“Addomesticare” i Big Data

«Che l'invenzione e lo sviluppo delle tecnologie digitali abbiano moltiplicato la complessità del mondo dei documenti è un fatto incontestabile»

L’invito rivolto ai Bibliotecari

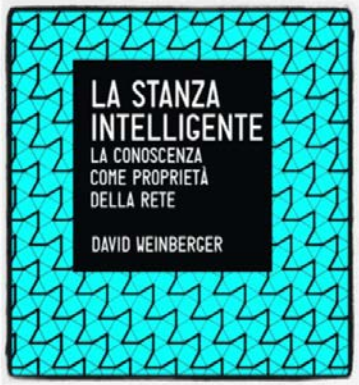
dal filosofo e tecnologo della comunicazione David Weinberger:

«entrate nella Stanza Intelligente»

"Quando la conoscenza entra a far parte di una rete, la persona più intelligente della stanza è la stanza stessa: la rete che unisce persone e idee presenti e le collega con quelle all'esterno"

"I bibliotecari sono impegnati a trovare una visione credibile e sostenibile per il futuro delle loro istituzioni: non solo discutendo i meriti delle nuove tecniche per accedere alle raccolte ma anche su come valutare le competenze della "folla" rispetto a quelle di studiosi con le opportune credenziali"

David Weinberger, *La stanza intelligente. La conoscenza come proprietà della rete*, Torino, Codice edizioni, 2012



Seguiamo Weinberger...

Metadatazione

- **Da diversi anni è in corso un tentativo di aggiungere metadati alla Rete in modo da poter raccogliere e usare i dati di tutto il web**

- **Creare questi metadati è difficile, soprattutto è difficile mettere a punto grandi e complesse rappresentazioni logiche dei vari domini, cioè scrivere ontologie particolareggiate per ogni settore e sotto-settore di sapere**

- **Se questa operazione si fa su quantità enormi di dati (Big Data) provenienti da domini diversi (finanza, medicina, cultura, governo, sociologia ecc.) si possono ottenere risorse senza precedenti per scoprire nuove idee basate su quello che già sappiamo sul nostro mondo**

Metadazione e luoghi «topici» per Selfie

- **Esempio di ricerca per scovare i luoghi dei Selfie, le varie fasi possibili:**
  - **Raccolta di autoscatti disponibili su i social media (ad esempio Flickr che mantiene i metadati delle foto)**
  - **Classificazione delle fotografie in base ai metadati ricavati dalle foto considerate come dati**
  - **Grazie al protocollo Exif che contiene i metadati delle foto (data, ora, marca dispositivo ecc.)**
  - **Se il dispositivo ha un sistema GPS, anche le coordinate geografiche**
    - **Possibile ricavare informazioni sulla geolocalizzazione degli scatti per verificare se esistono luoghi topici per i Selfie.**

Seguiamo Weinberger...

Web Semantico

Non è facile implementare un programma capace di riunire tutte le informazioni in Rete disponibili su uno specifico argomento

Se invece si seguissero le convezioni specificate dal Web Semantico

I programmi saprebbero in modo facile quali siti o sistemi si riferiscono allo stesso argomento

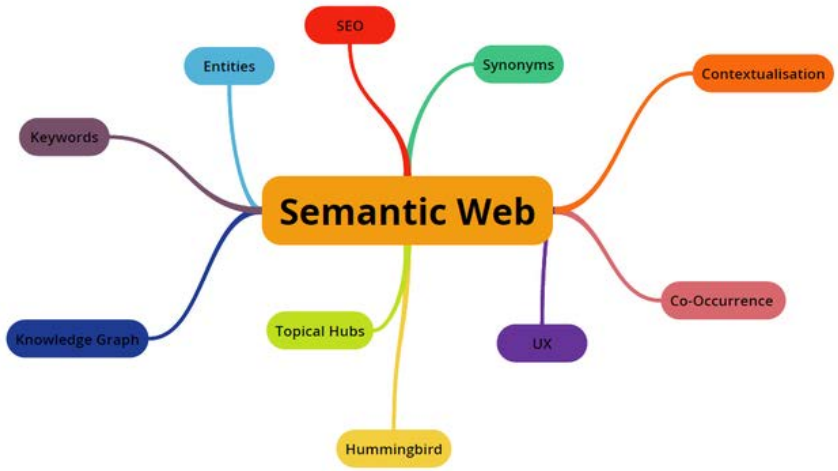
In questo modo la Rete sarebbe capace di esprimere più sapere di quanto non sia immesso, che è poi la definizione di Rete intelligente.

(**David Weinberger**)

Web Semantico

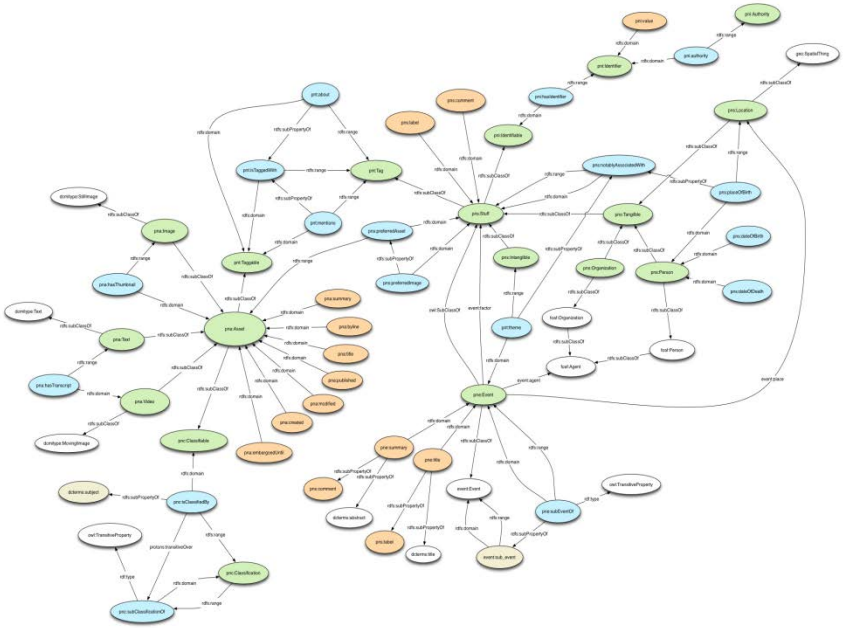
Si intende la trasformazione del Web in un ambiente dove i documenti pubblicati (pagine HTML, file, immagini, e così via) sono associati ad informazioni e dati (metadati) che ne specificano il contesto semantico in un formato adatto all'interrogazione e l'interpretazione (es. tramite motori di ricerca) e, più in generale, all'elaborazione automatica.

(Wikipedia)



La difficoltà: ontologie troppo complesse

Creare tutti questi metadati è difficile, soprattutto è difficile mettere a punto grandi e complesse rappresentazioni logiche dei vari domini, cioè scrivere ontologie particolareggiate per ogni settore e sotto-settore di sapere



Seguiamo Weinberger…

Senza aspettare accordi su ontologie troppo estese e complesse…

dati subito accessibili in una forma standardizzata ma imperfetta

La soluzione Linked Data

- **Tecnologia per**
  - **pubblicare**
  - **condividere**
  - **Collegare**

- **Singoli dati, informazioni, conoscenze**
  - **sul Web semantico**

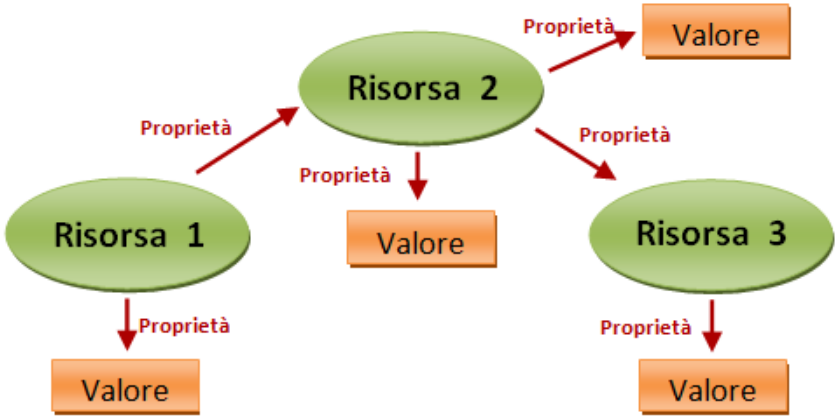
    - **Utilizzando**
      - **URI (Uniform Resource Identifier)**

Seguiamo Weinberger...
La soluzione dei Linked Data: semplificare per «addomesticare i Big Data»

- **Se, ad esempio, avete a disposizione delle informazioni di chimica, potete renderle disponibili sul Web attraverso una serie di affermazioni elementari chiamate Triple, che hanno forma di due oggetti uniti da una relazione**
- **Se questa operazione si fa su quantità enormi di dati (Big Data) provenienti da domini diversi (finanza, medicina, cultura, governo, sociologia ecc.) si ottengono risorse senza precedenti per scoprire nuove idee basate su quello che già sappiamo sul nostro mondo**
- **Una delle chiavi del successo dei LD è che si tratta di una procedura non troppo rigorosa rispetto ai suoi metadati**
- **Ad esempio come si dovrebbe indicare l'autore di un libro nei metadati LD? Come autore, scrittore, creatore?**
  - **Anziché rispondere a questa domanda i sostenitori degli LD vi diranno che potete usare uno qualsiasi di questi termini nelle vostre triple**
    - **ed esprimere il rapporto con un link che rimanda a qualche sito noto che abbia definito la relazione per voi**
    - **Così anziché usare la parola autore, mettere un link verso quello che lo Standard Dublin Core definisce "relazione autoriale".**
    - **A quel punto qualsiasi applicazione che voglia capire la vostra tripla saprà che il rapporto è quello definito sul sito Dublin Core.**

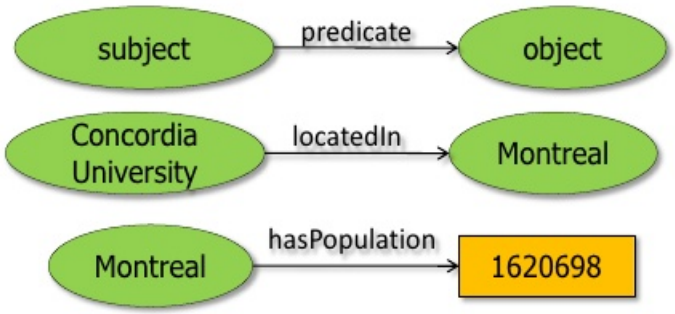
La grammatica dei Linked Data: RDF (Resource Description Framework)

- **Descrive le relazioni tra gli oggetti sotto forma di triple**
- **Una tripla è una dichiarazione, cioè un'affermazione nella quale si stabilisce che una proprietà di un certo soggetto assume un determinato valore**
- **La tripla è rappresentata da un grafo dove: gli ovali simboleggiano le risorse, i rettangoli i valori, e gli archi le proprietà**
- **Esempio: "Alessandro Manzoni" "è autore dei" "Promessi Sposi"**



RDF (2)

- Everything is a triple
  - **Subject** (resource), **Predicate** (relation), **Object** (resource or literal)
- The RDF graph is a collection of triples



Cambio di Paradigma: dal Web dei documenti al Web dei dati

- **È la visione di un nuovo scenario globale per il trattamento e la condivisione dell'informazione**

- **E' quello che Tim Berners-Lee definisce :l'evoluzione dal Web dei documenti al Web dei dati**
- **Web dei documenti**
  - **costituito da risorse HTML reciprocamente collegate tramite link 'non tipizzati' (untyped)**
  - **cioè link la cui unica funzione consiste, per l'appunto, nel rimandare da un documento all'altro del sistema**
  - **senza esprimere alcuna valutazione in merito alla natura concettuale del collegamento stesso**
- **Web dei dati**
  - **assume come punto di partenza l'adozione di uno schema (RDF)**
  - **che impone una destrutturazione del documento**
  - **individuandone, dal punto di vista concettuale, la natura, la tipologia e gli attributi degli elementi che lo costituiscono e dei legami che connettono questi elementi fra loro**

Vantaggi e limiti dei Linked Data

Vantaggi

- **Offrono la possibilità di correlare reciprocamente i contenuti recuperati da differenti archivi**
- **E' una tecnologia che può rendere interoperabili e quindi produttivi i grandi dataset del patrimonio culturale**

Limiti

- **Non è sufficiente ridurre un documento in triple per esplicitarne l'intero potenziale comunicativo**
- **In quanto un documento, non è costituito solo dai dati e dai legami fra loro intercorrenti ma anche da quelle caratteristiche progettuali che tengono questi dati uniti in un insieme sensato e coerente**
- **si tratta di una dimensione, questa, che i Linked Data non contemplano**

I limiti della Tripla (RDF)

- **In linea di principio tutti i documenti, intenzionali e non, possono essere triplizzati**
- **Tuttavia, in considerazione del fatto che l'architrave di una tripla è costituito da un 'predicato' (e cioè da una affermazione che indica una proprietà del soggetto)**
- **E’ fondamentale distinguere l'autorità di chi esprime tale giudizio di valore (ontologia)**
- **In relazione ai contenuti e alle caratteristiche formali che il documento**

L’auto usata di Tim Berners-Lee

- **C’è un tizio intenzionato a vendere un'automobile gialla usata**
  - **Questo tizio posta in Rete il relativo annuncio**

- **Ma i motori di ricerca non ne comprendono gli elementi essenziali (marca, modello, colore, stato di manutenzione ecc.) perché il messaggio è stato composto in testo semplice**

- **Invece se l'annuncio fosse stato composto tramite la compilazione di un modulo (RDF)**
  - **I dati sarebbero stati leggibili e interpretabili automaticamente dai computer**
  - **consentendo di creare un sistema di informazioni omogenee e correlate.**

- **Il problema è: chi stabilisce l'individuazione degli elementi essenziali (le entità) e le reciproche relazioni?**

L’auto usata di Tim Berners-Lee

- **Chi, insomma, disegna il modulo?**
  - **I responsabili di un portale di vendita online come Ebay?**
  - **L'autoconcessionario che viene incaricato della pratica? L'ACI o la FIA? Il PRA?**
  - **O, addirittura l'utente stesso, che supponiamo appassionato di informatica e, dunque, in grado di disegnarsi la propria personale ontologia?**
- **La risposta di Berners-Lee è: non importa.**
  - **Quello che conta è che esistano delle stanze di compensazione in grado di tradurre una categoria in un'altra rendendole reciprocamente trasparenti attraverso linguaggi inferenziali**

L’auto usata di Tim Berners-Lee

- **Affermare , come sostiene Tim Berners-Lee, che «non esiste nessuna autorità su alcunché», ovvero statuire il cosiddetto AAA principle: «Anyone can say Anything about Any topic»**

- **Significa esprimere un atto di fede pieno e assoluto nei confronti della capacità euristica dei computer connessi in rete e nell'intelligenza collettiva che scaturisce dalle menti dei loro utilizzatori.**

In realtà, per il buon funzionamento, i Linked Data, necessitano di alcuni ingredienti essenziali

- **L'autorità di chi individua le entità e ne stabilisce i relativi predicati**
- **La competenza di chi applica gli schemi**
- **In quanto contribuiscono a costruire quella 'fiducia'**
- **Considerata dallo stesso Berners-Lee come un «*prerequisito fondamentale di una società reticolare*»**

Riassumendo....

Se vogliamo che i Linked Data rappresentino un salto di qualità significativo nell'organizzazione della conoscenza, sono necessari:

specifici metadati in grado di identificare:

A. Provenienza dei dati

B. Coerenza dei dati

La “vocazione” delle biblioteche

Le biblioteche, a differenza della maggior parte di coloro che inseriscono dati in rete, hanno sempre prodotto dati fortemente strutturati sia del punto di vista della qualità che dell’autorità

- **Open**
  - **Sono tra le poche istituzioni che interpretano con vocazione primigenia l'idea di rendere aperti all'uso libero e gratuito della collettività i propri archivi e le notizie in essi contenuti**
- **Alla Standardizzazione**
  - **Hanno la cultura della standardizzazione descrittiva e dell'interoperabilità fra i sistemi**

I dati delle biblioteche

Ci sono milioni e milioni di dati bibliografici conservati negli OPAC delle biblioteche di tutto il mondo, ma non raggiungibili attraverso la Rete perché creati e registrati con formati non interoperabili nel Web

W3C: per l'interoperabilità dei dati delle biblioteche nel Web

Library Linked Data Incubator: gruppo di lavoro creato dal W3C per aumentare l'interoperabilità dei dati delle biblioteche nel Web

Gli scopi del Library Linked Data Incubator:

- **Importanza dei collegamenti tra dati bibliografici e informazioni prodotte all'esterno**
- **I dati di qualsiasi provenienza, compresi i dati bibliografici, devono essere condivisibili, modulari, riutilizzabili**

Le Biblioteche devono produrre: Library Linked Data

- **Non è necessario ricreare / riconvertire grandi masse di dati bibliografici**
- **Basta fornire i dati bibliografici di metadati adeguati per il Web semantico**
- **Sviluppando, ad esempio, una rappresentazione RDF dell'UNIMARC**
- **Come nel caso della sperimentazione in corso nell'ambito della rete SBN in modalità Linked Open Data (LOD)**

LOD di SBN:
una prima sperimentazione
Maria Cristina Mataloni
(ICCU)

30/03/2017

- **Prima fase di sperimentazione:**

  uno schema logico e operativo per la produzione e pubblicazione di dati SBN strutturati in LOD

  la mappatura di un set di record (set minimo di dati, senza specificità) estratto dall'OPAC SBN in formato UNIMARC utilizzando l'ontologia FRBRoo.

L’Ontologia presa a riferimento è FRBRoo
(FRBR “object oriented”)

Nata dallo sforzo di armonizzazione fra sistema FRBR (ambito bibliografico) e CIDOC CRM (ambito museale)

Sperimentazione su un campione di 300 record in formato UNIMARC
relativi a monografie moderne, antiche e periodici

Si è proceduto:

- **Alla mappatura dei campi UNIMARC SBN-FRBRoo**
- **Alla conversione in formato RDF**
- **Allo sviluppo di un prototipo di interfaccia online per la gestione,**
  ricerca e pubblicazione sul Web dei dati in formato LOD
- **Alla pubblicazione attraverso consultazione e download delle risorse**
  (file RDF/XML), endpoint SPARQL e API

Prima della pubblicazione i dati hanno subito processi di :

- **Normalizzazione, mirata all'analisi della coerenza e correttezza delle informazioni**
- **Arricchimento, per favorire l'interoperabilità e il riutilizzo attraverso il collegamento a fonti esterne autorevoli (Geonames, VIAF)**

La rivoluzione nel mondo delle biblioteche

- **L'adozione dei linked data comporterà una modifica radicale dell'intermediazione bibliotecaria**

- **Tutti i settori del servizio bibliotecario ne saranno interessati**

- **in particolare ne guadagneranno**
  - **la ricerca sul WEB,**
  - **il controllo bibliografico, con la descrizione, la creazione degli accessi,**
  - **il controllo d'autorità, la classificazione, la portabilità dei dati**

L'OPAC del futuro basato sui Library Linked Data

- **A partire da un determinato soggetto, può essere creata dinamicamente una pagina in grado di stabilire collegamenti a documenti, film, immagini, risorse multimediali e collegamenti esterni**
  - **Esempio: data.bnf.fr. progetto della Bibliothéque nationale de France**

Il mondo delle biblioteche e il mondo del Web sono 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 significanti**

Le biblioteche possono "addomesticare" i Big Data?

- **Si, ma restano delle questioni aperte:**
  - **dobbiamo cominciare a convertire massivamente nei nuovi formati le registrazioni già disponibili ?**
  - **il lavoro potrà essere automatizzato o richiederà un intervento umano di riqualificazione, record per record, dato per dato?**
  - **E dobbiamo abbandonare sin da subito gli attuali gestionali, strutturati in modo novecentesco?**
  - **E le risorse disponibili e gli strumenti attualmente in uso sono in grado di supportare tutto ciò?**