View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by AMS Tesi di Laurea

Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

Semantic Publishing: analisi di Linked Open Dataset

Relatore: Chiar.mo
Dott. Angelo Di Iorio

Presentata da:
Riccardo Giorgi

Sessione I
Anno Accademico 2013-2014

Indice

Introduzione

I recenti sviluppi nelle tecnologie Web hanno determinato una rivoluzione nell'*electronic publishing,* dove le pubblicazioni scientifiche vengono diffuse in un formato digitale, come il *PDF*, ovvero in una forma pensata per la consultazione da parte degli utenti.

L'avvento del *Semantic Web* ha rinnovato il *World Wide Web*, trasformandolo in un ambiente dove le risorse vengono accompagnate da una marcatura semantica in modo da renderle adatte per l'elaborazione automatica.
Questa evoluzione ha dato quindi origine al *Semantic Publishing,* ossia la pubblicazione di documenti, come appunto le pubblicazioni scientifiche, utilizzando i principi del Web Semantico.
Lo scopo principale è quello di arricchire i contenuti e strutturare le informazioni semantiche in modo che siano processabili da agenti software e che sia possibile sfruttare queste informazioni, ad esempio, per la ricerca e per l'analisi.

Le informazioni vengono rappresentate mediante le *ontologie*, ossia vocabolari condivisi utilizzati per concettualizzare un dominio di interesse.
Tra le più comuni troviamo *BIBO*, utilizzato per descrivere le informazioni bibliografiche di una pubblicazione, e *FOAF*, che si occupa solitamente di descrivere le persone.

Le informazioni semantiche vengono poi gestite attraverso i *Linked Dataset,* collezioni di dati collegati fra loro utilizzando il *formato RDF*, in modo da rendere i diversi *repository* interoperabili.

Ogni dataset gestisce le proprie entità in maniera differente, a seconda delle proprie esigenze.
Ad esempio, *J.UCS* utilizza principalmente FOAF per descrivere gli autori, includendo informazioni come l'indirizzo mail, mentre *DBLP RKB Explorer* usa l'ontologia AKT e per ogni persona viene specificato solo il nome completo.

Tuttavia, i Linked Dataset presentano una serie di difetti, per di più dovuti alla qualità dei dati, ovvero alla pubblicazione di informazioni che possono essere incomplete o, peggio, non corrette.

Molti dataset, infatti, dichiarano di pubblicare diverse informazioni che però vengono gestite in maniera non uniforme.
Come nel caso di *JISC Open Citations*, che si occupa di gestire le citazioni, ma, come vedremo successivamente, su molte entità che rappresentano gli articoli queste informazioni vengono omesse.

Il nostro lavoro si occupa, quindi, di fare una rassegna esaustiva di alcuni *Linked Open Dataset* nel contesto delle pubblicazioni scientifiche, cercando di inquadrare la loro eterogeneità ed identificando i principali pregi e difetti di ciascuno.
Per ciascun dataset è stata effettuata un'analisi, prendendo un campione delle entità presenti e contando la presenza di ogni attributo.

Nel primo capitolo, diamo una descrizione più dettagliata del Semantic Publishing, esponendo i principi del Linked Data.
In seguito, spieghiamo com'è stata effettuata l'analisi dei dataset, descrivendo quali aspetti abbiamo preso in considerazione.
Nel terzo capitolo vengono elencate le ontologie più comuni trovate nei vari dataset, dando una breve descrizione di ciascuna e citando le proprietà più riscontrate nelle entità esaminate.
Successivamente, passiamo alla rassegna dei dataset: vengono esposti i risultati delle analisi, evidenziando i difetti trovati e dando un giudizio personale per ognuno.

Nel quinto capitolo, inoltre, descriviamo come abbiamo cercato di risolvere un problema abbastanza comune, creando un tool per la *disambiguazione degli autori*.
All'interno di un dataset, infatti, può capitare che esistano diversi modi per fare riferimento ad uno stesso autore.
Questo capita soprattutto perché i dati vengono presi attraverso diversi tool di estrazione che prelevano le informazioni dagli articoli e ne consegue che una persona può essere associata a diverse entità.
GReAT (*Giorgi's Redundant Authors Tool*) è lo strumento da noi sviluppato per identificare in maniera univoca gli autori all'interno di un dataset, cercando di uniformare i vari casi.

Capitolo 1

Il Semantic Publishing

1.1 L'evoluzione del Web: dal Web 2.0 al Semantic Web

Il *World Wide Web* (abbreviato in *Web*) [1] è uno dei principali servizi di Internet che offre l'accesso di documenti, solitamente in formato *HTML,* connessi fra loro attraverso collegamenti ipertestuali, utilizzando il protocollo *HTTP*.

Durante la fase iniziale (denominata *Web 1.0*) le pagine web erano statiche, ossia puramente HTML.

La prima innovazione avvenne con l'introduzione di pagine dinamiche, utilizzando linguaggi di scripting come *Javascript* e *Php*.

Darcy DiNucci fu il primo parlare "*Web 2.0*", utilizzando questo termine per descrivere questa evoluzione nel suo articolo [2], raccontando come i contenuti interattivi si stiano diffondendo.

Per Web 2.0 si intende appunto un Web interattivo, dove vengono usate tecnologie come *Ajax* (*Asynchronous JavaScript and XML*), permettendo di inviare e ricevere dati in maniera asincrona.

Fu inoltre possibile integrare applicazioni all'interno delle pagine, introducendo strumenti come *Java* e *Flash*.

Grazie a questa evoluzione è stato possibile creare servizi come *Google Maps* [3], *Writely* [4] e *MySpace* [5].

Ma fin da quando l'idea del Web 2.0 comincio a diffondersi, alcune menti del W3 già introdussero parti di quello che venne poi chiamato *Web 3.0* [6].

In verità, con il termine "Web 3.0" non si intende un unico percorso evolutivo, ma diverse trasformazioni del Web; tra queste la più diffusa, ossia il "*Web Semantico*".

Tim Berners-Lee, co-inventore del World Wide Web assieme a Robert Cailliau, fu il primo ad introdurre il concetto di "*Semantic Web*", definendolo un estensione del Web attuale, in cui all'informazione viene dato un significato ben definito, permettendo ai computer ed alle persone di lavorare con una migliore collaborazione [7].

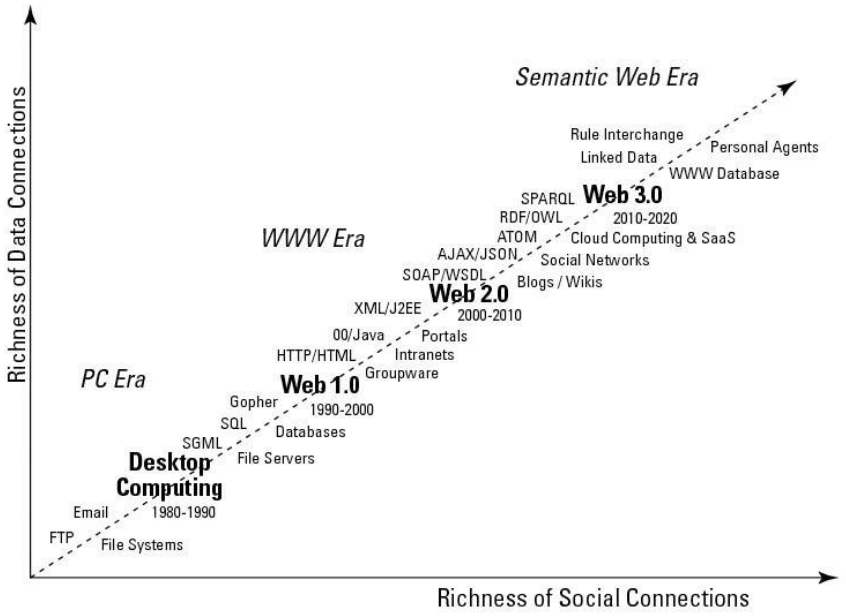

Figura 1 Evoluzioni del Web [F.1]

Vengono utilizzati metodi, come l'aggiunta di metadati, per includere un contenuto semantico alle risorse, permettendo ai dati di essere elaborati in maniera automatica (*machine-readable*).

Tra gli strumenti più importanti del Semantic Web ricordiamo lo standard *RDF* (*Resource Description Framework*), utilizzato per descrivere e modellare l'informazione, e *SPARQL* (*SPARQL Protocol and RDF Query Language*), linguaggio di interrogazione per i dati rappresentati tramite RDF.

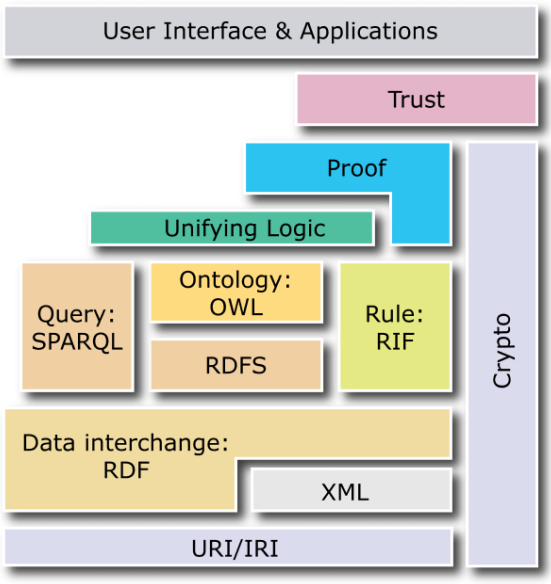

Figura 2 Strati del Semantic Web [F.2]

1.2 I principi del Linked Data

Per *Linked Data* intendiamo l'uso del Web per creare collegamenti tra dati di diversi sorgenti [8], ad esempio da diversi *dataset*, in modo tale da creare una certa interoperabilità tra sistemi eterogenei.
Per fare ciò, come nelle pagine HTML si usano i collegamenti ipertestuali, nei dati si utilizzano quelle che vengono definite *triple RDF*: ogni proprietà del dato viene descritta tramite la forma <*soggetto*>-<*predicato*>-<*oggetto*>.
Ma per poter collegare in modo efficiente i dati, bisogna seguire una serie di regole, ideate da Berners-Lee [9]:

- Assegnare un URI ad ogni entità, in modo tale che ognuna possa essere identificata in maniera univoca;
- Utilizzare http URI come nomi per le entità, in modo che le persone possano cercare quelle risorse;
- Fornire informazioni utili alle entità utilizzando strumenti come, ad esempio, le *ontologie*;
- Includere i collegamenti ad altre risorse, per poter creare una vera e propria rete di dati.

Per facilitarne lo sviluppo, sono stati creati diversi tool, ad esempio *D2R Server* [10] oppure *Virtuoso Universal Server* [11], in grado di pubblicare Linked Data e fornire servizi come l'interrogazione tramite SPARQL.

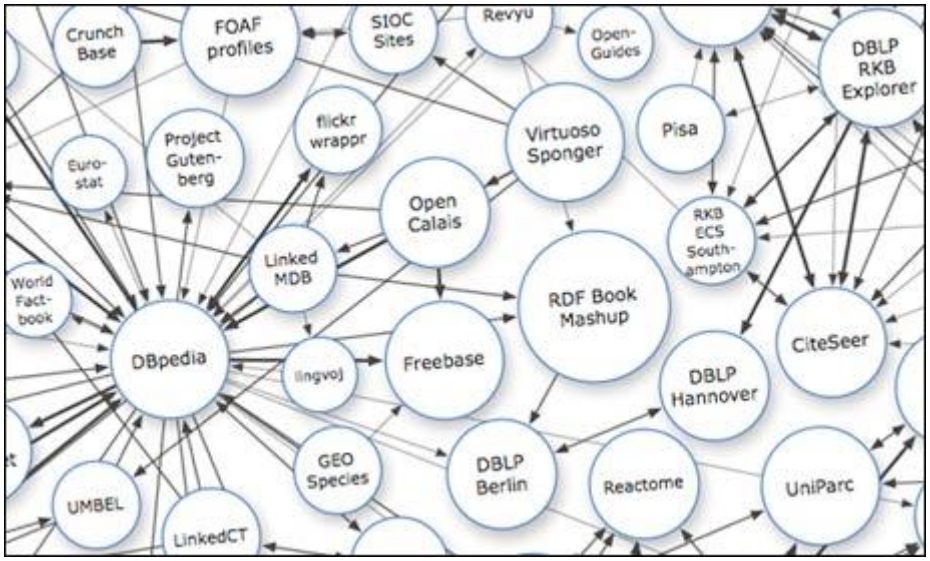

Figura 3 Parte del Linked Data dataset cloud diagram [F.3]

Inoltre, dato il successo di questa innovazione, sono state sviluppate diverse applicazioni che hanno la capacità di cercare dati da sorgenti diverse, tra cui i *Linked Data Browser* che, come i tradizionali browser permettono di navigare tra pagine HTML attraverso i collegamenti ipertestuali, essi consentono di navigare sui dati attraverso le triple RDF.

1.3 Una nuova frontiera per le pubblicazioni scientifiche: Semantic Publishing

Uno degli esempi d'uso del Linked Data è nel *Semantic Publishing*, ossia una serie di iniziative che mirano a migliorare come gli scienziati comunicano utilizzando le tecnologie semantiche [12].
David Shotton definì il Semantic Publishing come "qualcosa che migliora il senso di un articolo di giornale pubblicato, facilita la sua ricerca automatizzata, consente di collegare articoli correlati semanticamente, fornisce l'accesso ai dati all'interno dell'articolo in una forma processabile, o facilita l'integrazione dei dati tra documenti" [13].
L'avvento del Semantic Web portò una rivoluzione nel campo *dell'electronic publishing*, ossia la pubblicazione di documenti come articoli, giornali o libri in formato digitale.
Esso, infatti, presentava alcuni svantaggi, come la pubblicazione in formato *PDF*: anche se può sembrare più leggibile agli umani, questo formato è difficile da processare attraverso le applicazioni, rendendo più complessa la raccolta di informazioni.
Vengono effettuati quindi dei miglioramenti "semantici" alle pubblicazioni:
a seconda del tipo di informazione che bisogna descrivere, si utilizzano diverse *ontologie*, dove ognuna delle quali fornisce un vocabolario condiviso, che può essere utilizzato per modellare i domini, tra cui il tipo di entità, i concetti che sussistono, le loro proprietà e le loro relazioni [14].

Nell'articolo "*Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article*" [15] vengono presentati diversi esempi sui vantaggi di questi miglioramenti semantici, tra cui:

- l'integrazione dei dati geospaziali, dove, ad esempio, raccogliendo i dati da diverse locazioni, è possibile creare una *heatmap* (tradotto "mappa del calore");

- l’aggiunta di valore al testo, dove, per esempio, le entità nominate nel testo vengono evidenziate per poi essere collegate a sorgenti esterne;
- l’accessibilità delle informazioni, come i principali argomenti di un articolo, oppure le citazioni.

Il compito di chi pubblica è quello di arricchire il più possibile la pubblicazione dando più informazioni possibili, in modo da essere più facilmente trovato.
Anche gli autori, che conoscono meglio di tutti l’articolo ed il dominio del discorso, hanno il compito di aggiungere informazioni quali gli argomenti e gli articoli citati.

Come vedremo nei prossimi capitoli, il Semantic Publishing presenta una serie di problematiche: primo fra tutti il problema della qualità dei dati all’interno dei Linked Dataset, ossia la pubblicazione in maniera non uniforme delle informazioni.

Capitolo 2

Analisi dei Dataset su pubblicazioni scientifiche

Per *Linked Open Dataset* (che noi abbreviamo "*dataset*") si intende una collezione di dati collegati fra loro secondo i principi del Linked Data, elencati nel capitolo precedente, e accessibili liberamente sul Web.
Il nostro scopo è quello di analizzare alcuni dei tanti dataset che sono stati creati per il Semantic Publishing, come *JISC Open Citations* oppure *DBLP RKB Explorer*, in modo da fare chiarezza sui possibili punti in comune e sulle principali problematiche.
Infatti, la mancanza di uno standard fa sì che ogni sistema possa rappresentare i dati in maniera differente, utilizzando, ad esempio, modelli ontologici diversi oppure omettendo informazioni di una certa importanza, come le citazioni.

Per capire meglio la situazione, facciamo un esempio con due entità, prese dai due dataset sopracitati:

Property	Object/Value
akt:article-of-journal	ACM Southeast Regional Conference
akt:cites-publication-reference	Proceedings of the 42nd Annual Southeast Regional Conference, 2004, Huntsville, Alabama, USA, April 2-3, 2004
akt:has-author	Swathi Tanjore Gurumani
akt:has-author	Aleksandar Milenkovic
akt:has-date	2004
akt:has-title	Execution characteristics of SPEC CPU2000 benchmarks: Intel C++ vs. Microsoft VC++.
akt:has-web-address	http://doi.acm.org/10.1145/986537.986599 [Visit]
rdf:type	akt:Book-Section-Reference
rdf:type	_:genid3.871946836e0bc433
owl:sameAs	http://dblp.l3s.de/d2r/resource/publications/conf/ACMse/GurumaniM04

Figura 4 Esempio di dataset: DBLP

prism:doi	10.1016/j.neuron.2004.08.004
prism:publicationDate	2004-01-01 00:00:00
dcterms:creator	JD Richter
dcterms:published	2004-01-01
dcterms:title	RNA transport (partly) revealed!
http://purl.org/net/pingback/service	http://opencitations.net/pingback/xmlrpc/
http://purl.org/net/pingback/to	http://opencitations.net/pingback/rest/
fabio:hasPubMedId	15312643
pro:isRelatedToRoleInTime	authorship of JD Richter
frbr:embodiment	http://opencitations.net/id/manifestation:doi/10.1016/j.neuron.2004.08.004
rdf:type	fabio:JournalArticle

Figura 5 Esempio di dataset: JISC

Come si può vedere, DBLP utilizza unicamente l'ontologia AKT, mentre, invece, JISC Open Citations ne usa diverse, tra cui PRISM e Dublin Core.
Si può inoltre notare che l'entità presa da DBLP contiene un informazione riferita alle citazioni, dati che sono assenti nell'entità raccolta da JISC.

In questo capitolo spieghiamo come è stata effettuata l'analisi dei dataset, quali parametri sono stati presi in considerazione e come è strutturata la rassegna.

Durante l'esplorazione dei dataset, abbiamo notato che i dati hanno una certa eterogeneità, ossia vengono utilizzati per rappresentare informazioni di diversi ambiti.
Abbiamo quindi raggruppato i dati a seconda dell'informazione che essi identificano, individuando quali sono i principali *aspetti* che ricoprono.

Gli aspetti sono i seguenti:

- *Dati bibliografici*: con questo aspetto prendiamo in considerazione i dati come il titolo, la data di pubblicazione e le informazioni bibliografiche come l'intervallo di pagine oppure il numero di volume. Questi dati rappresentano le "informazioni generali" essenziali per il riconoscimento di una pubblicazione e quelle utili per arricchire le risorsa;
- *Identificatori*: gli identificatori univoci di una risorsa, che possono essere ID locali nel dataset oppure globali, come ad esempio i *DOI (Digital Object Identifier)*, gli *ISSN* oppure gli *URI*. Queste informazioni, se

presenti, possono aiutare a riconoscere una risorsa in maniera immediata;

- *Abstract*: le informazioni riguardanti gli abstract, ossia un breve riassunto del documento, all'interno di una pubblicazione. Quando vengono aggiunti questi dati alle pubblicazioni è possibile ricavare ulteriori informazioni su di esse, come gli aspetti fondamentali o i contenuti principali;
- *Autori*: i dati che rappresentano e descrivono le persone che hanno contribuito ad una pubblicazione. Queste informazioni sono importanti per il riconoscimento degli autori e sono utili soprattutto quando, ad esempio, vogliamo ricavare l'elenco degli articoli scritti da una determinata persona e presi da sorgenti diverse;
- *Enti – Affiliazioni*: le informazioni riguardo gli enti che hanno partecipato alla produzione o alla pubblicazione di una risorsa, oppure le affiliazioni di una certa persona. Questi dati sono importanti quando, ad esempio, vogliamo conoscere quali articoli sono stati prodotti da determinata organizzazione, oppure per ricavare un elenco di pubblicazioni di un determinato ente editoriale;
- *Classificazione*: informazioni utili nel classificare un documento, come l'attribuzione di un genere oppure l'elenco degli argomenti principali di una pubblicazione. Con questi dati è possibile, ad esempio, creare servizi come la suddivisione delle pubblicazioni per categoria;
- *Citazioni – Riferimenti*: le informazioni sui riferimenti e sulle citazioni di un documento. Questi dati sono fondamentali per la creazione di una vera e propria rete di articoli, dando la possibilità di navigare tra pubblicazioni correlate.

I dati presi da ogni dataset analizzato vengono sintetizzati su delle tabelle, suddivise attraverso gli aspetti sopracitati, che riassumono le informazioni trovate e le ontologie utilizzate per ogni tipologia di aspetto.
In questo modo diamo un quadro generale dei dataset, potendo facilmente cogliere quali informazioni, per ciascuno, sono state omesse e quali invece sono state inserite in maniera ridondante.

Oltre alla tabella riassuntiva, diamo una descrizione generale del dataset: com'è strutturato, la dimensione e come è possibile effettuare le ricerche.

Nel caso, elenchiamo anche gli eventuali problemi trovati del server.

Successivamente, entriamo nel dettaglio per ogni tipologia di risorsa analizzata, come appunto gli articoli o gli autori.
I risultati vengono organizzati attraverso altre tabelle, dove elenchiamo le informazioni trovate, secondo l'ordine degli aspetti sopra elencati, e con quanta frequenza state riscontrate, utilizzando una percentuale approssimata.

Per le risorse che rappresentano le persone o gli enti, suddividiamo le informazioni su altri due aspetti, ossia "Identificazione" e "Altri dati", che rappresentano rispettivamente le informazioni fondamentali come il nome e gli altri dati utili ad arricchire la risorsa, come la homepage o l'indirizzo di posta elettronica.
Infine, facciamo un breve riassunto per ogni dataset, descrivendo quali sono i principali pregi e difetti e dando un voto personale in decimi, cercando di far capire la qualità di ciascuno.

Essendo che, come si può notare dall'esempio all'inizio di questo capitolo, i dati sono organizzati secondo modelli ontologici diversi, prima di entrare nella rassegna diamo una descrizione delle ontologie più comuni che sono state trovate durante l'analisi.

Capitolo 3

Esempi di Ontologie per il Semantic Publishing

Esistono diverse ontologie che vengono utilizzate solitamente nel Semantic Publishing; in questo capitolo, descriviamo quelle che abbiamo riscontrato maggiormente.
La maggior parte delle ontologie esposte sono in grado di coprire più aspetti di quelli che abbiamo visto esprimere durante l'esplorazione dei dataset.
La seguente tabella, quindi, illustra e riassume le potenzialità di ognuna, distinguendo i tipi di informazione che ciascuna ontologia ha ricoperto nei dataset analizzati da quelli potrebbe ricoprire:

✓ = gli aspetti che l'ontologia in questione ha ricoperto nei dataset analizzati
X = gli aspetti che l'ontologia potrebbe ricoprire

	Dati bibliografici	**Identificatori**	**Abstract**	**Autori**	**Enti - Affiliazioni**	**Classificazione**	**Citazioni - Riferimenti**
SPAR	X	✓	✓	✓	✓		✓
BIBO	✓	✓	✓	✓	X		X
Dubiln Core	✓	✓	✓	✓	X	✓	X
FOAF				✓	✓	X	
SKOS						✓	
PRISM	✓	✓		X	✓	✓	
AKT	✓	✓	X	✓	✓	✓	✓
SWRC	✓	✓	✓	X	✓	✓	X

Tabella 1 Riassunto delle ontologie comuni

Per ogni ontologia, viene data una breve descrizione e l'elenco delle proprietà riscontrate durante l'esplorazione dei dataset, ordinate secondo gli aspetti presentati nel capitolo precedente.
Inoltre, descriviamo per quali altri aspetti l'ontologia potrebbe essere utile.

3.1 SPAR

Il *Semantic Publishing and Referencing Ontologies* (SPAR) [16] è un insieme di ontologie utili nel descrivere oggetti, record e riferimenti nel contesto bibliografico ed editoriale e nel descrivere documenti e citazioni.
Esso è diviso in moduli, ognuno dei quali rappresenta un diverso tipo di informazione.

I moduli sono:

- *CiTO:* utilizzata per descrivere la natura ed il tipo di citazione. Le citazioni possono essere dirette, indirette oppure implicite;
- *FaBiO:* utilizzata su entità che contengono riferimenti bibliografici. Può essere utile su pubblicazioni testuali come libri, riviste, giornali o articoli;
- *BiRO:* in conformità con FaBiO, questa ontologia si occupa di definire riferimenti e record bibliografici. Può essere utile per riconoscere riferimenti bibliografici in una lista di riferimenti che può essere contenuta, ad esempio, in un articolo pubblicato;
- *C4O:* permette il conteggio delle citazioni di un riferimento in una pubblicazione testuale. Inoltre, può definire il contesto di quel riferimento;
- *DoCO:* permette di descrivere i vari componenti di un documento. La descrizione può essere di tipo strutturale, per esempio i vari capitoli, oppure di tipo retorica, ad esempio i riconoscimenti e la lista dei riferimenti;
- *PRO:* questa ontologia è utile per specificare il ruolo di un agente all'interno di una pubblicazione, per esempio l'autore. Può specificare il periodo di tempo durante il quale il ruolo è tenuto;
- *PSO:* permette di caratterizzare lo stato di una qualsiasi pubblicazione. Ad esempio lo stato "bozza" per un documento;

- *PWO:* serve a descrivere i passi nel workflow associato ad un documento, o ad una qualsiasi entità pubblicata. Ad esempio “under review”, ossia sotto revisione;

Proprietà riscontrate:

Di seguito, riportiamo le proprietà di questa ontologia che sono state rilevate durante l’analisi dei dataset.

Identificatori:
- *fabio:hasPubMedCentralId* e *fabio:hasPubMedId* per l’identificatore (riferito a *PubMed*, ossia il più grande database contenente pubblicazioni in campo biomedico).

Abstract:
- *fabio:abstract* per l’abstract delle pubblicazioni.

Autori:
- *pro:holdsRoleInTime* e *pro:withRole* per definire il ruolo degli agenti.

Enti - Affiliazioni:
- *pro:withAffiliation* per specificare l’affiliazione.

Citazioni – Riferimenti:
- *cito:cites* e *biro:referenceList* per le citazioni e la lista dei riferimenti.

Oltre a questi aspetti, SPAR può essere utilizzato anche per inserire dati bibliografici alle pubblicazioni, usando, ad esempio, i moduli come BiRO e PSO.

3.2 BIBO

La *Bibliographic Ontology* (BIBO) [17] è un’ontologia utilizzata nel Semantic Web per descrivere un qualsiasi oggetto bibliografico, come i libri, le riviste e ovviamente gli articoli.
Questo vocabolario può essere usato per specificare citazioni, classificare documenti, dare informazioni bibliografiche come l’*issue* o il numero di volume oppure per dare semplici descrizioni ad un qualsiasi tipo di entità.

Questa ontologia è stata ispirata da molti vocabolari già esistenti, e può essere utilizzata come una base comune per la conversione di dati bibliografici provenienti da diverse fonti.

Proprietà riscontrate:

Dati bibliografici:
- *bibo:issue* per il numero di emissione;
- *bibo:pageStart*, *bibo:spage* e *bibo:pageEnd* per le indicazioni riguardo le pagine di un articolo;
- *bibo:volume* per il numero di volume;
- *bibo:status* per lo stato di un documento. Ad esempio "accepted", quando il documento è stato accettato dopo una revisione, oppure "published", quando il documento è stato pubblicato.

Identificatori:
- *bibo:identifier* per un identificatore locale nel dataset;
- *bibo:uri* per l'URI;
- *bibo:doi* per il DOI del documento.

Abstract:
- *bibo:abstract* per l'abstract.

Autori:
- *bibo:authorList* per ricondursi alla lista di autori di una pubblicazione;
- *bibo:editor* per l'editore di un articolo.

In questa ontologia esistono, inoltre, proprietà per dare informazioni riguardo agli enti, come il produttore o l'editore di una pubblicazione; e proprietà per aggiungere le citazioni ad un documento.

3.3 Dublin Core

Il *Dublin Core* [18] è un insieme di metadati utilizzati per poter descrivere una qualsiasi risorsa sul Web.

Esso è organizzato su due livelli:

- "*Simple DC*", che comprende quello che viene chiamato "nucleo" di quindici metadati "essenziali", che possono essere utilizzati in una grande varietà di contesti, come ad esempio "title", "date", "description", ed "identifier";
- "*Qualified DC*", ridefinisce il gruppo di elementi, perfezionandone la semantica, e include tre metadati in più, ossia "audience", "provenance" e "rightsholder".

Inoltre, è stato creato un nuovo vocabolario, "*DC Terms*", che include i quindici metadati del Dublin Core definiti come proprietà RDF, in modo da essere meglio integrati nel Semantic Web.
Grazie alla sua semplicità, esso è molto utilizzato in contesti differenti, anche se non può dare una descrizione dettagliata ad una risorsa.
Quest'ultimo punto, però, va a favore all'interoperabilità semantica, ossia poter integrare dati con semantiche diverse tra loro.

Proprietà riscontrate:

Dati bibliografici:
- *dc:title* per il titolo della pubblicazione;
- *dc:isPartOf* per sapere dov'è inclusa la risorsa;
- *dc:isVersionOf* e *dc:terms:isVersionOf* per conoscere la versione di un articolo;
- *dc:issued* per la data di pubblicazione;
- *dcterms:created* per la data di creazione.

Identificatori:
- *dc:identifier* per l'identificatore della risorsa.

Abstract:
- *dcterms:abstract* per l'abstract.

Autori:
- *dc:creator* e *dcterms:creator* per l'autore.

Classificazione:
- *dc:subject* per gli argomenti di una pubblicazione.

Oltre a queste proprietà, il Dublin Core può essere utilizzato per dare informazioni riguardo agli enti che hanno contribuito alla risorsa, oppure per inserire i riferimenti alle pubblicazioni.

3.4 FOAF

Il *Friend of a Friend* (FOAF) [19] è un vocabolario che ha lo scopo di creare pagine che descrivono entità come le persone, i gruppi, le aziende o un qualsiasi altro tipo di agente all'interno del Web.
Questa ontologia è utile per fornire informazioni quali le attività, le relazioni con altri agenti, gli interessi oppure i "dati personali" come il nickname o la mail box.
Attualmente, FOAF viene usata principalmente in alcuni blog, in plugin di browser come *Firefox* [20] e *Chrome* [21] ed in altri servizi come *FriendFeed* [22] e *WordPress* [23].

Come vedremo in seguito, uno dei problemi rilevati su alcuni dataset è la diversa modalità per fare riferimento ad uno stesso autore.
Per esempio, quando vengono inseriti più articoli scritti da una stessa persona può capitare che vengano create più istanze per il medesimo individuo.
Abbiamo quindi cercato di risolvere questo problema, creando un tool che tenti di unificare le entità che rappresentano uno stesso soggetto.
Le proprietà di questa ontologia sono servite al nostro tool per ricavare i nomi delle persone e le diverse informazioni da poter confrontare, come la mbox o la homepage, utili per la disambiguazione.
Nel capitolo 6 verrà spiegato in dettaglio il funzionamento del tool, che abbiamo battezzato "*GReAT*" (*Giorgi's Redundant Authors Tool*).

<u>*Proprietà riscontrate*</u>:

Autori:
(le proprietà elencate comprendono quelle utilizzate in GReAT)

- *foaf:name* per il nome completo di una persona;
- *foaf:firstName* e *foaf:givenName* per il nome;
- *foaf:lastName*, *foaf:familyName* e *foaf:surname* per il cognome;

- *foaf:homepage* per l'homepage della persona;
- *foaf:based_near* per la nazionalità della persona;
- *foaf:mbox* per la casella di posta elettronica.

Enti – Affiliazioni:
- *foaf:name* per il nome dell'organizzazione;
- *foaf:logo* per il logo dell'ente;
- *foaf:based_near* per sapere la nazionalità dell'ente;
- *foaf:homepage* per l'homepage dell'ente.

FOAF può essere usato, inoltre, per definire documenti, potendo coprire l'aspetto della classificazione di una pubblicazione, attraverso proprietà utilizzate per elencare gli argomenti principali.

3.5 SKOS

Il *Simple Knowledge Organization System* (SKOS) [24] è una famiglia di linguaggi utilizzata per rappresentare sistemi di organizzazione della conoscenza (KOS) come tesauri, tassonomie, schemi di classificazione o glossari.

Essa è strutturata su diversi moduli:

- "*SKOS Core*", utilizzato per rappresentare le caratteristiche comuni in un tesauro;
- "SKOS Mapping", un vocabolario per esprimere le connessioni fra diversi concetti;
- "SKOS Extensions", utilizzato per descrivere le relazioni tra concetti attraverso una semantica più precisa.

SKOS è stato utilizzato per descrivere molti vocabolari importanti come *EuroVoc* [25], *AGROVOC* [26] e *GEMET* [27].

Proprietà riscontrate:

Classificazione:
- *skos:Concept* per definire un concetto all'interno di una pubblicazione;
- *skos:exactMatch* e *skos:narrowMatch* per esprimere rispettivamente le corrispondenze esatte e quelle approssimative (riferite ai concetti).

Tutte le proprietà di SKOS sono concentrate sulla classificazione di risorse.

3.6 PRISM

Il *Publishing Requirements for Industry Standard Metadata* (PRISM) [28] è un insieme di metadati che hanno lo scopo di costruire e gestire contenuti digitali, tra cui le pubblicazioni editoriali.
È stata creata da *IDEAlliance* (*International Digital Enterprise Alliance*) [29], un'organizzazione no-profit leader nell'information technology e nell'editoria.
Con PRISM gli editori possono creare, gestire, produrre, distribuire e riusare contenuti digitali.
Questa ontologia offre proprietà come l'identificazione di un'entità, le informazioni relative alla pubblicazione e alle informazioni bibliografiche, e informazioni riguardanti le affiliazioni.

Proprietà riscontrate:

Dati bibliografici:
- *prism:startingPage* e *prism:endingPage*, equivalenti a *bibo:pageStart* e *bibo:pageEnd*, per le informazioni riguardo le pagine della pubblicazione;
- *prism:volume*, equivalente a *bibo:volume*, per il numero di volume;
- *prism:pubblicationDate* e per la data di pubblicazione.

Identificatori:
- *prism:elssn*, *prism:issn*, *prism:issueIdentifier* e *prism:number* per i numeri identificativi della pubblicazione.

Enti – Affiliazioni:
- *prism:pubblicationName per chi ha pubblicato l'articolo.*

Classificazione:
- *prism:genre* per identificare il genere del documento.

Questa ontologia può inoltre essere utilizzata per aggiungere informazioni riguardo alle persone coinvolte in una pubblicazione.

3.7 AKT

L'*Advanced Knowledge Technologies* (AKT) [30] è un'ontologia utilizzata per modellare e gestire le conoscenze scientifiche, realizzata con l'"Advanced Knowledge Technologies Project" [31].
Viene utilizzata per descrivere entità nel contesto scientifico, come le pubblicazioni, gli autori e le organizzazioni.
Essa è suddivisa in due parti:
- *AKT Support Ontology* serve per definire le entità del contesto, dando informazioni come la loro funzione e le loro relazioni;
- *AKTive Portal Ontology*, che utilizza AKT Support Ontology e serve per modellare i concetti del dominio, come i documenti, le persone o le organizzazioni.

Diversi Linked Dataset, come quelli di *ACM* e *DBLP* (che vedremo nel capitolo successivo), utilizzano unicamente questa ontologia per descrivere le entità.

Proprietà riscontrate:

Dati bibliografici:
- *akt:has-title* per il titolo della pubblicazione;
- *akt:has-volume*, equivalente a *bibo:volume*, per il numero di volume;
- *akt:has-date* per la data di pubblicazione.

Identificatori:
- *akt:has-web-address* per l'indirizzo web dell'articolo.

Autori:
- *akt:has-author* e *akt:edited-by* per specificare gli autori;
- *akt:full-name* per il nome completo di una persona.

Enti – Affiliazioni:
- *akt:article-of-journal* per le informazioni riguardo al giornale dell'articolo.

Classificazione:
- *akt:addresses-generic-area-of-interest* per descrivere gli argomenti all'interno di un articolo.

Citazioni – Riferimenti:
- *akt:cites-publication-reference* per inserire le citazioni.

Questa ontologia è in grado, inoltre, di inserire gli abstract all'interno delle pubblicazioni.

3.8 SWRC

Il *Semantic Web for Research Communities* (SWRC) [32] è un'ontologia usata per dare informazioni alle comunità di ricerca, tra cui persone, organizzazioni, pubblicazioni e le loro relazioni.
Questa ontologia è utilizzata per descrivere entità nel contesto scientifico come gli articoli e le conferenze.
Diversi servizi come *OntoWare* [33] o *SemIPort* [34] e applicazioni come *Bibster* [35] usano SWRC per modellare le risorse.

Proprietà riscontrate:

Dati bibliografici:
- *swrc:title* per il titolo della pubblicazione;
- *swrc:volume*, equivalente a *bibo:volume*, per il numero di volume;
- *swrc:pages* per le informazioni riguardo le pagine dell'articolo;
- *swrc:year* e *swrc:month* per l'anno ed il mese di pubblicazione.

Identificatori:
- *swrc:url* per l'URL del documento.

Abstract:
- *swrc:abstract* per inserire l'abstract.

Enti – Affiliazioni:
- *swrc:journal* per specificare il journal che ha pubblicato l'articolo;
- *swrc:affiliation* per specificare l'affiliazione di un agente.

Classificazione:
- *swrc:category* per specificare la categoria;
- *swrc:hasTopic* per specificare gli argomenti della pubblicazione.

SWRC offre, inoltre, proprietà per dare informazioni riguardo alle persone coinvolte in una pubblicazione, oppure per elencare le citazioni di un documento.

3.9 Le ontologie "proprietarie"

Oltre alle ontologie più comuni, diamo un piccolo sguardo anche alle ontologie che sono state create per un determinato sistema:

- ***NPG***: ontologia creata per *Nature Publishing Group* e utilizzata per descrivere ogni tipo di entità all'interno del proprio dataset.
Questo vocabolario viene utilizzato per dare informazioni come il titolo o la data di pubblicazione di un articolo, attribuire degli identificatori come il DOI, inserire dati come gli autori o i produttori e aggiungere le citazioni.

- ***SWC***: ontologia creata per *Semantic Web Conference*, usata principalmente per descrivere le conferenze all'interno del proprio dataset.
Tra le proprietà utilizzate di questa ontologia, troviamo quelle per collegare un articolo ad un evento (conferenza o workshop) e quelle per elencare gli argomenti principali di una pubblicazione.
Sono state utilizzate anche proprietà per dare informazioni ad un agente, come il ruolo e l'affiliazione.

- ***SWJ***: ontologia creata per *Semantic Web Journal*, usata per arricchire le entità del proprio dataset.
Questo vocabolario è stato utilizzato principalmente per collegare un articolo con le vecchie o nuove versioni e per dare informazioni riguardo le revisioni delle pubblicazioni.
È stato utilizzato, inoltre, per classificare un documento e dare un ruolo ad un agente all'interno del dataset.

Come si può notare, esistono diverse ontologie che vengono utilizzate nel Semantic Publishing, dove ognuna può essere utilizzata per descrivere un determinato tipo di entità.
Ci sono ontologie come FOAF e SKOS, che ricoprono solo pochi aspetti ma in modo esaustivo, e vocabolari come il Dublin Core, che cercano di coprire ogni aspetto delle pubblicazioni senza entrare troppo nei dettagli.
Nel prossimo capitolo, potremo vedere meglio come queste ontologie vengono usate e con quanta frequenza si possono incontrare nei dataset.

Capitolo 4

I Linked Open Dataset

In questo capitolo descriviamo i diversi *Linked Open Dataset* esaminati, esponendo i risultati delle analisi ed evidenziando i pregi ed i difetti trovati per ognuno.

Il seguente schema sintetizza quali aspetti, tra quelli citati nel capitolo 3, ogni dataset analizzato ricopre:

	Nature Linked Data	JISC Open Citations	ACM RKB Explorer	DBLP RKB Explorer	J.UCS Bibliography DB	Semantic Web Dog Food	Semantic Web Journal
Dati Bibliografici	✓	✓	✓	✓	✓	✓	✓
Identificatori	✓	✓		✓	✓	✓	✓
Abstract		✓			✓	✓	✓
Autori	✓	✓	✓	✓	✓	✓	✓
Enti – Affiliazioni	✓	✓			✓	✓	
Classificazione	✓	✓	✓		✓	✓	✓
Citazioni – Riferimenti	✓	✓	✓	✓			

Tabella 2 Riassunto dei dataset analizzati

Dalla tabella possiamo notare che quasi tutti i dataset, fra quelli esaminati, presentano una mancanza di dati, non permettendo di coprire ogni aspetto, e talvolta vengono omesse informazioni importanti.
Prendiamo, ad esempio, il caso di ACM: nelle entità del dataset non vengono prese in considerazione dati di una certa valenza come gli identificatori oppure gli abstract.
Altri dataset come J.UCS, SWDF e SWJ omettono informazioni altrettanto importanti come le citazioni delle proprie pubblicazioni.
Inoltre, come potremo vedere in seguito, su alcuni dataset i diversi aspetti non vengono gestiti in maniera ottimale.
Come in Nature, dove solo il 35% delle pubblicazioni esaminate contengono i riferimenti alle citazioni.

4.1 Nature Linked Data

NPG (*Nature Publishing Group*) [36] è un gruppo editoriale di grande impatto scientifico e medico.
Esso si occupa di pubblicare riviste accademiche, magazine, database online e altri servizi nel campo della scienza e della medicina.
Nature Linked Data [37] è il dataset che contiene le informazioni riguardo le proprie pubblicazioni.

Riepiloghiamo le dimensioni del dataset nel seguente schema:

Numero di triple	389 milioni
Limite massimo risultati query	1000
Tempo medio query	4 secondi

Tabella 3 Dimensioni di Nature Linked Data

La seguente tabella, invece, riassume quali proprietà sono state trovate all'interno di questo dataset, specificando quali ontologie sono state utilizzate per ogni aspetto:

	Dati bibliografici	**Identificatori**	**Abstract**	**Autori**	**Enti - Affiliazioni**	**Classificazione**	**Citazioni - Riferimenti**
proprietà	dc:title npg:title bibo:issue bibo:pageEnd bibo:pageStart bibo:volume bibo:spage prism:endingPage prism:startingPage prism:number prism:volume prism:pubblicationDate prism:pubblicationName npg:hasPubblication	dc:identifier npg:artid npg:doiHash prism:doi prism:url		npg:hasContributor foaf:familyName foaf:givenName foaf:name	npg:hasProduct prism:copyright foaf:homepage npg:pcode	prism:genre	npg:hasCitation npg:citeNum
ontologie	Dublin Core BIBO PRISM NPG	Dublin Core NPG PRISM		NPG FOAF	NPG PRISM FOAF	PRISM	NPG

Tabella 4 Proprietà utilizzate da Nature Linked Data

Come abbiamo visto, il dataset conta oltre 389 milioni di triple.
Al suo interno, oltre alle ontologie più comuni, viene utilizzato un proprio vocabolario, ossia *NPG*: una collezione di metadati attui a descrivere una qualsiasi entità all'interno di Nature Publishing Group.

Tramite il sito, è possibile effettuare le ricerche tramite una semplice form nella homepage, oppure scrivendo query col linguaggio SPARQL.
I risultati delle query possono essere al massimo 1000.
L'analisi di questo dataset è stata effettuata utilizzando entrambi i metodi di ricerca.
Il tempo di ricerca medio con le query, senza dare un limite al numero di risultati, è di circa 4 secondi.

Verso la fine di maggio 2014, l'accesso al dataset è stato chiuso, reindirizzando il sito sulla pagina di Nature che descrive il Linked Dataset; quindi non è stato possibile ottimizzare l'analisi.

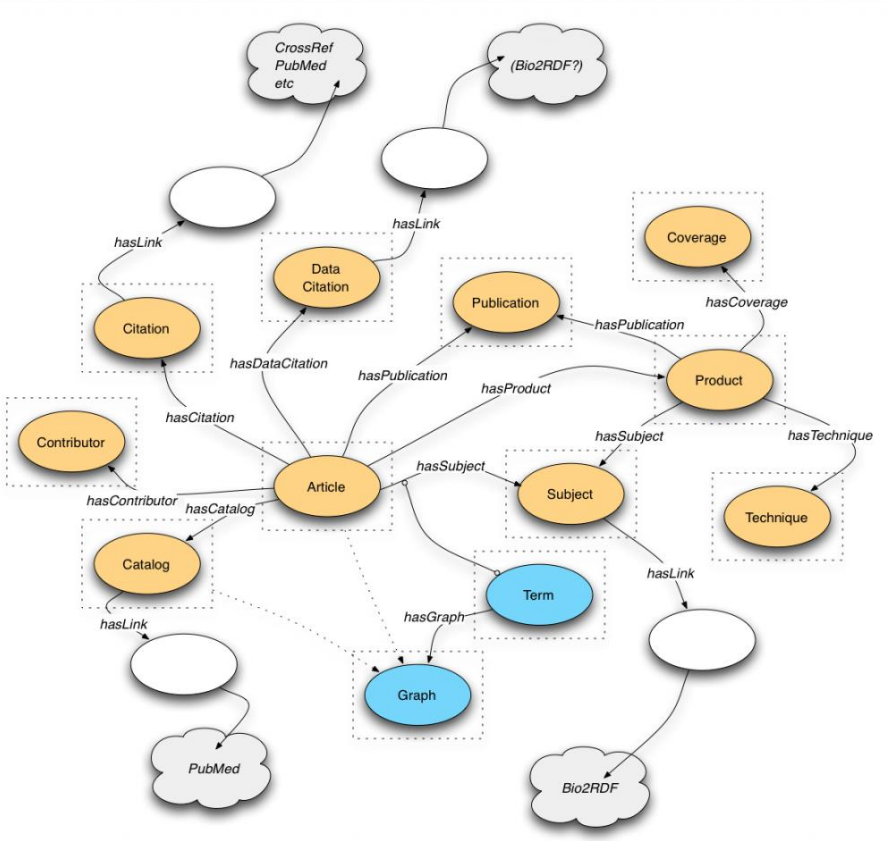

Figura 6 Grafo di Nature Linked Data [F.4]

Sono stati analizzati circa 150 articoli di contesti diversi, tra cui informatica, medicina e astronomia.

Tutte le proprietà trovate, per ogni tipo di entità, sono organizzate secondo i diversi aspetti, ed ordinate in base alla percentuale di ritrovamento, ossia la percentuale delle entità, relativamente a quelle analizzate, che possiedono quel determinato attributo.

Proprietà	**Percentuale**
Dati Bibliografici: npg:title; dc:title; dc:published; prism:pubblicationDate; prism:startingPage; prism:endingPage; prism:volume; prism:number. ***Identificatori:*** npg:artid; npg:doihash; dc:identifier; prism:doi; prism:url. ***Enti***: npg:hasProduct; prism:copyright.	100%
Classificazione: prism:genre.	85%
Dati bibliografici: bibo:issue; bibo:pageEnd; bibo:pageStart; bibo:volume; bibo:spage.	80%
Dati bibliografici: npg:hasPubblication;	65%
Autori: npg:hasContributor.	45%
Citazioni: npg:hasCitation.	35%

Tabella 5 Proprietà negli articoli di Nature Linked Data

Sono stati analizzati circa 50 produttori, ossia gli enti che hanno prodotto gli articoli:

Proprietà	**Percentuale**
Identificazione: dc:title; npg:pcode. ***Altri dati:*** foaf:homepage.	100%
Altri dati: npg:collection; npg:hasCoverage.	< 100% (non è stato riuscito stimare una percentuale poiché il sito ha chiuso prima del completamento dell'analisi)

Tabella 6 Proprietà nei produttori di Nature Linked Data

Sono stati analizzati oltre 100 contributori, ossia gli autori degli articoli:

Proprietà	**Percentuale**
Identificazione: foaf:familyName; foaf:givenName; foaf:name.	100%

Tabella 7 Proprietà nei contributori di Nature Linked Data

Sono stati analizzate circa 40 entità riferite alle citazioni:

Proprietà	**Percentuale**
Identificazione: npg:citeNum.	100%
Dati bibliografici: prism:pubblicationName; prism:pubblicationDate.	97,5% (una sola citazione è stata trovata senza queste proprietà)
Dati bibliografici: bibo:epage; bibo:spage; bibo:volume.	80%

Tabella 8 Proprietà nelle citazioni di Nature Linked Data

In conclusione:

Gli articoli contenuti nel dataset presentano molte informazioni bibliografiche, ed hanno tutti un DOI associato, oltre che ad un identificatore locale.
Pochi, invece, sono i riferimenti agli autori, dove a meno della metà delle pubblicazioni esaminate sono stati pervenuti questi dati.
In tutte le pubblicazioni analizzate nel dataset, inoltre, non è stato trovato un abstract.
Anche le citazioni vengono ignorate in gran parte dei casi: spesso, dall'articolo concreto alla pubblicazione nel dataset, vengono omessi alcuni riferimenti.
Ad alcune citazioni, invece, mancano le informazioni bibliografiche utili nello specificare da quale parte, all'interno dell'articolo citato, una pubblicazione ha raccolto le informazioni.
Complessivamente, mi è sembrato un buon dataset: alcuni aspetti sono ancora da migliorare, ma sicuramente è uno dei dataset più completi che abbia analizzato.
Quindi, il mio voto personale è un buon 8 su 10.

4.2 JISC Open Citations

JISC (*Joint Information Systems Committee*) [38] è un'istituzione britannica che si occupa di supportare l'educazione e la ricerca attraverso le tecnologie informatiche.
Open Citations [39] è uno dei progetti di JISC, che consiste in un database che raccoglie varie pubblicazioni tra quelle contenute in "*PubMed Central*" [40]

(archivio di pubblicazioni di natura biomedica) e gli articoli più citati in ogni campo della biomedicina.
Lo scopo principale è pubblicare gli articoli inserendo più informazioni possibili riguardo le citazioni in formato *RDF*, in modo tale da creare diversi collegamenti fra le varie pubblicazioni.

Articoli	140 mila
Journal	43 mila
Libri	28.5 mila

Tabella 9 Dimensioni di JISC

	Dati bibliografici	**Identificatori**	**Abstract**	**Autori**	**Enti - Affiliazioni**	**Classificazione**	**Citazioni - Riferimenti**
proprietà	frbr:embodiment dcterms:title prism:volume prism:issueIdentifier prism:endingPage prism:startingPage prism:pubblicationDate dcterms:published	fabio:hasPubMedCentralId fabio:hasPubMedId prism:doi dcterms:identifier prism:elssn prism:issn	frbr:part	dcterms:creator foaf:Person foaf:name pro:holdsRoleInTime pro:roleInTime pro:withRole pro:author ov:sortLabel	pro:withAffiliation	foaf:primaryTopicOf	cito:cites frbr:part
ontologie	FRBR Dublin Core PRISM	FaBiO PRISM Dublin Core	FRBR	Dublin Core FOAF OpenVocab	PRO	FOAF	CiTO FRBR

Tabella 10 Proprietà utilizzate da JISC

Nel sito è possibile effettuare ricerche in diversi modi:

- un link che ti porta all'elenco completo di tutti i journal a cui fanno riferimento gli articoli nel dataset, in ordine alfabetico;
- un link che ti porta ad una selezione pseudo casuale di articoli;
- una form dove poi cercare le pubblicazioni tramite parole chiavi;
- una pagina dove poter scrivere una query in SPARQL.

Le ricerche per l'analisi del dataset sono state effettuate attraverso le query.

È inoltre possibile scaricare il dataset in versione *BibJSON* (in vari formati *come Raw BibJSON* e *Sanitized BibJSON*) e nella serializzazione in RDF utilizzando il formato *N-Quads*.

Durante la ricerca delle entità, è capitato (circa 5 o 6 volte) un errore "Not Found", ossia che l'istanza trovata non è più presente nel dataset.
Per una volta, durante l'esecuzione di una query, è stato riscontrato un errore dal server, col messaggio: "Internal Server Error".

Gli articoli sono definiti principalmente mediante due classi (le dimensioni sono state trovate attraverso le query):

- "fabio:JournalArticle", dove troviamo oltre 307 mila entità di questo tipo;
- "fabio:Expression", dove ci sono quasi 33 mila entità di questo genere.

Sono stati analizzati circa 100 articoli di tipo "fabio:JournalArticle":

Proprietà	**Percentuale**
Dati bibliografici: dcterms:title; frbr:embodiment; prism:pubblicationDate; prism:published.	100%
Autori: dcterms:creator; pro:RelatedToRoleInTime.	99% (una solo entità è stata riscontrata senza questi attributi)
Identificatori: fabio:hasPubMedId.	97% (su 3 entità non è stata pervenuta questa proprietà)
Identificatori: prism:doi.	65%
Identificatori: fabio:hasPubMedCentralId.	25%

Tabella 11 Proprietà negli articoli 'fabio:JournalArticle' di JISC

Sono state analizzate circa 70 articoli di tipo “fabio:Expression”:

Proprietà	**Percentuale**
Dati bibliografici: dcterms:title; frbr:embodiment.	100%
Classificazione: foaf:primaryTopicOf.	60%
Dati bibliografici: prism:pubblicationDate; prism:published.	50%
Autori: dcterms:creator; pro:RelatedToRoleInTime.	45%
Identificatori: fabio:hasPubMedCentralId; fabio:hasPubMedId.	40%
Identificatori: prism:doi. ***Citazioni:*** cito:cites. ***Abstract, Riferimenti***: frbr:part.	30%

Tabella 12 Proprietà negli articoli 'fabio:Expression' di JISC

Sono stati analizzati circa 100 journal, identificati con la classe “fabio:Journal” (nel dataset se ne contano oltre 43 mila, utilizzando le query):

Proprietà	**Percentuale**
Identificatori: dcterms:identifier.	100%
Riferimenti: frbr:part.	96%
Dati bibliografici: dcterms:title; prism:issn.	16%
Identificatori: prism:elssn.	13%

Tabella 13 Proprietà nei journal di JISC

Sono stati analizzati circa 50 libri, descritti con la classe “fabio:book” (nel dataset ce ne sono circa 29.5 mila entità):

Proprietà	**Percentuale**
Dati Bibliografici: frbr:embodiment.	100%
Dati Bibliografici: dcterms:title.	99% (un’entità è stata trovata senza questa proprietà)
Dati Bibliografici: prism:pubblicationDate; prism:published.	90%
Autori: dcterms:creator; pro:RelatedToRoleInTime. ***Classificazione***: foaf:primaryTopicOf.	60%
Autori: dcterms:contributor.	25%
Identificatori: prism:doi.	6% (3 entità sono state trovate con questo attributo)
Identificatori: fabio:hasPubMedId.	4% (su 2 entità è stato riscontrato questo attributo)
Identificatori: fabio:hasPubMedCentralId.	2% (su una sola entità è stata trovata questa proprietà)

Tabella 14 Proprietà nei libri di JISC

Sono stati analizzati circa 100 autori, riconosciuti come “foaf:Person” (non sono riuscito a dare una dimensione a questo tipo di entità nel dataset, poiché durante la query è apparso l’errore “Error 403: Query timed out”):

Proprietà	**Percentuale**
Identificazione: foaf:name. ***Altri dati:*** pro:holdsRoleInTime.	100%
Identificazione: ov:sortLabel.	35%

Tabella 15 Proprietà negli autori di JISC

Sono state inoltre analizzate circa 100 entità “ruolo” (ogni persona nel dataset è associata ad una di queste entità), identificate come “pro:RoleInTime” (non sono riuscito a dare una dimensione per lo stesso motivo di ‘foaf:Person’):

Proprietà	**Percentuale**
Identificazione: pro:withRole.	100%
Affiliazione: pro:withAffiliation.	3%

Tabella 16 Proprietà nei ruoli di JISC

In Conclusione:

C’è un po’ di confusione tra gli articoli di tipo “fabio:Expression” e quelli di tipo “fabio:JournalArticle”: entrambi gestiscono i diversi aspetti in maniera differente.
Nelle istanze di tipo “fabio:Expression” vengono dati, anche se solo per un terzo tra quelle analizzate, anche l’abstract e la lista dei riferimenti, dati che vengono omessi nelle entità “fabio:JournalArticle”.
Queste ultime, invece, gestiscono meglio i riferimenti agli autori e le altre informazioni quali la data di pubblicazione e gli identificatori come il DOI.
L’affiliazione nelle entità ruolo (e quindi per ogni autore), si trova molto raramente.
Le istanze per i libri presentano molte informazioni; un po’ meno, invece, quelle per i journal.
Tutto sommato, è sembrato comunque un buon dataset, aggiudicandosi quindi un 7 su 10.

4.3 ACM RKB Explorer

ACM (*Association for Computing Machinery*) [41] è una società no-profit, considerata la più grande associazione dedicata al calcolo scientifico e all’informatica.
Tra i servizi che offre, troviamo l’*ACM Digital Library* [42], un grosso database contenente articoli scientifici (nel contesto matematico e informatico), incluse le pubblicazioni di ACM.
ACM RKB Explorer [43] è il semantic repository che contiene gli articoli pubblicati da ACM in formato RDF.

Numero di triple	12.3 milioni

Tabella 17 Dimensioni di ACM

	Dati bibliografici	Identificatori	Abstract	Autori	Enti - Affiliazioni	Classificazione	Citazioni - Riferimenti
proprietà	akt:has-title akt:has-pubblication-reference akt:has-date			akt:has-author akt:full-name		akt:addresses-generic-area-of-interest	akt:cites-publication-reference
ontologie	AKT			AKT			AKT

Tabella 18 Proprietà utilizzate da ACM

Il dataset utilizza unicamente l'ontologia AKT e conta oltre 12.3 milioni di triple.

Tramite il sito è possibile fare ricerche attraverso una semplice form nella homepage, cercando una parola chiave o un URI, oppure scrivendo query in SPARQL.
Le ricerche per l'analisi sono state fatte con entrambi i metodi.
Inoltre offre un servizio *CRS* (*Coreference Resolution Service*), per identificare e gestire gli URI duplicati.

Non è stato possibile contare ogni tipo di entità, poiché è stato riscontrato l'errore: "COUNT cannot be used with SPARQL" durante l'uso delle query.

Sono stati analizzati oltre 120 articoli, identificati con la classe “akt:Article-Reference”:

Proprietà	**Percentuale**
Dati bibliografici: akt:has-title; akt:has-publication-reference; akt:has-date.	100%
Autori: akt:has-author.	95%
Classificazione: akt:addresses-generic-area-of-interest.	85%
Citazioni: akt:cites-publication-reference.	60%

Tabella 19 Proprietà negli articoli di ACM

Sono stati analizzati circa 80 autori, identificati con le classi “akt:Person” e “akt:Generic-Agent”:

Proprietà	**Percentuale**
Identificazione: akt:full-name.	100%

Tabella 20 Proprietà negli autori di ACM

In conclusione:

Gli articoli in questo dataset non tengono conto di diversi aspetti tra cui l’abstract, gli enti associati e gli identificatori come il DOI.
Su quasi la metà delle pubblicazioni mancano i riferimenti alle citazioni.
Su altri, invece, vengono omessi alcuni riferimenti che invece si trovano nell’articolo concreto.
Gli autori hanno come unica informazione il nome completo, omettendo informazioni come l’affiliazione.
Complessivamente, questo dataset è molto povero: il mio voto è un 3 su 10, per la scarsa quantità di dati presenti nelle entità.

4.4 DBLP RKB Explorer

DBLP (*Digital Bibliography & Library Project*) [44] è un sito web contenente un database che raccoglie le informazioni sulle principali pubblicazioni di genere informatico, come libri, riviste, articoli o conference proceeding (collezioni di articoli pubblicati nel contesto di una conferenza).
DBLP RKB Explorer [45] è un semantic repository, uno dei dataset di DBLP che racchiude i dati presi da DBLP Computer Science Bibliography.

Numero di triple	43.1 milioni
Tempo medio query (con limite=250)	10 secondi

Tabella 21 Dimensioni di DBLP

	Dati bibliografici	**Identificatori**	**Abstract**	**Autori**	**Enti - Affiliazioni**	**Classificazione**	**Citazioni - Riferimenti**
proprietà	akt:has-title akt:article-of-journal akt:has-volume akt:has-date	akt:has-web-address		akt:has-author akt:edited-by akt:full-name			akt:cites-publication-reference
ontologie	AKT	AKT		AKT			AKT

Tabella 22 Proprietà utilizzate da DBLP

Il dataset usa esclusivamente l'ontologia AKT e conta oggi oltre 43.1 milioni di triple.
Come per ACM RKB Explorer, all'interno del sito è possibile effettuare ricerche sia tramite una form nella homepage, che attraverso query in SPARQL, ed è possibile usufruire del servizio *CRS*.
Per l'analisi, sono stati usati entrambi i metodi.
Su alcuni risultati di ricerca (circa una decina) ho riscontrato un errore 404, ossia "Not found", all'interno del dataset.
Non è stato possibile contare ogni tipo di entità, poiché è stato riscontrato l'errore: "COUNT cannot be used with SPARQL" nell'uso delle query, stesso problema trovato nel dataset di ACM.

Il tempo di risposta di una ricerca (tramite query) è relativo al limite dei risultati che inserisci; con un limite di 250 entità, il tempo medio è di circa 10 secondi.

Sono stati analizzati circa 100 articoli di tipo “akt:Book-Section-Reference”:

Proprietà	**Percentuale**
Dati bibliografici: akt:has-title; akt:article-of-journal; akt:has-date.	100%
Autori: akt:has-author.	99% (solo un entità è stata trovata senza questa proprietà)
Citazioni: akt:cites-publication-reference.	70%
Dati bibliografici: akt:has-volume.	1% (una sola entità è stata trovata con questo attributo)

Tabella 23 Proprietà negli articoli di DBLP

Sono stati analizzati circa 70 pubblicazioni di tipo “akt:Conference-Proceedings-Reference”:

Proprietà	**Percentuale**
Dati bibliografici: ak:has-title; akt:has-date.	100%
Dati bibliografici: akt:article-of-journal.	90%
Autori: akt:edited-by.	80%
Dati bibliografici: akt:has-volume.	50%
Identificatori: akt:has-web-address.	20%

Tabella 24 Proprietà nei conference proceeding di DBLP

Sono state analizzate circa 50 entità riferite ai libri, identificate con la classe "akt:Book-Reference":

Proprietà	**Percentuale**
Dati bibliografici: akt:has-title; akt:has-date. ***Autori***: akt:has-author.	100%
Dati bibliografici: akt:has-volume ***Identificatori:*** akt:has-web-address	30%
Dati bibliografici: akt:article-of-journal	1% (solo su 2 entità è stato trovato questo attributo)

Tabella 25 Proprietà nei libri di DBLP

Sono stati analizzati circa 70 journal, identificati con la classe "akt:Journal":

Proprietà	**Percentuale**
Dati bibliografici: akt:has-title.	100%

Tabella 26 Proprietà nei journal di DBLP

Sono stati analizzati circa 80 autori, definiti con le classi "akt:Person" e "akt:Generic-Agent":

Proprietà	**Percentuale**
Identificazione: akt:full-name.	100%

Tabella 27 Proprietà negli autori di DBLP

In conclusione:

Le citazioni vengono gestite in maniera più efficiente rispetto ai precedenti dataset.

Vengono, però, completamente omessi altri aspetti tra cui l'abstract, le informazioni riguardo gli enti ed i dati per la classificazione.

Le entità journal hanno solo l'informazione riferita al titolo, come anche gli autori, che hanno solamente il nome completo.
Anche questo dataset è abbastanza povero: il voto è 4 su 10, un po' di più rispetto ad ACM poiché le informazioni vengono gestite meglio.

4.5 J.UCS Bibliography Database

J.UCS (*Journal of Universal Computer Science*) [46] è uno dei primi giornali elettronici, apparso nel web nel 1994, che continua tutt'ora a pubblicare articoli.
Esso contiene una raccolta delle proprie pubblicazioni che occupano tutti gli aspetti di ogni campo dell'informatica.
All'interno del proprio dataset [47] , pubblicato con D2R Server, è possibile accedere alla lista di tutti gli articoli pubblicati e alla lista di tutti gli autori coinvolti.

Numero di Triple	125 mila
Articoli	2121
Autori	4478

Tabella 28 Dimensioni di J.UCS

	Dati bibliografici	**Identificatori**	**Abstract**	**Autori**	**Enti - Affiliazioni**	**Classificazione**	**Citazioni - Riferimenti**
proprietà	dc:isVersionOf dc:title sioc:title dc:SizeOrDuration dc:isPartOf swrc:volume dc:issued	dc:identifier	dcterms:abstract	dc:creator foaf:maker foaf:firstname foaf:surname foaf:name swrc:author foaf:mbox vcard:Country vcard:Locality foaf:topic_interest	swrc:journal iswc:has_affiliation	skos:concept foaf:primaryTopic skos:exactMatch skos:narrowMatch	
ontologie	Dublin Core SIOC SWRC	Dublin Core	Dublin Core	Dublin Core FOAF VCARD	SWRC ISWC	SKOS FOAF	

Tabella 29 Proprietà utilizzate da J.UCS

Il dataset può essere esplorato attraverso diversi Semantic Web Browser come *Tabulator* [48] o *Disco* [49], oppure scrivendo una query in SPARQL.
Le ricerche per questo dataset sono state effettuate tramite le liste.

Ad oggi, come abbiamo visto, si trovano 2121 articoli e 4478 autori, per un totale di oltre 125 mila triple (le conte sono state effettuate tramite le query).

Sono stati analizzati circa 100 articoli, identificati con le classi “swc:Paper”, “swrc:Article” e “foaf:Document”:

Proprietà	**Percentuale**
Dati Bibliografici: dc:SizeOrDuration; dc:isPartOf; dc:isVersionOf; swrc:volume. ***Identificatori:*** dc:identifier. ***Autori:*** dc:creator; foaf:maker. ***Enti:*** swrc:journal.	100%
Dati Bibliografici: dc:issued.	99% (una sola entità è stata trovata senza questo attributo)
Dati Bibliografici: dc:dateAcceptated; dc:dateSubmitted.	95%
Abstract: dcterms:abstract. ***Classificazione:*** skos:Concept.	85%
Classificazione: foaf:primaryTopic.	80%
Dati Bibliografici: sioc:title.	55% (chi non possiede “sioc:title” ha al suo posto “dc:title”)
Dati Bibliografici: dc:title.	45%
Classificazione: skos:exactMatch.	40%
Classificazione: skos:narrowMatch.	35%

Tabella 30 Proprietà negli articoli di J.UCS

Sono stati analizzati circa 100 autori, identificati con la classe “foaf:Person”:

Proprietà	**Percentuale**
Identificatori: foaf:firstname; foaf:surname; foaf:name. ***Affiliazioni:*** iswc:has_affiliation. ***Altri dati:*** foaf:mbox; swrc:author.	100%
Altri Dati: vcard:Country.	99% (ad una sola entità mancava questa proprietà)
Altri Dati: vcard:Locality.	90%
Altri Dati: foaf:topic_interest.	20%

Tabella 31 Proprietà nelle persone di J.UCS

In conclusione:

In questo dataset le entità degli articoli contengono una grande varietà di dati, cercando di coprire quasi tutti gli aspetti.
L’unica pecca, come si può notare dalla tabella riassuntiva, è la completa mancanza di citazioni.
Anche le entità degli autori vengono molto arricchite, dando diverse informazioni come l’affiliazione, la casella di posta ed il paese di provenienza.
In generale, questo dataset è molto buono, guadagnandosi un buon 8 su 10.

4.6 Semantic Web Dog Food Corpus

Semantic Web Dog Food Corpus (conosciuto anche come *Semantic Web Conference Corpus*) [50] è un database che raccoglie tutte le informazioni riguardo le conferenze ed i workshop sul Semantic Web, tra cui gli articoli pubblicati, le organizzazioni e le persone coinvolte.
Il nome deriva dall’espressione "*Eating your own dog-food*", ossia “mangia il tuo stesso cibo per cani”, usato per riferirsi ad uno scenario dove una compagnia utilizza il proprio prodotto per garantirne la qualità.

Numero di triple	246 mila
Articoli	4889
Persone	10982
Organizzazioni	3084
Conferenze	38
Workshop	235

Tabella 32 Dimensioni di SWDF

	Dati bibliografici	**Identificatori**	**Abstract**	**Autori**	**Enti - Affiliazioni**	**Classificazione**	**Citazioni - Riferimenti**
proprietà	dc:title swrc:title swc:isPartOf swrc:pages swrc:year swrc:month swc:relatedToEvent swc:hasRelatedEvent	swrc:url	swrc:abstract	foaf:member foaf:maker swrc:author dc:creator foaf:name foaf:made foaf:mbox_sha1sum foaf:mbox foaf:firstName swc:holdsRole foaf:based_near foaf:lastName foaf:familyName foaf:homepage foaf:page	swrc:affiliation swc:affiliation foaf:name foaf:homepage foaf:based_near swc:holdsRole foaf:logo foaf:page	dc:subject swrc:category swc:hasTopic	
ontologie	Dublin Core SWRC SWC	SWRC	SWRC	FOAF Dublin Core SWRC SWC	SWRC SWC FOAF	Dublin Core SWRC SWC	

Tabella 33 Proprietà utilizzate da SWDF

Ad oggi, come abbiamo già citato, il dataset raccoglie circa 4889 articoli, 10982 persone e 3084 organizzazioni in 38 conferenze e 235 workshop; per un totale di oltre 246 mila triple.
All'interno del dataset, oltre alle ontologie comuni, viene utilizzato un proprio vocabolario chiamato "SWC".

Nel sito è possibile effettuare ricerche tramite una semplice form dove inserire una parola chiave, oppure attraverso delle liste complete, una per ogni tipo di entità.
Inoltre, è possibile compiere query utilizzando il linguaggio SPARQL.

Le ricerche all'interno del dataset sono state effettuate utilizzando sia le liste complete che le query.

Sono stati analizzati circa 70 articoli, identificati con la classe "swrc:InProceedings":

Proprietà	**Percentuale**
Dati Bibliografici: dc:title.	95%
Dati Bibliografici: swc:isPartOf; swrc:year. ***Autori:*** foaf:maker.	90%
Abstract: swrc:abstract.	80%
Autori: swrc:author.	75%
Dati Bibliografici: swrc:month. ***Identificatori:*** swrc:url.	65%
Autori: dc:creator.	60%
Dati bibliografici: swrc:pages. ***Classificazione:*** dc:subject.	35%
Dati bibliografici: swc:relatedToEvent.	30%
Dati bibliografici: swrc:title. ***Classificazione:*** swrc:category.	20%
Classificazione: foaf:topic	15%
Dati bibliografici: swc:hasRelatedEvent. ***Classificazione:*** swc:hasTopic.	10%

Tabella 34 Proprietà negli articoli di SWDF

Sono state analizzate circa 100 persone, identificate con la classe “foaf:Person”:

Proprietà	**Percentuale**
Identificazione: foaf:name.	100%
Altri dati: foaf:made.	90%
Altri dati: swrc:affiliation; foaf:mbox_sha1sum.	70%
Identificazione: foaf:firstName.	55%
Altri dati: swc:holdsRole; foaf:based_near.	45%
Identificazione: foaf:lastName.	40%
Identificazione: foaf:familyName.	25%
Altri dati: foaf:homepage; foaf:page.	20%
Altri dati: foaf:mbox.	15%
Altri dati: swc:affiliation.	10%
Altri dati: foaf:status; foaf:depiction.	5%
Identificazione: foaf:givenName.	2%
Identificazione: foaf:surname. ***Altri dati:*** swc:givesKeynoteTalk; foaf:workplaceHomepage; swc:isPanelistAt; swc:moderates.	1% (queste proprietà sono state trovate una sola volta)

Tabella 35 Proprietà nelle persone di SWDF

Sono stati analizzati circa 60 organizzazioni, identificate con "foaf:Organization":

Proprietà	**Percentuale**
Identificazione: foaf:name. ***Altri dati:*** foaf:member.	100%
Altri dati: foaf:homepage.	17%
Altri dati: foaf:based_near.	10%
Altri dati: swc:holdsRole; foaf:logo; foaf:page.	5%

Tabella 36 Proprietà nelle organizzazioni di SWDF

In conclusione:

Nelle entità per gli articoli, vengono inserite una grande varietà di informazioni, ricoprendo quasi tutti gli aspetti.
Le uniche informazioni che mancano sono quelle relative alle citazioni, che vengono completamente omesse in ogni pubblicazione analizzata.
Anche negli autori vengono inseriti molti dati, ma in modo più sparso, come si può notare dalla tabella.
Tutto sommato, questo dataset è molto buono, anche se le informazioni dovrebbero essere gestite in maniera più efficiente.
Il voto complessivo è dunque 8 su 10.

4.7 Semantic Web Journal

Il *Semantic Web Journal* [51] è un rivista internazionale e interdisciplinare pubblicata da IOS Press.
Lo scopo di questa pubblicazione è riunire ricercatori provenienti da settori diversi che condividono la visione ed il bisogno di modi più efficaci e significativi per condividere le informazioni tra gli agenti ed i servizi nel futuro di Internet e altrove.

Numero di Triple	21335
Numero di entità	2440
Articoli	432
Persone	1139

Tabella 37 Dimensioni di SWJ

	Dati bibliografici	**Identificatori**	**Abstract**	**Autori**	**Enti - Affiliazioni**	**Classificazione**	**Citazioni - Riferimenti**
proprietà	dcterms:created swjterms:isEarliestVersion swjterms:isLatestVersion dcterms:isVersionOf bibo:status dcterms:title swjterms:updated bibo:issue bibo:pageEnd bibo:pageStart bibo:volume swjterms:editorDecisionSubmitted swjterms:reviewComment swjterms:reviewSubmitted swjterms:reviewerAccepted swjterms:hasPreviousVersion swjterms:reviewerInvited swjterms:hasNextVersion swjterms:editorAssigned	bibo:identifier bibo:uri bibo:doi	bibo:abstract	bibo:authorList dcterms:creator bibo:editor swjterms:reviewer foaf:name swjterms:isCreatorOf swjterms:isReviewerOf swjterms:editorAssigned		swjterms:submissionType	
ontologie	Dublin Core SWJ BIBO	BIBO	BIBO	BIBO Dublin Core SWJ		SWJ	

Tabella 38 Proprietà utilizzate da SWJ

SWJ è la prima rivista dedicata interamente alle ricerche per il Semantic Web, pubblicando i propri articoli online in modo da essere accessibili ovunque, senza il bisogno di registrazione.
All'interno del dataset di Semantic Web Journal [52] è possibile esplorare i dati unicamente effettuando query in SPARQL.

Ad oggi, conta 21335 triple e 2440 entità tra cui:
- 432 entità articolo;
- 1139 entità persona.

Nel dataset, oltre alle ontologie più comuni, viene utilizzato un proprio vocabolario, chiamato “SWJ”.

Sono stati analizzati circa 80 articoli, descritti utilizzando la classe “swjterms:AcademicArticleVersion”:

Proprietà	**Percentuale**
Dati bibliografici: dcterms:created; swjterms:isEarliestVersion; swjterms:isLatestVersion; dcterms:isVersionOf; bibo:status; dcterms:title; swjterms:updated. ***Identificatori:*** bibo:identifier; bibo:uri. ***Abstract:*** bibo:abstract. ***Autori:*** bibo:authorList; dcterms:creator; bibo:editor. ***Classificazione:*** swjterms:submissionType.	100%
Dati Bibliografici: bibo:issue; bibo:pageEnd; bibo:pageStart; bibo:volume. ***Identificatori:*** bibo:doi.	35%
Dati Bibliografici: swjterms:editorDecisionSubmitted.	23%
Dati Bibliografici: swjterms:reviewComment; swjterms:reviewSubmitted; swjterms:reviewerAccepted; swjterms:hasPreviousVersion; swjterms:reviewerInvited; swjterms:hasNextVersion; swjterms:editorAssigned. ***Autori:*** swjterms:reviewer.	20%
Dati Bibliografici: swjterms:hasPreviousVersion.	14%
Dati Bibliografici: swjterms:reviewerInvited.	13%
Dati Bibliografici: swjterms: hasNextVersion.	10%
Dati Bibliografici: swjterms:editorAssigned.	5%

Tabella 39 Proprietà negli articoli di SWJ

Sono state analizzate circa 120 persone, identificate con la classe “foaf:Person”:

Proprietà	**Percentuale**
Identificatori: foaf:name.	100%
Altri Dati: swjterms:isCreatorOf.	87%
Altri Dati: swjterms:isReviewerOf.	15%
Altri Dati: swjterms:isEditorOf.	3%

Tabella 40 Proprietà nelle persone di SWJ

In conclusione:

In questo dataset, le informazioni per le pubblicazioni vengono gestite in maniera più organizzata rispetto agli altri dataset che abbiamo analizzato.
Ogni articolo esaminato ricopre quasi tutti gli aspetti che abbiamo preso in considerazione.
Mancano le informazioni riferite agli enti, ma questo perché è Semantic Web Journal stesso che produce e pubblica gli articoli nel proprio giornale.
Anche in questo caso, sono completamente assenti le informazioni riguardo le citazioni delle pubblicazioni, non essendo pervenute in nessuna entità esaminata.
In generale, questo è il miglior dataset che ho analizzato, guadagnandosi un voto di 9 su 10.

Capitolo 5

Identificazione degli autori: il prototipo GReAT

5.1 Introduzione al problema

Un problema abbastanza comune, tra quelli riscontrati nei dataset esaminati, è sicuramente la corretta identificazione di un autore, dovuto alle diverse entità che fanno riferimento ad una stessa persona.
Solitamente, quando un dataset viene creato, vengono utilizzati diversi tool di estrazione che prendono in input gli articoli e producono le informazioni che verranno poi pubblicate.
Tutto questo porta ad avere Linked Dataset in cui una stessa persona può essere associata a molte entità e, quindi, può essere rappresentata in modi diversi.
Infatti può capitare che, per esempio, ad un'entità venga tralasciata l'informazione sul secondo nome, oppure che vengano fatti errori di sintassi come l'omissione degli accenti.

Prendiamo due entità da *Journal of Web Semantics*, un dataset che raccoglie gli articoli in ambito del Semantic Web e che, come vedremo in seguito, abbiamo usato per il testing del tool:

Entità 1:

```
<http://www.essepuntato.it/semantic-lancet/1-s2.0-S1570826803000039/author/katia-sycara>
  a foaf:Person ;
  foaf:givenName "Katia" ;
  foaf:familyName "Sycara" ;
  rdfs:label "Katia Sycara" ;
  pro:holdsRoleInTime <http://www.essepuntato.it/semantic-lancet/1-s2.0-
S1570826803000039/author/katia-sycara-as-author> .
```

Entità 2:

<http://www.essepuntato.it/semantic-lancet/1-s2.0-S1570826808000838/author/katia-sycara>
 a foaf:Person ;
 foaf:givenName "Katia" ;
 foaf:familyName "Sycara" ;
 rdfs:label "Katia Sycara" ;
 pro:holdsRoleInTime <http://www.essepuntato.it/semantic-lancet/1-s2.0-S1570826808000838/author/katia-sycara-as-author> .

Queste due entità rappresentano la stessa persona, ossia "Katia Sycara", ma vengono gestite come se facessero riferimento a due individui diversi.
Ciò accade perché quando sono stati caricati nel dataset i due articoli da lei scritti, sono state create due istanze per lo stesso autore.

Abbiamo così creato un *tool* in grado di individuare e cercare di risolvere questo genere di problema.
GReAT (*Giorgi's Redundant Authors Tool*) è uno strumento che può essere utilizzato la *disambiguazione degli autori*, identificando così univocamente le persone all'interno di un dataset.

5.2 La logica del tool

Per stabilire se due entità fanno riferimento ad uno stesso autore, GReAT prende in considerazione diversi aspetti:

- il nome ed il cognome possono essere stati aggiunti separatamente, potendo dare un informazione più precisa riguardo al nominativo di una persona;
- i nomi possono essere stati inseriti in maniera differente, ad esempio utilizzando accenti diversi oppure omettendo caratteri come l'apostrofo;
- alcune parti dei nominativi, come i secondi nomi, potrebbero essere stati puntati. In altri casi, invece, potrebbero essere stati omessi;
- altre proprietà, oltre al nome completo, possono essere utili per la disambiguazione degli autori.

Tenendo presente i punti sopracitati, abbiamo ideato un algoritmo per la disambiguazione degli autori che, confrontando due entità, cerca di riconoscere se esse facciano riferimento alla stessa persona.

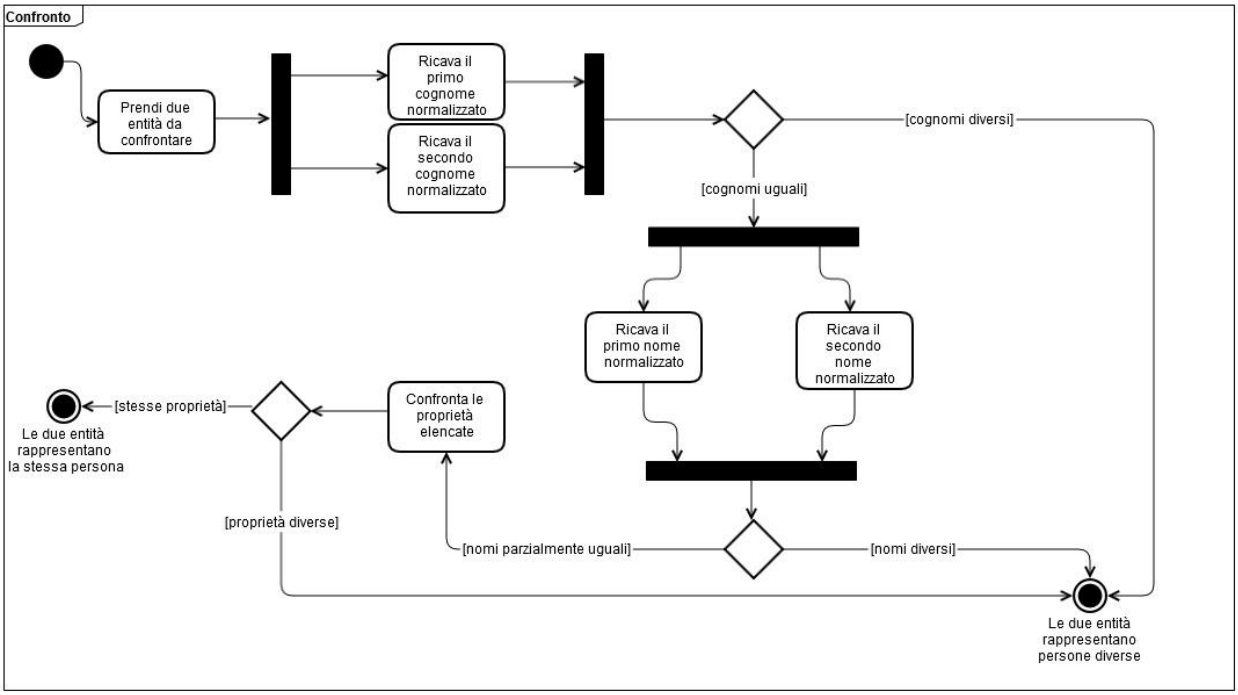

Figura 7 La logica di GReAT

La prima comparazione che effettua è quella dei cognomi, poiché se sono differenti possiamo già dedurre che due persone siano diverse.
La seconda operazione è il controllo dei nomi: verifichiamo se uno dei due nomi è sottoinsieme dell'altro, tenendo conto di quelli puntati.
Sia i nomi che i cognomi vengono "normalizzati", ossia vengono eliminati eventuali accenti o caratteri speciali.
Infine, facciamo i confronti con le proprietà che abbiamo deciso di comparare.
Il controllo delle proprietà è utile perché se due individui hanno lo stesso nome non è detto che siano la stessa persona, come nel caso degli omonimi.
Se, invece, riusciamo a comparare informazioni come l'indirizzo e-mail oppure la homepage possiamo dare la certezza che le due entità fanno riferimento alla stessa persona.

Poiché eravamo indecisi sul grado di "bontà" con cui due entità possono essere considerate equivalenti o meno, abbiamo fatto in modo che il tool potesse essere eseguito su due diverse modalità:

- modalità “Loose” (buono), in cui basta che il nome corrisponda per poter già dire che due entità facciano riferimento alla stessa persona;
- modalità “Strict” (pedante), in cui deve esserci almeno una proprietà in comune, oltre al nome, per poter dire due entità siano equivalenti.

5.3 L’implementazione del Tool

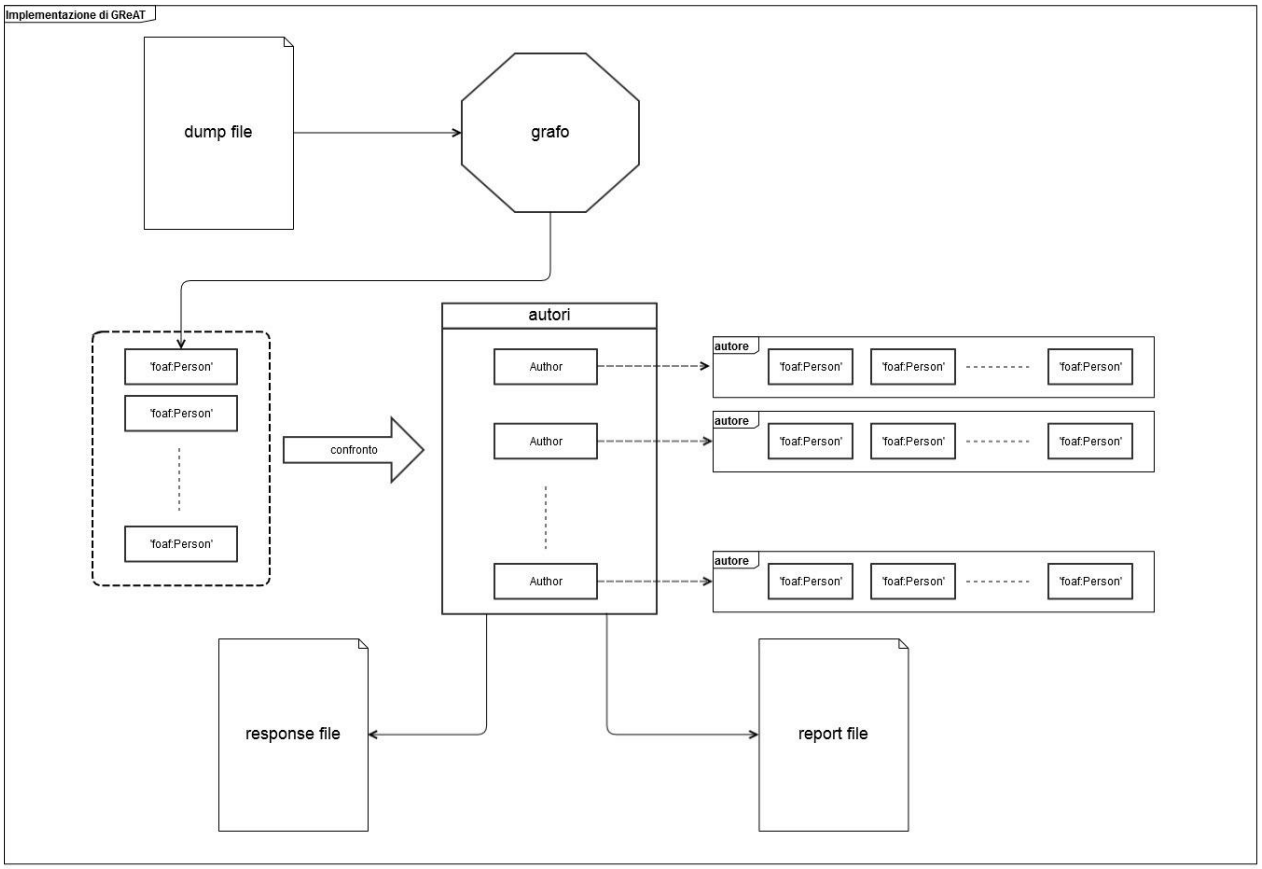

Figura 8 L’implementazione di GReAT

Per effettuare la disambiguazione degli autori in un dataset, GReAT ha bisogno del *dump file*, ossia un file dove viene salvato il contenuto del dataset.
Il tool è stato implementato con il linguaggio *php* e usa la libreria *Graphite*, che a sua volta utilizza *ARC2,* per analizzare il file.
Inoltre, come già accennato nel capitolo sulle ontologie, viene utilizzato il vocabolario *FOAF* sia per riconoscere le entità che fanno riferimento ad una persona che per ricavare il nome.

L'implementazione del software è sintetizzata nei seguenti punti:

- vengono presi in input due parametri: il formato ("rdf" oppure "turtle") ed il percorso del file da analizzare;
- viene creata una classe "*Author*" contenente un array con le entità delle persone che abbiamo riconosciuto equivalenti;
- creiamo un vettore chiamato "final_authors" di oggetti Author, che rappresenterà gli autori effettivi del dataset;
- istanziamo il grafo del dataset con la classe "*Graphite*", dove aggiungiamo il contenuto del file;
- vengono prelevate tutte le entità del grafo di tipo '*foaf:Person*' che verranno inserite in una lista;
- ogni entità nella lista viene aggiunta nel vettore finale, controllando che l'autore a cui fa riferimento non sia già presente, attraverso una funzione "is_already_in()". Questa funzione compara l'entità per ogni elemento nel vettore di ogni oggetto Author presente, utilizzando una funzione chiamata "are_same_people()" che confronta le varie entità utilizzando la logica spiegata nel paragrafo precedente;
- finiti i confronti, vengono aggiunti gli opportuni "*owl:sameAs*" per collegare le entità che fanno riferimento allo stesso autore;
- infine, serializziamo il grafo in un *file di risposta* col formato "turtle", creando anche un *file di report* che riassume le disambiguazioni effettuate.

Viene, inoltre, utilizzato un file ini per elencare le proprietà che vogliamo confrontare e per settare il grado di bontà del tool.

Nel prossimo paragrafo controlleremo l'efficienza di GReAT, testando il tool su diversi dump file ed esaminando i risultati.

5.4 Testing

Per verificare l'efficacia di GReAT, abbiamo raccolto diversi dump file e abbiamo lanciato il tool per ciascuno di essi.
Attraverso i file di report, abbiamo poi verificato quante disambiguazioni sono state sicuramente corrette, facendo ricerche su internet e controllando

manualmente che le entità considerate equivalenti dal tool facciano effettivamente riferimento alla stessa persona.
Non siamo, invece, riusciti a verificare i *falsi negativi*, ossia le disambiguazioni che sarebbero dovute essere fatte, ma che il tool non ha eseguito.

Tutte le prove sono state effettuate a riga di comando utilizzando *Php 5.5*.

Di seguito, riassumiamo i vari risultati ottenuti.

GReAT su Semantic Web Dog Food

Abbiamo testato il tool su alcuni dump file di diverse conferenze prese da SWDF ed esponiamo i risultati nei casi in cui ci sono state delle disambiguazioni. Ogni file contiene tutte le informazioni sugli articoli che fanno riferimento alla relativa conferenza, comprese le liste degli autori e le entità per ogni persona.
Tutte le prove su questo dataset sono state effettuate avendo come proprietà da confrontare '*foaf:homepage*', '*foaf:mbox*', '*foaf:mbox_sha1sum*', '*foaf:page*' e '*foaf:based_near*', ossia quelle che sono state riscontrate maggiormente durante l'analisi.

- *European Semantic Web Conference 2014*:

Test sul dump file che contiene i dati relativi alla conferenza europea sul Semantic Web tenutasi nel 2014.

Modalità	Loose
Tempo	152 secondi
Persone trovate	893
Autori effettivi identificati	892
Totale disambiguazioni	1
Disambiguazioni corrette	1

Tabella 41 Risultati di GReAT su ESWC2014

Con il tool in modalità "Loose", abbiamo avuto i risultati dopo 152 secondi, trovando 893 persone e riconoscendo 892 autori, per un totale di una sola disambiguazione.
I due autori unificati avevano lo stesso nome e stesso cognome, ma erano identificati con due URI diversi.

Facendo le opportune ricerche, abbiamo confermato che la disambiguazione era corretta.

- *International Semantic Web Conference 2013*:

Test sul dump file che contiene i dati relativi alla conferenza internazionale sul Semantic Web, avvenuta nel 2013.

Modalità	Loose
Tempo	170 secondi
Persone trovate	1013
Autori effettivi identificati	1008
Totale disambiguazioni	5
Disambiguazioni corrette	5

Tabella 42 Risultati di GReAT su ISWC2013

Con il tool in modalità "Loose", abbiamo ottenuto 1008 autori identificati fra 1013 entità, per un totale di 5 disambiguazioni.
In tutti e 5 i casi il problema era nel secondo nome, un'informazione che veniva inserita su un'entità e omessa in un'altra.
Controllando manualmente su internet, abbiamo avuto la certezza che tutte e 5 le disambiguazione erano corrette, poiché le informazioni come l'affiliazione combaciavano.

Modalità	Strict
Tempo	174 secondi
Persone trovate	1013
Autori effettivi identificati	1012
Totale disambiguazioni	1
Disambiguazioni corrette	1

Tabella 43 Risultati di GReAT su ISWC2013 in modalità "Strict"

Con il tool in modalità "Strict", invece, abbiamo ottenuto una sola disambiguazione dopo 174 secondi.
Questa disambiguazione, già avvenuta in modalità "Loose", è stata confermata grazie all'attributo in comune "foaf:mbox_sha1sum".

- *International World Wide Web Conference 2012*:

Test sul dump file con i dati relativi alla conferenza internazionale sul World Wide Web, avvenuta nel 2012.

Modalità	Loose
Tempo	508 secondi
Persone trovate	1627
Autori effettivi identificati	1614
Totale disambiguazioni	13
Disambiguazioni corrette	12

Tabella 44 Risultati di GReAT su WWW2012

Dopo più di 8 minuti, il tool in modalità "Loose" ha effettuato 13 disambiguazioni.
12 tra queste sono state confermate senza fare ricerche su internet, poiché l'attributo "foaf:mbox_sha1sum" combaciava.
Una disambiguazione, invece, era sbagliata: oltre ad avere la proprietà "foaf:mbox_sha1sum" diversa, il nome completo era chiaramente differente, ossia il primo era "Su Myeon Kim" ed il secondo era "Su Nam Kim".

Modalità	Strict
Tempo	517 secondi
Persone trovate	1627
Autori effettivi identificati	1616
Totale disambiguazioni	11
Disambiguazioni corrette	11

Tabella 45 Risultati di GReAT su WWW2012 in modalità "Strict"

Con il tool in modalità "Strict", abbiamo confermato 11 fra le 13 disambiguazioni trovate con la modalità "Loose".
Tutte e 11 sono state confermate senza la necessità di ricerche grazie all'attributo "foaf:mbox_sha1sum", come nella prima prova.

GReAT su Journal of Web Semantics

Test sul dump file del dataset di *Journal of Web Semantics*, già descritto all'inizio del capitolo, contenente tutte le entità che rappresentano le persone presenti, comprese le entità per i ruoli e le liste di autori di ogni pubblicazione.
Per ogni autore nel dataset sono specificati il nome, il cognome ed il suo ruolo.

Modalità	Loose
Tempo	249 secondi
Persone trovate	1226
Autori effettivi identificati	892
Totale disambiguazioni	334
Disambiguazioni controllate	60
Disambiguazioni corrette (tra quelle controllate)	50

Tabella 46 Risultati di GReAT su JWS

Non avendo trovato altre proprietà nelle entità, la disambiguazione è stata effettuata unicamente con il tool in modalità "Loose" e senza un elenco di attributi.
Con un tempo di 256 secondi, su 1226 entità persona all'interno del file, GReAT ha riconosciuto 892 autori effettivi, per un totale di 334 disambiguazioni.
Ne abbiamo controllato circa 60, facendo le opportune ricerche su internet:

- 50 erano sicuramente corrette, dando così una percentuale di successo (riferita al campione) dell'83% circa;
- 8 disambiguazioni sono state considerate errate, poiché le informazioni come l'email o l'affiliazione non combaciavano;
- 2 disambiguazioni, oltre alle informazioni errate, avevano i nomi chiaramente diversi, ossia "Philip A. Bernstein" - "Abraham Bernstein", e "Neo D. Martinez" - "David Martinez".

5.5 Possibili Sviluppi di GReAT

GReAT è il primo passo verso un sistema molto più articolato, che possa essere in grado di effettuare le disambiguazioni con un margine di errore minimo.
Il tool, infatti, presenta diversi limiti: primo tra tutti è il confronto puramente sintattico fra gli attributi.
L'identificazione univoca degli autori è un problema molto complesso, in cui ci sono molti più casi da gestire rispetto a quelli di cui si occupa il tool.

Una possibile estensione di GReAT, che abbiamo tentato di implementare, può essere il confronto dei co-autori, ossia verificare che due soggetti abbiano collaborato con uno stesso individuo per poter avere un'ulteriore conferma della loro equivalenza.

Conclusioni

Con il nostro lavoro, abbiamo cercato di fare chiarezza sulla situazione attuale dei Linked Open Dataset per le pubblicazioni scientifiche.

I dataset analizzati, come abbiamo potuto osservare, presentano una certa eterogeneità: le entità vengono gestite attraverso ontologie differenti e le pubblicazioni vengono descritte utilizzando diversi aspetti, a seconda delle proprie esigenze.
Basti confrontare il dataset di ACM con quello di Semantic Web Journal: mentre il primo tiene conto solo delle informazioni essenziali di un articolo come il titolo e la data di pubblicazione, il secondo prende in considerazione diversi aspetti come le versioni precedenti o successive, lo stato del documento e le informazioni riguardo le eventuali revisioni.
L'obiettivo del Linked Data è quello di collegare i dati presi da sorgenti differenti, rendendo interoperabili dataset con strutture diverse.

La nostra rassegna ha voluto evidenziare, inoltre, come il problema della qualità dei dati sia abbastanza comune tra i vari dataset.
Infatti, come si può notare, le entità di uno stesso tipo all'interno del medesimo dataset possono essere rappresentate in maniera non uniforme.
Prendiamo, ad esempio, il caso di Semantic Web Dog Food: nelle diverse pubblicazioni abbiamo trovato molte proprietà che però non sono state pervenute in gran parte delle entità esaminate.
Oppure in J.UCS, dove il 55% delle entità analizzate riferite agli articoli descrivono il titolo attraverso l'attributo "sioc:title", mentre il rimanente 45% utilizza "dc:title".

La nostra rassegna va considerata comunque come un primo sguardo verso un ambiente molto più vasto, dove dovrebbero essere presi in considerazione molti più parametri di valutazione.
Gli aspetti di cui abbiamo tenuto conto, per esempio, possono essere estesi, dando un ulteriore livello di specifica rispetto ai tipi di informazione presenti nelle entità dei dataset.

Riferimenti

[1] World Wide Web Consortium (W3C). http://www.w3.org/

[2] DiNucci, Darcy (1999). "*Fragmented Future*". http://darcyd.com/fragmented_future.pdf

[3] Google Maps. https://www.google.com/

[4] Writely. http://www.writely.com/

[5] Myspace. https://myspace.com/

[6] Shannon, Victoria (2006). "*A 'more revolutionary' Web*". The New York Times. http://www.nytimes.com/2006/05/23/technology/23iht-web.html

[7] James Hendler, Tim Berners-Lee, Eric Miller (2002). "*Integrating Applications on the Semantic Web*". Journal of the Institute of Electrical Engineers of Japan, Vol 122(10) p. 676-680. http://www.w3.org/2002/07/swint

[8] Christian Bizer, Tom Heath, Tim Berners-Lee (2009). "*Linked Data - the story so far*". International Journal on Semantic Web and Information Systems, 5, (3), 1-22. http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

[9] Tim Berners-Lee (2006). "Linked Data - Design Issues". http://www.w3.org/DesignIssues/LinkedData.html

[10] D2R Server. http://d2rq.org/

[11] Virtuoso Universal Server. http://virtuoso.openlinksw.com/

[12] A. de Waard, (2010). "*From Proteins to Fairytales: Directions in Semantic Publishing, Intelligent Systems*". Intelligence Systems, IEEE. http://lpis.csd.auth.gr/mtpx/sw/material/IEEE-IS/IS-25-2.pdf

[13] David Shotton (2009). "*Semantic publishing: The coming revolution in scientific journal publishing*". Learned Publishing 22 (2): 85–94. doi:10.1087/2009202. http://delos.zoo.ox.ac.uk/pub/2009/publications/Shotton_Semantic_publishing_evaluation.pdf

[14] Fredrik Arvidsson, Annika Flycht-Eriksson (2008). "*Ontologies I*". http://www.ida.liu.se/~janma/SemWeb/Slides/ontologies1.pdf

[15] David Shotton, Katie Portwin, Graham Klyne, Alistair Miles (2008). "*Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article*". In Bourne, Philip E. PLoS Computational Biology 5 (4). http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000361

[16] SPAR Ontologies. http://sempublishing.sourceforge.net/
[17] BIBO. http://bibliontology.com/
[18] Dublin Core Metadata Initiative. http://dublincore.org/
[19] FOAF Project. http://www.foaf-project.org/
[20] Firefox FOAF plug-in. https://addons.mozilla.org/en-us/firefox/tag/foaf
[21] Chrome FOAF plug-in. https://chrome.google.com/webstore/search/foaf
[22] FriendFeed. http://friendfeed.com/
[23] WordPress. http://wordpress.org/
[24] SKOS. http://www.w3.org/2004/02/skos/
[25] EuroVoc. http://eurovoc.europa.eu/
[26] AGROVOC. http://aims.fao.org/standards/agrovoc/about
[27] GEMET. http://www.eionet.europa.eu/gemet/
[28] PRISM Metadata Initiative. http://www.idealliance.org/specifications/prism-metadata-initiative
[29] IDEAlliance. http://www.idealliance.org/
[30] AKT Reference Ontology. http://projects.kmi.open.ac.uk/akt/ref-onto/
[31] Advanced Knowledge Technologies. http://www.aktors.org/
[32] SWRC Ontology. http://ontoware.org/swrc/
[33] OntoWare. http://ontoware.org/
[34] SemIPort. http://km.aifb.uni-karlsruhe.de/semiport
[35] Bibster. http://bibster.semanticweb.org/
[36] Nature Publishing Group. http://www.nature.com/
[37] Nature Linked Data. http://data.nature.com/
[38] Joint Information Systems Committee. http://www.jisc.ac.uk/
[39] JISC Open Citations. http://opencitations.net/
[40] PubMed Central. http://www.ncbi.nlm.nih.gov/pmc/
[41] Association for Computing Machinery. http://www.acm.org/
[42] ACM Digital Library. http://dl.acm.org/
[43] ACM RKB Explorer. http://acm.rkbexplorer.com/
[44] Digital Bibliography & Library Project. http://dblp.uni-trier.de/
[45] DBLP RKB Explorer. http://dblp.rkbexplorer.com/
[46] Journal of Universal Computer Science. http://www.jucs.org/
[47] J.UCS Bibliography Database. http://jucs.org:8181/d2rq/
[48] Tabulator. http://www.w3.org/2005/ajar/tab
[49] Disco. http://wifo5-03.informatik.uni-mannheim.de/bizer/ng4j/disco/
[50] Semantic Web Dog Food Corpus. http://data.semanticweb.org/

[51] Semantic Web Journal. http://www.semantic-web-journal.net/
[52] SWJ Dataset. http://semantic-web-journal.com:3030

Riferimenti nelle Figure:

[F.1] Evoluzioni del Web. http://paanchiweb.blogspot.it/2012/12/getting-essence-of-semantic-web.html
[F.2] Strati del Semantic Web. http://www.w3.org/2007/03/layerCake.png
[F.3] Parte del Linked Data dataset cloud diagram. http://www.theguardian.com/open-platform/blog/linked-data-open-platform
[F.4] Grafo di Nature Linked Data. http://www.nature.com/developers/documentation/linked-data-platform/