reso ancora più complesso dal fatto che ogni sistema si basa tipicamente su infrastrutture eterogenee: diversi linguaggi, formati, protocolli, ecc. 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 (dove si trova il vero valore dell'operazione)

2.3 Normativa

Riguardo all'ambito europeo è importante richiamare le disposizioni contenute nella direttiva sull'informazione del settore pubblico (PSI). La Direttiva 2003/98/CE del Parlamento europeo e del Consiglio, approvata il 17 novembre 2003 e pubblicata nella GUCE n. L 345 del 31 dicembre 2003, costituisce il primo passo in tema di riutilizzo dell'informazione del settore pubblico. Essa nasce come raccomandazione e non obbligo per gli Enti pubblici, affidandogli il compito di favorire il riuso e rendere disponibili i documenti attraverso indici on-line e licenze standard.
La direttiva è stata recepita in Italia con il Decreto legislativo 24 gennaio 2006, n. 36, pubblicato nella G.U. del 14 febbraio 2006, n. 37 e successivamente modificato dalla L. 96/2010, ed è stata recentemente emendata accogliendo alcuni dei principi basilari dell'Open Data.

Negli ultimi tempi la tematica risulta fortemente di interesse perché i dati delle amministrazioni sono visti come elemento infrastrutturale che costituisce una ricchezza per il Paese, un'opportunità di sviluppo economico, di crescita occupazionale, di riduzione degli sprechi e di aumento dell'efficienza.
In particolare, l'art. 9 del DL n. 179/2012, convertito in Legge n. 221/2012, ha interamente riscritto l'art. 52 del CAD sull'accesso telematico e riutilizzo dei dati delle pubbliche amministrazioni. Esso stabilisce che i soggetti pubblici o a maggioranza pubblica "*pubblicano nel proprio sito web, all'interno della sezione Trasparenza, valutazione e merito, il catalogo dei dati, dei metadati e delle relative banche dati in loro possesso, fatti salvi i dati presenti in Anagrafe tributaria.".*
In ottemperanza a tale articolo del CAD, come modificato, nella prima parte del 2013 è stato elaborato il documento di linee guida per "*l'individuazione degli standard tecnici, compresa la determinazione delle ontologie dei servizi e dei dati, le procedure e le modalità di attuazione delle disposizioni del Capo V del Codice dell'Amministrazione Digitale con l'obiettivo di rendere il processo omogeneo a livello nazionale, efficiente ed efficace."*

A luglio 2013 è stata pubblicata sul sito di AgID la prima versione di tali linee guida. Il documento è stato redatto dal "*Gruppo di lavoro dell'Agenzia per l'Italia Digitale per le linee guida sulla valorizzazione del patrimonio informativo pubblico*", cui hanno partecipato numerose pubbliche amministrazioni centrali e locali.
Le linee guida sono intese a supporto delle amministrazioni nel processo di valorizzazione del proprio patrimonio informativo pubblico, definendo gli interventi principali da compiere per l'attuazione della strategia dettata dall'agenda nazionale. Propongono schemi operativi e organizzativi, identificano standard tecnici e "best practice" di riferimento, suggeriscono aspetti di costo e di licensing da tenere in considerazione, ecc. Sono soggette a revisione periodica annuale.

Infine tra le novità di rilievo si segnala che, secondo la nuova versione della direttiva, pubblicata in gazzetta ufficiale lo scorso 26 giugno 2013, la diffusione della PSI non è più una raccomandazione ma diventa obbligatoria.

2.4 Linked Open Data

Mentre in generale gli Open Data abbattono le barriere culturali, legali ed economiche per favorire il riuso, il movimento Linked Data si concentra piuttosto sulla messa a punto di strumenti che permettono di dare ai dati un'identità e di renderli collegati ed interoperabili tra di loro. Il problema dell'interoperabilità deriva in parte proprio dall'identità: se non si sa come "*risolvere*" l'identità di una "*cosa*" fra i diversi sistemi che ne "*parlano*", è molto difficile aggregare le informazioni ad essa relative. È' possibile, infatti, che a una stessa entità sistemi differenti assegnino identità differenti. Pertanto, queste andrebbero allineate per continuare a garantire l'interoperabilità tra sistemi eterogenei.

I Linked Data sono stati proposti nel 2006 da Tim Berners-Lee come metodo elegante ed efficace per semplificare e omogeneizzare le soluzioni proprio ai problemi di identità e interoperabilità. Il metodo consiste dei seguenti quattro principi base:

1. Usare URI per identificare oggetti.
2. Usare HTTP URI in modo che questi oggetti possano essere referenziati e cercati da persone e user agent.
3. Fornire informazioni utili sull'oggetto quando la sua URI è deferenziata, usando formati standard come RDF.
4. Includere link ad altre URI relative ai dati esposti per migliorare la ricerca di altre informazioni relative nel Web.

Adottare modelli, tecnologie e standard aperti di Linked Data (e.g., RDF), sfruttando le esperienze maturate nell'ambito del Web Semantico, offre benefici di sicuro interesse per utenti e sviluppatori. I primi acquisiscono la possibilità di riferirsi a entità specifiche, anziché a posizioni all'interno di un database, e di navigare tra i dati; i secondi possono realizzare applicazioni, anche complesse, che combinano i dati della pubblica amministrazione con altri, aprendoli anche

alla possibilità di arricchimento automatico attraverso il cosiddetto ragionamento automatico (inferenza).

Da qualche tempo un numero crescente di amministrazioni, rendendosi conto del valore aggiunto dei Linked Data (anche in termini di interoperabilità interna alla PA stessa), si sta facendo carico non solo del lavoro necessario per pubblicarli come dati aperti ma anche di quello di pubblicarli direttamente in modalità "linked". L'esempio per eccellenza in Europa in questo settore è quello del governo inglese, che è stato un precursore nel pubblicare i suoi dati come Linked Data (http://data.gov.uk/).
Prende corpo dunque uno scenario nel quale la PA, anche in virtù della sua possibilità di definire norme, può svolgere un ruolo da protagonista in un sistema complesso di aziende, comunità e cittadini.

2.5 Standard per i Linked Open Data

I LOD ereditano gli standard definiti dal W3C e impiegati nel contesto del Web Semantico. Il Web Semantico è l'evoluzione del Web dei documenti (un grande contenitore di documenti collegati tra loro), verso il Web delle entità.

2.5.1 RDF (Resource Description Framework)

RDF è un linguaggio relativamente semplice che permette di rappresentare dati e metadati attraverso la definizione di asserzioni, dette triple, secondo lo schema <*soggetto*> <*proprietà*> <*oggetto*>. Gli elementi fondamentali del linguaggio sono le risorse, identificate univocamente per mezzo di URI, che possono comparire in una delle tre posizioni di una tripla. Una risorsa in posizione di proprietà mette in relazione due risorse in posizione <*soggetto*> e <*oggetto*>. Una proprietà può anche mettere in relazione una risorsa e un "literal", cioè un'espressione simbolica: numero, stringa, ecc. RDF genera così un grafo di nodi interconnessi, chiamato anche grafo RDF.

Esempio:
“Mario Rossi” “è autore di” “Mio libro”

In RDF/XML:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:au="http://description.org/schema/">
  <rdf:Description about="http://www.biblioteca.it/Mio_libro/">
    <au:author>Mario_Rossi</au:author>
  </rdf:Description>
 </rdf:RDF>
```

In triple RDF:
<http://www.raccoltautori.it/Mario_Rossi>
<http://description.org/schema#author>
<http://www.biblioteca.it/Mio_libro>

RDF ha anche un'estensione, chiamata RDF Schema (RDFS) che permette di definire semplici schemi.

2.5.2 OWL (Web Ontology Language)

RDF permette di assegnare un tipo a qualsiasi entità. Esse usate come tipi (classi) o proprietà formano il vocabolario (chiamato anche schema o ontologia) usato da un dataset. I vocabolari migliori sono scritti in OWL, una famiglia di linguaggi per la rappresentazione della conoscenza mediante ontologie. OWL è diventato lo standard W3C per la rappresentazione di ontologie su Web e può essere rappresentato come un’estensione di RDFS. OWL permette di esprimere le ontologie in maniera più dettagliata e precisa di RDFS, garantendo anche la possibilità di verificare automaticamente la correttezza logica di ciò che si rappresenta. OWL permette inoltre l’uso di ragionatori automatici per le cosiddette logiche descrittive, per derivare inferenze logiche dalla struttura dei dati.

2.5.3 SPARQL (Sparql Protocol And RDF Query Language)

È un linguaggio con una sintassi simile a quella SQL per l'interrogazione di dati RDF e un protocollo di comunicazione basato su HTTP. Un client SPARQL può quindi interrogare un endpoint SPARQL con interrogazioni (query) riguardanti un grafo RDF. Le query esprimono le caratteristiche che un sottografo (un insieme di connessioni tra risorse di un certo tipo e con certe caratteristiche) del dataset RDF deve avere. Le risposte alla query sono tutti quei sottografi del grafo RDF che soddisfano le caratteristiche volute.

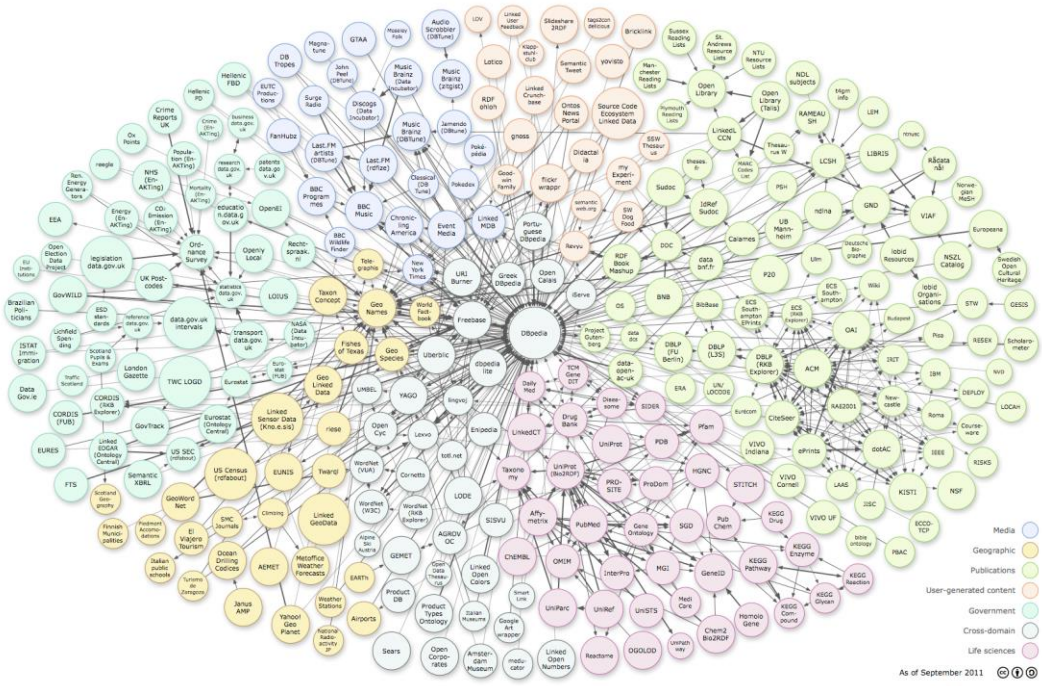


Figura 4: Basi di dati pubblicate nella LOD cloud (fonte: http://linkeddata.org)

2.6 Il modello a cinque stelle

Tim Berners-Lee suggerisce, per gli Open Data, uno schema a livelli che chiama "Modello a cinque stelle".

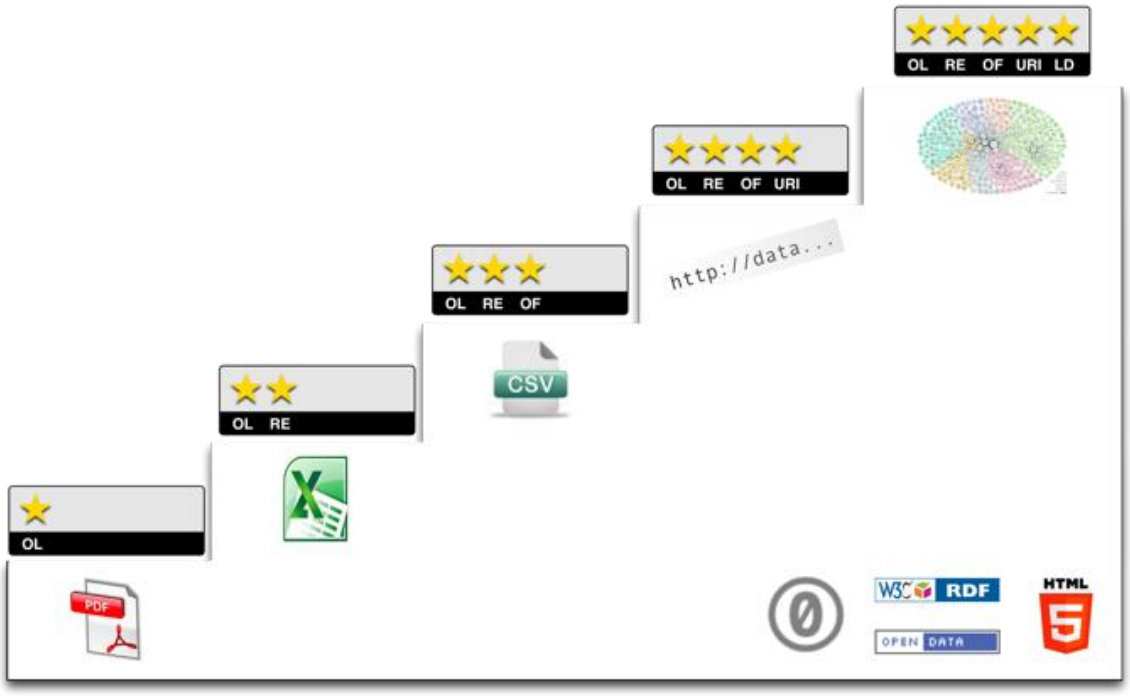


Figura 5: Schema delle caratteristiche degli open data (fonte: 5stardata.info)

Ecco nel dettaglio la suddivisione:

★ Dati disponibili in qualunque formato, ma con una licenza aperta.

★★ Dati disponibili in un formato leggibile da un agente automatico. Tipicamente, rientrano in questo livello dati in formati proprietari (e.g., excel).

★★★ Dati con caratteristiche del livello precedente ma con un formato non proprietario.

★★★★ Dati con caratteristiche del livello precedente ma esposti usando gli standard W3C RDF e SPARQL.

★★★★★ Dati con caratteristiche del livello precedente ma collegati a dati esposti da altre persone e organizzazioni.

Per definire meglio i livelli, possiamo analizzare il loro comportamento in tre differenti dimensioni: informazione, accesso e servizi implementabili.

2.6.1 Informazione

Questa dimensione descrive la qualità dell'informazione fornita insieme ai dati aperti. In questo senso si possono avere:
• documenti: i dati sono incorporati all'interno di documenti senza struttura e quindi leggibili e interpretabili solo da umani (livello 1);
• dati grezzi: i dati sono leggibili anche da un programma ma l'intervento umano è fortemente necessario per una qualche elaborazione degli stessi (livelli 2 e 3);
• dati arricchiti semanticamente: i dati sono descritti semanticamente tramite metadati e ontologie (livello 4);
• dati arricchiti semanticamente e collegati: i dati sono descritti semanticamente tramite metadati e ontologie (livello 5). L'intervento umano si può ridurre al minimo e talvolta addirittura eliminare.

2.6.2 Accesso

Questa dimensione descrive la facilità con cui utenti umani e agenti automatici riescono ad accedere ai dati e considera quindi anche lo sforzo di comprensione della struttura dei dati al fine di poterli interrogare e utilizzare in modo corretto. I gradi individuati sono:
• solo umano: solo gli umani sono in grado di leggere i documenti senza struttura e quindi dare un senso ai dati in esso presenti (livello 1);
• umano e semi-automatico: gli agenti automatici possono elaborare i dati ma non sono in grado di interpretarli; pertanto si rendere necessario un intervento umano al fine di scrivere programmi adhoc per il loro utilizzo (livelli 2 e 3);
• umano e automatico: gli agenti automatici che conoscono l'ontologia di riferimento possono elaborare i dati senza un ulteriore intervento umano (livelli 4 e 5).

2.6.3 Servizi

Questa dimensione descrive la tipologia di servizi che possono essere progettati e implementati con i dati aperti. Dalla tipologia ne deriva il grado di efficienza e capacità con cui un servizio riesce a sfruttare informazioni anche provenienti da sorgenti diverse. In particolare, si distinguono i seguenti gradi:

• nessun servizio: nessun servizio può essere abilitato a partire dai dati contenuti nei documenti, a meno di significativi interventi umani di estrazione ed elaborazione dei possibili dati (livello 1);

• servizi non efficienti: applicazioni ad-hoc che usano i dati. Queste applicazioni devono incorporare al loro interno i dati (livelli 2 e 3);

• servizi e apps efficienti: applicazioni, anche per dispositivi mobili, che sfruttano accessi diretti a Web per reperire i dati di interesse (livello 4);

• servizi efficienti e con mashup di dati: applicazioni, anche per dispositivi mobili, che sfruttano sia accessi diretti a Web sia l'informazione ulteriore catturata attraverso i "link" dei dati di interesse (livello 5).

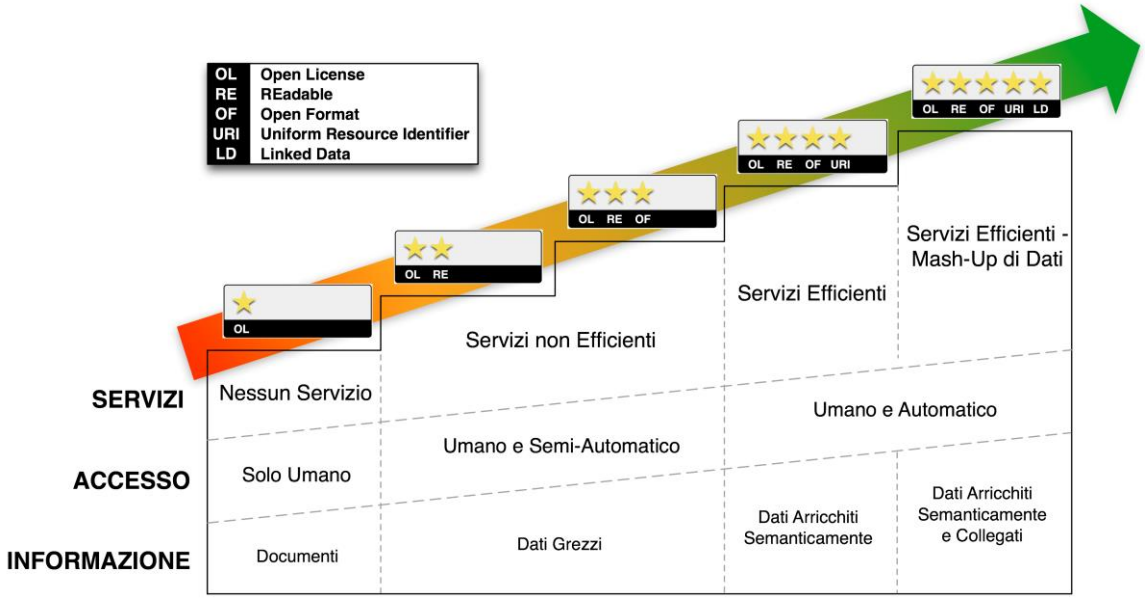


Figura 6: Schema riassuntivo dei tipi di open data (fonte: documento Interoperabilità)

3. Progettazione di un Portale Open Data

Il processo di progettazione si deve porre l'obiettivo di pianificare le azioni da intraprendere per raggiungere la pubblicazione dei dati a 5 stelle. Tale obiettivo non è detto si possa raggiungere nella totalità dei casi ma – in generale – il processo si sviluppa per passi, individuando come obiettivo minimo la pubblicazione dei dati a 3 stelle, per poi passare gradualmente alle 5 stelle.

I passi seguenti si basano sulle metodologie descritte nelle "Linee Guida per l'Interoperabilità Semantica attraverso i Linked Open Data" pubblicate dall'agenzia per l'Italia Digitale ed individua le azioni specifiche che occorre attuare perchè i dati della pubblica amministrazione possano essere pubblicati.

3.1 Individuazione e selezione dei dataset

L'individuazione e selezione dei dataset è il punto di partenza del processo di apertura dei dati. In primo luogo è bene selezionare i dati sulla base di ciò che potrebbe essere utile per il territorio, ovvero quei dataset che possono avere un impatto significativo o sull'amministrazione o sulla collettività del territorio. Secondariamente vanno individuati i vincoli all'apertura dei dati presenti, sia da un punto di vista normativo, sia organizzativo. Infine è opportuno effettuare una pre-analisi del dominio di riferimento in modo da ottenere informazioni preliminari utili per le fasi successive di trasformazione e identificare altri elementi-chiave per la misura della complessità.

3.2 Bonifica

La bonifica, come è facilmente intuibile, si rende necessaria in quanto i dati all'interno dei sistemi informativi o degli archivi sono spesso imperfetti (nascono con finalità legate all'operatività interna, non sono di solito creati per una pubblicazione esterna) e pertanto non immediatamente pronti per l'esposizione o eventuali elaborazioni. La qualità del dato in entrata rappresenta quindi un aspetto particolarmente importante perché un dataset non "*pulito*" può rendere inefficienti o infattibili alcune operazioni di confronto, di similitudine e di aggregazione sui dati. Questi sono alcuni dei problemi più diffusi: dati incompleti, ambigui, in formati diversi, mancanza di corrispondenza fra arricchimento tramite nomi usati negli schemi fisici e dati effettivamente contenuti, ecc.

Alla fine di questa fase, i dati sono già pronti per essere pubblicati rispettando lo standard entro le tre stelle. Per pubblicare Linked Open Data invece, servono alcuni passi successivi.

3.3 Analisi e modellazione

Durante il processo di analisi e modellazione viene definito il modello concettuale e viene data una rappresentazione coerente con esso del dataset di riferimento. Si tratta quindi di ripensare concettualmente i dati; in altre parole è come se la base dati venisse re-ingegnerizzata. Sia il modello concettuale sia il dataset alla fine di questa fase saranno rappresentati in RDF. In questo formato, i dati e le relazioni tra essi sono rappresentate attraverso delle proposizioni ("triple") della forma <soggetto> <predicato> <oggetto>.
Gli elementi devono avere dei nomi appropriati, devono seguire semplici convenzioni e devono rispettare gli schemi URI dei Linked Data. In pratica, ogni elemento deve avere un indirizzo Web univoco che dev'essere "dereferenziabile".

3.4 Arricchimento e metadatazione

I dati precedentemente bonificati e modellati sono arricchiti inserendo informazioni di contorno (metadatazione) che rendono il riutilizzo più semplice e/o facendo derivare contenuto informativo aggiuntivo mediante l'estrazione automatica dell'informazione o di ragionamento automatico (inferenza).
I metadati arricchiscono il contenuto informativo dei dati esplicitandone delle proprietà che semplificano il processo di fruizione dei dati stessi facilitandone la ricerca, il recupero, la composizione e di conseguenza il riutilizzo. Così come i dati, anche i metadati sono espressi attraverso RDF che nasce proprio come standard per l'annotazione semantica di pagine Web per poi divenire lo strumento base per la creazione di Linked Data. Alcuni dei metadati più utili ai fini dell'interoperabilità e del riutilizzo dei dati sono: le informazioni sulla semantica dei dati, le informazioni di contesto e le informazioni di provenienza.

3.5 Linking esterno

Il contenuto informativo viene legato ad altre informazioni (linking), che possono essere altri dataset della stessa amministrazione, oppure dataset presenti nel Web dei Dati. Questo consente un accesso a un più ampio insieme di dati e di ottenere dati attestati a livello 5 della classifica di qualità (Linked Open Data). In pratica, si tratta di allineare entità di diversi dataset.

3.6 Validazione e pubblicazione

In generale, possono essere eseguite tre tipi di validazione: quella sintattica, quella logica e quella concettuale. Nella validazione sintattica viene verificato che il contenuto dei dati rispetti i formati standard del W3C. La validazione logica, individua un insieme di casi di test che devono essere soddisfatti. I casi di test possono essere costituiti da un insieme di domande di cui si vorrebbe conoscere la risposta (chiamate come "*competency questions*") e di cui si sa

esserci i dati. Tali domande, tradotte in interrogazioni espresse in un linguaggio per l'interrogazione dei dati (tipicamente SPARQL) sono processate e i risultati comparati con quelli attesi. Nel caso in cui alcuni risultati non diano le conferme aspettate è possibile dover ricontrollare i passi di analisi e modellazione per identificare eventuali errori commessi. Infine, nella validazione concettuale viene verificata l'adeguatezza dell'ontologia ai requisiti e all'intuizione degli esperti. In generale, se uno solo dei tre tipi di analisi fallisce, è necessario ricontrollare le fasi precedenti.

La pubblicazione deve essere vista come un processo graduale che giunge a rilasciare nel tempo dati potenzialmente sempre più raffinati. Si deve infatti pensare che da un primo grezzo processo di reingegnerizzazione si possano estrarre i primi dati aperti, su cui nel tempo si adeguano versioni più ricche e collegate tra loro.
Fondamentale in questa fase è la selezione della piattaforma tecnologica di pubblicazione, poiché è molto importante che essa metta a disposizione funzionalità che facilitino il riutilizzo e l'interoperabilità dei dati.

4. Analisi dei Software

Questo capitolo intende affrontare l'analisi dei software e delle soluzioni disponibili al fine della creazione di un portale per la pubblicazione di open data. Verranno analizzati programmi al fine di ottenere:

- Un sito web con un discreto livello di personalizzazione che consente l'accesso da parte di utenti.
- L'unicità di ogni URI attraverso il quale ogni dato del portale deve essere accessibile;
- L'upload dei dati sul portale ed un servizio di linking per i file situati su host esterni.
- Un'interfaccia per l'inserimento di metadati al momento del caricamento degli open data sul portale;
- La garanzia di poter sempre scaricare i dati presenti sul portale.
- La compatibilità della soluzione sia per i dati di livello 1, 2 e 3 che per quelli di livello 4 o 5, quindi deve essere presente, o integrabile, un'interfaccia per l'interrogazione dei Linked Open Data tramite SPARQL endpoint.

Alcune delle soluzioni analizzate in questo capitolo sono state prese dal documento "*Linee guida per l'interoperabilità semantica attraverso i linked open data*" pubblicato dalla DigitPA. In questo documento vengono elencati alcuni software già largamente utilizzati nel mondo degli open data.

4.1 Soluzioni che offrono un portale per l'accesso agli Open Data

Per assolvere alla creazione di un portale su cui pubblicare gli open data sono possibili molteplici soluzioni. E' consuetudine scegliere una piattaforma su cui sono già implementati molti dei servizi base richiesti dal progetto e che permetta l'eventuale integrazione di ulteriori moduli sviluppati ad hoc. Non è quindi concettualmente sbagliato lo sviluppo da zero di un portale, magari se si ha una certa praticità con i CMS, ma il riuso di progetti che già assolvono i compiti primari comporta:

- Un notevole risparmio di tempo in fase di sviluppo e di riuso di funzioni già implementate;
- La possibilità per gli sviluppatori di utilizzare la ricca documentazione di un progetto già usato su larga scala;
- Se il progetto scelto ha una community molto attiva, si presume maggior garanzia di futuri aggiornamenti della piattaforma;
- Un effetto di standardizzazione dei portali.

In quest'analisi non vengono considerate soluzioni che non si appoggiano a software già esistenti, o che non assolvono ai compiti citati nella precedente pagina.
L'analisi delle piattaforme si basa sul seguente schema:

4.1.1 Requisiti funzionali

I requisiti funzionali sono le caratteristiche fondamentali per il corretto funzionamento dell'applicazione. Esse sono: i sistemi operativi supportati, i database interfacciabili, il linguaggio di sviluppo (che comporta certi vincoli). Sono i pilastri del software, e sono le prime dipendenze che si prendono in considerazione al momento della scelta del portale.

I punti su cui si basa quest'analisi sono:

- Sistema Operativo richiesto
- Database supportati
- Linguaggio di sviluppo principale
- Installazione

4.1.2 Il supporto del progetto

Il supporto del progetto delinea il supporto che viene dato agli sviluppatori dalla community del progetto, la presenza e la qualità della documentazione fornita, e la frequenza d'aggiornamento del portale. Inoltre viene effettuato un piccolo censimento di quanto quella soluzione è adottata rispetto alle altre per lo scopo preposto, facendo esempi di siti web che utilizzano quel tipo di soluzione. Si discute anche l'eventuale presenza di API per sviluppatori.

I punti su cui si basa quest'analisi sono:

- Documentazione
- Stato del progetto
- Presenza di API per sviluppatori

4.1.3 In generale

In questa sezione vengono elencate le funzionalità peculiari che mette a disposizione il portale e viene dato un giudizio generale con i principali vantaggi e svantaggi di utilizzarlo.

4.2 CKAN

Ckan è sviluppato da Open Knowledge Foundation ed è un prodotto open source che in primis offre la catalogazione di risorse aventi la natura di file tramite URL (URI). E' una piattaforma ben integrata, altamente personalizzabile, con cui si possono realizzare tutti gli elementi di un sistema di gestione di Open Data, dalla loro memorizzazione fisica, organizzazione logica, metadatazione e, infine, esposizione su un sito Web.

4.2.1 Requisiti funzionali

La parte della documentazione riguardante la guida all'installazione è ben strutturata, e guida passo passo lo sviluppatore sino al primo avvio del portale. L'installazione da pacchetto richiede Ubuntu 12.04 64-bit ed è complessivamente veloce. Per un'installazione indipendente dal sistema operativo, è possibile effettuare quella da sorgente. Le principali dipendenze di CKAN sono: Python, PostgreSQL, libpq, pip, virtualenv, Git, Apache Solr, Jetty, e JDK versione 6.

Sistemi operativi supportati: Tutti i sistemi operativi che supportano Python.
Database supportati: L'unico tipo di database supportato è PostgreSQL.
Linguaggio di programmazione: Python.

4.2.2 Il supporto del progetto

E' il progetto leader nel settore di pubblicazione open data. E' ampiamente supportato da una vasta community e da periodiche release (a febbraio è stata rilasciata la versione 2.2). La documentazione per lo sviluppatore è molto curata in ogni dettaglio: la guida ai primi passi dell'amministratore, l'esposizione delle API (di tipo RPC. Esse espongono tutte le funzionalità del core di Ckan ai client

API). , la guida alla creazione di estensioni personalizzate sono solo una piccola parte di quello che offre.
Per quanto riguarda l'adozione da parte di siti web, si possono citare fonti di grande rilievo, come il portale dati del governo inglese, del governo statunitense (catalog.data.gov, passato da Socrata a Ckan nel 2013), il portale dati europeo e del governo italiano, ma anche numerosi portali in territorio italiano, quali il portale open data del Trentino, della Toscana e dell'Umbria.
Documentazione: http://docs.ckan.org/en/latest/maintaining/installing/

4.2.3 In generale

Ckan è un DMS, data management system che rende facile pubblicare, condividere ed usare i dati. Consente agli sviluppatori di creare estensioni con funzioni personalizzate che affiancano o sostituiscono le funzioni base di Ckan. La guida al "templating" è molto approfondita, e aiuta lo sviluppatore a prendere dimestichezza con il core di Ckan.

Vantaggi:
++Documentazione ricca
++E' la soluzione più diffusa a livello mondiale
+API potenti

Svantaggi:
-Può essere necessario un corso di Python
-PostreSQL unico database supportato

Figura 7: Pagina di ricerca dei dataset. A sinistra vi sono i filtri da aggiungere alla ricerca (organizzazioni, tag, licenze), mentre al centro vi sono i dataset. Vengono mostrate anche le etichette dei formati se nel dataset è presente almeno una risorsa di quel formato (fonte: dati.gov.it)

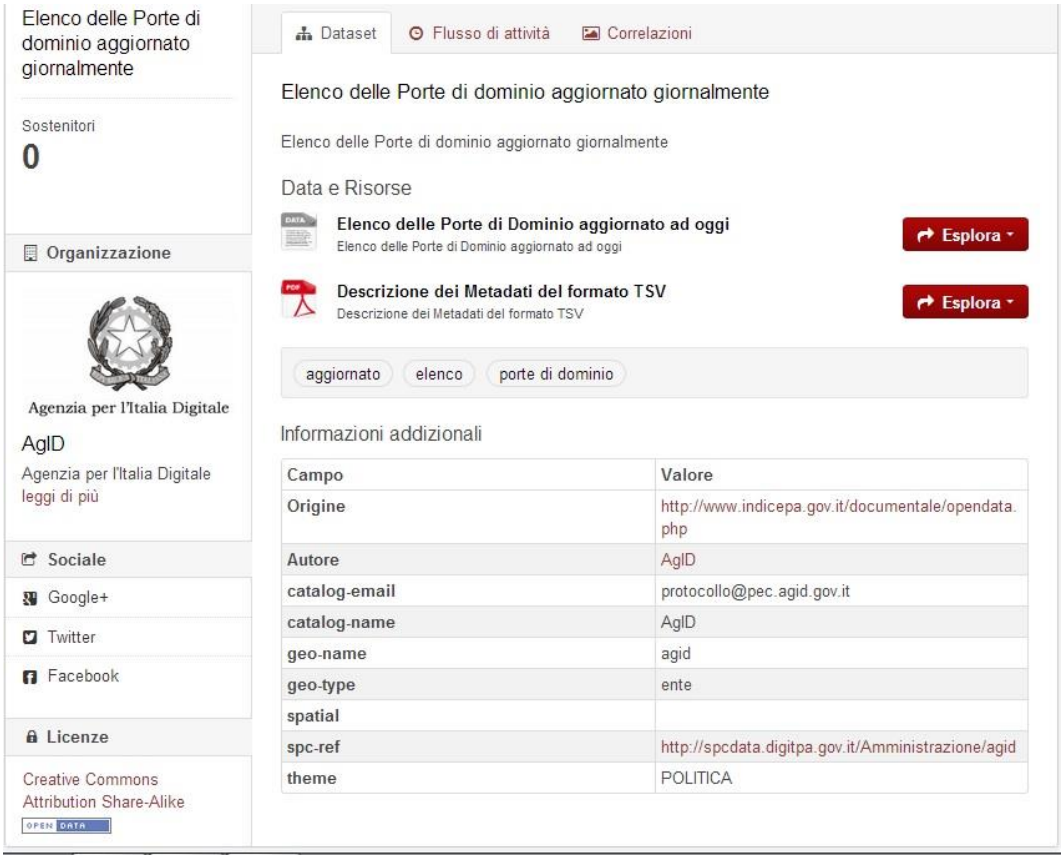


Figura 8: La pagina del dataset: in basso vi sono i metadati associati, mentre i file accedibili sono al centro in primo piano (fonte dati.gov.it)

4.3 DKAN

DKAN è un portale open data basato su Drupal e sostenuto da Nuams. Nasce dopo CKAN integrando tutte le sue funzionalità, ma fornendo la personalizzazione di un sito web tramite un CMS. Come CKAN anche questa soluzione è altamente personalizzabile, ma solo aggiungendo moduli al "core" del sistema, ossia alle funzioni base di Drupal.
I moduli di Dkan possono essere presi separatamente ed inseriti all'interno di un sito Web già esistente, rendendo la lavorazione con gli open data semplice ed immediata.

4.3.1 Requisiti funzionali

L'installazione richiede la presenza di un portale Drupal. Dopo aver installato Drupal si può seguire la guida su GitHub per rendere operativo il modulo Dkan.

Sistema Operativo Richiesto: Windows, Mac OS X, Linux e qualsiasi piattaforma software che supporti i web server Apache (versione 1.3 o superiore) o IIS (versione 5 o superiore) ed il linguaggio PHP (versione 4.3.3 o superiore).
Database Supportati: Raccomandato: MySQL. Supportati: PostgreSQL, SQLite. Supportati grazie a moduli aggiuntivi: Microsoft SQL Server e Oracle.
Linguaggio Di Sviluppo Principale: PHP 5.

4.3.2 Il supporto del progetto

Dkan è ancora in fase di sviluppo, ma si sta modellando sull'esempio di Ckan. La maggior parte delle funzioni di Ckan sono state già replicate, ma prima che venga raggiunta la stessa completezza dovrà ancora passare del tempo. La documentazione del progetto è molto scarsa rispetto a quelle disponibili per gli altri portali. Anche le API sono in via di completamento, e sono basate su quelle di Ckan.

Non vi sono siti di rilievo che utilizzano questa soluzione, probabilmente perché ancora non è stata rilasciata una release completa. Documentazione: http://docs.getdkan.com/

4.3.3 In generale

Come CKAN anche questa soluzione è altamente personalizzabile, ma solo aggiungendo moduli al "core" del sistema, ossia alle funzioni base di Drupal. In particolare si stanno sviluppando e sono già stati sviluppati moduli Drupal completamente integrabili.

Vantaggi:
++Per chi è familiare con l'ambiente Drupal e PHP è davvero semplice iniziare a pubblicare open data.
++Integrabile in un sito già basato su Drupal

Svantaggi:
-API non complete
--Beta
--Numero di utilizzi molto basso al momento

4.4 DATATANK

The Datatank è una piattaforma open source di pubblicazione di open data che rende i dati leggibili da API. La particolare caratteristica di Datatank è quella della gestione del "content negotiation": i dati richiesti vengono mostrati secondo il formato scelto dall'utente (se supportatato) semplicemente interpretando i metadati associati al dato. Dalla nuova versione (4.0) può anche essere usato come estensione della piattaforma Ckan.

4.4.1 Requisiti funzionali

L'installazione richiede composer e la clonazione del repositor tramite Git. La guida è di facile comprensione e l'installazione non è particolarmente complicata.

Dipendenze di The Datatank: Apache2 o Nginx, modulo mod rewrite abilitato, PHP 5.4 o superiore, Git.
Sistema operativo richiesto: Windows, Mac OS X, Linux e qualsiasi piattaforma software che supporti i web server Apache (versione 1.3 o superiore) o IIS (versione 5 o superiore) ed il linguaggio PHP (versione 4.3.3 o superiore).
Database Supportati: Tutti i database supportati da Laravel 4.
Linguaggio Di Sviluppo Principale: PHP 5.4.

4.4.2 Il supporto del progetto

Stato Del Progetto: Aggiornato recentemente alla versione 4.0. E' stato sviluppato un contratto di Service Level Agreement.
Documentazione: Documentazione non particolarmente ricca, ma ampliata in modo significativo nell'ultima release. E' possibile provare una demo del portale.
Documentazione: http://docs.thedatatank.com/4.1

4.4.3 In generale

La gestione del content negotiation è un aspetto molto interessante del portale. I file non sono memorizzati localmente, vengono memorizzati solo i metadati, che vengono poi interpretati al fine di restituire in output le informazioni nel formato richiesto. Al momento i formati supportati includono: JSON, XML, CSV, XLS, tutti disponibili tramite API di tipo REST.

Ecco un esempio di come Datatank gestisce il content-negotiation: l'URL del dato che raccoglie le informazioni degli alberghi di una città potrebbe essere "www.dati.citta.com/alberghi/info", dove **alberghi** è una categoria che comprende tutti i dataset relativi agli alberghi, ed **info** rappresenta il dato in formato tabulare. Ora inserendo, ad esempio, ".json" alla fine dell'URL, accediamo al file in quel formato, mentre scrivendo ".csv" scarichiamo il file nel formato relativo.

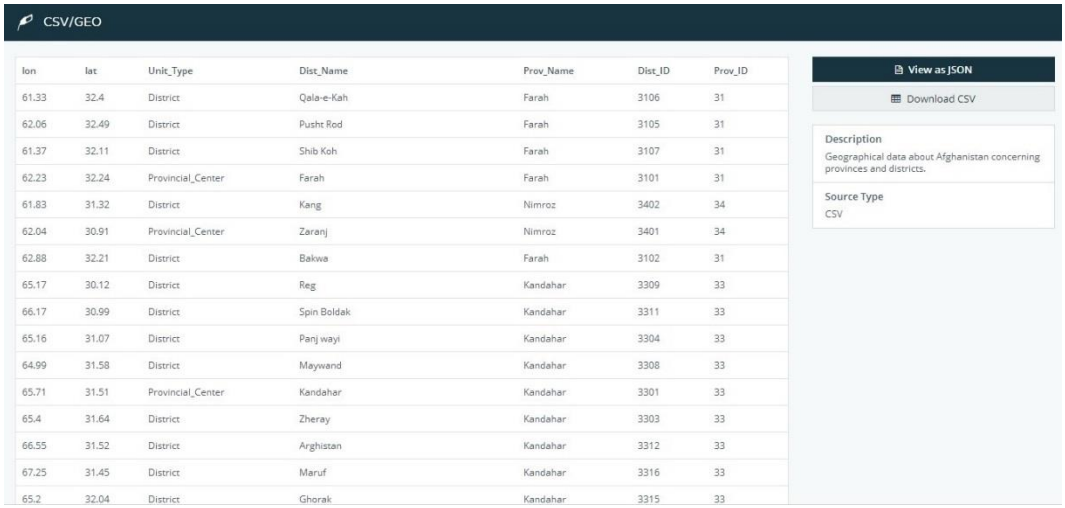
Vantaggi:

++Content Negotiation

+ Integrabile in Ckan grazie ad un'estensione

Svantaggi:

--Numero di utilizzi registrati come modulo stand-alone pressoché nullo.



lon	lat	Unit_Type	Dist_Name	Prov_Name	Dist_ID	Prov_ID
61.33	32.4	District	Qala-e-Kah	Farah	3106	31
62.06	32.49	District	Pusht Rod	Farah	3105	31
61.37	32.11	District	Shib Koh	Farah	3107	31
62.23	32.24	Provincial_Center	Farah	Farah	3101	31
61.83	31.32	District	Kang	Nimroz	3402	34
62.04	30.91	Provincial_Center	Zaranj	Nimroz	3401	34
62.88	32.21	District	Bakwa	Farah	3102	31
65.17	30.12	District	Reg	Kandahar	3309	33
66.17	30.99	District	Spin Boldak	Kandahar	3311	33
65.16	31.07	District	Panj wayi	Kandahar	3304	33
64.99	31.58	District	Maywand	Kandahar	3308	33
65.71	31.51	Provincial_Center	Kandahar	Kandahar	3301	33
65.4	31.64	District	Zheray	Kandahar	3303	33
66.55	31.52	District	Arghistan	Kandahar	3312	33
67.25	31.45	District	Maruf	Kandahar	3316	33
65.2	32.04	District	Ghorak	Kandahar	3315	33

Figura 9: Un dataset in formato tabulare. Con il bottone "View as JSON" è possibile accedere al file in formato JSON generato runtime (fonte: demo.thedatatank.com)

4.5 OGDI Data Lab

OGDI (Open Government Data Initiative) è un'iniziativa targata Microsoft per il settore pubblico composta da diversi moduli. OGDI DataLab (attualmente aggiornato alla versione 6.0) è un catalogo di open data su cloud rilasciato sotto licenza Microsoft Public License (Ms-PL). Utilizza la piattaforma Windows Azure per la pubblicazione e l'utilizzo degli open data da parte di entità pubbliche.
I principali componenti di OGDI DataLab sono tre:

- Data Service è un Web service per esporre i dati ad interrogazioni di tipo REST. I protocolli usati sono AtomPub ed Odata, mentre i formati renderizzati includono KML, JSON, e JSONP.
- Data Browser è una web application che utilizza jQuery ed altri componenti open source per la navigazione dei dati. I dati possono essere visualizzati in tabelle, mappe, grafi a barre o grafi a torta, e sono navigabili anche da anteprima (come Ckan).
- Data Loader è uno strumento che aiuta gli sviluppatori del portale ad usufruire velocemente delle funzioni di OGDI. Esiste un'interfaccia grafica ed una console per il caricamento dei dati.

4.5.1 Requisiti funzionali

Installazione: Abbastanza impegnativa. Dopo aver creato un account per Windows Azure, scaricato la piattaforma, eseguito il setup del servizio in cloud ed installato le dipendenze, mancano ancora alcuni passi per iniziare la distribuzione di open data. La guida su Github per l'installazione è aggiornata alla versione 5.0.

Sistema Operativo Richiesto: Microsoft Windows Az+ure
Database Supportati: SQL Azure Database.
Linguaggio Di Sviluppo: C#.

4.5.2 Il supporto del progetto

Questa soluzione è meno conosciuta rispetto a Ckan e Socrata, che sono ormai portali affermati in campo open data. Il progetto Microsoft sta portando avanti una collaborazione con il governo Francese. Ecco alcuni siti web che utilizzano questo software:
http://ogdifrance.cloudapp.net/
http://opendata.bordeaux.fr/
http://data.medicinehat.ca/
http://openregina.cloudapp.net/
La documentazione presente su GitHub è più ricca di quella di The Datatank, e la guida all'installazione è ben strutturata, anche se complessa.
Documentazione: https://github.com/openlab/OGDI-DataLab/wiki

4.5.3 In generale

API: Sono presenti API di tipo REST che utilizzano il protocollo OData. La vera differenza introdotta è il catalogo di open data in cloud, che mette a disposizione un servizio "*open data as a service*". Windows Azure è già integrato con il framework di sviluppo .NET. Su questa base si innestano inoltre una serie di strumenti che realizzano un sistema integrato di interfacce per l'accesso e la gestione dei dati presenti in Azure.
Di particolare rilevanza sono le funzioni di anteprima dei dati. Esse offrono maggiori servizi rispetto a quella offerte da Ckan.

Vantaggi:
+Servizio in cloud
+Funzioni di anteprima dei dati particolarmente curate

Svantaggi:

--Acquisto della licenza di Windows Azure (e pagamento continuato del servizio)

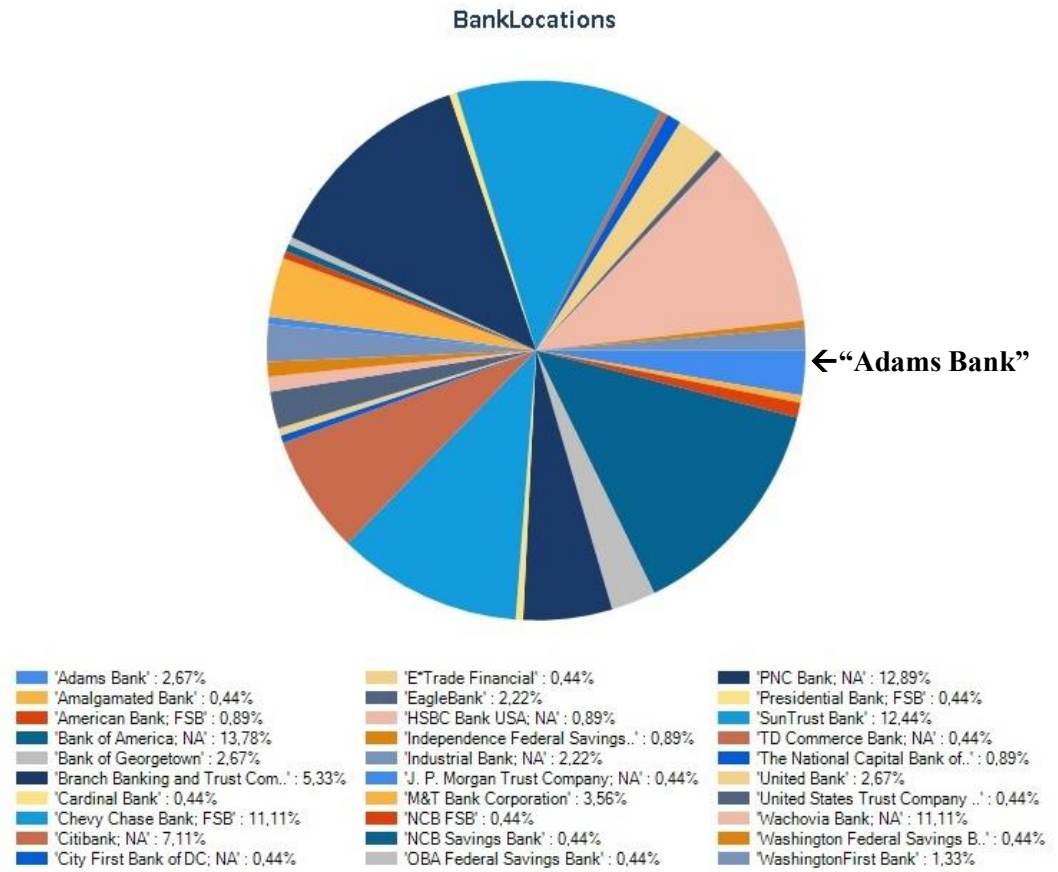


Figura 10: Esempio di grafico ottenuto come servizio di anteprima del dato contenente la tabella delle filiali delle banche di un particolare distretto. In figura il grafico è stato ottenuto filtrando le colonne della tabella per nome della banca. Il grafico è da leggere in senso orario (fonte: ogdifrance.cloudapp.net/)

4.6 Altre soluzioni

4.6.1 Socrata

Socrata è una soluzione proprietaria Open Data molto diffusa in America che offre un servizio "opendata as a service". Dal portale è possibile la gestione di più organizzazioni (come Ckan), in modo che si venga a creare un portale regionale con più enti che pubblicano i propri dati su di esso. È composta da: un sistema semantico di archiviazione dell'informazione; un'interfaccia web dinamica per l'accesso ai dati mediante maschere parametriche; una Socrata Open data API, per esporre interfacce applicative; una serie di strumenti di indagine statistica e di visualizzazione grafica con cui realizzare semplici attività di data mining; un sistema di social networking integrato, con cui gestire il feedback degli utenti; un sistema integrato di metadatazione e classificazione dell'informazione. Offre la possibilità di caricare dataset su di un sistema esterno, e di utilizzare una serie di funzionalità avanzate che permettono ad un ente anche sprovvisto di un proprio asset IT interno di avere esposti i propri Open Data. Infatti il servisto offerto da Socrata è quello di creazione del portale oltre che di hosting del servizio.
Esiste anche una versione open source chiamata "Socrata Open Data Server: Community Edition", ma attualmente si trova ancora in fase beta.
E' utilizzato dal portale dati della regione Lombardia.

4.6.2 Utilizzo di Geoportali

I geoportali sono siti Web realizzati in modo tale da costituire un punto di accesso unico ai servizi relativi a dati e risorse spaziali, che non devono necessariamente risiedere all'interno del sito stesso ma che possono invece essere distribuiti. Un geoportale offre servizi di consultazione, scambio e

cessione dei dati territoriali, nonchè accesso diretto alle informazioni e alla documentazione di riferimento, visualizzazione e scarico del materiale informativo e cartografico.

Alcuni software di riferimento sono: GEOServer e GEONetwork.

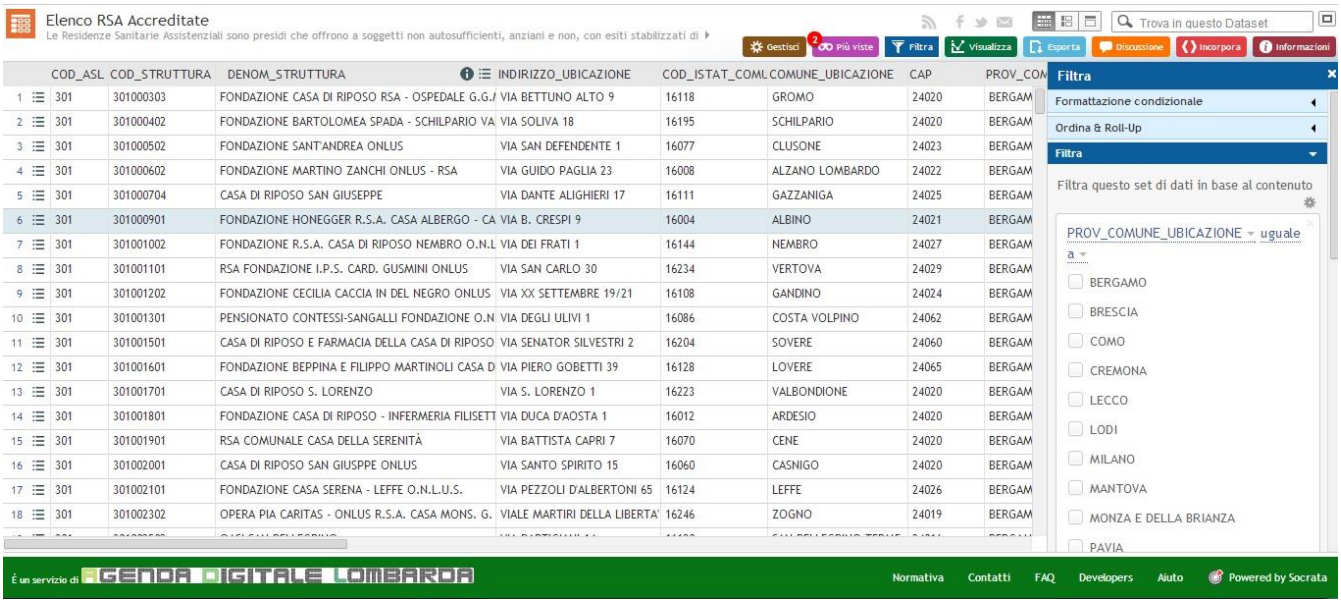


Figura 11: Ecco le interessanti funzioni di anteprima offerte da Socrata (opzioni filtro, incorpora, visualizza come grafico, esporta ecc.) (fonte: dati.lombardia.it)

4.7 Soluzioni per la gestione dei dati di livello 4 e 5

I portali analizzati precedentemente offrono un esempio della struttura sulla quale si basa il sito per la pubblicazione degli open data. Per avere una soluzione completa bisogna però associare un modulo per la gestione dei dati di livello 4 o 5, in quanto nessuna delle soluzioni appena analizzate fornisce un supporto nativo alla gestione di dati di livello 4 o 5.
Il modulo che verrà integrato nel progetto dovrà garantire le seguenti funzioni:

- Deve disporre di un motore SPARQL per effettuare interrogazioni sui dati;
- Deve permettere l'implementazione dell'interfaccia in una pagina html che sarà integrata nel portale scelto come soluzione nel paragrafo precedente;
- Deve poter gestire la navigazione dei dati di livello 4 e 5 sia tramite interfaccia grafica, sia tramite librerie per l'accesso automatizzato

Ecco due esempi equivalenti per assolvere alla gestione dei dati di livello 4 e 5.

4.7.1 Primo approccio

Nel primo caso vi è la produzione di file RDF come centro delle funzionalità

1) Analisi dei dati
2) Creazione file RDF
3) Esposizione dati su apposito portale
4) Eventuale esportazione del file RDF

Il file RDF dovrà essere memorizzato su un Triplestore (base di dati per la memorizzazione di triple) e successivamente potrà essere interrogato tramite un motore SPARQL. Infine sarà necessario implementare un'interfaccia web che

permetta all'utente non tecnico che vuole esplorare i dati la navigazione per hyperlink.
I componenti necessari per tale architettura sono i seguenti:

- Triplestore per memorizzare le triple RDF
- Motore SPARQL
- Portale per l'accesso ai dati

Le soluzioni identificate sono date da Virtuoso Open Server e Sesame OpenRDF per quanto riguarda la memorizzazione e l'interrogazione dei dati, mentre la navigazione user-friendly può essere delegata a soluzioni come Pubby, già ampiamente usato in contesti di comprovata efficienza (e.g.: DBPedia.org).

4.7.2 Sesame Open RDF

Sesame è un framework per processare dati RDF. Questo include parsers, storage solutions (i triplestore) e un motore di query basato su SPARQL. Il programma offre un modo semplice e flessibile di usare le API java che possono connettersi a tutte le maggiori soluzioni di storage RDF.

Requisiti:
Sistemi operativi che supportano Java 6 o superiore ed un Java servlet container.

Vantaggi:
-Deploy semplice rispetto ad altre soluzioni ed una maggiore facilità di utilizzo.
-Disponibilità di API (è possibile interrogare il server tramite REST).
-Documentazione intuitiva.

Svantaggi:
-Nessuna interfaccia di pubblicazione, è necessario aggiungere Pubby

4.7.3 Virtuoso Open Source Edition

Virtuoso offre funzionalità di RDBMS, ORDBMS, XML Database, RDF Store, Web Service Server. Inoltre consente molto semplicemente la configurazione e l'esposizione di uno SPARQL endpoint.

Requisiti:
Sistemi operativi supportati: Windows, Linux, Unix (AIX, HP-UX, Solaris, etc.), Mac OS X.

Vantaggi:
-Presenza di un motore SPARQL.
-E' usato in molti portali Open Data che permettono interrogazioni con SPARQL orientate alle prestazioni.
-Caratteristiche vicine all'ambiente enterprise (e.g. gestione degli utenti).
-Documentazione ampia.

Svantaggi:
-Alcune funzioni, quali la mappatura da RDBMS esterno a RDF, sono disponibili solo nella versione commerciale.
-Più complesso degli altri strumenti, mette a disposizione una gran varietà di funzioni.
-Nessuna interfaccia di pubblicazione.

4.7.4 Pubby

Pubby non è uno strumento sostitutivo di altri, ma si appoggia come layer aggiuntivo a strumenti di storage quali Sesame o Virtuoso. Visto che verrebbe utilizzato come strumento integrativo per le mancanze dei due sistemi, non c'è la voce "limiti" nell'analisi, mentre compare "funzionalità".
Pubby può essere usato per aggiungere ad endpoint SPARQL le interfacce dei linked data trasformando uno SPARQL endpoint in un server di Linked Data. E' implementato come una Java web application.

Requisiti:
Java e relativo servlet container.

Vantaggi:
-Il programma è incorporato in D2RQ.
-E' integrabile ai software Sesame e Virtuoso.
- Pubby permette la navigazione tramite URI umanamente leggibili (i cool URI, che sono un requisito secondo le linee guida della DigitPA)

4.7.5 Secondo approccio

La seconda comporta il mapping da database con la produzione di file RDF opzionale

1) Mapping database
2) Esposizione dati su apposito portale
3) Eventuale esportazione del file RDF

La seconda metodologia analizzata differisce notevolmente dalla prima, eliminando la necessità di creazione e manutenzione dei file RDF, a favore dell'accesso diretto al database attraverso meccanismi di mapping dei dati. In questo caso alla richiesta del dato segue l'accesso al database, l'incapsulamento delle informazioni secondo la struttura ontologica definita e la restituzione del dato lavorato.
I componenti necessari per tale architettura sono i seguenti:

- Strumento di mapping database
- Portale per l'accesso ai dati

Le soluzioni identificate per il mapping del database sono D2RQ, DartGrid e SquirrelRDF. D2RQ è la soluzione dal maggior numero di funzionalità, con una documentazione adeguata e numerosi casi d'uso di successo, pertanto la si preferisce agli altri prodotti.

A differenza dei sistemi per la gestione di RDF analizzati in precedenza, come comportamento predefinito D2RQ include tra i suoi componenti Pubby e dunque assolve a tutti i requisiti senza necessità di software esterni o moduli sviluppati appositamente.
D2RQ di default non crea file RDF con i dati lavorati, ma prevede la possibilità di produrli tramite esportazione. Questa operazione è del tutto opzionale e qualora si rendesse necessaria è possibile provvedere degli script che ne agevolino l'esecuzione periodica.

4.7.6 D2RQ

La piattaforma D2RQ è un sistema per l'accesso a database relazionali visti come grafi RDF. Offre un servizio di mapping RDF del contenuto di un database senza dover replicare i dati in Triplestore. Usando D2RQ è possibile:

- Effettuare query ad un database relazionale usando SPARQL;
- Accedere al contenuto del database come Linked Open Data;
- Esportare i dati in formato RDF.

Requisiti:
Sistemi operativi che supportano Java (Windows, Linux, Mac OS X, Solaris)
Database supportati: Oracle, MySQL, PostgreSQL, SQL Server, HSQLDB, Interbase/Firebird.

Vantaggi:
-Accesso diretto al database senza dover memorizzare le triple rdf in un Triplestore.
-Mapping condizionale tramite espressioni regolari.
-Una soluzione basata su D2RQ richiede un minor dialogo tra componenti

Svantaggi:
-D2RQ non può mappare schemi differenti all'interno della stessa istanza. Ad ogni istanza di d2rq-server corrisponde un solo schema.
-La funzionalità di mapping è supportata solo per i database.

5. Conclusioni

Per quanto riguarda l'analisi dei portali, quelli che hanno riscosso più successo nel mondo degli open data sono la migliore soluzione adottabile al momento. Le piattaforme Dkan e The Datatank non offrono una soluzione stand-alone sicura, anche se Dkan ha molte buone opportunità di sviluppo, portando le funzioni di Ckan in un sito basato su Drupal. Le piattaforme che restano, ossia Ckan, Socrata e OGDI DataLab, sono da valutare in base alla situazione di deploy. (Server disponibili, familiarità con certi sistemi operativi, disponibilità di fondi, ecc.)
Per quanto riguarda i costi delle soluzioni OGDI DataLab e Socrata, la licenza di Microsoft Windows Azure, comporta un costo minimo di €372,35/mese, mentre la regione Lombardia ha speso €9300 per commissinoare la creazione del portale, ed i costi di gestione per l'anno 2012 sono stati €17.700. Da valutare è l'adozione della piattaforma Socrata Socrata Open Data Server Community Edition.

Il processo di trasformazione delle informazioni in documenti conformi agli standard open data di livello 4 e 5 abbiamo visto che è dipendente dagli strumenti software usati e possono essere identificati due approcci operativi differenti:

- Analisi dei dati, creazione file RDF, esportazione su portale
- Mapping database, link su portale

La prima soluzione prevede un maggior numero di componenti, la necessità di un'area per la memorizzazione fisica dei file RDF prodotti e un accesso ai dati indiretto. Il vantaggio principale di tale approccio è dato dall'universalità del

procedimento e dalla scarsa dipendenza dal tipo di dato, in quanto in fase di sviluppo si ha il controllo completo del processo di trasformazione.
La seconda soluzione proposta ha il vantaggio dell'accesso diretto ai dati e una maggiore semplicità nella distribuzione ma è fortemente relazionata alle funzionalità messe a disposizione dagli strumenti di mapping.
Entrambi gli approcci, sono comprovati e stabilmente usati a livello internazionale nella gestione di Open Data di livello 4 e 5.

6. Sitografia:

Documenti rilasciati pubblicamente dalla DigitPa :

CdC-SPC-GdL6-InteroperabilitaSemOpenData_v2.0.doc

LG_Val_PSI_v1.0.doc

Siti Web:

http://opendefinition.org

http://www.opendatahandbook.org

http://5stardata.info

https://github.com

http://www.opendata71.fr

http://docs.thedatatank.com/4.1

http://docs.ckan.org/en/latest/

http://www.wikipedia.org

http://www.socrata.com

http://wifo5-03.informatik.uni-mannheim.de/pubby/

http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/

http://www.openrdf.org/

http://d2rq.org/

http://opendatahub.it

http://datiopen.it

http://linkedopendata.it