JLIS.it

OpLiDaF
Open Linked Data Framework: una piattaforma per la creazione e la pubblicazione di linked data

Tiziana Possemato

Il progetto ITACH@

Il progetto ITACH@, Innovative Technologies And Cultural Heritage Aggregation, ha l'obiettivo di proporre strumenti innovativi per la valorizzazione dell'industria turistica e culturale italiana. Il sistema proposto nel progetto, in corso di studio e di strutturazione, ha come campo di applicazione l'intero complesso informativo prodotto da enti e istituzioni afferenti all'ambito dei beni culturali (le biblioteche, gli archivi, i musei e gli enti turistici); è pensato, cioè, per essere esteso ad ambiti affini, contigui o relazionati. L'idea progettuale si innesta in un contesto in cui si individuano:

- un'esigenza poco riconosciuta e non soddisfatta di un accesso integrato ai dati, al di là della loro eterogeneità, quantità, distribuzione, proprietà;
- la necessità di condivisione e utilizzo (e ri-utilizzo) dei dati: organizzazioni e individui che scelgono di condividere i

Traduzione italiana a cura dell'Autore.

dati e che traggono vantaggio dalla creazione di 'ecosistemi' organizzati e fruibili.

E che sollecitano alcune domande fondamentali:

- qual è il modo migliore per fornire l'accesso ai dati così che possano essere facilmente riutilizzati?
- come rendere possibile la scoperta di dati pertinenti all'interno della moltitudine di insiemi di dati disponibili?
- come consentire alle applicazioni di integrare dati provenienti da fonti eterogenee e sconosciute?

La riflessione su queste tematiche pone il progetto ITACH@ all'interno della più ampia tematica del web semantico: la pubblicazione dei dati secondo gli standard e le buone pratiche previste da questo ambito, e le sue declinazioni tecnologiche, come i linked data.

La piattaforma OpLiDaF

In particolare ci soffermeremo su una componente tecnologica del sistema, l'OpLiDaF , che è concepita come un framework per la creazione, strutturazione e visualizzazione di dati in formato Resource Description Framework (RDF)/XML. Intende essere una piattaforma ideata per il trattamento (inteso come mappatura, conversione, pulizia e pubblicazione) di linked data a partire da dati in formati eterogenei, attraverso tool e procedure scritti ad hoc, oppure sistemi open source opportunamente integrati, e attraverso l'utilizzo di standard e linguaggi riconosciuti nell'ambito del web semantico. Le funzioni principali previste nella piattaforma OpLiDaF sono:

- selezione di ontologie;

- mappatura tra i dati di origine e l'ontologia o le ontologie selezionate;
- creazione di specifiche ontologie a partire da un set di dati;
- produzione di file RDF/XML;
- bonifica del dato (data cleaning).

Il sistema OpLiDaF ha origine dall'osservazione della composizione e tipologia dei dati, eterogenei, sia in termini di contenuti sia di formati, che costituiscono il blocco informativo delle biblioteche, degli archivi, dei musei, dei contesti turistici, territoriali e di altre realtà. Potremmo sostenere che l'elenco qui proposto rispetta un andamento e un ordine decrescente rispetto all'uso di formati standard riconosciuti: dalle biblioteche, che sono certamente gli enti che più di altri hanno storicamente utilizzato standard per la strutturazione e la pubblicazione dei propri dati, ad ambiti in cui i dati sono raccolti in tabelle access, excel, CSV etc. Le biblioteche stesse, detentrici di un primato di standardizzazione soprattutto nei diffusi formati Machine Readable Cataloguing (MARC), affiancano a questi dati, relativi soprattutto alle descrizione bibliografiche e di authority, altri dati in formati eterogenei: pensiamo a quelli più tipicamente gestionali (anagrafiche degli utenti, dati su prestiti e prenotazioni, dati sugli acquisti, etc.) come ai dati descrittivi e gestionali dei periodici e dei seriali (per riportare esempi diffusi), molto spesso gestiti, per comodità e facilità o tradizione, al di fuori del database bibliografico centralizzato. In alcune realtà, non sporadiche e isolate, per lo meno in ambito italiano, anche tipologie di risorse differenti (archivi fotografici, materiali musicali, raccolte digitali) sono gestite su database diversi rispetto a quello bibliografico centrale. Questa composizione eterogenea e sfaccettata di fonti informative diventa tanto più evidente quanto più ci si allontana

da contesti tradizionalmente biblioteconomici per arrivare a quelli quali archivi e musei.

La pubblicazione dei linked data a partire da database relazionali

L'analisi condotta su questa eterogenea varietà di dati, molti dei quali di grande interesse pubblico, è accompagnata dalla consapevolezza che una loro conversione in linked data, secondo i principi, gli standard e le pratiche ormai riconosciute e diffuse, non comporterebbe l'abbandono dei rispettivi sistemi nativi di gestione, né delle applicazioni di business, ma semplicemente l'aggiunta di uno strato tecnologico supplementare per collegare questi dati nel web semantico. Analizziamo nella figura 1 a fronte un possibile workflow di pubblicazione di dati eterogenei in linked data.
Senza scendere nel dettaglio delle differenti ipotesi di workflow, vogliamo attirare l'attenzione sull'alto potenziale di trasformazione, attraverso percorsi e strumenti differenti, di dati per il web semantico (sia dati strutturati sia dati testuali, altra enorme ricchezza poco sfruttata nel web tradizionale rispetto all'alto potenziale informativo), con l'interessante scenario che così si viene ad aprire in termini di utilizzo (e riutilizzo) dei dati, senza necessariamente intervenire sui sistemi legacy in uso presso gli enti (legacy sono definiti i sistemi[1] informatici esistenti o un'applicazione che continua a essere usata poiché l'utente non vuole o non può sostituirla). Le politiche e le pratiche di pubblicazione dei dati nel web semantico cambiano in relazione a differenti fattori, tra cui:

- il formato di origine del dato (dato strutturato o dato testuale);

[1] http://it.wikipedia.org/wiki/Sistema_informatico

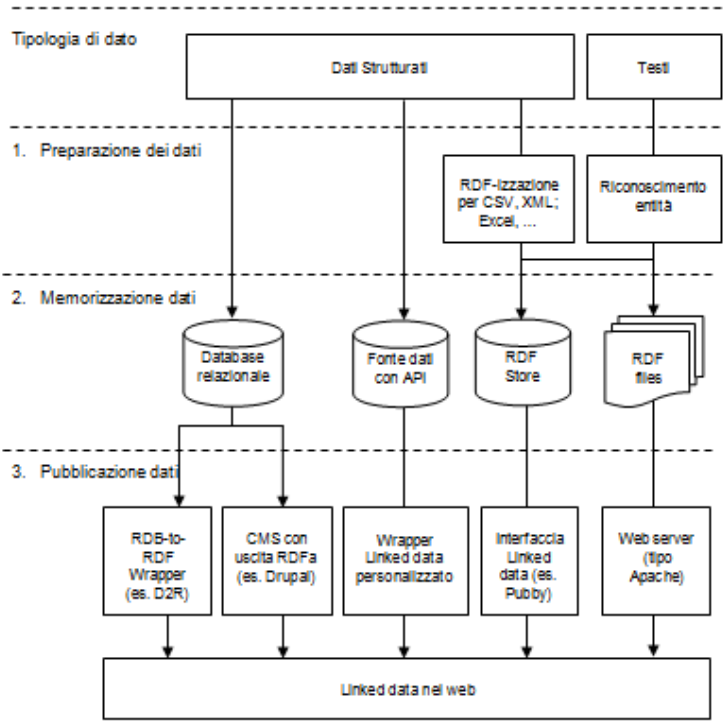

Figura 1: Workflow di pubblicazione di dati eterogenei in linked data.

- la quantità di dati che si intenda includere in un data set;
- la periodicità di aggiornamento dei dati.

OpLiDaF si concentra in particolare sul primo e sul terzo dei fattori citati, quelli relativi alla differente struttura dei dati originari e alla loro necessità di aggiornamento, puntando su una metodologia tecnologica che produca uno strato trasversale unico finalizzato a convogliare e coordinare le differenti esigenze gestionali di questi eterogenei dati. Se concentriamo la nostra attenzione sull'ambito delle biblioteche, non possiamo fare a meno di pensare al trattamento di dati dal formato MARC (in particolare dal MARC21 all'RDF/XML). Si tratta di un processo noto, supportato da un'ampia letteratura; possiamo considerarlo come il primo passo per una biblioteca per pubblicare i propri dati nel web semantico. Preferiamo, dunque, concentrarci su un ambito meno frequentato di quello della conversione dal MARC21, soffermandoci sulle procedure e tecniche di trattamento dei dati contenuti in database relazionali, per analizzare il potenziale del sistema OpLiDaF che utilizza linguaggi di mappatura e standard riconosciuti. Molti dati bibliografici strutturati in MARC21 sono memorizzati in database relazionali, consentendo in fase di export o in caso di accessi esterni al dato (per esempio da client Z39.50) una ricomposizione del dato in formato MARC. L'esercizio e lo studio su come tradurre dati da database relazionali in linked data risulta di particolare interesse anche per i dati bibliografici e di authority, essendo la rappresentazione relazionale del dato disgiunta dalla rappresentazione in MARC. I dati conservati in database relazionali possono essere facilmente pubblicati come linked data attraverso l'uso di un insieme di tool oggi disponibili, che partono da processi di mappatura dai database relazionali in grafi RDF, poi pubblicati sul web secondo i principi dei linked data. Questa possibilità diventa tanto più interessante se si pensa all'enormità di dati gestionali interni, prodotti e conservati su sistemi

legacy, non destinati necessariamente al web inteso come spazio aperto e pubblico, ma destinati, per esempio, alle intranet aziendali: la stessa tecnologia dei linked data destinata a un uso interno ma altrettanto utile e necessario per la diffusione controllata delle informazioni esistenti. Il W3C RDB2RDF Working Group sta lavorando alla elaborazione di linguaggi standard per la mappatura di dati relazionali e di schemi di database relazionali in RDF e Web Ontology Language (OWL): i due principali linguaggi a oggi disponibili sono il Direct Mapping (DM) e il RDB2RDF Mapping Language (R2RML). Da un punto di vista tecnologico uno dei tool più diffusi e utilizzati per la pubblicazione di database relazionali nel web semantico è il D2R Server, che consente a browser RDF e HTML di navigare i contenuti del database utilizzando SPARQL come linguaggio di ricerca. Si tratta di standard e tecnologie ampiamente riconosciute nel semantic web, ma quello che a noi interessa è mostrare il potenziale di un altro linguaggio di mappatura tra schemi di database relazionali e ontologie implementate in RDF(S) o OWL, utilizzato nella piattaforma OpLiDaF : l'R2O (Relational to Ontology), che consente di produrre un set estensibile di primitive con una semantica esplicita e ben riconosciuta. L'R2O è un linguaggio di alto livello indipendente dall'RDBMS (nel nostro caso Oracle) che opera con database che utilizzino l'SQL standard. L'R2O è ispirato al D2R, ma cerca di superarne due principali limiti:

- R2O è più potente ed elastico, e dunque più adatto allo sviluppo di mappature complesse, colmando un certo grado di espressività che manca al D2R;

- R2O è un linguaggio dichiarativo (consente cioè di specificare cosa vogliamo ottenere, senza descrivere come ottenere il risultato), al contrario del D2R che non lo è completamente.

Un presupposto di utilizzo dell'R2O è che il database e l'ontologia (implementata in OWL/RDF) condividano un elevato grado di somiglianza nella struttura, presupponendo che sia il database sia l'ontologia utilizzata siano pre-esistenti e non necessitino di modifiche per essere utilizzate. Per mostrare il flusso di generazione di data set RDF a partire da un database relazionale all'interno della piattaforma OpLiDaF abbiamo utilizzato l'approccio di selezionare, ai fini della mappatura, delle ontologie esistenti, invece di generare in modalità semi-automatica l'ontologia a partire dal database relazionale (altra strategia possibile, utilissima nei contesti in cui non fossero disponibili ontologie utilizzabili). Le ontologie utilizzate per questo studio, a titolo di test, sono:

- Bibtex = http://bibotools.googlecode.com/svn/bibo-ontology/tags\/1.3/bibo.xml.owl
- Bibo = http://zeitkunst.org/bibtex/0.2/bibtex.owl

Il database relazionale è Oracle, che contiene dati bibliografici ed è strutturato, per ottenere una semplificazione necessaria ai fini dimostrativi, in due differenti viste realizzate per poter mappare le due differenti ontologie: BOOK (mappata sull'ontologia bibtex) e PARTI (mappata sull'ontologia bibo). Altri strumenti utilizzati per questo studio:

- Neon ToolKit: un ambiente di progettazione per ontologie, open source e multipiattaforma. È basato sulla piattaforma di sviluppo Eclipse e mette a disposizione decine di plug-in utili a coprire una notevole varietà di funzioni legate al ciclo di vita delle ontologie. Uno dei plug-in utilizzati è ODEMapster. Neon ToolKit è stato sviluppato come parte del progetto NeOn [2] e supportato dalla Fondazione NeOn [3];

[2] http://www.neon-project.org.
[3] http://www.neon-foundation.org.

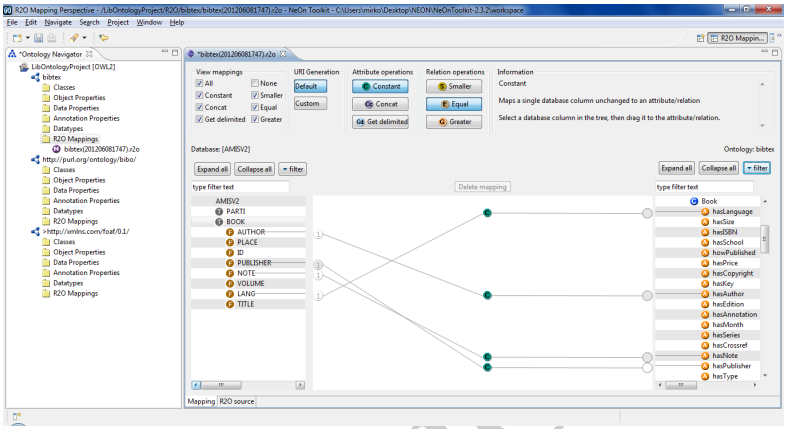

Figura 2

- ODEMapster: plug-in per Neon ToolKit: consente le operazioni di mappatura tra le tabelle del database relazionale e l'ontologia selezionata in una modalità estremamente semplificata e soprattutto guidata, come mostrato nella figura successiva, relativa alla fase di mappatura della vista BOOK sull'ontologia bibtex Ciascun dato presente in una colonna della tabella selezionata può essere mappato con una classe o attributo opportunamente individuati nella ontologia utilizzata.

La sezione sinistra dell'immagine 2 riporta l'elenco delle ontologie utilizzate o utilizzabili (si veda tra queste anche il vocabolario FOAF (Friend of A Friend), inserito ma non utilizzato nella sperimentazione). La sezione sinistra della parte centrale dello schermo riporta i campi della tabella che si intende mappare sull'ontologia selezionata, visibile nella sezione di destra della parte centrale dello schermo, dove Book rappresenta la classe e i pallini in giallo gli attributi. La selezione dei campi del database da mappare con gli attributi del-

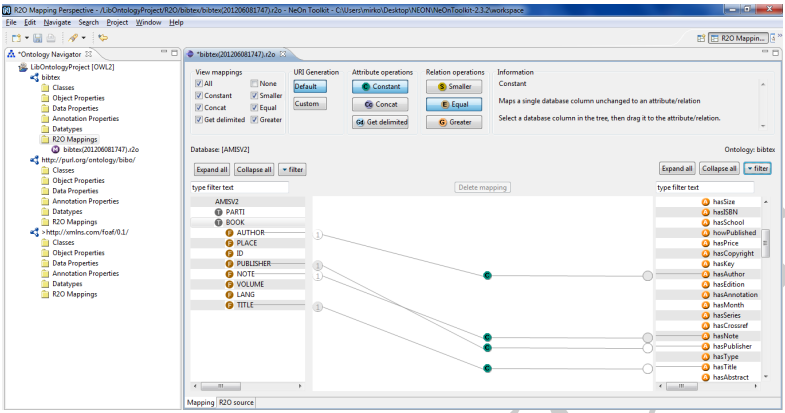

Figura 3

l'ontologia dipende dalla volontà di pubblicazione e condivisione del dato da parte dell'ente. Nel nostro caso abbiamo realizzato una mappatura semplice:

- Campo AUTHOR Bibtex.hasAuthor
- Campo TITLE Bibtex.HasTitle
- Campo PUBLISHER Bibtex.hasPublisher
- Campo NOTE Bibtex.hasNote
- Campo LANG Bibtex.hasLanguage

La fase successiva alla mappatura tra il database relazionale e l'ontologia è la produzione del file R2O, che è l'XML che descrive in forma di linguaggio la mappatura grafica tra database e ontologia. Serve poi a ODEMapster per generare l'RDF (il listato 1 a fronte ne riporta solo una sezione).

Listing 1: Sezione di file RDF generato da ODEMapster

```
<?xml version="1.0" encoding="UTF-8"?>
  <r20>
    <dbschema-desc name="AMISV2">
      <has-table name="PART1">
      <has-table name="BOOK">
        <nonkeycol-desc name="AUTHOR" />
        <nonkeycol-desc name="PLACE" />
        <nonkeycol-desc name="ID" />
        <nonkeycol-desc name="PUBLISHER" />
        <nonkeycol-desc name="NOTE" />
        <nonkeycol-desc name="VOLUME" />
        <nonkeycol-desc name="LANG" />
        <nonkeycol-desc name="TITLE" />
      </has-table>
    </dbschema-desc>
    <conceptmap-def name="http://purl.org/net/nknouf/ns/bibtex
        #Book">
      <uri-as type="DEFAULT">
        <operation oper-id="concat">
          <arg-restriction on-param="string1">
            <has-value>http://purl.org/net/nknouf/ns/bibtex#
                Book</has-value>
          </arg-restriction>
          <arg-restriction on-param="string2">
            <has-column>AMISV2.BOOK.AUTHOR</has-column>
          </arg-restriction>
        </operation>
      </uri-as>
      <default_uri-as>
        <operation oper-id="concat">
          <arg-restriction on-param="string1">
            <has-value>http://purl.org/net/nknouf/ns/bibtex#
                Book</has-value>
```

```
        </arg-restriction>
        <arg-restriction on-param="string2">
          <has-column>AMISV2.BOOK.AUTHOR</has-column>
        </arg-restriction>
      </operation>
    </default_uri-as>
    <described-by>
      <attributemap-def name="http://purl.org/net/nknouf/ns/
          bibtex#hasLanguage">
        <selector>
          <aftertransform>
            <operation oper-id="constant">
              <arg-restriction on-param="const-val">
                <has-column>AMISV2.BOOK.LANG</has-column>
              </arg-restriction>
            </operation>
          </aftertransform>
        </selector>
      </attributemap-def>
      <attributemap-def name="http://purl.org/net/nknouf/ns/
          bibtex#hasAuthor">
```

Come terzo passo, il sistema interroga la base dati, estrae i record e li mappa nel formato RDF, secondo gli schemi impostati nelle fasi precedenti.
Si riporta un estratto del file RDF solo per semplificarne la lettura:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns\#"
    xmlns:j.0="http://purl.org/net/nknouf/ns/bibtex\#" >
  <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
      bibtex\#BookGoni\%2C_Enrico">
    <j.0:hasVolume> </j.0:hasVolume>
    <j.0:hasPublisher>All'insegna del Veltro</j.0:hasPublisher
        >
```

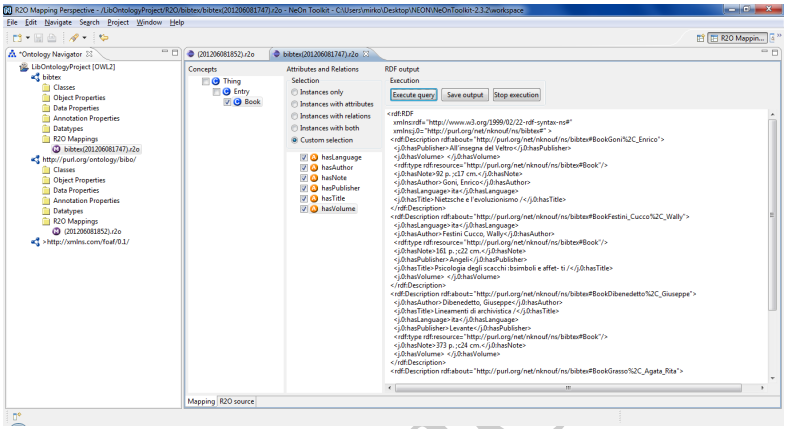

Figura 4

```
    <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
        bibtex\#Book"/>
    <j.0:hasLanguage>ita</j.0:hasLanguage>
    <j.0:hasAuthor>Goni, Enrico</j.0:hasAuthor>
    <j.0:hasNote>92 p. ; c17hcm.</j.0:hasNote>
    <j.0:hasTitle>Nietzsche e l'evoluzionismo /</j.0:hasTitle>
  </rdf:Description>
  <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
      bibtex\#BookFestini_Cucco\%2C_Wally">
    <j.0:hasLanguage>ita</j.0:hasLanguage>
    <j.0:hasAuthor>Festini Cucco, Wally</j.0:hasAuthor>
    <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
        bibtex\#Book"/>
    <j.0:hasNote>161 p. ; c22 cm.</j.0:hasNote>
    <j.0:hasPublisher>Angeli</j.0:hasPublisher>
    <j.0:hasTitle>Psicologia degli scacchi : bsimboli e affet-
        ti /</j.0:hasTitle>
    <j.0:hasVolume> </j.0:hasVolume>
```

```
</rdf:Description>
<rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
    bibtex\#BookDibenedetto\%2C_Giuseppe">
  <j.0:hasAuthor>Dibenedetto, Giuseppe</j.0:hasAuthor>
  <j.0:hasTitle>Lineamenti di archivistica /</j.0:hasTitle>
  <j.0:hasLanguage>ita</j.0:hasLanguage>
  <j.0:hasPublisher>Levante</j.0:hasPublisher>
  <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
      bibtex\#Book"/>
  <j.0:hasNote>373 p. ; c24 cm.</j.0:hasNote>
  <j.0:hasVolume> </j.0:hasVolume>
</rdf:Description>
<rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
    bibtex\#BookGrasso\%2C_Agata_Rita">
  <j.0:hasLanguage>ita</j.0:hasLanguage>
  <j.0:hasTitle>Le difficolta di apprendimento: guida
      bibliografica : testi per gli alunni e volumi per gli
      insegnanti/</j.0:hasTitle>
  <j.0:hasAuthor>Grasso, Agata Rita</j.0:hasAuthor>
  <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
      bibtex\#Book"/>
  <j.0:hasNote>94 p. ; 24 cm.</j.0:hasNote>
  <j.0:hasVolume> </j.0:hasVolume>
  <j.0:hasPublisher>Edizioni del cerro</j.0:hasPublisher>
</rdf:Description>
```

Il file RDF può essere messo a confronto con il contenuto del database relazionale illustrato nella figura sottostante:

Il data cleaning

Da una successiva analisi del file RDF prodotto possiamo verificare alcuni limiti ed errori del risultato rispetto a quanto previsto dai

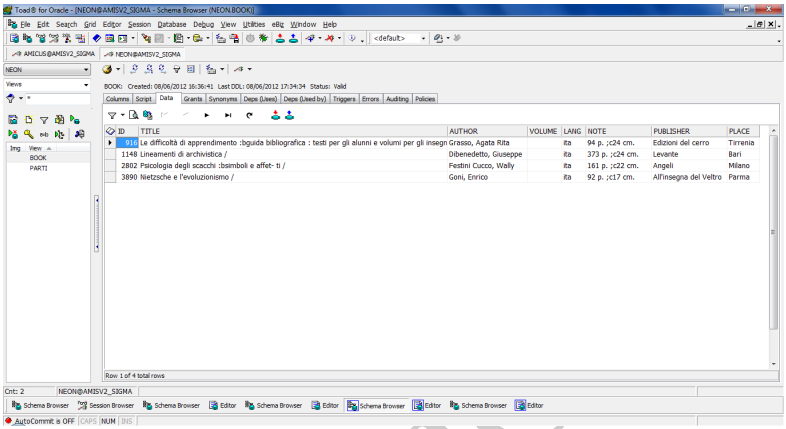

Figura 5

princìpi di produzione dei linked data. Riportiamo casi di errori a titolo esemplificativo, tenendo in considerazione che nel nostro caso alcuni di questi errori sono dovuti alla povertà contenutistica dei dati utilizzati, e quindi alla bassa espressività raggiungibile:

- Il file presenta un numero relativamente basso di asserzioni e relazioni tra entità ed entità (nell'esempio riprodotto l'unica relazione è con l'entità Type);

- la maggior parte delle asserzioni ha dei letterali come oggetto, rendendo povere e isolate le risorse RDF: l'autore del nostro esempio dovrebbe essere un'entità dotata di propria autonomia e referenziata da un URI, e non un letterale, dunque: non `<j.0:hasAuthor>Goni, Enrico</j.0:hasAuthor>` ma `<j.0:hasAuthor rdf:resource= http://atcult.it/autori/283235467/>`

- sono presenti in alcuni casi i caratteri separatori con relativi codici di sottocampo, ereditati dalla strutturazione del dato salvato nelle tabelle Oracle (si tratta dei codici di sottocampo presenti nel record MARC21 prodotto in fase di catalogazione): si veda l'esempio `<j.0:hasNote>94 p. ;24 cm.</j.0:hasNote>`

- dove è presente il codice di sottocampo $c del tag 300 del record, prima del campo relativo alle dimensioni della risorsa;

- alcune asserzione sono invalide non avendo un oggetto e pertanto risultando non esprimibili come triple (che devono essere composte da soggetto-predicato-oggetto).

Listing 2: Esempio di tripla RDF non valida: chi è l'autore?

```
<rdf:Description
rdf:about=http://purl.org/net/nknouf/ns/bibtex\#
    BookDibenedetto\%2C Giuseppe>
  <j.0:hasAuthor></j.0:hasAuthor>
</rdf:Description>
```

Sulla base dei risultati di questa analisi condotta sul file RDF prodotto in OpLiDaF è possibile attivare una serie di procedure per giungere a quella che viene definita la fase di data cleaning o bonifica dei dati, tra cui:

- l'utilizzo di tool di pulizia per l'eliminazione di caratteri sporchi ben identificabili, come i codici di sottocampo dei tag MARC21;

- l'identificazione di processi di scansione delle triple per verificarne la validità;

- la scrittura di procedure di verifica ed individuazione di triple letterali a fronte di triple RDF;
- la creazione automatica di entità identificabili da Uniform Resource Identifier (URI) attraverso l'utilizzo, per esempio, di identificatori univoci, nella maggior parte dei casi già presenti nei database relazionali, o creati secondo criteri stabiliti.

Quando si parla di condivisione dei dati la qualità rimane sempre un obiettivo importante da perseguire, perché diventa caratteristica fondamentale per la selezione del data set prodotto anche da parte di terzi, che vogliano condividere e collegare a esso i propri dati.

OpLiDaF e il ciclo di vita dei linked data

A questo punto proviamo a verificare, partendo dal ciclo di vita dei linked data che può essere suddiviso in vari step e che noi assumiamo diviso in sette passi («Methodological Guidelines for Publishing Government Linked Data»), quanto la piattaforma OpLiDaF può coprire:

1. identificazione della fonte dati;
2. modellizzazione del vocabolario;
3. generazione dei dati in formato RDF, tramite i diversi linguaggi di mappatura disponibili;
4. pubblicazione dei dati in RDF;
5. bonifica dei dati prodotti;
6. creazione di collegamenti tra data set differenti;

7. messa a disposizione dei dati, con differenti passi, tra cui la pubblicazione del data set ottenuto dal processo sul CKAN Registry (Comprehensive Knowledge Archive Network).

I passi da 2 a 5 sembrano poter essere pienamente soddisfatti dalla piattaforma, che si propone, dunque, come uno strumento utile per chiunque voglia arrivare alla produzione di linked data (a prescindere dal sistema di gestione, dal formato dei dati, dalla grandezza del data set e dalla modalità e ciclicità di aggiornamento), risolvendo una parte di ostacoli e problematiche che il passaggio dal web tradizionale al web semantico pone.

Riferimenti bibliografici

Villazòn-Terrazas, Boris, Luis M. Vilches-Blàsquez e Kristin Gòmez-Pérez Asunciòn. «Methodological Guidelines for Publishing Government Linked Data». (2011): 27–49. (Cit. a p. 17).

Ai fini di una corretta indicizzazione, si invitano i lettori a citare esclusivamente il testo in lingua inglese; l'unico, infatti, che presenta l'indicazione del numero di pagina, l'abstract, le keywords e le date del processo redazionale.

Possemato, T. "OpLiDaF - Open Linked Data Framework: a platform for the creation and publication of linked data". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6313. DOI: 10.4403/jlis.it-6313. Web.