Open Data: la piattaforma di dati aperti per il Linked Data

GIUSEPPE RIZZO[*]

FEDERICO MORANDO[#]

JUAN CARLOS DE MARTIN[+]

SOMMARIO: *1. Introduzione – 2. Interoperabilità tra i dati – 3. Tecnologie per favorire l'interoperabilità: RDF e SPARQL – 3.1. Resource Description Framework (RDF) – 3.2. SPARQL Protocol and RDF Query Language (SPARQL) – 4. Linee guida per la pubblicazione dei dati – 5. Pubblicare i Linked Data – 5.1. Pubblicazione di dati strutturati – 5.2. Pubblicazione di dati non strutturati: testo – 6. Open Data incontra il Linked Data – 6.1. Open Data e Open Licenses – 7. Creare i link verso archivi esterni – 8. Conclusioni*

1. INTRODUZIONE.

Il *World Wide Web*, o semplicemente *Web*, ha radicalmente alterato il modo di condividere i contenuti ed informazioni di ogni tipo, facilitando la pubblicazione degli stessi da parte degli autori e l'accesso per i lettori. Questi ultimi hanno potuto contare sull'abbattimento delle barriere di accesso, potendo così navigare liberamente tra una molteplicità di oggetti immateriali creati dall'uomo ("artefatti") utilizzando il *Web* come uno spazio di informazione globale. Il *World Wide Web* si presenta dunque come strumento atto a veicolare informazioni eterogenee, le quali possono essere raccolte e catalogate all'interno di quello che può essere definito il più vasto archivio della conoscenza umana. Le informazioni sono intrecciate tra

[*] L'autore è dottorando di ricerca in Ingegneria Informatica e dei Sistemi presso il Dipartimento di Automatica ed Informatica del Politecnico di Torino.

[#] L'autore è reserch fellow presso il centro NEXA del Politecnico di Torino.

[+] L'autore è co-fondatore e co-direttore del centro NEXA e professore associato al Politecnico di Torino.

loro, mediante dei collegamenti ipertestuali, chiamati *hyperlink*. L'indicizzazione degli *hyperlink* permette di creare delle associazioni tra le informazioni e le parole chiave, favorendo al lettore un punto di accesso ad ulteriore conoscenza.

Il *Web*, dalla sua nascita, ha conservato queste peculiarità, che lo hanno reso una tecnologia di grande successo ed in continua crescita[1]. Tuttavia, sino ad ora, esso è stato utilizzato principalmente per condividere artefatti come i documenti complessi. Sfruttare adeguatamente i principi che il Web ha seguito per anni, ma al fine di condividere dati elementari, è uno dei principali temi attuali di ricerca nella comunità degli studiosi di Internet. L'unità di base di scambio dell'informazione passa così dal documento al dato grezzo, *raw*. Ciò permette di separare in modo embrionale il contenuto di un artefatto in tante parti, potendo collegare tra loro o con altre informazioni i dati presenti all'interno dell'artefatto stesso, al fine di inferire nuove informazioni o creare nuovi artefatti. Adottando l'immagine proposta dall'inventore del *Web*, Tim Berners-Lee, possiamo affermare che l'artefatto è qualcosa che possiamo semplicemente leggere, mentre un dato può essere processato in differenti modi per creare nuova informazione. In questo modo, il *Web* diventa uno spazio di condivisione globale dei dati, il *Web of Data*. La specializzazione introdotta dal *Web of Data* permette di strutturare i dati della risorsa, in modo che essi possano essere letti separatamente e all'occorrenza, per rispondere alle richieste dell'utente, aggregati con altri. Tipicamente un insieme di dati contiene una conoscenza circa un dominio particolare, come film, musica o viaggi. Se questi insiemi di dati sono tra loro interconnessi, una macchina può navigare attraverso le relazioni create, senza subire l'effetto del rumore.

Nel Web a cui siamo abituati, le risorse sono descritte in HTML (*Hypertext Markup Language*) e pubblicate in spazi di aggregazione, quale i siti. La domanda da parte dei lettori di informazioni processabili dalle macchine ha spinto i creatori a pubblicare dati strutturati: esempi di questo sono i dati che spesso provengono da tabelle di vario tipo e sono poi esposti sul Web in formato CSV (C*omma Separated Value*), XML (*eXtensible Markup Language*) o

[1] Bizer C. , Heath T. and Berners-Lee Tim, *Linked Data - The Story So Far*, in “The International Journal on Semantic Web and Information Systems”, pp. 1-22, vol. 5, issue 3, 2009.

tramite delle tabelle nelle pagine HTML, sacrificando gran parte della semantica della risorsa iniziale. La natura delle relazioni tra due risorse collegate è implicita, ossia l'informazione di collegamento è annegata all'interno della risorsa stessa, poiché l'HTML non è un linguaggio sufficientemente espressivo per gestire le inferenze tra differenti risorse. Da qui l'esigenza di poter descrivere i dati in maniera strutturata sfruttando formati in grado di gestire la relazione tra gli stessi. Per rispondere a questa esigenza, il *Web* sta evolvendo da spazio di raccolta e collegamento di documenti ad uno in cui i dati grezzi contenuti nelle risorse sono collegati o collegabili facilmente dalle macchine. Questa evoluzione è conosciuta con il nome di *Linked Data*. Il *Linked Data* è una metodologia che permette di aggregare e collezionare dati provenienti da fonti distribuite[2]. Per rendere completamente accessibili questi dati al mondo del Web, i dati stessi devono essere pubblicati sotto condizioni d'uso "aperte" (o "libere"), che ne consentano consultazione, navigazione (con qualsiasi mezzo e anche tramite *deep-linking*[3]) e aggregazione.

L'*Open Data*, letteralmente "dati aperti", è la corrente di pensiero (e il relativo "movimento") che cerca di rispondere all'esigenza di poter disporre di dati legalmente "aperti"[4], ovvero liberamente (ri-)usabili da parte del fruitore, per qualsiasi scopo[5]. L'obiettivo dell'Open Data può essere raggiunto per legge, come negli USA dove l'informazione generata dal settore pubblico federale è in pubblico dominio[6], oppure per scelta dei detentori dei diritti, tramite opportune

[2] Bizer, C., Heath, T., Ayers, D., Raimond, Y., *Interlinking Open Data on the Web in* "The 4th European Semantic Web Conference", 2007.

[3] Navigazione esaustiva lungo tutto l'albero delle relazioni.

[4] Per una definizione di dati aperti (o, più genericamente, di conoscenza aperta) si veda l'Open Knowledge Definition: http://www.opendefinition.org/. Una definizione sostanzialmente equivalente (ma che pone l'accento sull'aggettivo "free", anziché su "open") è proposta tramite la definizione di Free Cultural Works: http://freedomdefined.org/Definition.

[5] "Per qualsiasi scopo", beninteso, compatibile con la normativa vigente (ad esempio, un elenco di indirizzi email non potrà mai essere utilizzato per l'invio di spam) e nel rispetto di eventuali condizioni, le quali – secondo l'approccio Open Data – non possono prevedere che un diritto di attribuzione all'autore o citazione della fonte e, eventualmente, l'obbligo di utilizzare la stessa licenza del dataset originario per la pubblicazione di eventuali dataset modificati (cosiddetta clausola "virale" o "copyleft").

[6] Cf. US Copyright Act, § 105.

licenze[7]. Per motivare la necessità di avere dei dati in formato aperto, possiamo usare una comparazione del tipo: l'*Open Data* sta al *Linked Data*, come la rete *Internet* sta al *Web*. L'*Open Data*, quindi, è l'infrastruttura (o la "piattaforma") di cui il *Linked Data* ha bisogno per poter creare la rete di inferenze tra i vari dati sparsi nel Web. Il *Linked Data*, in altre parole, è una tecnologia ormai abbastanza matura e con grandi potenzialità, ma ha bisogno di grandi masse di dati tra loro collegati, ossia "linkati", per diventare concretamente utile. Questo, in parte, è già stato ottenuto ed è in corso di miglioramento, grazie a progetti come *DBpedia*[8] o *FreeBase*[9]. In parallelo ai contributi delle community online, un altro tassello importante – una sorta di "bulk upload" molto prezioso – potrebbe essere dato dalla disponibilità di grosse masse di dati pubblici, idealmente anche già linkati dalle istituzioni stesse – o comunque messi a disposizione in modo strutturato – che aiutino a raggiungere una "massa critica" di dati collegati. A partire dal substrato, rappresentato dalla disponibilità di fatto dei dati e dalla loro piena riutilizzabilità (in modo legale), il *Linked Data* può offrire una potente rappresentazione degli stessi, in termini di relazioni (collegamenti): in questo senso, *Linked Data* ed *Open Data* convergono e raggiungono la loro piena realizzazione nell'approccio *Linked Open Data*.

2. Interoperabilità dei dati.

L'interoperabilità tra i dati è uno dei vantaggi più importanti del modello *Open Data*. I dati, se isolati, hanno poco valore; viceversa, il loro valore aumenta sensibilmente quando differenti archivi di dati, anche detti dataset, prodotti e pubblicati in modo indipendente da diversi soggetti, possono essere incrociati o aggregati liberamente dal fruitore. Questo processo di aggregazione è generato da applicazioni prodotte per creare delle "viste" (presentazioni) sul modello dei dati a disposizione. Facendo un'analogia con il mondo del Web, questa tecnica permette di discernere la forma dal contenuto, di fatto implementando la separazione degli interessi (o SoC: S*eparation of Concerns*). Le applicazioni, di valore sociale e/o economico (come

[7] Si veda, anche per riferimenti ad eventuali approfondimenti, il § 6.1.

[8] http://dbpedia.org/

[9] http://www.freebase.com/

modelli di business), sfruttano quello che può essere immaginato come un grande archivio aperto e distribuito per offrire servizi, partendo da particolari viste. Tali viste possono essere di diverso tipo e non univoche, data la natura stessa della vista scelta dal lettore o consumatore del dato. Uno dei maggiori vantaggi dell'approccio *Linked Open Data* sta nella possibilità di ricostruire un proprio punto di vista (eventualmente, un punto di vista che ambisca ad essere il più possibile "imparziale").

Benché la tecnica dell'aggregazione dei dati al fine di trovare l'informazione desiderata non risulti essere innovativa nel settore dell'*Information Retrieval* (IR)[10], essa rappresenta un'innovazione se inserita nel contesto del *Linked Open Data*. In esso, le linee guida, *best practises*, della pubblicazione delle risorse hanno spinto i creatori o fornitori dell'informazione a pubblicare insiemi di dati grezzi. La forma del dato non viene decisa dal fornitore dell'informazione, ma dal lettore. Il fornitore, dalla sua, decide solo il contenuto. Per consentire il riuso dei dati occorre, quindi, poter combinare e mescolare liberamente i *dataset*. Occorre cioè collegare i dati tra loro, stabilendo un collegamento, *link*, diretto quando i dati (possibilmente provenienti da diverse sorgenti) si riferiscono a oggetti identici o comunque relazionati tra loro. Tale collegamento diretto si manifesta come la possibilità di "saltare" da un dataset all'altro, ad esempio quando si vuole accedere ai dati (come i dettagli su una particolare entità) che non si posseggono all'interno. È proprio grazie a questo scenario che le possibilità offerte dall'aggregazione dei dati risultano espanse, in modo che non ha precedenti.

Uno degli scenari di uso e consumo degli Open Data è quello delle pubbliche amministrazioni. Esse detengono una grande quantità di dati, che potrebbe (e a detta di molti commentatori dovrebbe, in quanto raccolta con risorse e per finalità pubbliche[11,12]) essere messa a disposizione del pubblico (al netto – ovviamente, e solo per fare un esempio di cautela necessaria – di informazioni che potrebbero

[10] Scienza che si occupa di studiare le tecniche per il recupero mirato dell'informazione in formato elettronico.

[11] G. Aichholzer, H. Burkert, *Public Sector Information in the Digital Age. Between Markets, Public Management and Citizen's Right*, In EE, Celtenham, 2004.

[12] B. Ponti, *Il regime dei dati pubblici. Esperienze europee e ordinamento nazionale*, In Maggioli Editore, Sant'Arcangelo di Romagna, 2008.

mettere a rischio la privacy dei cittadini). Gli attori di questo scenario sono: la pubblica amministrazione, creatore o collettore delle informazioni, il cittadino fruitore e ri-utilizzatore dell'informazione (più eventuali intermediari, di tipo associativo e/o imprenditoriale, che possono realizzare le applicazioni e le "viste" di cui si è accennato). La sottosezione successiva esporrà i vantaggi derivanti dal possesso di dati strutturati per abilitare l'interoperabilità.

3. Tecnologie per abilitare l'interoperabilità: RDF e SPARQL.

L'interoperabilità, che è alla base del *Linked Data*, ha bisogno di un insieme di strumenti in grado di agevolare la navigazione ed il collegamento tra le risorse da parte delle applicazioni. La definizione di standard di comunicazione e di scambio di dati è necessaria per rendere il *Linked Data* funzionante. RDF e SPARQL sono i due modelli sui quali il *Linked Data* si basa. In questa sezione discuteremo più in dettaglio sugli aspetti tecnici di entrambi i modelli.

3.1. Resource Description Framework (RDF)

RDF definisce un modello per rappresentare le relazioni tra le informazioni descrittive di risorse *Web* senza perdita di semantica. Benché RDF permetta di definire un modello dei dati, esso non definisce come i dati siano identificati. In questo scenario, l'uso delle URI (*Uniform Resource Identifier*), identificatori globali dello spazio del Web, diviene parte integrante della tecnologia RDF. Ricapitolando, il dato viene rappresentato mediante delle URI mentre i dati stessi vengono tra loro collegati mediante il modello RDF. La definizione delle relazioni tra le URI viene indicata con il termine di dichiarazione, *statement*, e si compone di tre parti: soggetto, predicato ed oggetto. Essendo una relazione ternaria, la dichiarazione prende il nome di tripla e viene rappresenta usando il formalismo *(s,p,o)*. L'analogia con la linguistica è forte, infatti il concetto di soggetto, predicato ed oggetto sono propri del modo di comunicare degli umani. Per capire questa analogia, proviamo a modellare in RDF l'espressione: "Giacomo Leopardi scrisse Il sabato del villaggio". Il soggetto di questa frase è "Giacomo Leopardi", l'oggetto è "Il sabato del villaggio", infine il predicato è "scrisse". Questa espressione è

comprensibile agli umani, ma occorre renderla comprensibile alle macchine utilizzando una dichiarazione RDF.

Come abbiamo accennato prima, RDF descrive il modello, ossia come la dichiarazione deve essere formattata. Compito invece delle URI è quello di identificare in modo univoco nello spazio del *Web* gli elementi della tripla *(s,p,o)*. Occorre quindi definire quali sono gli identificatori dei tre elementi della tripla (come vedremo in seguito, l'uso dell'URI è fondamentale per il soggetto ed il predicato, mentre l'oggetto può essere rappresentato sia da una URI, ma anche da un valore testuale, chiamato "literal"). La scelta degli identificatori non è casuale, anzi richiede particolare attenzione in fase di definizione. Supponiamo di definire le seguenti URI:

- http://www.esempio.it/autore/Leopardi: soggetto,
- http://www.esempio.it/scrisse: predicato,
- http://www.esempio.it/poesia/Il_sabato_del_villaggio: oggetto.

Le URI riportate sono solo a titolo di esempio e non hanno nessun validità all'interno dello spazio del Web. Proviamo adesso a modellare il seguente esempio usando uno schema, grafo, come quello riportato in Figura 1.

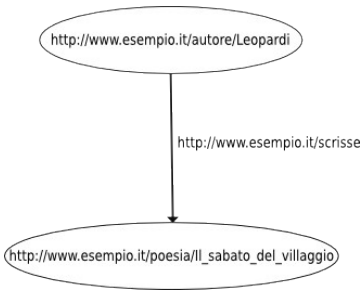

Figura 1: Una semplice dichiarazione RDF in cui il nodo http://www.esempio.it/autore/Leopardi inferisce mediante l'arco http://www.esempio.it/scrisse la relazione con http://www.esempio.it/poesia/Il_sabato_del_villaggio.

RDF modella le dichiarazioni mediante nodi ed archi in un grafo. Il soggetto e l'oggetto sono rappresentati mediante dei nodi, mentre il predicato è rappresentato dall'arco che unisce le due entità. Molteplici collegamenti tra dichiarazioni, "triple linkate", creano grafi complessi. Riprendiamo l'espressione modellata in Figura 1: "Giacomo Leopardi

scrisse Il sabato del villaggio" ed aggiungiamo le dichiarazioni "Giacomo Leopardi nacque il 29 Giugno del 1978" e Giacomo Leopardi conobbe Pietro Giordani". Il grafo equivalente è riportato in Figura 2.

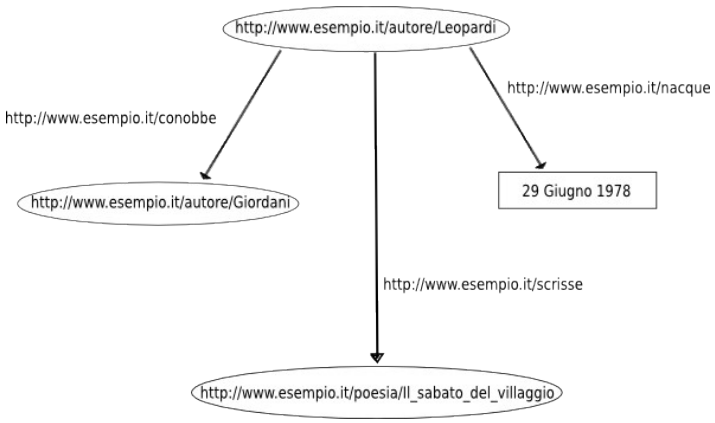

Figura 2: Esempio di query complessa, in cui possiamo distinguere nodi di tipo URI e un nodo di tipo "literal" che rappresenta la data di nascita di Leopardi.

Un grafo o un insieme di grafi creano un archivio di dati collegati mediante dichiarazioni RDF. Al fine di condividere questo archivio, il grafo prodotto viene serializzato[13] usando uno dei seguenti formati: XML/RDF, N-Triples, Turtle. Sebbene l'enumerazione non sia esaustiva, i tre formati coprono gran parte della formattazioni maggiormente usate. Essi differiscono solo sul modo con cui rappresentano le triple e non sul contenuto trasportato. In Figura 3 riportiamo la serializzazione dell'esempio precedente, usando il formato Turtle. Nella sottosezione successiva discuteremo come fare ad inferire i dati da archivi sparsi nella rete al fine di ricostruire un grafo di relazioni in grado di rispondere a specifiche domande, dette query.

3.2. SPARQL Protocol and RDF Query Language (SPARQL)

SPARQL è lo standard per interrogare archivi conformi al formalismo RDF sparsi nel *Web*[14]. Una delle specifiche di SPARQL

[13] Operazione che porta alla salvataggio di una descrizione di un oggetto in un file mediante un opportuno formato.

[14] http://www.w3.org/TR/rdf-sparql-query/

prevede che l'archivio sia consultabile mediante il protocollo di comunicazione HTTP[15], ovvero che sia in grado di rispondere a determinate richieste fatte attraverso HTTP. Un archivio di dati può fornire un punto di accesso SPARQL sul Web; questi punti di accesso prendono il nome di SPARQL "endpoint".

```
@prefix es: http://www.esempio.it/ .
@prefix aut: URI: http://www.esempio.it/autore/ .
@prefix poesia: URI: http://www.esempio.it/poesia/ .

aut:Leopardi
            es:nacque "29 Giugno 1978" ;
            es:conobbe aut:Giordani ;
            es:scrisse poesia:Il_sabato_del_villaggio .
```

Figura 3: Esempio di serializzazione in Turtle del grafo riportato in Figura 2.

Prendiamo l'esempio riportato in Figura 2, in cui il nodo http://www.esempio.it/autore/Leopardi è il nodo radice del grafo ed è collegato agli altri nodi mediante degli archi. Fare una inferenza in SPARQL, vuol dire navigare il grafo lungo gli archi ed i nodi, fino a soddisfare il vincolo dell'inferenza. Le richieste vengono serializzate mediante il formato Turtle, mentre il risultato della richiesta è generalmente una tabella. Importante notare, come la richiesta può inferire i valori dei nodi, così come i valori degli archi.

Si prenda come esempio il grafo rappresentato in Figura 2. Supponiamo di voler "inferire tutte le poesie delle quali Leopardi è stato l'autore". Figura 4 mostra un esempio di query serializzata in base al formato Turtle, mentre in Figura 5 è riportata la risposta.

```
PREFIX es: <http://www.esempio.it/>

SELECT ?poesia
WHERE {     http://www.esempio.it/autore/Leopardi
            es:scrisse
            ?poesia
        }
```

Figura 4: Esempio di query SPARQL serializzata mediante il formato Turtle. In essa, il valore cercato (parametro) è indicato dalla coppia punto interrogativo ed etichetta che si vuol dare all'arco o nodo "?poesia".

[15] E' il protocollo sul quale si basa l'intera infrastruttura del Web.

poesia
http://www.esempio.it/poesia/Il_sabato_del_villaggio

Figura 5: Risposta alla query riportata nella precedente figura. La risposta è rappresentata in una tabella, in cui l'unico valore presente è il nome del nodo cercato.

Una richiesta SPARQL ci permette di navigare l'albero delle relazioni e di ottenere una tabella il cui numero di colonne è dipendente dal numero di parametri richiesti. Cerchiamo di rappresentare in dettaglio cosa vuol dire navigare l'albero delle relazioni, facendo un esempio più complesso del precedente. Supponiamo di voler: "inferire tutti gli archi e tutti i nodi che hanno come nodo padre Leopardi". Figura 6 mostra l'esempio di query ed infine Figura 7 la risposta ottenuta.

```
PREFIX aut: <http://www.esempio.it/autore/>

SELECT ?p  ?o
WHERE {    aut:Leopardi
           ?p
           ?o
       }
```

Figura 6: Esempio di query complessa, nella quale i parametri richiesti sono tutti gli archi (predicati) e tutti i nodi (oggetti) che hanno un'inferenza di primo livello con il nodo "aut:Leopardi".

p	o
http://www.esempio.it/nacque	29 Giugno 1978
http://www.esempio.it/conobbe	http://www.esempio.it/autore/Giordani
http://www.esempio.it/scrisse	http://www.esempio.it/poesia/Il_sabato_del_villaggio

Figura 7: Esempio di risposta alla query riportata in Figura 7. La colonna a sinistra "p" racchiude tutte le informazioni sugli archi che sono collegati al nodo "act:Leopardi", mentre la colonna a destra, "o", i valori dei nodi che sono collegati mediante "?p" ad "act:Leopardi".

La sezione successiva esporrà le linee guida per la pubblicazione dei dati in modo che essi siano interoperabili, usando gli standard

elencati in questa sezione.

4. Linee guida per la pubblicazione dei dati.

Le linee guida per la pubblicazione dei dati sul Web sono state proposte da Tim-Berners Lee, il quale ha enunciato una serie di principi atti a favorire la scoperta e l'utilizzo dei dati, definendo delle grammatiche di comunicazione comuni[16]. Inoltre, queste regole si basano sull'idea della durabilità del dato: esso infatti può anche essere tempo-variante (ossia assumere nuovo valore informativo), ma deve essere sempre raggiungibile al medesimo punto. Definito un punto dello spazio del *Web*, esso deve rimanere tale, al fine di favorire le aggregazioni o associazioni con altri dati. Le linee guida possono essere enunciate dai seguenti principi:

1. uso di URI per l'identificazione delle risorse;
2. uso del protocollo HTTP come protocollo di comunicazione per lo scambio di dati nel *Web*;
3. uso degli standard RDF e SPARQL per la presentazione dei dati e per la loro pubblicazione nel *Web*;
4. creazione di collegamenti tra le varie URI al fine di scoprire nuovi dati.

Il primo principio elegge l'URI a identificativo per tutto quello che è referenziato nel *Web*, quindi non solo documenti o contenuti digitali, ma anche oggetti del mondo reale (persone, luoghi) e concetti astratti (relazioni, stati d'animo). Il *Web* è lo spazio di tutte le cose (*Web of things*), le quali sono rappresentate mediante dei dati (*Web of Data*). Nell'attuale architettura del *Web*, le URI sono usate per combinare identificativi globali univoci con un meccanismo di conversione dei nomi centralizzato, il *Domain Name System* (DNS). Sebbene questo presenti notevoli problemi, per la natura centralizzata del sistema, esso permette di navigare il *Web* tramite l'uso delle URI.

Il secondo principio, invece, definisce che l'unica grammatica di comunicazione valida per il *Web* è il protocollo HTTP. Uno degli obiettivi del *Linked Data* è quello di fornire l'interoperabilità tra i vari archivi di dati, permettendo a tutte le applicazioni di accedere ai contenuti in maniera standard: HTTP rappresenta questo standard di comunicazione e scambio di dati.

[16] http://www.w3.org/DesignIssues/LinkedData

Il terzo principio definisce il modello di presentazione dei dati da usare RDF e il modello di pubblicazione basato sui grafi SPARQL, in merito ai quali rimandiamo al § 3.

Infine, il quarto principio sostiene l'uso degli *hyperlink* non solo tra i documenti del *Web*, come avveniva nel *Web* dei documenti, ma tra qualsiasi tipo di oggetto, persone, luoghi, documenti digitali. Inoltre, i collegamenti ipertestuali usati nel *Web* erano degli atomici puntatori a punti dello spazio dei documenti. Non avevano nessuna descrizione, semantica, riguardo il collegamento. Nel *Linked Data*, i collegamenti assumono un importante ruolo di trasporto di semantica. Per distinguersi dal *Web* dei documenti, il collegamento nel *Linked Data* viene sempre definito dalla termine *RDF link*.

5. Pubblicare i Linked Data.

La pubblicazione dei dati sul Web richiede l'adozione delle linee guide riportate nella sezione precedente. Esse definiscono il modello dei dati, di rappresentazione e di serializzazione. L'aspetto critico di queste linee guida è rappresentato dalla retro-compatibilità con i modelli sviluppati per archivi di dati già esistenti. Questo è lo scenario che la maggior parte delle aziende private e delle pubbliche amministrazioni si trovano ad affrontare. Esse possiedono un grosso patrimonio di dati, conservati perlopiù in tabelle relazionali o in file tabulari, ad esempio file *Comma Separated Value* (CSV), *eXtensible Markup Language* (XML), *Calculator* (CALC) o semplice testo[17]. Quindi, da un lato l'adozione delle linee guida può risultare semplice per chi costruisce nuovi archivi e decide di pubblicarli sul *Web*, dall'altro richiedono un ulteriore elaborazione per archivi già esistenti. In questa sezione analizzeremo alcune strategie per la presentazione e serializzazione di dati di archivi già esistenti.

5.1. Pubblicazione di dati strutturati

Generalmente, i dati strutturati sono quelli archiviati in database relazionali (SQL) o in file tabulari (CSV, XML, CALC, etc.). Essi trasportano l'informazione di Entità, per esempio nel caso

[17] Per quanto non sia infrequente l'uso di formati più o meno ad hoc per immagazzinare dati, l'esportazione in uno o più dei formati qui menzionati è normalmente agevole.

dell'anagrafica esse trasportano informazioni associate ad una Persona (entità) e per ogni entità trasportano le informazioni di domicilio, residenza, etc.; queste ultime sono attributi dell'entità. In generale, un dato strutturato è facilmente interpretabile da un'applicazione e questo rappresenta un passo fondamentale per l'interoperabilità. Prima della pubblicazione il dato, seppur strutturato, deve essere convertito usando le linee guida enunciate in precedenza. Il processo di pubblicazione si può effettuare secondo due differenti metodologie: conversione di dati presi da database SQL e conversione partendo da fogli di calcolo. Nel primo caso si effettua un processo di conversione mediante strumenti chiamati RDBtoRDF (*Relational Database to RDF*) come ad esempio Virtuoso[18], D2R[19] Mapper, mentre nel secondo caso si usano strumenti di trasformazione da tabelle in RDF come ad esempio RDFizers[20]. Alla fine del processo di conversione, il dato può essere pubblicato sul Web usando sia lo schema di presentazione del dato "raw" (mediante SPARQL endpoint) sia mediante un'interfaccia Web. Figura 8 riassume l'intero processo, partendo dal dato strutturato fino al dato "linkable".

5.2. Pubblicazione di dati non strutturati: testo

Una delle sfide più importanti che la comunità del *Machine Learning* ha affrontato è rendere le applicazioni abili ad interpretare il testo. Quest'ultimo per sua natura, è ricco di semantica la quale, però, è implicitamente espressa. Rendere un testo strutturato richiederebbe l'intervento di un umano, il quale, se è cultore della materia, può sintetizzarlo usando la forma (chiave, valore)[21]. Solo dopo questo processo il testo può essere comprensibile alle applicazioni. In questa direzione, gli estrattori di Entità (Named Entitiy extractor) svolgono un ruolo fondamentale: essi sono in grado di annotare il testo mediante dei concetti e di associarli a delle risorse. Se questi strumenti sono combinati con l'uso del Web, le risorse associate hanno delle URI. Pertanto, data una chiave, ossia una parola che riassume il testo,

[18] http://virtuoso.openlinksw.com/

[19] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[20] http://openstructs.org/resources/rdfizers

[21] In cui la chiave è l'entità, ad esempio il nome di Leopardi ed il valore è associato ad un valore testuale che descrive sinteticamente il testo (e.g. biografia).

si associa un URI. Sistemi in grado di svolgere queste funzioni sono DBpedia Spotlight[22], OpenCalais[23]. Ritornando all'obbiettivo di partenza, il testo così convertito può essere pubblicato utilizzando gli stessi strumenti enunciati nella sottosezione precedente.

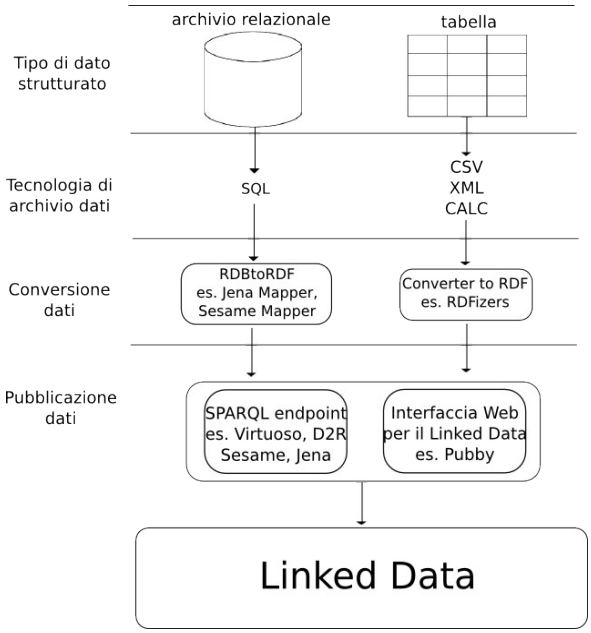

Figura 8: Dal dato strutturato al dato "raw" che entra a far parte del Linked Data. Dal grafico si evince come il Linked Data si basa sull'infrastruttura dei tanti dati "raw" presi da differenti archivi.

6. Open Data incontra il Linked Data.

Il tema dell'Open Data è di estrema attualità per le pubbliche amministrazioni di tutto il mondo. In questa direzione, importanti sforzi sono stati fatti da molteplici governi per dar vita a spazi di informazione quanto più possibile oggettiva (ovvero, non sottoposta a filtri interpretativi arbitrari). Questa tendenza può essere frutto di una precisa decisione relativa alla trasparenza dell'attività politica; ma può

[22] http://dbpedia.org/spotlight

[23] http://www.opencalais.com/

anche essere semplicemente legata al desiderio di sviluppare il potenziale, anche economico, relativo al ri-uso dei dati in nuovi contesti. In ogni caso, aprire i dati significa scegliere una tecnologia atta a veicolarli. La scelta è ricaduta sul *Web* perché è una tecnologia largamente diffusa. Quasi la totalità degli abitanti del pianeta sono dei "web-nauti" che navigano con diversi strumenti: calcolatori e strumenti portatili come notebook, tablet e smartphone.

Una delle priorità di questo movimento però è fornire l'informazione al cittadino, senza filtri. Questa informazione può essere definita come *raw* e può essere usata per essere "linkata" con altre informazioni. Questa necessità di unire le informazioni ha spinto le istituzioni coinvolte in questo processo ad abbracciare l'idea del *Linked Data*. I dati hanno notevole valore, ad esempio perché raccontano fedelmente l'attività di un ministero o di un'agenzia statale, ma assumono ancor più importanza se sono aggregati tra loro. Di primaria importanza è, quindi, l'interoperabilità[24]. In questo scenario possiamo citare il governo Americano e il governo Inglese, che prima di altri si sono messi in grado di rendere questo possibile. Il governo Americano ha dato alla luce il portale data.gov[25], mentre il governo Inglese data.gov.uk[26]. Entrambi i portali sono delle viste di cataloghi di dati aperti. Essi si basano su due differenti tecnologie per gestire gli archivi di dati (Figura 9).

portale	catalogo di dati
data.gov	non dichiarato
data.gov.uk	ckan.org

Figura 9: Il portale data.gov usa un catalogo di dati che non ha rilasciato, mentre data.gov.uk usa il catalogo di dati ckan.org.

6.1. Open Data e Open Licenses

Il *Linked Data* ha bisogno di grandi masse di informazioni

[24] http://www.w3.org/DesignIssues/GovData.html

[25] http://www.data.gov/

[26] http://data.gov.uk/

“linkate” (o almeno “linkabili”), per diventare effettivamente utile. Come accennato, questo traguardo è sempre più vicino, grazie in particolare a progetti come DBpedia, che linkano contenuti aperti generati dagli utenti (di Wikipedia), li rappresentano secondo il formalismo RDF e li rimettono a disposizione della collettività. Tuttavia, la disponibilità di grandi moli di dati già strutturati potrebbe dare un altro contributo cruciale al *Linked Data*: i dati messi a disposizione dalle pubbliche amministrazioni, dunque, potrebbero rappresentare una risorsa di immenso valore, sia direttamente, sia strumentalmente all'affermazione del *Linked Data*. Perché questi dati siano davvero utili, tuttavia, devono essere disponibili come *Open Data*.

Qualora dei dati siano già rappresentati tramite il formalismo RDF ed esposti su piattaforme con punti di accesso adeguati (SPARQL endpoint), il fatto che i dati stessi siano “aperti/liberi” dal punto di vista legale – ovvero che siano legalmente modificabili e ripubblicabili – potrebbe apparire superfluo. In realtà, è facile argomentare che non sia questo il caso, sia perché RDF e SPARQL potrebbero, in futuro, essere soppiantati da o affiancati ad altre tecnologie, così condizionando l'utilità dei dataset in questione ad un tempestivo aggiornamento da parte dei detentori dei diritti, sia perché la possibilità di rappresentare dati in altri formati e/o di renderli accessibili su altre piattaforme (o semplicemente di realizzarne una copia su un'altra macchina) è spesso tecnicamente molto utile, se non indispensabile. Tutto ciò senza menzionare il rischio che alcuni dati possano semplicemente scomparire dalla Rete, insieme all'organismo che li ha messi originariamente a disposizione.

Ciò detto, quel che è più rilevante è che spesso i dati disponibili online non sono affatto rappresentati secondo le metodologie raccomandate dal W3C[27] e descritte in questo articolo. Inoltre, è frequente che i dataset così come sono stati pubblicati contengano dati spuri ed errori o, ancora, abbiano semplicemente la possibilità di essere ulteriormente arricchiti, magari tramite un'attività di *editing* manuale. Dunque, siccome per trasformare un dataset strutturato in un dataset *linked* si crea un'opera derivata, è necessario avere l'autorizzazione del detentore dei diritti e vi sono evidenti vantaggi in

[27] http://www.w3.org/

termini di costi di transazione, qualora tale autorizzazione sia concessa automaticamente ed *ex ante*, tramite una licenza standard.

Non a caso, i dati presenti su portali quali data.gov e data.gov.uk sono liberamente riutilizzabili. I primi, come accennato, sono in pubblico dominio grazie alle previsioni del Copyright Act statunitense[28]. Mentre i secondi sono disponibili ai sensi della Open Government Licence[29], che include semplicemente una clausola di attribuzione ("acknowledge the source of the Information"). Ciò implica, in particolare, che i dati disponibili sui portali data.gov e data.gov.uk siano mescolabili con dataset licenziati tramite le licenze standard offerte da Creative Commons ed Open Data Commons[30] ed utilizzate da molte *communities* online.

7. Creare i link verso archivi esterni.

Nelle sezioni precedenti abbiamo descritto il processo di conversione e pubblicazione dei dati, alla fine del quale siamo in grado di esporre i nostri dati sul *Web* e di renderli disponibili alla navigazione e all'aggregazione con dati esterni, provenienti da altri archivi. Riportiamoci all'esempio del grafo in Figura 1: potremmo voler dire che il nodo Leopardi, contenuto nel nostro archivio la cui autorità è www.esempio.it, è "identico" a quello contenuto nell'archivio di dbpedia.org. Proviamo a modellare questo esempio con un nuovo grafo (Figura 10). L'inferenza creata suggerisce di navigare lungo i dati contenuti all'interno di DBpedia al fine di avere altre informazioni relative a Leopardi.

[28] Negli USA, lo scenario è assai più variegato per quanto riguarda i dati statali e delle città. Ad esempio, lo Stato della California offre numerosi dati sul portale data.ca.gov, che "salvo dove indicato diversamente, sono da considerarsi in pubblico dominio". "Furthermore, the US Government cannot vouch for any analyses conducted with data retrieved from Data.gov."

[29] http://www.nationalarchives.gov.uk/doc/open-government-licence/

[30] Nel caso del portale data.gov, i cui dati sono in pubblico dominio, questo si applica effettivamente a tutte le licenze offerte da CC ed ODC, incluse le "dediche al pubblico dominio" e/o rinunce ai diritti, quali Creative Commons Zero o la ODC Public Domain Dedication and Licence. Nel caso di data.gov.uk, invece, la compatibiltà riguarda solo le "licenze" in senso stretto, le quali includono tutte una clausola di attribuzione.

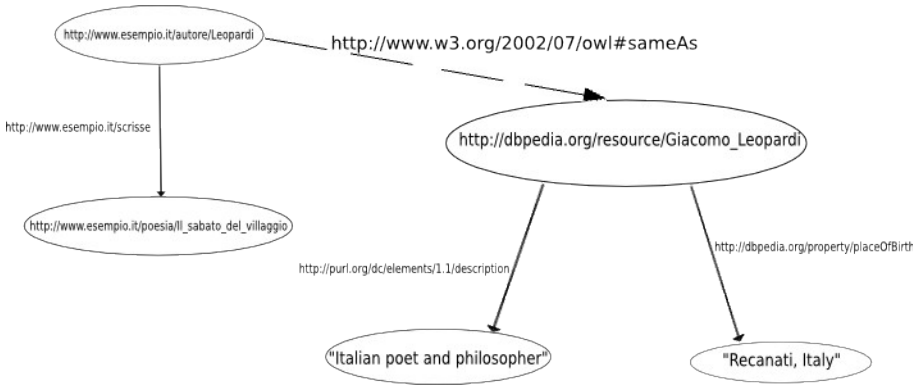

Figura 10: Il predicato http://www.w3.org/2002/07/owl#sameAs crea una relazione di identità tra il nodo definito nel nostro archivio di dati con quello presente in DBpedia.

In questo scenario un ruolo importante è svolto dall'uso della terminologia con la quale si definiscono le relazioni tra i nodi stessi. Infatti, sempre nella Figura 10, possiamo notare come il verbo *description* è modellato usando un verbo rappresentato all'interno del vocabolario Dublin Core. Mentre il verbo *placeOfBirth* è modellato usando una relazione espressa nel vocabolario definito da DBpedia stessa. Il ruolo dei vocabolari è quello di poter descrivere in maniera universale e non ambigua cosa significano nodi o predicati. Benché lo scopo di questo lavoro non sia quello di descrivere il ruolo dell'ontologia o del vocabolario, suggeriamo l'adozione di alcuni dei concetti rappresentati all'interno di quelli che sono stati individuati dalla comunità come vocabolari ed ontologie universali. A tal fine si riporta in Figura 11 una tabella riassuntiva.

nome	URI	Dettagli
dc	http://purl.org/dc/elements/1.1/	Dublin Core definisce un insieme di proprietà quali creator, title, description, le quali possono essere usate per descrivere qualsiasi tipo di risorsa (fisica o astratta).
foaf	http://xmlns.com/foaf/0.1/	FOAF definisce entità (rectius: Classi) quali Organization, Person, o Image e relazioni (rectius: Proprietà) quali familyName, firstName, birthday o depicts.
gr	http://purl.org/goodrelations	GoodRelations definisce entità quali BusinessEntity, Offering o ProductOrService e relazioni quali condition, serialNumber.

Figura 11: Tabella riassuntiva dei vocabolari e delle ontologie più usate dalla comunità.

8. Conclusioni.

Il tema dell'*Open Data* assume un'importanza strategica nel *Web*, che diventa sempre più un enorme archivio pubblico distribuito di *Linked Data*, in cui le applicazioni sono in grado di navigare automaticamente le relazioni tra le risorse ed inferire nuova conoscenza. Le linee guida di pubblicazione dei dati sul Web, sintetizzate in questo articolo, definiscono puntualmente il modello dei dati e come essi debbano essere serializzati al fine di creare uno spazio interoperabile. L'applicazione delle linee guide, tuttavia, resta un punto critico, perché richiede un processo di conversione che è più complesso della semplice pubblicazione di strutture tabulari. Benché la pubblicazione di Open Data in qualsiasi formato sia utile, come affermato dagli autori dello "Open Data Manual"[31], la stessa comunità dei fruitori è sempre più attenta alla qualità dei dati pubblicati. Essa è inversamente proporzionale allo sforzo che l'utente deve compiere per diventare fruitore dei dati. Gli autori di "Goverment Data and the Invisible Hand"[32], inoltre, sottolineano l'importanza di pubblicare i dati seguendo regole condivise dalla comunità dei fruitori, affinché quest'ultimi siano in grado di consumare i dati senza barrire all'acceso.

L'uso di licenze libere, inoltre, traduce l'interoperabilità tecnica in libertà di mescolare e migliorare la qualità dei dati. In particolare, quando un *data holder* non possiede le risorse finanziarie o tecniche per esporre direttamente *Linked Data*, offrire ad altri i diritti necessari ad operare questa conversione allevia, in parte, i limiti di una pubblicazione sub-ottimale.

In questo dibattito, un commento autorevole è quello di Tim Berners-Lee, il quale, citando una sua famosa frase "data is relationship, "we need more data", afferma che il *Web* ha sempre più bisogno di dati collegabili per poter crescere. Questa affermazione tende a sancire senza ombra di dubbio che la disponibilità di dati collegabili può abilitare di fatto la tecnologia del *Linked Data*.

[31] Dietrich D., Gray J., McNamara T., Poikola A., Pollock R., Tait J., Zijlstra T., "Open Data Manual, a Guide to open data", http://opendatamanual.org/.

[32] Robinson D. G., Yu H., Zeller W. P. and Felten E. W., "Goverment Data and the Invisible Hand", The Yale Journal of Law & Technology, vol. 11, Fall issue, pp . 160-176, 2009.

Ricorrendo all'analogia della vita di una pianta, i dati sono le radici che si trovano nel sottosuolo, il sottosuolo è il luogo ove i dati vengono inseriti, *Open Data*, mentre la pianta nasce dall'unione delle radici, *Linked Data*. L'incontro tra le radici crea nuove piante, dalle quali nasce un nuovo dato. Ma il dato più fertile è il dato "raw". Ovvero, spogliato di ogni tipo di decorazione, il dato deve essere reso pubblico al netto di altre informazioni che possono arricchire, ma allo stesso tempo condizionare la vista sul dato stesso. Solo in questo modo le pubbliche amministrazioni possono rendere trasparente il loro operato, ma – ancor più importante – il cittadino può creare le proprie viste, le quali possono essere usate per creare crescita sociale e partecipazione democratica, ma anche profitto[33]. Questioni aperte in questo dibattito restano il ruolo degli enti certificatori dei dati e l'uso di dizionari ed ontologie standard. Attualmente, la reputazione dell'autorità che condivide il dato è anche un'importante proxy della qualità del dato stesso. Inoltre, la comunità di ricerca sta affrontando le questioni legate alle modalità di condivisione dei dati, i detentori dei quali dovrebbero utilizzare quanto più possibile le stesse definizioni usate da altri fornitori di dati. Sebbene questo richieda un massiccio uso di vocabolari o ontologie distribuite e condivise da diverse comunità, e quindi il processo sia di difficile scalabilità, una maggiore uniformità potrebbe risolvere il problema dell'ambiguità nell'uso dei termini.

Vi è uno strato nascosto di dati sparsi per il *Web*, dati che generalmente sono usati per determinati, specifici obbiettivi. L'idea che è alla base del *Linked Data* è quella di aggregare questi dati nascosti, combinandoli e rendendoli accessibili ed interoperabili, al fine di creare nuova conoscenza. Grazie alla pubblicazione di *Open Data* secondo i precetti del *Linked Data*, numerosi individui e comunità hanno contribuito alla costruzione di un Web nel quale siamo già oggi e saremo sempre più in grado di supportare la fruizione, il riutilizzo e l'integrazione di dati originati da fonti distribuite ed eterogenee.

[33] Lynn A. Streeter, Robert E. Kraut, Henry C. Lucas Jr., Laurence Caby, "How open data networks influence business performance and market structure", Communications of the ACM, pp. 62-73, vol. 39, issue 7, 1996, doi: 10.1145/233977.233998. R. POLLOCK, *The Economics of Public Sector Information*, University of Cambridge, Cambridge, 2008.