Linked data: un nuovo alfabeto del web semantico

MAURO GUERRINI
Università di Firenze
mauro.guerrini@unifi.it

TIZIANA POSSEMATO
@Cult, Firenze
tiziana.possemato@atcult.it

Nei giorni 18 e 19 giugno si terrà presso l'Aula magna dell'Università di Firenze il seminario "Global interoperability and linked data in libraries" (per informazioni <www.linkedheritage.org/linkeddataseminar/>). Uno dei temi al centro del seminario è stato affrontato da Mauro Guerrini e Tiziana Possemato in occasione del convegno "I nuovi alfabeti della biblioteca" (Milano, Palazzo delle Stelline, 15-16 marzo 2012). Ci è sembrato utile pertanto proporre ai lettori di "Biblioteche oggi" il testo della loro relazione come anticipazione del seminario di Firenze e "ponte" tra le due iniziative.

1. Cosa sono i linked data

La formulazione *linked data* sta entrando nel vocabolario della comunicazione e, per quello che ci interessa in questa sede, nello specifico della terminologia biblioteconomica. Il concetto è complesso, ma potremmo sintetizzarlo in quell'insieme di buone pratiche che servono per pubblicare e collegare dati sul web *a uso di una macchina*. È un'espressione impiegata per descrivere un metodo di esporre, condividere e connettere dati tramite URI dereferenziabili. Per dereferenziazione s'intende l'accesso a una rappresentazione di risorse identificate da un URI. Con linked data, in altre parole, ci si riferisce a dati pubblicati sul web in una modalità *leggibile e interpretabile da una macchina*, il cui significato sia esplicitamente definito tramite una stringa costituita da parole e marcatori. Si costruisce così un reticolo di *dati collegati* (linked data, appunto) appartenenti a un *dominio* (che costituisce il contesto di partenza), collegato a sua volta ad altri set di dati esterni, ovvero fuori dal dominio, in un contesto di relazioni sempre più estese.

È presentata di seguito la Linking Open Data cloud (LOD), che raccoglie i dataset open messi a disposizione sulla rete, e il paradigma della sua crescita esponenziale avvenuta in pochissimi anni quale dimostrazione del livello di interesse che i linked data riscuotono in enti e istituzioni di differente natura (cfr. figure 1, 2 e 3).
Il concetto di linked data è strettamente connesso al web semantico, seppure il web semantico non si risolva nel solo *tecnicismo* dei linked data, ma richieda, per la sua costruzione, il rispetto di alcune importanti regole finalizzate alla creazione di uno strato di contenuti accessibili a processi automatizzati. Essi rendono espliciti i significati e le connessioni implicitamente contenuti (o in alcuni casi, assenti) nelle risorse del web (dati, pagine, programmi ecc.).
Le due espressioni – linked data e web semantico – attengono al medesimo ambito semantico e applicativo. I linked data sono una tecnologia adoperata per la realizzazione del web semantico. Per capire meglio il concetto ci aiuta la definizione che Tim Berners-Lee, idea-

Figura 1 – Diagramma della Linking Open Data cloud (LOD) nel 2007

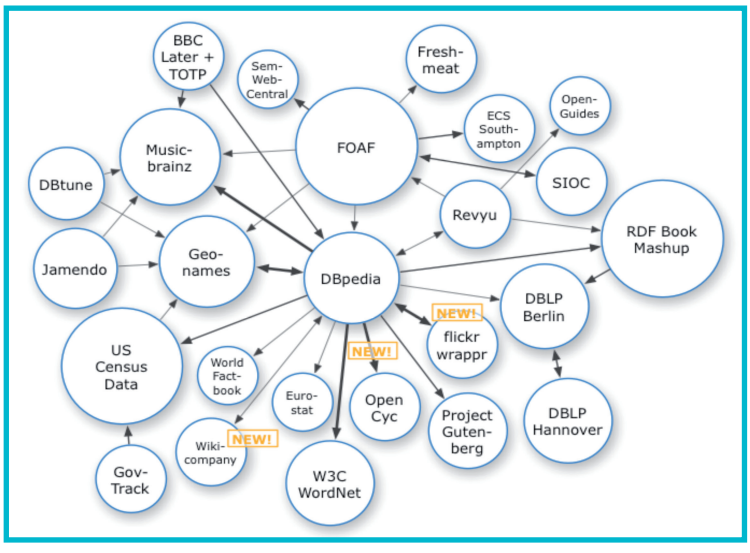
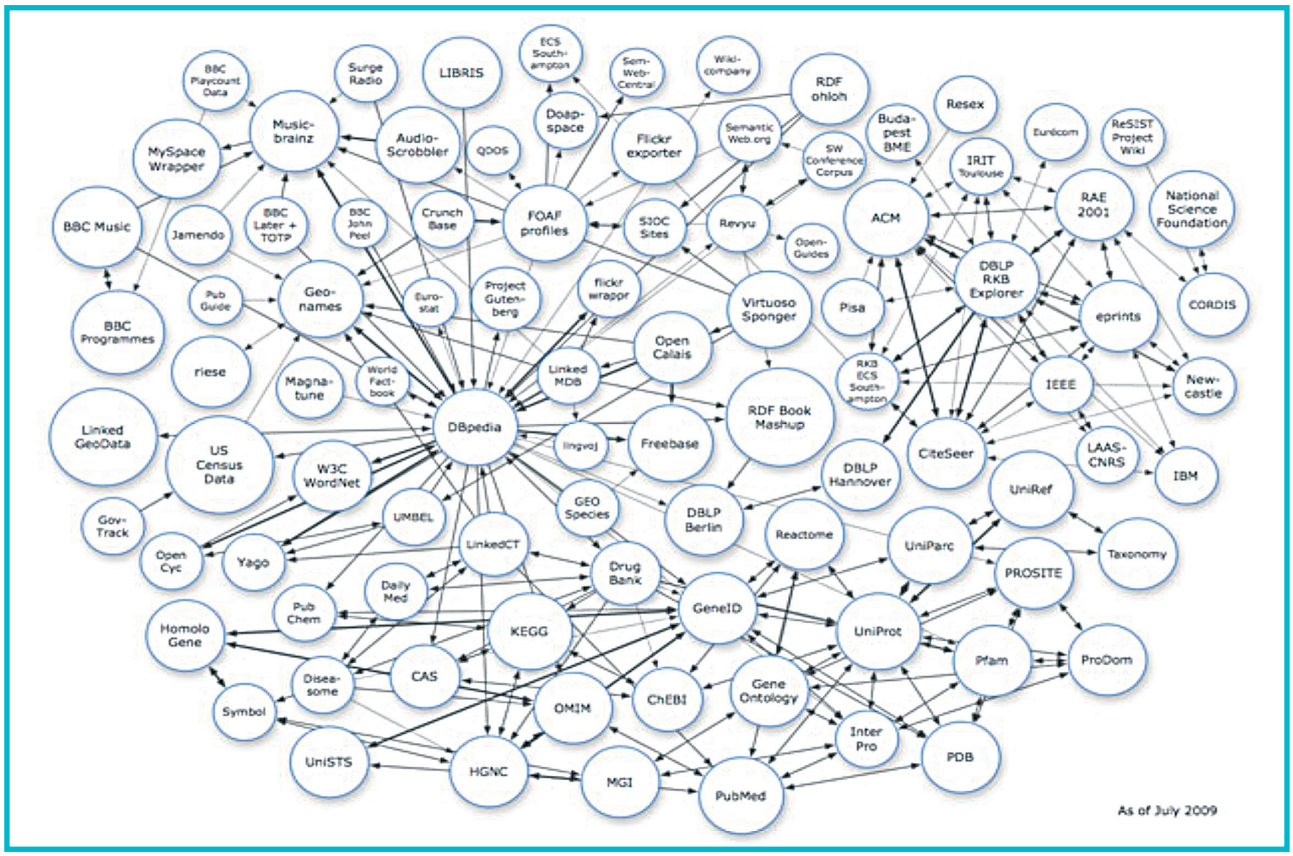


Figura 2 – Diagramma della Linking Open Data cloud (LOD) nel 2009



tore del world wide web (www), fornisce di web semantico: "A web of things in the world, described by data on the web", formulazione non facilmente traducibile, che potremmo rendere in italiano con "una rete di cose del mondo, descritta dai dati nel web". Il concetto è generico, ma contiene riferimenti importanti: la rete (il reticolo), le cose (gli oggetti relazionati), i dati (non più un record ma singoli elementi, atomi). Esso differenzia il web tradizionale (l'hypertext web) – costituito da documenti, da oggetti HTML, connessi tramite hyperlink non classificati – dal web costituito di "cose reali" (le entità esistenti) descritte tramite dati. Comincia a definirsi un'immagine più precisa:

- il web ipertestuale o *web di documenti* come rappresentazione piatta, lineare, degli oggetti; la concretezza del web semantico si oppone all'astrattezza del web tradizionale;
- il web semantico o *web di dati* come un contenitore di cose, di oggetti, piuttosto che un contenitore di rappresentazioni di oggetti: un'idea di concretezza, nel senso che i dati afferiscono alla risorsa e partecipano alla sua natura, ovvero ne sono parte integrante perché la risorsa non sarebbe rappresentabile senza questi dati.

Il web semantico non nasce, dunque, per sostituire il web tradizionale, bensì per estenderne il potenziale, realizzando quanto Tim Berners-Lee descrive come un mondo in cui "i meccanismi quotidiani del commercio, della burocrazia, e delle nostre vite quotidiane saranno gestiti da macchine che interagiscono con altre macchine, lasciando agli umani il compito di fornire l'ispirazione e l'intuizione".

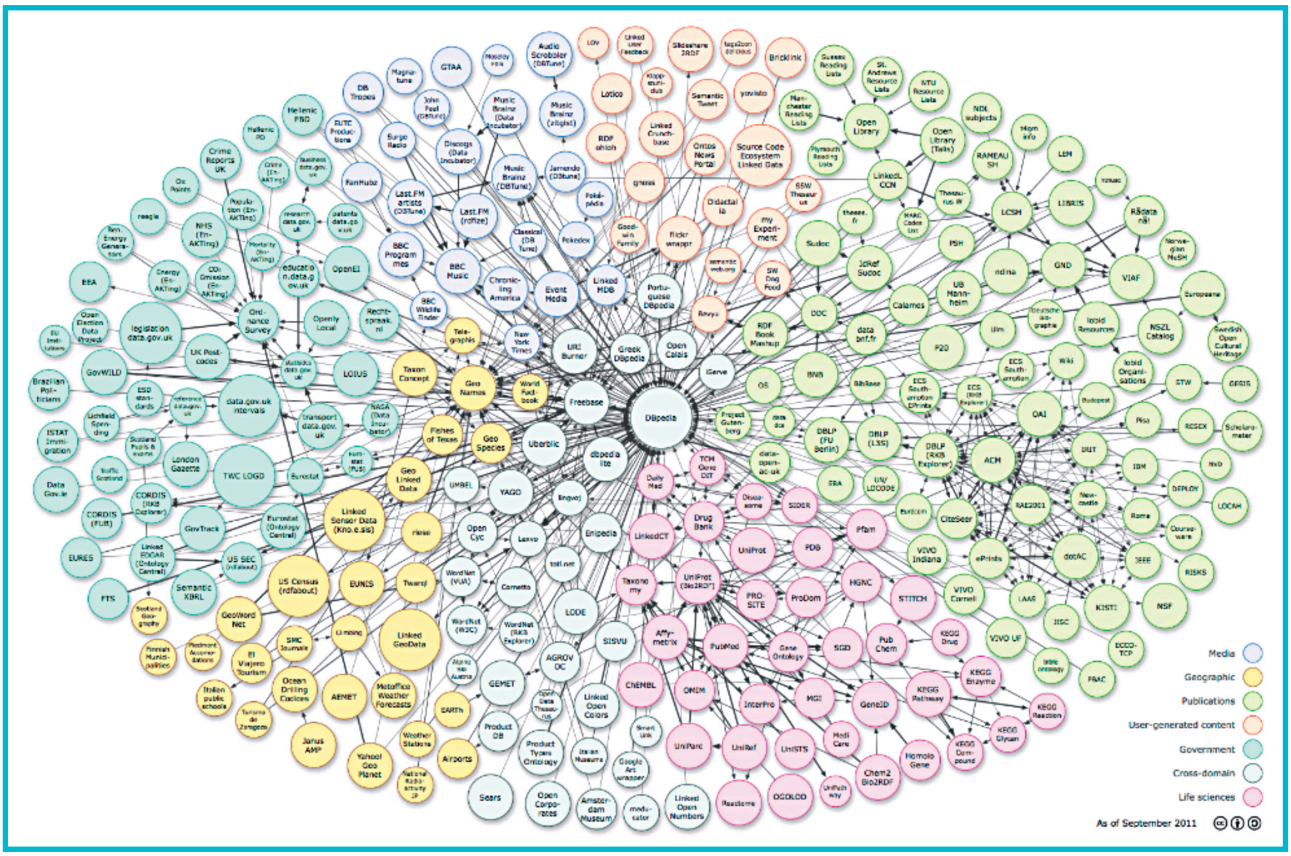
Il web di dati è, pertanto, la naturale evoluzione del web di documenti.

Cerchiamo di individuare le peculiarità distintive di ciascuno dei due, confrontandone le caratteristiche:

A. *web di documenti* (web ipertestuale):

- analogie con un file system globale, espressione di estrema ricchezza ma anche di particolare monoliticità;
- descrizione piatta di oggetti, di documenti;
- rete di relazioni presente tra gli oggetti costituita da relazioni tra i documenti non connaturate o strutturate negli oggetti stessi; di conseguenza:
- semantica del contenuto e dei legami tra documenti empirica, associata agli oggetti, ovvero non è parte dell'oggetto stesso, creata da un operatore umano;

Figura 3 – Diagramma della Linking Open Data cloud (LOD) nel 2011



- grado di struttura degli oggetti basso;
- oggetti rappresentati sul web creati per essere utilizzati dagli umani, non interpretabili dalle macchine.

Il web ipertestuale è semplice nella struttura, ha scarse connessioni tra i dati. Possiamo immaginarlo come un enorme bloc-notes, in cui le informazioni sono appuntate in modo lineare, cioè poco strutturato e poco relazionato, e in cui i documenti sono leggibili e fruibili solo dall'uomo.

B. *web di dati* (web semantico):

- database globale analogo al concetto di database relazionale, costituito da singoli oggetti ben relazionati tra di loro, che a loro volta formano entità più ampie;
- descrizione articolata dell'oggetto, descrizione che diventa essa stessa oggetto nel web, perché riutilizzabile;
- rete di relazioni tra gli oggetti connaturata agli oggetti stessi; di conseguenza:
- semantica del contenuto e di relazioni esplicita, parlante;
- grado di struttura degli oggetti alto;
- entità progettate principalmente per la macchina e secondariamente per l'umano.

Il paragone con i database relazionali è un concetto basilare della letteratura sull'argomento. Leggiamo sul sito del W3C:

> *Il web semantico e i database relazionali.* Il modello dati del web semantico è direttamente connesso col modello dei

Figura 4 – Rappresentazione del web di documenti, 17th International World Wide Web Conference W3C Track @ WWW2008, Beijing, China 23-24 April 2008 - Linked data: principles and state of the art

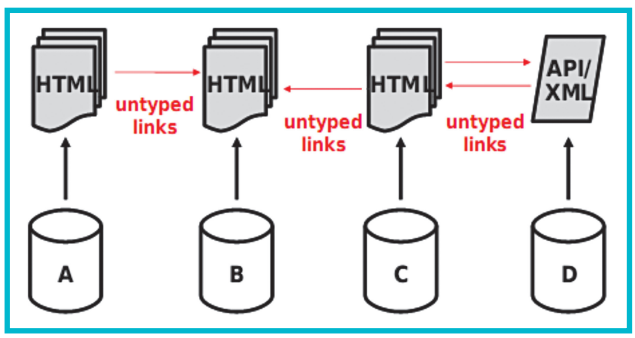
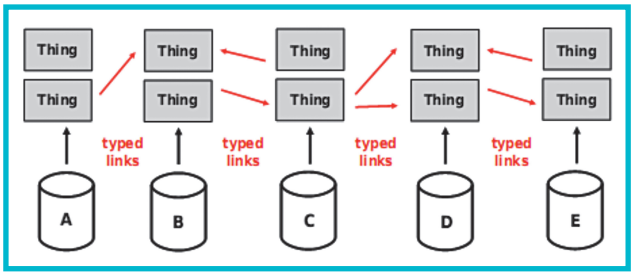


Figura 5 – Rappresentazione del web di dati, 17th International World Wide Web Conference W3C Track @ WWW2008, Beijing, China 23-24 April 2008 - Linked data: principles and state of the art



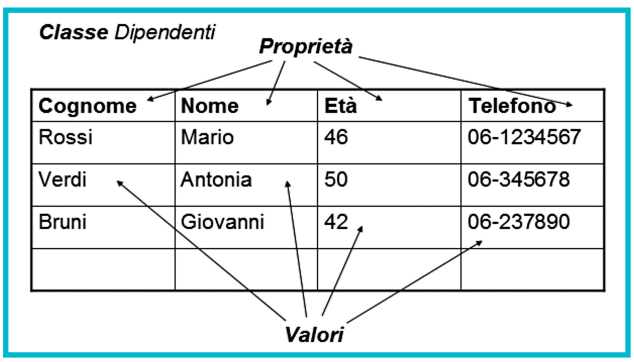
database relazionali. Un database relazionale è costituito da tabelle, realizzate da righe o record. Ogni record è costituito da una serie di campi. Il record non è altro che il contenuto dei suoi campi, proprio come un nodo RDF non è altro che i suoi collegamenti: i valori delle proprietà. La mappatura è diretta:
– un record è un nodo RDF;
– il nome del campo (colonna) è il tipo di proprietà RDF;
– il campo (la singola cella) è il valore.
Un punto di forza principale del web semantico è sempre stata l'espressione, sul web, della grande quantità di informazioni del database relazionale formulate in una modalità processabile da una macchina. Il formato di serializzazione RDF – con la sua sintassi XML – è un formato funzionale a esprimere le informazioni di database relazionale.

L'analogia risulta appropriata poiché il punto centrale dei linked data sono proprio i "predicati" che esprimono tipi di relazione mediante i quali si possono rappresentare ontologie e reti.
L'*atomizzazione* della struttura dell'informazione esprime le caratteristiche del web di dati; non si ha più un oggetto monolitico, bensì un insieme di singoli dati, delle particelle minime – atomi – riaggregabili con modalità e finalità differenti; ogni attributo dell'oggetto ha un valore in sé, partecipa alla sua natura, tramite relazioni parlanti, auto-esplicanti. Le entità costituite dall'insieme degli atomi sono articolate in un insieme di *dati strutturati*, ciascuno in sé autonomo, combinabile logicamente con altri dati per produrre nuove entità. Se abbiamo fatto l'esempio del bloc-notes per illustrare il web di documenti, possiamo assumere adesso come esempio il meccano (di ranganathiana memoria), in cui ogni elemento, in sé autonomo, può essere combinato e riusato in una molteplicità di soluzioni infinite. Il web di dati è, dunque, un network globale di *asserzioni* (o frasi) collegate tramite link *qualificati* e autoparlanti che diventano una collezione di conoscenza leggibile e utilizzabile da una macchina, prima che da una persona.

2. Linked data: il mondo di internet e il ruolo delle biblioteche, degli archivi e dei musei

Perché il mondo dell'informazione in rete è così interessato al patrimonio dei dati prodotti dalle biblioteche, dagli archivi e dai musei? Perché ugualmente le biblioteche, gli archivi e i musei sono interessati ai linked data (l'interesse è infatti reciproco)? Le biblioteche hanno sempre prodotto dati di qualità in record bibliografici e di autorità fortemente strutturati, rispondenti a regole condivise e diffuse, una quantità enorme di dati. Il mondo delle biblioteche e il mondo di internet sono entrambi interessati all'integrazione in rete; il primo per garantire la visibilità e l'usabilità dei dati, il secondo per sfruttare informazioni e creare reticoli sempre più ampi e significativi.
La *quantità* e *qualità* delle informazioni che viaggiano in rete sono due aspetti spesso inversamente proporzionali: molta informazione e bassa qualità. L'aumento dell'informazione in rete (tramite strumenti di pubblicazione sempre più diffusi e utilizzati, quali per esempio, il *self-publishing*, i social network) non è, infatti, sempre sinonimo di qualità.
La crescita e l'uso esponenziale dell'informazione disponibile in rete non coincide nemmeno con la crescita di fiducia nelle notizie: il grado di loro affidabilità è basso. Gli utenti devono selezionare tra un mare di informazioni restituite per arrivare a una notizia attendibile. Sulla base di quale criterio scegliere? L'autorevolezza della fonte diventa l'elemento discriminante, la selezione avviene a monte, preferendo scegliere la risorsa sulla base dell'autorevolezza di chi l'ha creata, anziché a valle, scegliendo acriticamente sulla base del *ranking* delle notizie che appaiono sulla pagina. La qualità della fonte, la certezza della provenienza diventano, dunque, elementi determinanti nel percorso esplorativo del ricercatore.

Figura 6 – Rappresentazione di un database relazionale



Il ruolo delle biblioteche, degli archivi e dei musei diviene pertanto rilevante, per la tradizione di attenzione alla qualità delle informazioni da loro prodotte. Queste istituzioni assumono il ruolo di generatrici di informazione di qualità per la rete. È per questo motivo che i loro dati sono ambiti.

3. I metadati storici delle biblioteche: ancora funzionali?

La storia dei cataloghi delle biblioteche mostra un impiego antico e diffuso dei metadati, intesi come informazioni vicarie della risorsa. L'evoluzione dei dati in notizie sempre più strutturate e dettagliate ha coinciso con la rinnovata centralità del catalogo su cui oggi si basano tutti i servizi della biblioteca, col moltiplicarsi dei formati delle risorse bibliografiche e con il ruolo costitutivo dell'informatica che ha imperniato il sistema biblioteca. Le caratteristiche principali dei metadati sono:
- la *natura*: è artefatta, costruita sulla risorsa;
- la *finalità*: descrive un oggetto;
- l'*utilizzo*: dev'essere strutturato in modo che sia processabile (cioè utilizzabile) da una macchina, da un computer.

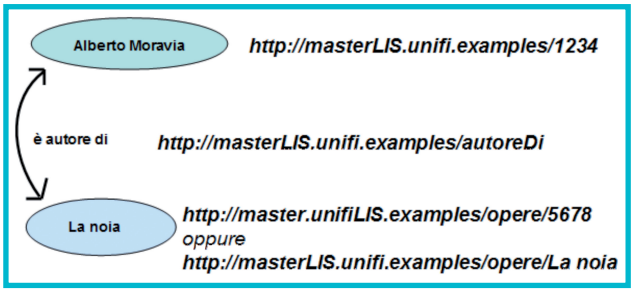
Le biblioteche hanno perseguito l'obiettivo costante e coerente di condividere le informazioni tramite i metadati, e hanno sempre assegnato importanza alla loro qualità.
I metadati finora usati sono ancora funzionali? Rispondono alle attuali esigenze informative dell'utenza? È sufficiente esporre sul web i dati che le biblioteche hanno prodotto nel corso dei secoli? Questa esposizione (per esempio, in formato MARC) è comprensibile e usabile al di fuori del contesto strettamente bibliotecario? Non rischia di essere un'esposizione di nicchia, circoscritta a un ambito ristretto, a un dominio chiuso e altamente professionalizzato?

4. Il catalogo del futuro: *del web* e non solo *sul web*

Constatiamo, con amarezza, che i dati prodotti dalle biblioteche – i cataloghi –, la cui redazione ha richiesto elaborazione di normative, competenze professionali e finanziamenti, non sono *sul* web, ma *isolati* dal web. I cataloghi non sono, infatti, integrati nel web, non sono interrogabili, pur essendo il web il luogo in cui la maggior parte degli utenti lavora, gioca, opera e crea altra informazione. La questione, dunque, è: "Come modificare i cataloghi e i dati perché siano *del web* e non solo *sul web*?".
È proprio la filosofia che sottende la tecnologia dei linked data che può offrire un interessante punto di partenza per il raggiungimento di questo obiettivo strategico, pena la morte dei cataloghi, abbandonati dagli utenti a favore di altri strumenti di reperimento dell'informazione, come i motori di ricerca. Si tratta di un passaggio fondamentale: l'adozione inevitabile dei linked data comporterà una nuova rivoluzione, ancora più radicale rispetto a quella che avvenne negli anni Sessanta del secolo scorso, quando si passò dal catalogo cartaceo al catalogo automatizzato prima e informatizzato poi, una rivoluzione che corona il ruolo che l'informatica ha assunto nella gestione dei processi di comunicazione e, dunque, per ciò che ci riguarda più da vicino, nella creazione degli strumenti di mediazione tra universo bibliografico e utente.
On the record, il *report* della Library of Congress Working Group on the Future of Bibliographic Control, traccia una buona guida per raggiungere l'obiettivo; il cambiamento implica:
- la trasformazione della *descrizione testuale* in *set di dati* usabili per processi ed elaborazioni automatiche da parte di macchine;
- la necessità di rendere gli elementi di dati *univocamente identificabili* all'interno del contesto informativo del web;
- la necessità che i dati siano *compatibili con le tecnologie* e gli *standard* del web;
- la necessità, in sintesi, di usare un *linguaggio* trasversale e interoperabile nella realtà del web.

Il concetto di *identificazione univoca degli oggetti* è di particolare interesse: l'oggetto identificato, caratterizzato dall'essere la medesima cosa a prescindere dalla sua espressione testuale (di avere dunque il medesimo significato) dovrebbe avere un identificativo univoco, in modo da essere utilizzabile in contesti differenti (biblioteche, editori, librai, distributori, produttori di biografie online...), anche tramite l'utilizzo di valori testuali differenti.
Tim Berners-Lee individua quattro regole per la creazione dei linked data sul web:
1. *usare URI (Uniform Resource Identifiers) per identificare cose (oggetti)*: l'URI è un sistema di identificazione globale, valido cioè per tutte le risorse contenute nell'intero web. L'URI è una pietra miliare dell'architettura del web, in quanto costituisce un meccanismo di identificazione delle risorse comune a tutto il web. Ciascuna risorsa sul web (un sito, una pagina di

Figura 7 – Rappresentazione di una tripla (nodi ed archi) in RDF



un sito, un documento, un qualsiasi oggetto) dev'essere identificata da un URI se vuole essere ricercata da altri sistemi, utilizzata, collegata, ecc.;

2. *usare HTTP URI in modo che gli oggetti possano essere individuati da persone e da user agent (browser, programmi...)*: lo schema utilizzato per la costruzione di un URI è dichiarato nell'URI stesso prima dei due punti (:); per esempio, http://weather.example.com/). L'http che utilizza l'HyperText Transfer Protocol come protocollo è precisamente lo schema prescritto per il web semantico;
3. *fornire informazioni utili sull'oggetto (quando si individua un URI), usando formati standard come RDF, SPARQL (linguaggio d'interrogazione che nasce per i linked data)*: è necessario definire il contesto e le caratteristiche della risorsa, tramite l'attribuzione della risorsa stessa a una *classe*, l'identificazione di *proprietà* e l'assegnazione di *valori*;
4. *includere link ad altri URI relativi ai dati esposti per migliorare la ricerca nel web di altre informazioni affini a quella di partenza*: più i dati sono collegati, più sarà possibile il loro utilizzo nell'ottica di arricchimento e deduzione delle informazioni.

5. Linked data: RDF (resource description framework)

Produrre linked data significa, dunque, esprimere i *significati* delle informazioni, renderle *condivisibili* fra differenti applicazioni e *utilizzabili* da applicazioni diverse da quelle per cui erano state originariamente create.
Il *data model* utilizzato per la strutturazione di linked data è RDF, uno standard flessibile proposto dal W3C per caratterizzare *semanticamente* le risorse e le *relazioni* che intercorrono tra esse.
Abbiamo definito la *realtà del web* come un network globale di asserzioni (o frasi) collegate tramite link qualificati. Il modello RDF codifica i dati in forma di asserzioni costituite da:

- *soggetto*: la parte della frase che identifica la cosa descritta;
- *predicato*: la proprietà della cosa specificata dalla frase;
- *oggetto*: il valore della proprietà della cosa (le *triple* RDF).

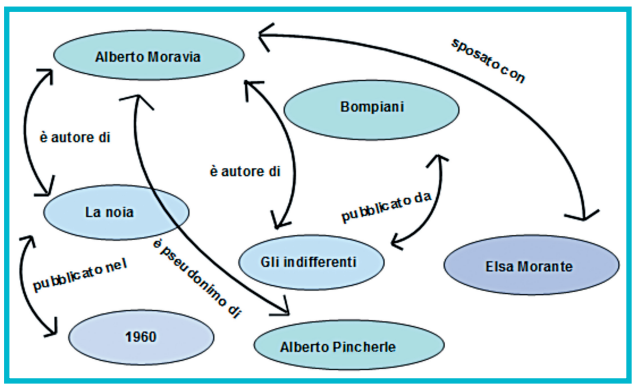
Ogni asserzione è un concetto atomico e significativo: *l'unità significativa atomica*.
Esempio:

Alberto Moravia *è autore de* La noia
Bompiani *ha pubblicato* Il nome della rosa
Alberto Moravia *è pseudonimo di* Alberto Pincherle

Ciascun elemento della tripla, ci ricorda Tim Berners-Lee, può, o meglio, dev'essere, tecnicamente, rappresentata tramite URI dereferenziabili. Più URI sono utilizzati più l'informazione risulta riusabile; ciò non è obbligatorio ed elementi della tripla possono essere espressi anche in modalità testuale.
Le asserzioni, o triple, sono espresse da RDF in forma di *grafi* (nodi e archi) che rappresentano le risorse, le loro proprietà e i rispettivi valori.
Le triple sono codificate tramite sintassi *XML-based* (RDF/XML) affinché siano leggibili e utilizzabili da una macchina, che può essere quella per cui il dato è nato (dal sistema nativo) o un sistema differente (esterno) a quello per il quale era nato. Questa è la caratteristica più importante, che apre i dati alla comunità informativa globale.
Proviamo a osservare le seguenti asserzioni:
– Marco è figlio di Gianni;
– Susanna è figlia di Gianni;
– Gianni è figlio di Chiara.

Da queste semplici asserzioni è possibile ricavarne per lo meno altre tre, seppur non esplicitate da triple:
– Marco e Gianni sono maschi;
– Susanna è femmina;
– Chiara è nonna di Marco e di Susanna.

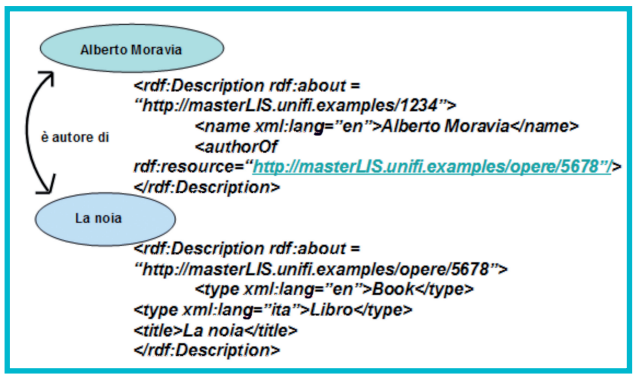
Figura 8 – Rappresentazione di un reticolo di asserzioni o triple



E potremmo *dedurne* ancora, per esempio:
– Marco e Susanna *sono nipoti* di Chiara;
– *Marco è fratello* di Susanna;
– Susanna *è sorella* di Marco.

Questo meccanismo, definito *inferenza* – il processo con il quale da una proposizione accolta come vera si passa a una seconda proposizione la cui verità è dedotta dal contenuto della prima – è il principio dei motori che stanno dietro il web semantico, che *deducono conoscenza tramite percorsi*.
Ogni nuova asserzione, espressa in forma di tripla e dunque di grafo, diventa a sua volta generatrice di nuova informazione; più gli ambiti di appartenenza di queste asserzioni (databaset) crescono e *si intersecano*, più la rete semantica presente e disponibile sul web si arricchisce e diventa informazione classificata. Il meccanismo dell'inferenza è ben conosciuto nella logica e nella matematica (calcolo inferenziale) ed è largamente utilizzata nelle applicazioni informatiche. Esso acquisisce un valore specifico applicato al mondo delle biblioteche; il meccanismo esplicita, infatti, le relazioni presenti nei dati bibliografici, ma non sempre evidenti e di cui si è avuto consapevolezza piena con la sistematizzazione teorica compiuta da FRBR: una sistematizzazione di concetti esistenti nella tradizione catalografica, almeno da Cutter in poi, e resi sempre più dichiarati.
Perché questo meccanismo funzioni è necessario utilizzare un'infrastruttura tecnologica in cui i concetti siano univocamente identificati e in cui agenti software riconoscano questi oggetti e realizzino associazioni ed equivalenze tra essi, tramite il riferimento a *ontologie*, rappresentazioni formali, condivise ed esplicite di specifici domini della conoscenza. Le ontologie consentono di rappresentare le entità tramite la descrizione delle loro caratteristiche e l'identificazione delle relazioni esistenti tra esse, e dunque della semantica che lega tali entità, utilizzata soprattutto per realizzare *categorizzazioni* e ragionamenti deduttivi.
Esempi di vocabolari e ontologie diffusi nel mondo delle biblioteche sono:
– FOAF (Friend Of A friend): un'ontologia utilizzata per descrivere le persone, le loro attività, le loro relazioni con altre persone o cose, utilissima per strutturare authority file in linked data;
– SKOS (Simple Knowledge Organization System): una famiglia di linguaggi formali creati per rappresentare thesauri, schemi di classificazione, tassonomie, soggettari e ogni tipologia di vocabolario controllato.

Figura 9 – Rappresentazione di una tripla in RDF/XML



L'IFLA sta concentrandosi sulla pubblicazione di propri standard in RDF con la creazione di vocabolari e ontologie per FRBR, FRAD, FRSAD e ISBD, pubblicate sull'*Open Metadata Registry* (precedentemente *NSDL Registry*), spazio che il W3C ha creato per supportare sviluppatori e utilizzatori di vocabolari controllati, ospitando ontologie di differenti ambiti, tra cui i vocabolari per RDA (*Resource Description and Access*), il nuovo standard di catalogazione che sostituisce le AACR2 (*Anglo-American Cataloging Rules, 2nd edition*) creato dalla comunità bibliotecaria angloamericana, ampliato a realtà europee (in particolare alla Francia) e proposto alla comunità bibliografica e bibliotecaria internazionale.
Le ontologie sono dunque necessarie per creare e pubblicare un *dataset,* il quale esprime un *dominio di appartenenza* rappresentando una sorta di collezione di risorse (o di grafi), accomunate da una qualche caratteristica, e identificate tramite URI dereferenziabili. Esempi di dataset disponibili sul web sono:
- Dbpedia: dataset che contiene i dati estratti da Wikipedia;
- LinkedMDB: dataset sul mondo del cinema;
- VIAF: Virtual International Authority File.

Proviamo a elaborare *inferenze* possibili combinando i dati presenti in questi dataset:
– Eduardo De Filippo è vissuto dal 1900 e al 1984 (da VIAF);
– Eduardo De Filippo è autore di Filumena Marturano (da VIAF);
– Eduardo De Filippo è nato a Napoli (da Dbpedia);
– Napoli è capoluogo della Regione Campania (Dbpedia);
– Questi fantasmi è un film diretto da Eduardo De Filippo (da linked MDB);
– Massimo Troisi è regista di Ricomincio da tre (da Dbpedia);

– Massimo Troisi è nato a Napoli (Dbpedia);
– Ricomincio da tre è un film del 1981 (da linked MDB);
– Scusate il ritardo è un film diretto da Massimo Troisi (da linked MDB).

Se volessimo creare un dataset relativo alle *personalità campane che si sono distinte nella letteratura e nel cinema* potremmo utilizzare le triple sopra dette, estratte da differenti dataset per alimentare il nostro insieme e *dedurre* così nuova informazione:
– Eduardo De Filippo e Massimo Troisi sono personalità campane del secolo XX, autori di opere letterarie e registi.

6. Open Linked Data Project

Quanto sono *accessibili* questi dataset, e quali sono le modalità per renderli davvero fruibili a comunità più aperte? Ciascuna istituzione potrebbe produrre propri linked data, secondo quanto definito dai criteri e le regole sopra citate, ma non renderli *aperti* all'utilizzo nel web. Perché un dataset sia *open* (e dunque non condizionato da licenze commerciali o restrizioni d'uso) dev'essere pubblicato secondo quanto definito dall'Open Linked Data Project, che prevede la conversione di dataset già esistenti o la produzioni di nuovi, secondo i principi del linked data, ma con licenze *open*. Il progetto, inizialmente partito con la partecipazione di piccoli enti, ricercatori e sviluppatori in ambito universitario, ha ottenuto nel tempo numerose adesioni di enti e istituzioni autorevoli e di grandi dimensioni, tra cui la BBC, la Thomson Reuters e la Library of Congress.
Questa adesione e diffusione in ambiti noti, riconosciuti e prevalenti ha prodotto una crescita e un'espansione notevolissima del progetto, agevolata dalla sua *natura aperta*: ognuno può partecipare pubblicando un set di dati che rispettino i principi dei linked data e creando collegamenti incrociati (*interlinking*) con altri dataset già esistenti.

7. Library Linked Data Project

Il W3C Library Linked Data Incubator Group nasce per supportare e favorire lo sviluppo e la crescita della interoperabilità dei dati di biblioteca, archivi e musei sul web. Segue i principi dei linked data e del web semantico, e i lavori del gruppo sono stati condotti in stretta collaborazione con gli operatori di questi ambiti.

Casi interessanti per la stesura del *Final report* dell'Incubator Group sono stati i progetti sostenuti da enti piccoli, medi, fino a quelli delle grandi biblioteche nazionali. Il *Final report* parte dall'analisi dei progetti in atto e definisce un quadro d'insieme, che può essere così riassunto:
- analisi dei benefici possibili nell'applicazione dei principi dei linked data in ambito biblioteconomico;
- discussione sui problemi aperti con particolare riferimento ai dati tradizionali;
- analisi e censimento dei progetti e delle iniziative di linked data in ambito biblioteconomico;
- discussione dei problemi dei diritti legali e di pubblicazione;
- elaborazione di raccomandazioni per i prossimi passi nel processo di applicazione dei principi dei linked data al settore.

8. Ciclo di vita dei linked data

Quali sono i passi che un ente deve percorrere per trattare i propri dati e arrivare alla loro pubblicazione come linked data? Un buon riferimento metodologico è fornito da Boris Villazón-Terrazas in *Methodological guidelines for publishing linked data*, che riproduce il ciclo di vita per la produzione di linked data in sette passi:
– *identificazione della fonte dati*;
– *modellizzazione del vocabolario*, con l'adozione di ontologie esistenti, espresse in OWL, *Web Ontology Language*, o RDF(S) o con la creazione (più complessa) di nuove ontologie;
– *generazione dei dati in formato RDF*, tramite diversi linguaggi di mappatura disponibili, anche in relazione al formato di origine del dato. In questa fase l'operazione più delicata è la creazione di URI, poiché essi sono la chiave per allineare risorse eterogenee provenienti da fonti differenti;
– *pubblicazione dei dati in RDF*;
– *bonifica dei dati prodotti*, per individuare eventuali e possibili errori di conversione e rendere il dato qualitativamente usabile;
– *creazione di collegamenti tra dataset differenti*, con l'identificazione di dataset di interesse che possano diventare *linking target*, identificando relazioni tra i singoli dati, validando le relazioni individuate;
– *rendere concreto l'utilizzo dei dati*, con differenti passi, tra cui la pubblicazione del dataset ottenuto dal processo sul CKAN Registry (Comprehensive Knowledge Archive Network), un registro per la pubblicazione di

dati e pacchetti open, che rende possibile la loro scoperta, la condivisione e il riutilizzo.

9. Le 5 stelle degli open linked data

Il dataset ottenuto con i sette passi suggeriti da Boris Villazón-Terrazas può essere poi valutato tramite un sistema di *rating* definito da Tim Berners-Lee per assegnare un punteggio ai siti che espongono dati sul web, definito le *5 stelle per gli open linked data*:

- make your stuff available on the web (whatever format);
- make it available as structured data (e.g. excel instead of image scan of a table);
- non-proprietary format (e.g. csv instead of excel);
- use URLs to identify things, so that people can point at your stuff;
- link your data to other people's data to provide context.

La valutazione sugli open linked data prodotti dev'essere realizzata considerando dunque cinque punti fondamentali:

1) i propri dati siano disponibili sul web (in qualsiasi formato);
2) il materiale messo sul web sia disponibile come dato strutturato (per esempio, in excel anziché come scansione dell'immagine di una tabella);
3) siano stati scelti formati non proprietari (per esempio, in csv invece che excel);
4) siano stati utilizzati URL per identificare gli oggetti, in modo che gli utenti possano puntare a questi oggetti;
5) i propri dati siano stati collegati a dati prodotti da altri in modo da definire un contesto.

L'indicazione di Tim Berners-Lee per la valutazione degli open linked data è stata seguita da una serie di raccomandazioni, suggerimenti, modalità per definire norme e regole di valutazione sempre più precise, per arrivare a uno standard quanto più possibile partecipato e condivisibile.

Conclusioni

L'applicazione del concetto dei linked data costituisce un'evoluzione considerevole (una rivoluzione?) nel mondo della comunicazione globale. Occorre pertanto comprenderlo e divenirne protagonisti, chiamando le istituzioni, le agenzie bibliografiche e le biblioteche in particolare, a investirvi in termini di risorse economiche e di competenze professionali.

ABSTRACT

The statement *linked data* is entering the vocabulary of communication and terminology of LIS. The concept implies the best practices used to publish and link data on the web using machine. Semantic web and linked data related to the same semantic context and application. The linked data are a technology used for the realization of web semantic. The semantic web differentiates the traditional web (hypertext web) – made up of HTML objects, connected by the data. The role of libraries, archives and museums become relevant for the tradition of attention to quality of information produced by them. Data produced by libraries are not on the web, but isolated from the web. How to change the catalogues and data for the web and not only on the web? Applying the concept of linked data is a considerable evolution (a revolution?) in the world of global communication.