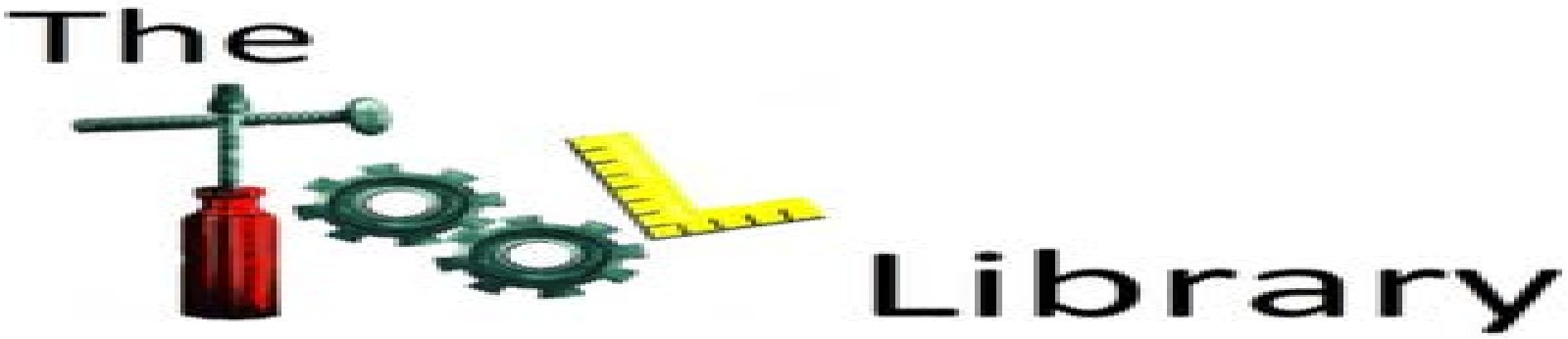
Big Data Tools

Biblioteche alla prova

nella gestione di grandi masse di dati



Big Data e biblioteche: impatto e nuove opportunità

Impatto

- **L'impatto dei "big data" necessita grossi impegni economici, tecnologici , operativi e potrebbe determinare un rivolgimento degli attuali assetti bibliotecari**
- **Una rivoluzione che non sembra alla portata delle piccole biblioteche e potrebbe significare un declino dei sistemi distribuiti con il ritorno a modelli più centralizzati**

Nuove opportunità

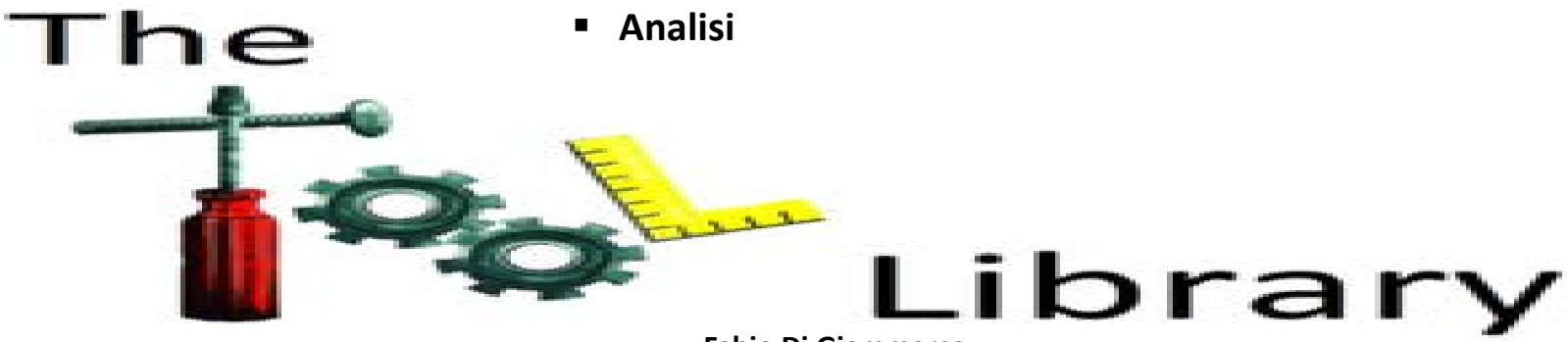
L'utilizzo dei Big Data in biblioteca, potrebbe:

- **Aprire scenari per nuovi ruoli e servizi delle biblioteche nell'ambito dell'innovazione tecnologica**
- **Potenziare e ottimizzare i flussi e le performance interne delle biblioteche**

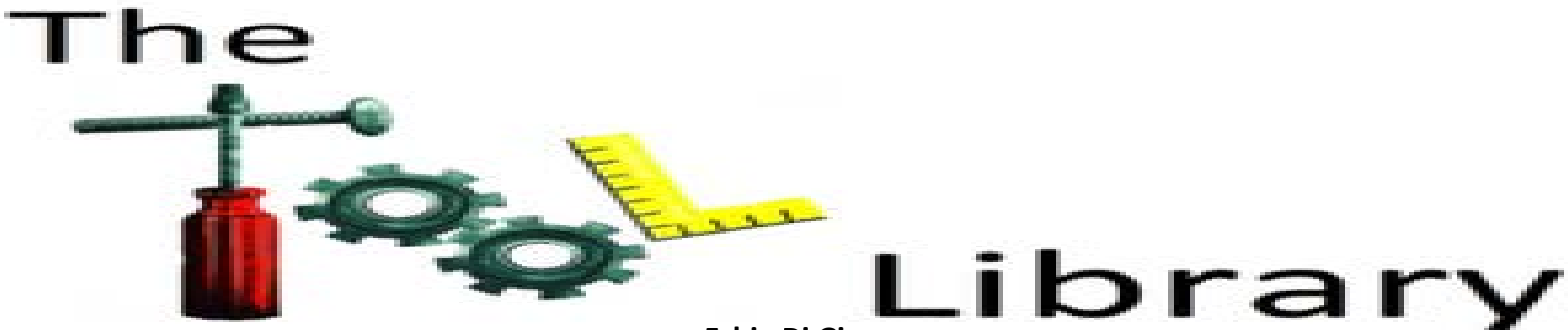
La sfida: nuovi servizi per Big Data

- **Lo spostamento del «focus» dalle collezioni ai "Big Data", non significa solo nuovi costi infrastrutturali**

- **Significa anche progettare e supportare nuovi servizi per la gestione e accesso ai dati:**
  - **Archiviazione**
  - **Ricerca**
  - **Condivisione**
  - **Trasferimento**
  - **Visualizzazione**
  - **Analisi**

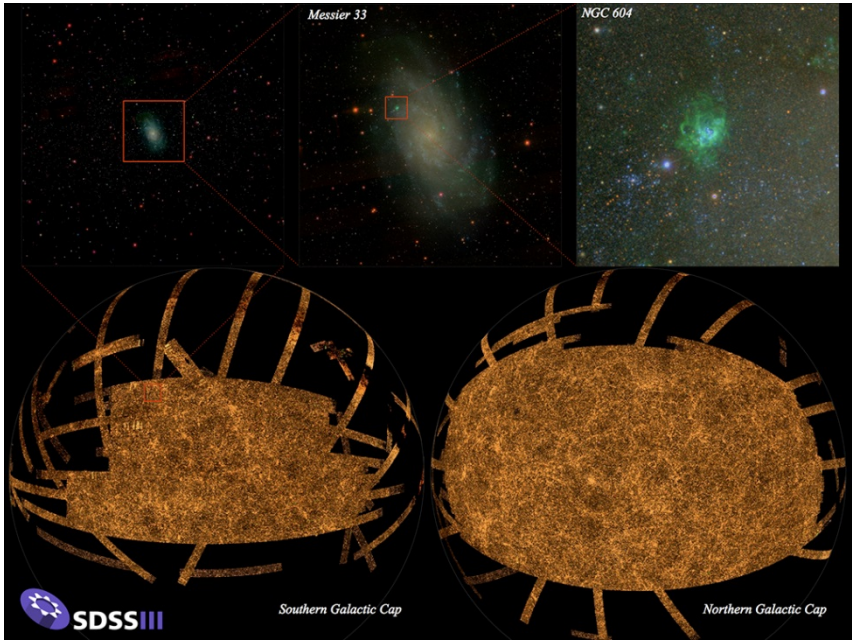


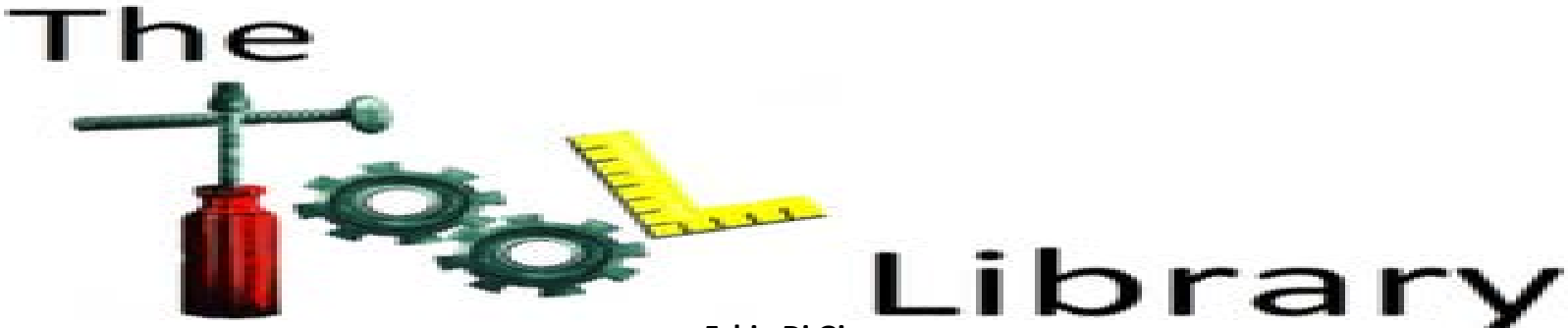
La sfida:
un “nuovo” bibliotecario?



Biblioteche e dati della ricerca

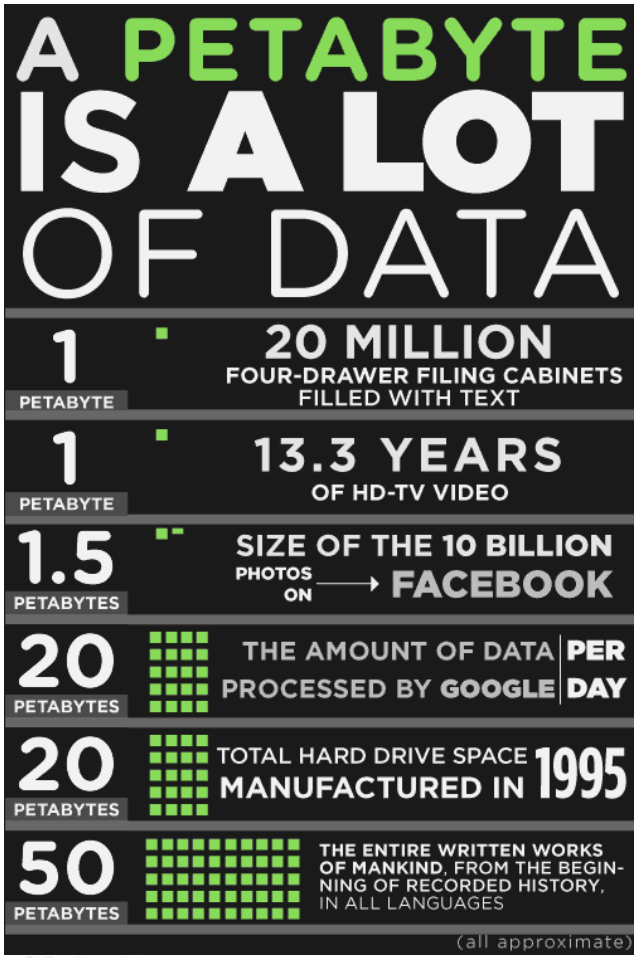
- **Lo “Sloan Digital Sky Survey ” è un progetto astronomico molto importante che prevede la mappatura dello spazio profondo.**
- **Finora sono stati catalogati circa 100 milioni di stelle, 1 milione di galassie e 100 mila quasar.**
- **Un’impresa titanica con conseguente produzione di un’enorme massa di dati.**
- **Per gestire la complessità di questi “big data” sono arrivate in aiuto le biblioteche della John Hopkins University di Baltimora**





Biblioteche e dati della ricerca

- **Da una recente inchiesta risulta che i ricercatori di Oxford – solo nel 2012 - avrebbero generato almeno 3 petabyte (3 milioni di gigabyte) di dati.**
- **In pratica, il doppio delle capacità attuali del sistema centrale dell'università.**
- **Dell'organizzazione di questa mole di dati se ne stanno occupando anche le biblioteche della Bodleian**



Biblioteche e dati della ricerca

- **L'UC3 Curation Center della California Digital Library ha cominciato a fornire servizi d'assistenza e supporto per l'intero ciclo dei dati.**
- **Tra questi l'implementazione di un sistema di storage (a pagamento) modulato in base alle diverse esigenze: da livelli completamente chiusi per le informazioni sensibili come quelle di carattere medico fino a data sets aperti con accesso pubblico e utilizzo di metadati.**

Biblioteche e dati della ricerca

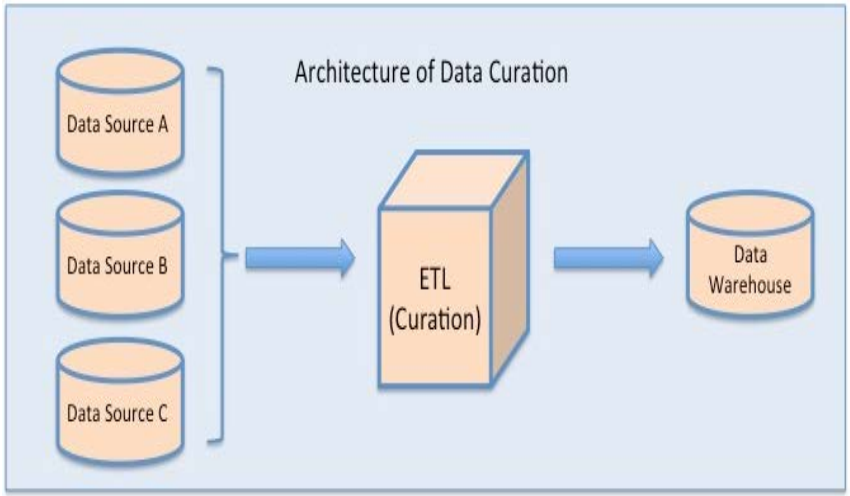
- **La Biblioteca accademica UC di San Diego partecipa a HathiTrust, una partnership internazionale di 52 biblioteche accademiche e di ricerca impegnati a conservazione digitale a lungo termine.**

- **È partner in Cyberinfrastucture Research Initiative dell'università (RCI), che offre ai ricercatori l'elaborazione, la rete, e l'infrastruttura necessaria per creare, gestire e condividere i dati**

- **Ha implementato il Research Data Curation Program**

Biblioteche e dati della ricerca

Secondo uno studio dell'Università del Tennessee su oltre 100 biblioteche universitarie, almeno il 40% è impegnato a sviluppare programmi per supportare gli scienziati nelle procedure di "big data curation"



Biblioteche e dati della ricerca

- ***"Le biblioteche hanno gestito i dati per secoli"*, ha affermato Marcy Strong, responsabile del servizio di metadati presso le biblioteche River Campus**

- **Il nuovo progetto di Data Science da 50 milioni di dollari dell'Università di Rochester, New York, si baserà anche sulle competenze dei bibliotecari del "River Campus Libraries"**

Settori dove è più richiesto il supporto delle biblioteche

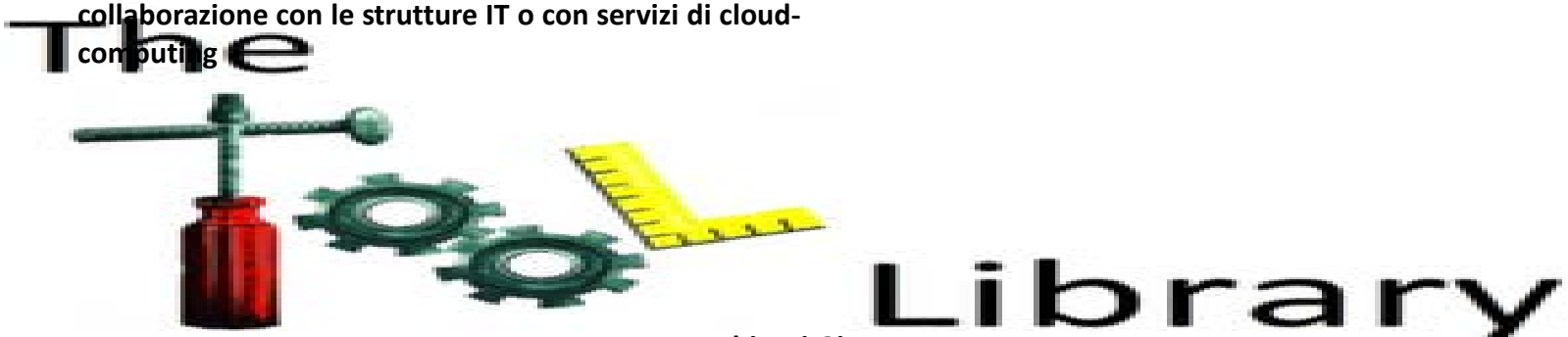
- **Consulenza su tipologia e formato dati**
- **Stime per lo storage**
- **Supporto tecnico e legale (copyright) per i ricercatori**

(Indagine università austriache e britanniche 2015: dati in linea con indagini anche in altri Paesi)

Dieci raccomandazione per le biblioteche per gestire di dati della ricerca

1. **Offrire assistenza nella gestione dei dati**
2. **Contribuire allo sviluppo dei metadati e standard dei dati e fornire servizi di metadatazione**
3. **Creare le figure professionali dei data librarian**
4. **Partecipare attivamente nelle creazione di policy sui dati della ricerca delle istituzioni**
5. **Collaborare con i ricercatori e gruppi di ricerca per la creazione di infrastruttrure interoperabili per l'accesso ai dati e alla condivisione dei dati**
6. **Sostenere il ciclo di vita dei dati fornendo servizi di archiviazione, discovery e accesso permanente**
7. **Promuovere l'utilizzo di identificatori persistenti per l'accesso permanente ai dati**
8. **Fornire un catalogo dei dati**
9. **Familiarità con la gestione di dati nelle varie discipline**
10. **Offrire o mediare l'archiviazione sicura in collaborazione con le strutture IT o con servizi di cloud-computing**

[Final report of the LIBER working group on E - Science / Research Data Management]



Biblioteche: supporto ai (Big)Open Data

I dati prodotti con soldi pubblici devono essere liberamente accessibili ai cittadini

The Tool Library

Non solo “Open” Data, ma “Intelligently Open”

- **Rendere i dati soltanto accessibili non è sufficiente**
- **I dati devono essere “intelligently open”, nel senso che devono poter essere accuratamente esaminati e opportunamente riutilizzati**
- **I dati devono essere di alta qualità per quanto possibile, affidabili, autentici, e di rilevanza scientifica**
- **Devono essere intelligibili - ci devono essere abbastanza informazioni di base per rendere chiara la pertinenza dei dati al problema specifico oggetto di indagine**
- **Devono essere arricchiti con un'adeguata descrizione tramite metadati**

Legge 221/2012
“L’Open Italia”

- **Il Legislatore italiano con la Legge 17 dicembre 2012, n. 221 ha formalizzato una definizione di dati aperti (formalmente "dati di tipo aperto") inserendola all'interno dell'art. 68 del Codice dell'Amministrazione Digitale.**
- **Secondo tale definizione, sono dati di tipo aperto, i dati che presentano le seguenti tre caratteristiche:**
- **a) sono disponibili secondo i termini di una licenza che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato;**
- **b) sono accessibili attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, in formati aperti ai sensi della lettera a), sono adatti all'utilizzo automatico da parte di programmi per elaboratori e sono provvisti dei relativi metadati;**
- **c) sono resi disponibili gratuitamente attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, oppure sono resi disponibili ai costi marginali sostenuti per la loro riproduzione e divulgazione.**
- **Tale definizione, in coordinamento con quanto disposto dall'articolo 52 dello stesso codice, rappresenta la base per il cosiddetto *principio open by default* ora presente nell'ordinamento italiano.**

(Wikipedia)

Open Data: le licenze

- **Licenze Creative Commons sono le più diffuse e hanno il vantaggio di essere utilizzate a livello internazionale.**
  - **Nell'ambito delle licenze Creative Commons, per i dati aperti solitamente si utilizzano queste combinazioni:**
    - **CC0, cioè completamente libero da diritti, e i dati possono essere utilizzati da terzi anche per fini commerciali senza obbligo di citare la fonte;**
    - **CC BY, i dati possono essere utilizzati da terzi anche per fini commercia, ma con obbligo di citare la fonte;**
    - **CC BY SA, i dati possono essere utilizzati con obbligo di citare la fonte, e debbono essere condivisi allo stesso modo, cioè se il materiale viene remixato o trasformato o ci si basa su di esso, è necessario distribuire i propri contributi con la stessa licenza del materiale originario.**

Licenze CC Open Data

- **Le licenze CC sono 5 + 1 (pubblico dominio) e rappresentano una via di mezzo tra copyright completo (*full-copyright*) e pubblico dominio (*public domain*): da una parte la protezione totale realizzata dal modello *all rights reserved* ("tutti i diritti riservati") e dall'altra *no rights reserved* ("assenza totale di diritti").**
- **La filosofia su cui si fonda lo strumento giuridico delle licenze CC si basa sul concetto *some rights reserved* ("alcuni diritti riservati"): in questo senso è l'autore di un'opera che decide quali diritti riservarsi e quali concedere liberamente.**
  (Wikipedia)

- Attribuzione

- Attribuzione
- Condividi allo stesso modo

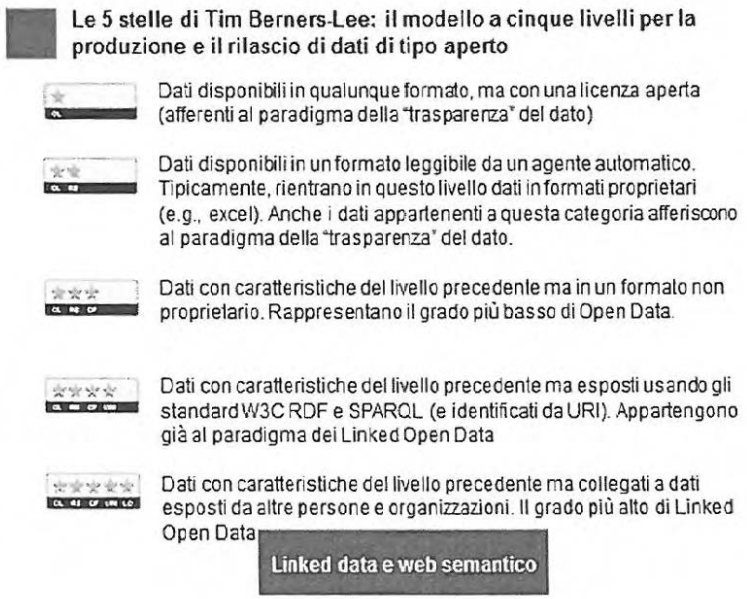
- Attribuzione
- Non opere derivate

- Attribuzione
- Non commerciale

- Attribuzione - Non commerciale
- Condividi allo stesso modo

- Attribuzione - Non commerciale
- Non opere derivate

Le 5 stelle di Tim



Un modello di Government's open data

DATA.GOV

Biblioteche partner di sistemi pubblici basati su i (Big) Open Data

James R. Jacobs, bibliotecario di Stanford sostiene:

- **che le biblioteche a partire dagli standard di metadati e dalle strategie di conservazione delle informazioni**
- **possono candidarsi a partner fondamentali nei processi di costruzione di sistemi pubblici basati sugli “open data” e “big data”**

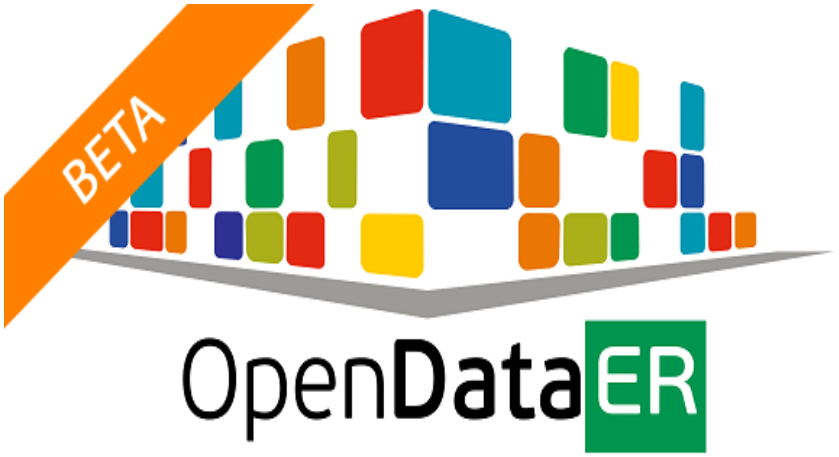
Biblioteche partner di sistemi pubblici basati su i (Big) Open Data

William Michener – coordinatore e-science - del sistema bibliotecario dell'Università di New Mexico

- **Riguardo ai data sets relativi alla ricerca finanziata dal governo federale, ha fatto presente l'importanza di conservarli correttamente per poi renderli disponibili a tutta la comunità scientifica**
  - **Compito che secondo Michener solo le biblioteche sono in grado di assolvere al meglio**

Le Biblioteche pubbliche degli Enti Locali emiliano romagnoli

- **Raccolgono, organizzano e diffondono informazioni e documenti**
- **Concorrendo, con le loro specifiche funzioni, a promuovere le condizioni che rendono effettivo il diritto all'informazione, allo studio, alla cultura, alla formazione e all'impiego del tempo libero di tutti i cittadini**
- **Il dataset raccoglie informazioni sulle Biblioteche di ente locale, i Conservatori degli archivi storici di ente locale e di interesse locale e i Musei presenti sul territorio dell'ambito provinciale di Reggio Emilia**



Un progetto Open: i 25 milioni di records della Library of Congress

Carla Hayden, bibliotecaria della LOC, ha dichiarato:

"La Biblioteca del Congresso è un monumento alla conoscenza della nostra nazione, e dobbiamo assicurarci che le porte siano aperte a tutti, non solo fisicamente ma anche digitalmente…rendere accessibili i dati bibliografici del catalogo on line è un grande passo in avanti. Sono impaziente di vedere come le persone utilizzeranno queste informazioni"

Un progetto di (Linked)Open Data culturali

Il Sistema Archivistico Nazionale (SAN) - aggregatore nazionale di risorse archivistiche e digitali statali e non statali, pubbliche e private - rende disponibile secondo il modello Linked Open Data le proprie schede descrittive delle entità archivistiche: soggetti conservatori, soggetti produttori, profili istituzionali, complessi archivistici, strumenti di ricerca.

La pagina http://www.san.beniculturali.it/web/san/dati-san-lod permette l'accesso a :

» Ontologia SAN LOD in formato OWL e ad una sua rappresentazione HTML;

» Tesauro SAN in formato SKOS ed in formato grafico navigabile;

» 97 datasets disponibili per il download nei formati turtle, rdf/xml e csv per un totale complessivo di 1,978 GB

» Endpoint SPARQL per l'interrogazione diretta sui Linked Open Data;

» Esplorazione diretta delle entità archivistiche (Soggetti Conservatori, Soggetti Produttori, Complessi Archivistici) con la loro rappresentazione come Linked Oped Data arricchita dei dati derivanti dalle attività di riconciliazione.

Contenuti: 5.312.474 triple rdf e 128 datasets disponibili in download nei formati turtle, rdf/xml e csv relative a schede descrittive delle entità archivistiche (dati al 3 Febbraio 2015).