Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Scienze di Internet

ESTRAZIONE E ANNOTAZIONE DI UNA RETE CITAZIONALE DA ARTICOLI SCIENTIFICI

Relatore:
Chiar.mo Prof.
ANGELO DI IORIO

Presentata da:
ALESSANDRO DE TROIA

Sessione II
Anno Accademico 2013/2014

Alla mia famiglia,
a mio zio Pasquale.

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Sono passati ormai più di vent'anni dalla pubblicazione del primo sito web[1] di Tim Berners-Lee del 6 agosto 1991 e da quell'insieme di pagine statiche lo stesso informatico britannico è arrivato a teorizzare ciò che oggi sono definiti *Web Semantico*[20], ovvero la trasformazione del Word Wide Web in un ambiente dove le informazioni pubblicate siano in un formato adatto all'interrogazione e interpretazione e, più in generale, all'elaborazione automatica, e i cosiddetti *Linked Data*[21], cioè l'utilizzo del Web al fine di creare collegamenti tra informazioni di diverse sorgenti in maniera da offrire interoperabilità tra sistemi eterogenei.
Tra le caratteristiche più importanti del Semantic Web c'è sicuramente il formato RDF (Resource Description Framework), uno standard creato per descrivere e modellare le informazioni e SPARQL (SPARQL Protocol and RDF Query Language) per l'interrogazione dei dataset annotati in RDF.
Uno degli esempi dell'uso dei Linked Data è il cosiddetto *Semantic Publishing* definito da Shotton[1] come qualsiasi cosa che accresca il significato di un articolo di una rivista, faciliti la sua scoperta automatica, fornisca accesso ai dati all'interno dell'articolo in maniera processabile, o faciliti l'integrazione tra dati e pubblicazioni. Novità in questo settore, visto che fino a quel momento il formato standard *de facto* dell'*Electronic Publishing* era il PDF, con tutti i suoi limiti e svantaggi per l'elaborazione automatica, è l'utilizzo delle ontologie, ovvero dei vocabolari condivisi che vengono utilizzati per modellare un certo dominio.

[1] http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html

Peculiarità del rendere in Linked Data le pubblicazioni scientifiche è la creazione di metadati utili al collegamento e alla creazione di relazioni tra articoli, tra articoli e autori, tra organizzazioni e autori e via discorrendo. In questo ambito, quindi, è fondamentale che vengano rilasciate anche le citazioni e i riferimenti bibliografici in maniera da poter ottenere una rete citazionale tra pubblicazioni su cui poter applicare gli strumenti messi a disposizione dal Semantic Web (es. Reasoner).
É stato presentato anche un manifesto[22] per rendere fruibili gratuitamente le liste dei riferimenti ma per la creazione di dataset aperti bisogna comunque fare alcune considerazioni: si dovrebbe cercare di armonizzare lo stile con cui vengono descritti i documenti referenziati e tener conto degli errori di battitura per tutti gli elementi che compongono una citazione (titolo, autori, data di pubblicazione, ecc.). In aggiunta l'esportazione delle entità bibliografiche non sempre è consentita per formati semantici e le ontologie che modellano questo dominio devono tener conto di fattori come la posizione della citazione nel testo, della lista dei riferimenti, e la semantica della citazione (es. citazione positiva o negativa).
Il seguente elaborato si propone di presentare il lavoro svolto su articoli scientifici codificati in XML al fine di creare un software che permetta di convertirli, attraverso delle opportune euristiche di conversione, in rete citazionale in RDF da poter interrogare tramite query SPARQL. I documenti, provenienti dall'archivio di PubMed Central, sono stati forniti in formato JATS, un particolare standard per le pubblicazioni scientifiche.
Nel primo capitolo verranno illustrate le caratteristiche del Semantic Publishing con particolare riferimento ai vocabolari e alle ontologie più utilizzate nonché alcuni Dataset che forniscono pubblicazioni scientifiche in Linked Open Data analizzando nel dettaglio quali sono le proprietà utilizzate per i riferimenti bibliografici.
Nel secondo capitolo verranno approfonditi il formato JATS con la sua evoluzione e storia e le regole di trasformazione dall'XML al linguaggio per il Semantic Web.

Nel terzo capitolo sarà presentato il prototipo CiNeX (Citation Network eXtractor) sia sotto l'aspetto dell'analisi sia per quanto riguarda l'implementazione.
Nel quarto capitolo saranno elencati i risultati ottenuti prima su una serie di file scelti per la Semantic Publishing Challenge da cui è stato possibile effettuare un'analisi qualitativa grazie al rilascio di alcuni golden standard, ovvero dei risultati attesi. In seguito è stata effettuata una valutazione quantitativa su un insieme di documenti più ampio per verificare se CiNeX riesce ad analizzare e restituire tutte le citazioni presenti in una pubblicazione, tenendo conto del fatto che potrebbero essere in formati diversi.
Infine le conclusioni riguarderanno gli sviluppi futuri e il riassunto delle problematiche riscontrate durante lo studio e lo sviluppo del tool.

Capitolo 1

Semantic Publishing

In questo capitolo verrà illustrato il background scientifico relativo al Semantic Publishing e alle ontologie di questo ambito. Inoltre verranno presentati e analizzati sotto l'aspetto dell'annotazione dei riferimenti bibliografici alcuni dei principali Linked Open Data Dataset contenenti articoli scientifici.

1.1 Pubblicazioni scientifiche e Semantic Web

Il Semantic Publishing è la possibilità di migliorare la comunicazione scientifica con l'utilizzo delle tecnologie semantiche. Shotton in [1] lo definisce come qualsiasi cosa che accresca il significato di un articolo di una rivista, faciliti la sua scoperta automatica, fornisca accesso ai dati all'interno dell'articolo in maniera processabile, o faciliti l'integrazione tra dati e pubblicazioni. Tra le altre cose, riguarda l'arricchimento degli articoli con metadati appropriati che ne permettano l'analisi automatica. L'utilizzo degli strumenti semantici aumenta il valore intrinseco della pubblicazione influenzando positivamente l'estrazione di informazioni e conoscenza.
In [5] viene illustrata la difficoltà del coniugare, nella stesura di un testo scientifico, la facilità di lettura per gli esseri umani e la comprensibilità per gli elaboratori in quanto hanno un diverso modo di comunicare. Ci sono stati diversi approcci a questo problema, data anche l'immensa mole di dati gene-

rati dagli scienziati negli ultimi anni. La sfida reale è quella di permettere a coloro che producono materiale scientifico di non nascondere la conoscenza in testi talvolta scritti male spendendo energie nel cercare di capire cosa l'autore ha realmente scoperto. Gli autori hanno dimostrato, comparando i pregi e i difetti delle tecnologie esistenti per la modellazione di articoli (XML, PDF, RDF, Testo, XHTML) come non esista ancora un formato che riesca a risolvere tutti i problemi per tutti gli scopi prefissati (Memorizzazione, Attività umana e attività computerizzata).

Esistono diversi vantaggi nell'adozione di tecnologie semantiche: per gli editori, se il manoscritto è online non c'è più bisogno dei tradizionali scritti stampati; se la rivista è totalmente online non esiste la necessità di una distribuzione fisica; il peer reviewing è accelerato perchè non si verificano i ritardi relativi alle poste mentre la produzione elettronica abbatte i costi ed è più veloce. Inoltre permette alternative commerciali alla semplice sottoscrizione ad una rivista con modalità come la pay-per-view, riviste virtuali e fornisce l'opportunità di una ricerca più accurata. Per i lettori c'è un più facile accesso alle riviste online rivoluzionando le abitudini di lettura che non devono per forza essere legate agli orari di apertura delle biblioteche o delle librerie.

Allo stesso modo esistono degli svantaggi e tra questi i costi. Gli editori necessitano di una componente ICT che sia sempre aggiornata sulle ultime tecnologie. Per i lettori, si trovano di fronte ad una vastità di riviste mai vista prima mentre per le librerie, il cui ruolo è ormai stato bypassato dalle risorse online, devono trovare un modo per reinventarsi nel mondo dell'era digitale.

Attualmente siamo ben lontani da un'armonizzazione delle tecnologie e da un'accettazione vasta della digitalizzazione delle riviste scientifiche anche perché gli editori non adottano standard condivisi, utilizzando invece XML e DTD (Document Type Definition) perlopiù proprietari costringendo la creazione di tool per mapparli e convertirli da uno all'altro. D'altro canto standard come RDF e OWL 2 si stanno affermando per permettere alle macchine

di integrare informazioni web-based e di interrogare i metadati ed è molto importante che gli editori adottino questi standard al fine di effettuare inferenze sull'intera mole di pubblicazioni nelle riviste, nei libri e nei proceedings delle conferenze.
Peroni in [4] ha individuato 8 aree di ricerca in questo dominio:

1. lo *sviluppo di tecnologie di markup* che facilitino la creazione di documenti complessi e semantici così da poter avere una descrizione semantica formale della loro struttura (capitoli, introduzione, paragrafi) così come il loro contenuto;

2. lo *sviluppo di modelli semantici* (ontologie, vocabolari) che rispettino i requisiti dell'editoria scientifica;

3. lo *sviluppo di strumenti di visualizzazione e documentazione* che permettano alle ontologie di essere facilmente intese dagli utenti;

4. lo *sviluppo di strumenti di annotazione* che permettano a questi modelli di essere usati dagli utenti finali (editori, autori) per inserire asserzioni semantiche rilevanti;

5. lo *sviluppo di nuovi algoritmi* che possano creare vantaggi a questo nuovo livello semantico di annotazioni, per esempio nella ricerca di grandi dataset di documenti online;

6. lo *sviluppo di nuovi modelli di business* che organizzino i processi editoriali per la creazione e l'uso delle asserzioni semantiche;

7. lo *studio e realizzazione di stime* che provino i benefici e/o gli svantaggi del Semantic Publishing sia per gli autori che per gli editori;

8. l'*organizzazione di eventi* come conferenze, workshop, progetti per pubblicizzare e promuovere i principi del Semantic Publishing.

1.2 Vocabolari e ontologie per l'editoria

In questa sezione verranno illustrate le principali ontologie e i vocabolari utilizzati per il Semantic Publishing.

1.2.1 Dublin Core

Nati a seguito di una conferenza tenuta nel 1995 negli USA a Dublin, Ohio, le versioni di Dublin Core (DC) Metadata Elements [8] e il DC Metadata Terms [9] sono i vocabolari più usati per descrivere e catalogare risorse. Nonostante sia molto utile per la creazione di metadati, la più grande limitazione di DC consta nel fatto che i suoi termini sono troppo generici. Per esempio, DC Terms identifica una risorsa bibliografica ma non un articolo di rivista, un identificatore ma non un ISSN e così via.

1.2.2 PRISM

Il Publishing Requirements for Industry Standard Metadata (PRISM) [10] è una specifica che definisce un insieme di termini per descrivere lavori editoriali e oggi coinvolge alcuni tra i maggiori editori con le relative aziende. I termini in PRISM possono essere espressi sia in XML, secondo uno specifico DTD, sia in RDF. PRISM ha un insieme di termini più ricco di DC ma la sua limitazione sta nella struttura piatta, in quanto mancante di gerarchie. Per esempio in PRISM manca il concetto di volume come una classe distinta dalle altre facente parte di una gerarchia ben precisa (Articolo, Issue, Volume, Rivista).

1.2.3 BIBO

La Bibliographic Ontology (BIBO) [11] è un'ontologia in OWL che permette di descrivere documenti (*bibo:Document* è la classe principale del modello) per la pubblicazione nel Semantic Web. Include sia DC che PRISM e aggiunge ulteriori classi e proprietà per meglio descrivere il dominio editoria-

le. L'unica pecca di questa ontologia è che non rispetta le specifiche di OWL 2 DL il che limita l'applicabilità di reasoner e altri tool.

1.2.4 FRBR

Il Functional Requirements for Bibliographic Record (FRBR) [12] è un modello generale, proposto dall'International Federation of Library Association (IFLA), per descrivere documenti e la loro evoluzione. Funziona sia con risorse fisiche sia digitali ed è molto flessibile e potente. Uno dei più importanti aspetti di FRBR è il fatto che non è associato ad un particolare tipo di metadata schema o implementazione.
FRBR descrive tutti i documenti da 4 differenti ma correlati punti di vista ciascuno dei quali è un FRBR *Endeavour*. Questi sono:

- **Opera** (work), una specifica creazione intellettuale;
- **Espressione** (expression), una realizzazione dell'opera;
- **Manifestazione** (manifestation), la materializzazione di un'espressione;
- **Unità** (item), un singolo esemplare di una manifestazione.

Un esempio che si può fare per capire meglio questi concetti riguardano l'opera *The Body in the Library* di Agatha Christie, l'espressione è, ad esempio, la traduzione italiana *C'è un cadavere in biblioteca* di Alberto Tedeschi, una manifestazione è la traduzione dell'opera nella collana Oscar gialli della casa editrice Mondadori del 1985 e l'unità è ad esempio la copia della manifestazione in una biblioteca (si veda inoltre l'esempio nella figura 1.1). Nonostante FRBR accresca l'espressività attraverso i suoi diversi livelli pecca nella mancanza di termini in linguaggio adatto alle pubblicazioni (es. rivista scientifica, capitolo di un libro, ecc.) in quanto utilizza dei termini più astratti come *work* o *expression*. L'ultima versione è stata sviluppata nel 2010 da Paolo Ciccarese e Silvio Peroni in OWL 2 DL[1].

[1]http://purl.org/spar/frbr

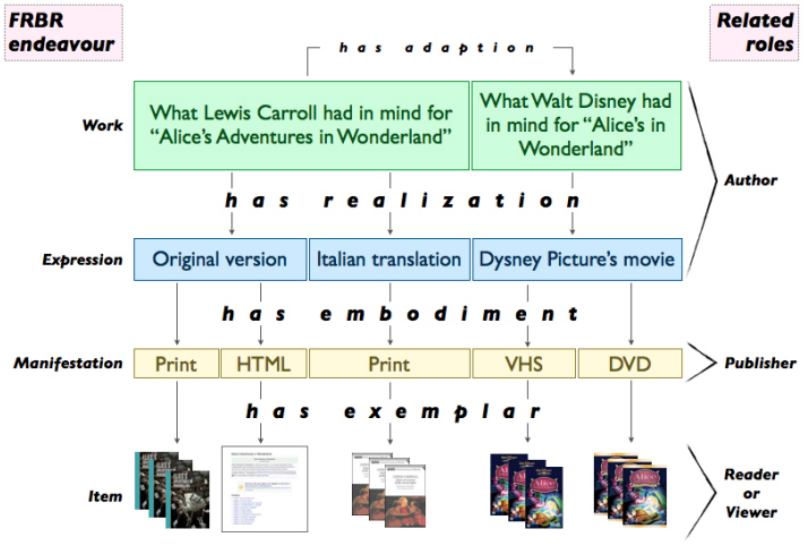

Figura 1.1: Diagramma dei livelli di FRBR con gli specifici ruoli per ciascuno di essi. In [4], p. 18

1.2.5 SWAN Citations Ontology

Un ulteriore modello utilizzato per definire risorse bibliografiche è il Citation Ontology incluso nelle ontologie SWAN (Semantic Web Applications in Neuroscience) [13]. La super classe di questa ontologia è la classe *Citation* di cui tutte le altre risorse (*Web Article, Book* etc.) sono sotto classi. Al contrario di BIBO questa ontologia è *compliant* con il OWL 2 DL.

1.2.6 SKOS

Il Simple Knowledge Organization System (SKOS) [14] è un modello RDFS che supporta l'uso di thesauri, sistemi di classificazione, tassonomie all'interno del framework del Semantic Web. É strutturato secondo tre moduli: SKOS Core, SKOS Mapping e SKOS Extensions.

1.2.7 SWRC

Il Semantic Web for Reasearch Communities (SWRC) [19] è un'ontologia utilizzata per fornire informazioni alle comunità di ricerca, tra cui organizzazioni, persone e le relazioni che vi intercorrono. Servizi come OntoWare[2] utilizzano SWRC per modellare le proprie risorse.

1.2.8 SPAR

SPAR (Semantic Publishing and Referencing Ontologies) è un insieme di ontologie per la creazione di RDF su tutto ciò che riguarda il semantic publishing. Le ontologie che formano SPAR sono descritte nella figura 1.1 ovvero nel diagramma a fiore creato da Benjamin o'Steen. Ciascuna di esse è codificata in OWL 2.0. Insieme vanno oltre la semplice descrizione di risorse bibliografiche perché utilizzano metodologie che permettono di gestire metadati RDF su citazioni, parti di documenti e vari aspetti del processo di pubblicazione delle riviste scientifiche.

CiTO

La Citation Typing Ontology[3] (CiTO) è un'ontologia che permette la caratterizzazione delle citazioni che possono essere dirette ed esplicite, indirette o implicite. Questa ontologia contiene la proprietà `cito:cites` e la sua inversa `cito:isCitedBy` che dalla versione 2.0 sono state integrate con ulteriori sottoproprietà dell'ontologia SWAN.

FaBiO

La FRBR-aligned Bibliograpgic Ontology è un'ontologia per la descrizione semantica delle entità che sono pubblicate o potenzialmente pubblicabili.

[2] http://ontoware.org/
[3] http://purl.org/spar/cito

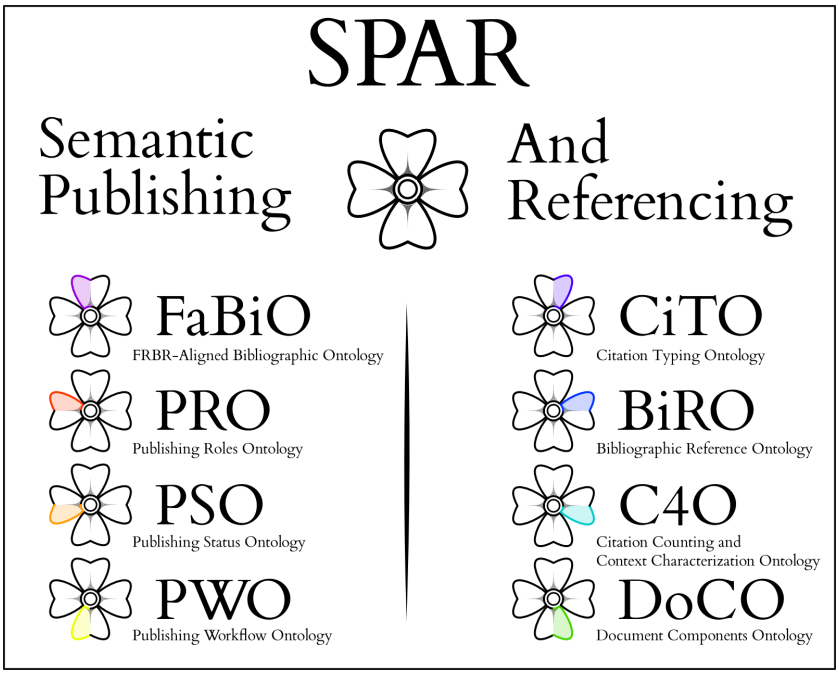

Figura 1.2: Il diagramma, creato da Benjamin o'Steen, che mostra le ontologie di SPAR

PRO

La Publishing Roles Ontology è un'ontologia per la caratterizzazione dei ruoli di agenti (autori, editori, ecc.) del processo di pubblicazione.

BiRO

La Bibliographic Reference Ontology è un'ontologia strutturata secondo il modello dati FRBR per la definizione di record bibliografici (come sottoclassi di *frbr:Work*), di riferimenti bibliografici (come sottoclasse di *frbr:Expression*) e la loro compilazione in collezioni e liste.

C40

La Citation Counting and Context Characterization Ontology è un'ontologia che permette la caratterizzazione di citazioni bibliografiche in base al

loro numero e al loro contesto.

PSO

La Publishing Status Ontology è un'ontologia per la descrizione dello stato di pubblicazione di un documento.

PWO

La Publishing Workflow Ontology è un'ontologia per la descrizione del flusso di lavoro associato con la pubblicazione di un documento o di altre tipologie di pubblicazioni (es. in fase di review, pubblicato sul Web).

DoCO

La Document Components Ontology è un'ontologia per la caratterizzazione delle varie parti, sia strutturali (es. linea, paragrafo, sezione) che retoriche (es. introduzione, discussione, riconoscimenti), le quali compongono un documento.

1.3 LOD su articoli scientifici

Con il seguente lavoro si intende dimostrare come sia possibile costruire una rete citazionale in linguaggio semantico a partire da documenti XML JATS. Un approccio simile è stato utilizzato da Castro, McLaughin e Garcia in Biotea[16] processando dei documenti di PubMed Central fornendo un insieme di file RDF arricchiti con annotazioni semantiche, un Web Service per l'interrogazione del dataset, un endpoint SPARQL contenente un sottoinsieme dei file RDF da poter interrogare e uno strumento di visualizzazione grafica per la ricerca di termini biologici (es. proteine, enzimi). Il dataset comprende circa 270000 articoli distribuiti su 2401 journal. Peculiarità di Biotea è l'annotazione del testo con l'individuazione dei termini grazie a strumenti di text-mining come Whatizit e NCBO Annotator. Attualmente

dalle API messe a disposizione[4] è possibile accedere a 20000 documenti, con 105778 termini, 550657 topic e 18 vocabolari. Tra le ontologie utilizzate (si veda la figura 1.4) spiccano per l'annotazione dei metadati DoCO, DCMI, FOAF e BIBO. Quest'ultima è stata utilizzata in larga parte per la lista dei riferimenti bibliografici utilizzando la proprietà `bibo:cites` e la sua inversa `bibo:isCitedBy` ma, come abbiamo visto in precedenza, presenta il difetto di non essere compatibile con OWL 2 DL e, di conseguenza, non è possibile applicare dei reasoner e altri tool specifici di questo linguaggio. Uno dei pri-

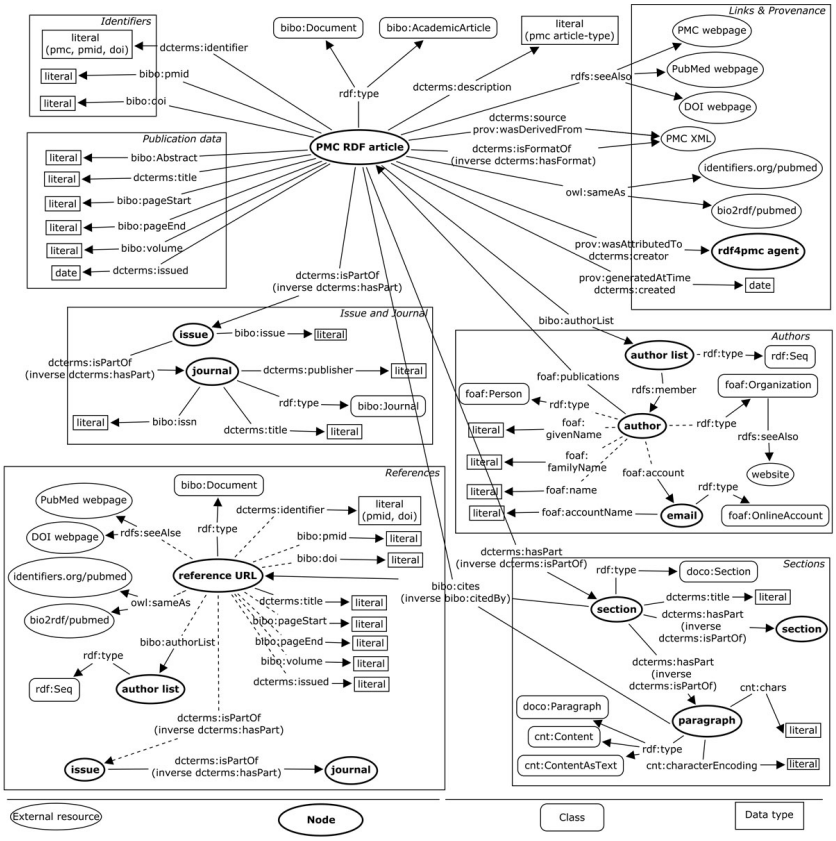

Figura 1.3: Diagramma RDF di Biotea[16]

mi dataset per il Semantic Publishing è stato l'NPG Linked Data Platform[5]

[4]http://biotea.idiginfo.org/api

[5]http://www.nature.com/developers/documentation/linked-data-platform/

che include articoli pubblicati dalla rivista Nature dal 1845 e racchiude circa 400 milioni di triple, strutturate secondo le ontologie Dublin Core, FOAF, PRISM e BIBO. Inoltre il Nature Publishing Group ha creato un proprio vocabolario, NPG appunto, con cui ha annotato diverse informazioni come il titolo, la data di pubblicazione di un articolo, il DOI e i riferimenti bibliografici con le proprietà `npg:hasCitation` e `npg:citeNum`.
Sul sito sono implementati due endopoint:

- */query* per uso interattivo
- */sparql* per uso remoto

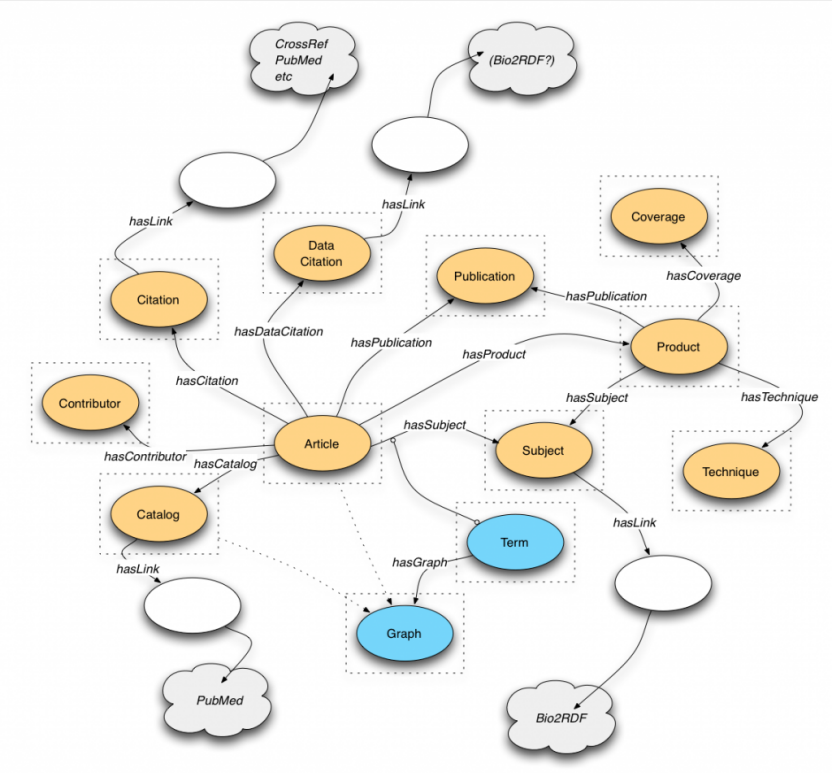

Figura 1.4: Diagramma RDF di NPG Linked Data Platform

Una caratteristica interessante di questo dataset è la sua connessione con servizi esterni, come CrossRef[6], per la gestione delle citazioni.

Un repository di citazioni bibliografiche è l'Open Citations Corpus[7], finanziato dal JISC[8] comprendente oltre 40 milioni di riferimenti da una lista di 600000 articoli dell'Open Access Subset di PubMed Central. Tutte le citazioni sono codificate in Linked Open Data utilizzando le ontologie SPAR. Per le citazioni utilizza le proprietà `cito:cites` e `frbr:part`. Benefici del raggruppare così tanti articoli in un unico corpus sono , ad esempio, poter ottenere informazioni sulle relazioni, tra articolo e articolo, tra articolo e database, tra gli autori, tra le istituzioni, tra le fonti di finanziamento, ecc. Inoltre possono essere sviluppati servizi di analisi per la ricerca e la scoperta di trend, o la visualizzazione su timeline. Alcuni di questi strumenti sono già stati implementati sul sito e sono tuttora in fase di espansione. Infine l'OCC fornisce il supporto per la correzione per quei riferimenti che sono citati più volte analizzando il contenuto del testo, ad esempio del titolo o dell'autore, armonizzandolo con quello delle altre citazioni[17].
Sul sito è possibile interagire con i contenuti in diversi modi: attraverso i journal, gli articoli, un motore di ricerca per identificatore o frase, un endpoint SPARQL. É inoltre possibile effettuare il download dell'intero dataset in diversi formati tra cui BibJSON (Raw, Sanitized e Unified) e N-Quads RDF.
Caratteristica di OCC è la possibilità di visualizzare la rete citazionale attraverso un apposito diagramma (vedi figura 1.5).

Per quanto riguarda l'informatica ci sono diversi dataset disponibili, tra questi ricordiamo DBLP++[9] che utilizza l'ontologia SWRC e ACM[10] conte-

[6]http://www.crossref.org
[7]http://opencitations.net
[8]http://www.jisc.ac.uk/about
[9]http://dblp.l3s.de/dblp++.php
[10]http://acm.rkbexplorer.com/

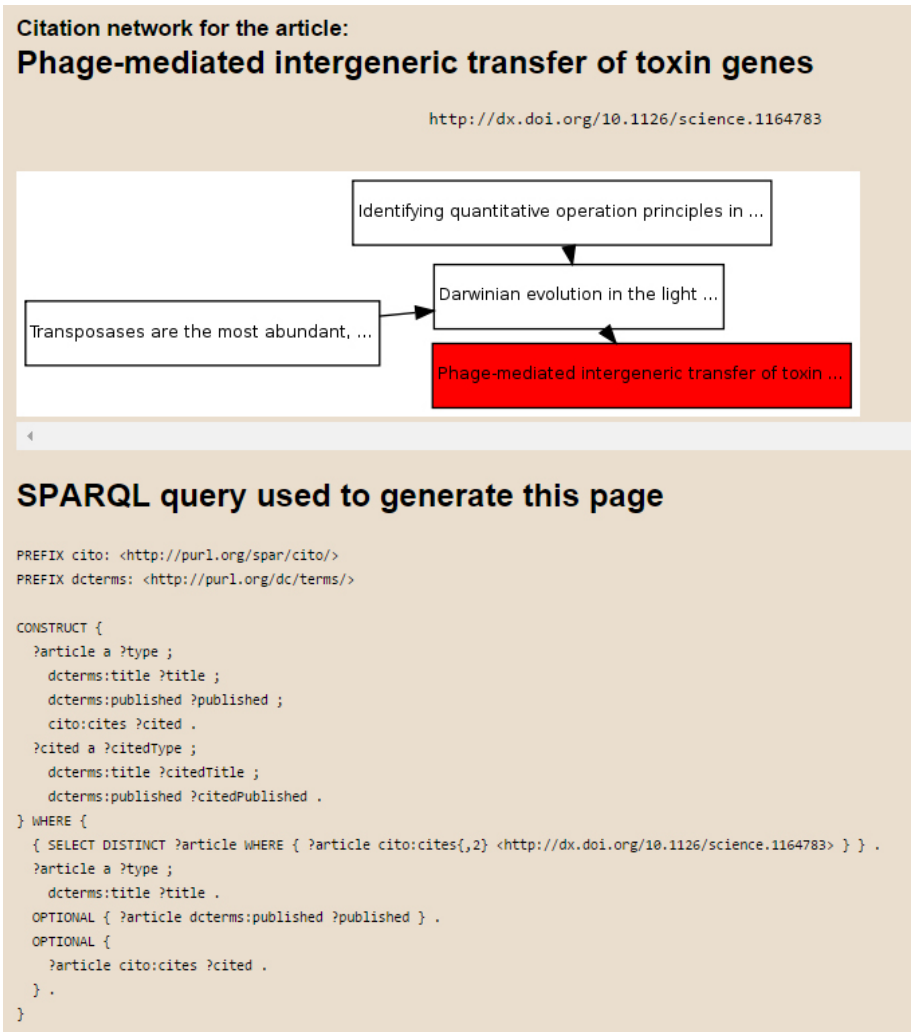

Figura 1.5: Esempio di visualizzazione di citazioni nell'Open Citations Corpus

nente circa 12 milioni di triple ma che non viene aggiornato dal 2006. Da una prima analisi delle pubblicazioni messe a disposizione non si tiene conto dei riferimenti bibliografici (vedi la figura 1.6).

Un ulteriore dataset sempre riguardante l'informatica è l'RKB Explorer dell'Assosiation for Computing Macinery (ACM)[11] che offre oltre 12 milioni

[11]http://acm.rkbexplorer.com

Property	Value
dcterms:bibliographicCitation	<http://dblp.uni-trier.de/rec/bibtex/series/acvpr/JillelaR13>
dc:creator	<http://dblp.l3s.de/d2r/resource/authors/Arun_Abraham_Ross>
dc:creator	<http://dblp.l3s.de/d2r/resource/authors/Raghavender_R._Jillela>
foaf:homepage	<http://dx.doi.org/10.1007%2F978-1-4471-4402-1%5F13>
foaf:homepage	<http://dx.doi.org/10.1007/978-1-4471-4402-1_13>
dc:identifier	DBLP series/acvpr/JillelaR13 (xsd:string)
dc:identifier	DOI 10.1007%2F978-1-4471-4402-1%5F13 (xsd:string)
dcterms:issued	2013 (xsd:gYear)
rdfs:label	Methods for Iris Segmentation. (xsd:string)
foaf:maker	<http://dblp.l3s.de/d2r/resource/authors/Arun_Abraham_Ross>
foaf:maker	<http://dblp.l3s.de/d2r/resource/authors/Raghavender_R._Jillela>
swrc:pages	239-279 (xsd:string)
dcterms:partOf	<http://dblp.l3s.de/d2r/resource/publications/series/acvpr/978-1-4471-4402-1>
owl:sameAs	<http://bibsonomy.org/uri/bibtexkey/series/acvpr/JillelaR13/dblp>
owl:sameAs	<http://dblp.rkbexplorer.com/id/series/acvpr/JillelaR13>
rdfs:seeAlso	<http://dblp.uni-trier.de/db/series/acvpr/iris2013.html#JillelaR13>
rdfs:seeAlso	<http://dx.doi.org/10.1007/978-1-4471-4402-1_13>
swrc:series	<http://dblp.l3s.de/d2r/resource/collections/acvpr>
dc:title	Methods for Iris Segmentation. (xsd:string)
dc:type	<http://purl.org/dc/dcmitype/Text>
rdf:type	swrc:InCollection
rdf:type	foaf:Document

Figura 1.6: Esempio di pubblicazione nel dataset DBLP

di triple utilizzando l'ontologia AKT. Sul sito è possibile effettuare ricerche cercando parole chiavi / letterali o gli URI e inoltre fornisce un Coreference Resolution Service per la gestione degli URI duplicati.
Per le citazioni il dataset utilizza la proprietà `akt:cites-publication-reference`.

Un dataset contenente informazioni su workshop e conferenze sul Semantic Web è il Semantic Web Dog Food[12] conosciuto anche come Semantic Web Conference Corpus con circa 5000 articoli descritti secondo le ontologie SWRC, SWC (una ontologia propria del dataset) e Dublin Core. Anche in questo caso non vengono prese in considerazione i riferimenti bibliografici e le citazioni (vedi la figura 1.8).

Infine il Semantic Web Journal dataset[13] è un contenitore di pubblicazioni dove si possono trovare riferimenti bibliografici a tutti gli articoli dell'omonima rivista. Presenta oltre 20000 triple con circa 430 articoli utilizzando come ontologie Dublin Core, BIBO e un vocabolario proprio intitolato SWJ.

[12] http://data.semanticweb.org/
[13] http://semantic-web-journal.com:3030/

Generation as method for explorative learning in computer science education

Showing combined information from the following coreferenced URIs ...
http://acm.rkbexplorer.com/id/1008019 [View this URI only]

The CRS managing coreference data for acm.rkbexplorer.com knows of 1 additional equivalent URIs in other repositories.
You can also view the global equivalence closure across all repositories.

Results in repository acm.rkbexplorer.com...

Subject	Property	Object/Value
Generation as method for explorative learning in computer science education	akt:addresses-generic-area-of-interest	K.3.2. Computer and Information Science Education
Generation as method for explorative learning in computer science education	akt:addresses-generic-area-of-interest	K.3.2. Computer and Information Science Education
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Visualizing principles of abstract machines by generating interactive animations
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Visualizing principles of abstract machines by generating interactive animations
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Levels of exploration
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Levels of exploration
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	id:574901
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	id:574901
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Animation der semantischen Analyse
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Animation der semantischen Analyse
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Animation of the Generation and Computation of Finite Automata for Learning Software
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Animation of the Generation and Computation of Finite Automata for Learning Software
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Visual Exploration of Generation Algorithms for Finite Automata on the Web
Generation as method for explorative learning in computer science education	akt:cites-publication-reference	Visual Exploration of Generation Algorithms for Finite Automata on the Web
Generation as method for explorative learning in computer science education	akt:has-author	Andreas Kerren
Generation as method for explorative learning in computer science education	akt:has-author	Andreas Kerren
Generation as method for explorative learning in computer science education	akt:has-date	2004-01-01

Figura 1.7: Esempio di pubblicazione nel dataset ACM

Description of http://data.semanticweb.org/conference/iswc/2002/poster-proceedings/paper-13:

property	hasValue	isValueOf
swc:isPartOf	<http://data.semanticweb.org/conference/iswc/2002/poster-proceedings>	-
dc:title	"Six Challenges for the Semantic Web"	-
dcterms:title	"Six Challenges for the Semantic Web"	-
<http://purl.org/ontology/bibo/authorList>	<http://data.semanticweb.org/conference/iswc/2002/poster-proceedings/paper-13/author_list>	-
swrc:author	<http://data.semanticweb.org/person/asuncion-gomez-perez>	-
swrc:author	<http://data.semanticweb.org/person/jesus-contreras>	-
swrc:author	<http://data.semanticweb.org/person/oscar-corcho>	-
swrc:author	<http://data.semanticweb.org/person/v-richard-benjamins>	-
swrc:category	"Poster Paper"	-
swrc:link_open_access	<http://iswc2002.semanticweb.org/posters/VR_Benjamins.pdf>	-
swrc:listAuthor	"V. Richard Benjamins,Jesus Contreras,Oscar Corcho and Asuncion Gomez-Perez"	-
swrc:url	<http://iswc2002.semanticweb.org/posters/VR_Benjamins.pdf>	-
swrc:year	"2002"	-
rdf:type	swrc:InProceedings	-
rdfs:label	"Six Challenges for the Semantic Web"	-
foaf:maker	<http://data.semanticweb.org/person/asuncion-gomez-perez>	-
foaf:maker	<http://data.semanticweb.org/person/jesus-contreras>	-
foaf:maker	<http://data.semanticweb.org/person/oscar-corcho>	-
foaf:maker	<http://data.semanticweb.org/person/v-richard-benjamins>	-
swc:hasPart	-	<http://data.semanticweb.org/conference/iswc/2002/poster-proceedings>
foaf:made	-	<http://data.semanticweb.org/person/asuncion-gomez-perez>
foaf:made	-	<http://data.semanticweb.org/person/jesus-contreras>
foaf:made	-	<http://data.semanticweb.org/person/oscar-corcho>
foaf:made	-	<http://data.semanticweb.org/person/v-richard-benjamins>

Figura 1.8: Esempio di pubblicazione nel dataset Semantic Web Conference Corpus

Anche questo dataset non presenta proprietà per la gestione delle citazioni.

Possiamo concludere questa rassegna notando come esista una certa eterogeneità nella gestione delle citazioni da parte dei dataset analizzati e, in alcuni casi, queste non vengono nemmeno prese in considerazione. Il dataset che,

a nostro avviso, mette al centro delle attenzioni i riferimenti bibliografici è l'Open Citations Corpus che nasce proprio con il proposito di rendere fruibili le citazioni e i collegamenti tra i diversi articoli scientifici. Nella tabella 1.1 abbiamo riassunto le proprietà utilizzate dai vari dataset per la gestione dei riferimenti bibliografici. Metà di questi non ne prevede l'annotazione mentre una di queste utilizza un vocabolario creato ad hoc per il dataset.

Dataset	**Proprietà**
Biotea	`bibo:cites`
NPGLDP	`npg:hasCitation`
OCC	`cito:cites`
DBLP++	
SWC	
SWJD	

Tabella 1.1: Tabella riassuntiva delle proprietà RDF utilizzate dai vari dataset per le citazioni

Capitolo 2

Estrazione di una rete citazionale da documenti XML

Come visto nel capitolo 1 i riferimenti bibliografici sono marcati in maniera eterogenea nei vari dataset analizzati e, in alcuni casi, non vengono neppure presi in considerazione o resi disponibili.
Scopo del nostro lavoro è quello di analizzare il database di PubMed Central[1], contenente oltre 3 milioni di articoli, e creare un tool che permetta di ottenere una rete citazionale a partire dalle informazioni contenute nei file marcati nel formato XML JATS.
Ciò che caratterizza il database di PubMed Central è l'eterogeneità dei documenti. In altre parole, vi sono articoli ben strutturati e altri dove sono presenti tag, come ad esempio `<mixed-citation>` o `<element-citation>`, che necessitano di un'analisi automatica più strutturata e complessa.
Inoltre bisogna tener conto delle informazioni relative al documento analizzato, come il DOI[2], il PubMedId, la rivista, il volume, la issue, la pagina iniziale e finale e tutto ciò che permetta di ottenere dei dati rilevanti sull'articolo. Questo ovviamente vale anche per la lista dei riferimenti bibliografici.

[1] http://www.ncbi.nlm.nih.gov/pmc/

[2] Il Digital Object Identifier è uno standard che consente l'identificazione duratura, all'interno di una rete digitale, di qualsiasi entità che sia oggetto di proprietà intellettuale e di associarvi i relativi dati di riferimento

Di seguito verrà introdotto il formato JATS e le regole fissate, attraverso un esempio, per la traduzione in RDF utilizzando l'ontologia SPAR.

2.1 Journal Article Tag Suite

Il Journal Article Tag Suite (JATS), pubblicato il 22 Agosto 2012 come ANSI/NISO Z39.96-2012, è il successore del National Library Medicine (NLM) DTD. È lo standard *de facto* per le pubblicazioni scientifiche largamente usato da diversi editori accademici nonchè per le pubblicazioni di PubMed Central. Definisce un vocabolario di elementi e attributi XML che descrivono contenuti e metadati di articoli di riviste, dove con articoli di riviste si intendono sia gli articoli di ricerca sia quelli non di ricerca. Quindi ricerche originali, review, lettere all'editore, editoriali e libri sono tutti definiti come articoli di riviste. Questo insieme di Tag contiene elementi che descrivono i contenuti narrativi di un articolo, i suoi componenti grafici e i metadati dell'intestazione.
JATS si basa su un precedente lavoro della National Library of Medicine e un accurato studio di DTD e schema. Nel 2003, il National Centre for Biotechnology Information[3] (NCBI), una divisione del National Institutes of Health (NIH) rilasciò le National Library of Medicine[4] (NLM) Archiving and Interchange Tag Suite, e due Document Type Definition (DTDs) per gli articoli di riviste: l'Archiving and Interchange DTD e il Journal Publishing DTD. Nel 2005, fu aggiunto nella versione 2.1 un terzo DTD, l'Article Authoring. La versione 3.0 fu rilasciata nel 2008.
Questo NLM DTD divenne uno standard *de facto* per le pubblicazioni scientifiche e per i futuri sviluppi fu incaricato un gruppo di lavoro del National Information Standards Organization[5] (NISO) che poco dopo rilasciò una versione provvisoria (la 0.4) dello standard, rinominato Journal Article Tag Suite (JATS), come un minore aggiornamento della versione 3.0 della NLM

[3] http://www.ncbi.nlm.nih.gov/
[4] http://www.nlm.nih.gov/
[5] http://www.niso.org/

Tag Suite che restava comunque totalmente retrocompatibile con la versione NLM 3.0.
Nel 2012 è stata pubblicata la versione 1.0 di JATS raggiungibile all'url *http://www.niso.org/apps/group_public/document.php?document_id=8975*.
Come l'NLM DTD, JATS contiene tre set di tag, il Journal Archiving and Interchange Tag Set, il Journal Publishing Tag Set e l'Article Authoring Tag Set, destinati a differenti scopi. Il Journal Publishing Tag set a un insieme di tag moderatamente proscrittivo ottimizzato per regolarizzare e controllare la sequenza dei contenuti XML, quindi non accetta qualunque adattamento fornito da ogni particolare editore.

2.2 Da JATS a RDF

In questa sezione verrà discusso come si è deciso di tradurre in RDF i tag XML di JATS.
Come punto di partenza si faccia riferimento a [6]. Si è deciso inoltre di non tradurre in RDF l'intero contenuto della pubblicazione in JATS ma solo alcune parti necessarie al raggiungimento della creazione della rete citazionale. Tra i metadati supportati ci sono il nome dell'editore, l'ISSN, il volume, la issue, la pagina iniziale e quella finale, il DOI, il PubMedCentralId e il PubMedId, la data di pubblicazione, il titolo e gli autori.
Tra le ontologie analizzate si è deciso di utilizzare SPAR, in quanto OWL 2 DL compliant e comprendente ontologie come FRBR, FOAF, PRISM, Dublin Core. In sostanza SPAR sembra essere l'ontologia più completa attualmente nella comunità del Semantic Publishing.

Innanzitutto per i local name degli URL di una qualsiasi classe si sono utilizzate le lettere minuscole separate dal carattere -. Un esempio di local name per il work di un articolo è il seguente: *Hovy, Eduard, Roberto Navigli, and Simone Paolo Ponzetto. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence, Volume 194,*

January 2013, Pages 2-27 si traduce in `:hovy-2013-collaboratively-built-semi-structured-work`.

Prendiamo ad esempio il file `Am_J_Epidemiol_2008_Dec_15_168(12)_1343-1352.nxml` e vediamo come dalla notazione XML si è passati al linguaggio RDF/Turtle.

Partiamo dal tag `<journal-meta>`

```
<journal-meta>
        <journal-id journal-id-type="nlm-ta">Am J Epidemiol</journal-id>
        <journal-id journal-id-type="hwp">amjepid</journal-id>
        <journal-id journal-id-type="publisher-id">aje</journal-id>
        <journal-title>American Journal of Epidemiology</journal-title>
        <issn pub-type="ppub">0002-9262</issn>
        <issn pub-type="epub">1476-6256</issn>
        <publisher>
                <publisher-name>Oxford University Press</publisher-name>
        </publisher>
</journal-meta>
```

Si traduce in:

```
:american-journal-of-epidemiology-volume168-issue12 a fabio:JournalIssue ;
prism:issueIdentifier "12" ;
frbr:partOf :american-journal-of-epidemiology-volume168 .

:american-journal-of-epidemiology-volume168 a fabio:JournalVolume ;
prism:volume "168" ;
frbr:partOf :american-journal-of-epidemiology .

:american-journal-of-epidemiology a fabio:Journal ;
dcterms:title "American Journal of Epidemiology" ;
prism:issn "0002-9262" ;
prism:eIssn "1476-6256" ;
frbr:realizer :oxford-university-press .

:oxford-university-press a foaf:Organization ;
foaf:name "Oxford University Press" ;
pro:holdsRoleInTime :oxford-university-press-publisher .
```

```
:oxford-university-press-publisher a pro:RoleInTime ;
pro:withRole pro:publisher ;
pro:relatesToDocument :american-journal-of-epidemiology .
```

Per quanto riguarda il tag `<article-meta>` :

```
<article-meta>
        <article-id pub-id-type="pmid">18974084</article-id>
        <article-id pub-id-type="pmc">2638553</article-id>
        <article-id pub-id-type="doi">10.1093/aje/kwn259</article-id>
        ...
        <title-group>
                <article-title>Estimating Influenza Vaccine Efficacy
From Challenge and Community-based Study Data</article-title>
        </title-group>
        <contrib-group>
                <contrib contrib-type="author">
                        <name>
                                <surname>Basta</surname>
                                <given-names>Nicole E.</given-names>
                        </name>
                 </contrib>
                        ...
                        <pub-date pub-type="ppub">
                                <day>15</day>
                                <month>12</month>
                                <year>2008</year>
                        </pub-date>
                        <pub-date pub-type="epub">
                                <day>29</day>
                                <month>10</month>
                                <year>2008</year>
                        </pub-date>
                        <pub-date pub-type="pmc-release">
                                <day>29</day>
                                <month>10</month>
                                <year>2008</year>
                        </pub-date>
```

```
                    <volume>168</volume>
                    <issue>12</issue>
                    <fpage>1343</fpage>
                    <lpage>1352</lpage>
                    ...
                    ...
</article-meta>
```

è mappato come:

```
:basta-2008-estimating-influenza-vaccine-efficacy-work frbr:creator
:nicole-basta , :elisabeth-halloran , :laura-matrajt , :ira-longini .

:nicole-basta a foaf:Person ;

foaf:givenName "Nicole E." ;
foaf:familyName "Basta" ;

pro:holdsRoleInTime :nicole-basta-author .

:nicole-basta-author a pro:RoleInTime ;
pro:withRole pro:author ;
pro:relatesToDocument :basta-2008-estimating-influenza-vaccine-efficacy-work .

#Lo stesso per gli altri autori

:basta-2008-estimating-influenza-vaccine-efficacy-expression a
 fabio:JournalArticle ;

dcterms:title "Estimating Influenza Vaccine Efficacy From Challenge and
Community-based Study Data" ;
prism:doi "10.1093/aje/kwn259" ;
fabio:hasPubMedCentralId "2638553" ;
fabio:hasPubMedId "18974084" ;
frbr:partOf :american-journal-of-epidemiology-volume168-issue12 ;
frbr:embodiment
:basta-2008-estimating-influenza-vaccine-efficacy-manifestation-ppub ,
:basta-2008-estimating-influenza-vaccine-efficacy-manifestation-epub ,
:basta-2008-estimating-influenza-vaccine-efficacy-manifestation-pmc-release .
```

```
:basta-2008-estimating-influenza-vaccine-efficacy-manifestation-ppub
a fabio:PrintObject ;
prism:startingPage "1343" ;
prism:endingPage "1352" ;
prism:publicationDate "2008-12-15"^^xsd:date ;
frbr:producer :oxford-university-press .

:basta-2008-estimating-influenza-vaccine-efficacy-manifestation-epub
a fabio:DigitalManifestation ;
prism:startingPage "1343" ;
prism:endingPage "1352" ;
prism:publicationDate "2008-10-29"^^xsd:date ;
frbr:producer :oxford-university-press .

:basta-2008-estimating-influenza-vaccine-efficacy-manifestation-pmc-release
a fabio:Manifestation ;
prism:startingPage "1343" ;
prism:endingPage "1352" ;
prism:publicationDate "2008-10-29"^^xsd:date ;..
frbr:producer :oxford-university-press .
```

Il tag `<ref-list>` e i relativi riferimenti bibliografici vengono mappati come:

```
:basta-2008-estimating-influenza-vaccine-efficacy-expression
frbr:part
:basta-2008-estimating-influenza-vaccine-efficacy-reference-list .

:basta-2008-estimating-influenza-vaccine-efficacy-reference-list
a biro:ReferenceList ;
co:firstItem :basta-2008-estimating-influenza-vaccine-efficacy-reference-item-1 ;
co:item
:basta-2008-estimating-influenza-vaccine-efficacy-reference-item-2 ,
:basta-2008-estimating-influenza-vaccine-efficacy-reference-item-3 ,
:basta-2008-estimating-influenza-vaccine-efficacy-reference-item-4 ,
...
:basta-2008-estimating-influenza-vaccine-efficacy-reference-item-27 ;
co:lastItem
```

```
:basta-2008-estimating-influenza-vaccine-efficacy-reference-item-28 .

:basta-2008-estimating-influenza-vaccine-efficacy-reference-item-1
a co:ListItem ;
co:itemContent :basta-2008-estimating-influenza-vaccine-efficacy-reference-1 ;
co:index "1" ;
co:nextItem :basta-2008-estimating-influenza-vaccine-efficacy-reference-item-2 .

:basta-2008-estimating-influenza-vaccine-efficacy-reference-1
a biro:BibliographicReference ;
biro:references :halloran-1999-design-interpretation-vaccine-field-expression .
```

Poi per ciascun riferimento bisogna definire tutti i dati a disposizione dell'entità come fatto in precedenza quindi per il seguente riferimento bibliografico:

```
<ref id="bib1">
        <label>1.</label>
        <citation citation-type="journal">
                <person-group person-group-type="author">
                        <name>
                                <surname>Halloran</surname>
                                <given-names>ME</given-names>
                        </name>
                                ...
                                ...
                </person-group>
                <article-title>
                        Design and interpretation of vaccine field
                        studies
                </article-title>
                <source>Epidemiol Rev.</source>
                <year>1999</year>
                <volume>21</volume>
                <issue>1</issue>
                <fpage>73</fpage>
                <lpage>88</lpage>
                <pub-id pub-id-type="pmid">10520474</pub-id>
```

```
        </citation>
</ref>
```

Definiamo innanzitutto:

```
:halloran-1999-design-interpretation-vaccine-field-work cito:isCitedBy
:basta-2008-estimating-influenza-vaccine-efficacy-work

:basta-2008-estimating-influenza-vaccine-efficacy-work cito:cites
:halloran-1999-design-interpretation-vaccine-field-work
```

Poi ci sarà da definire:

```
:halloran-1999-design-interpretation-vaccine-field-work
:halloran-1999-design-interpretation-vaccine-field-expression
:halloran-1999-design-interpretation-vaccine-field-manifestation
:epidemiol-rev-volume21-issue1
:epidemiol-rev-volume21
:epidemiol-rev
```

Infine nella figura 3.1 mostriamo una vista parziale, quindi senza tutte le istanze, le proprietà e le classi descritte in precedenza, del risultato della trasformazione di un file JATS in RDF utilizzando le ontologie del Semantic Publishing. Notiamo come il documento preso in considerazione abbia diversi riferimenti bibliografici (annotati dalla proprietà `cito:cites`) e che l'autore Richard Appleton è creatore (`frbr:creator`) dell'articolo principale e di un articolo facente parte delle citazioni.
Il grafo è stato generato utilizzando il tool TopBraid Composer Maestro Edition[6].

[6]http://www.topquadrant.com/tools/ide-topbraid-composer-maestro-edition/

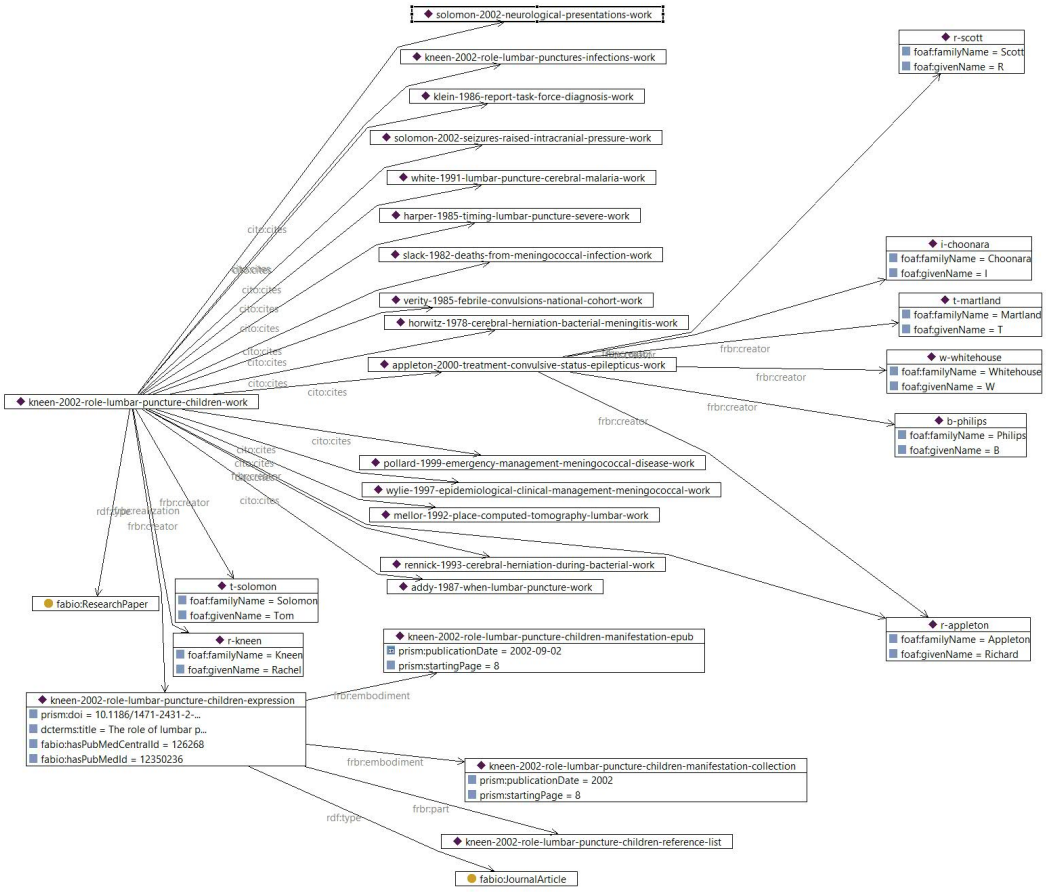

Figura 2.1: Una vista parziale della traduzione RDF di un file JATS

2.3 Citazioni

La parte centrale del lavoro, come detto in precedenza, consiste nel creare una rete citazionale. Per fare questo bisogna individuare i riferimenti bibliografici all'interno del file JATS e tradurli in RDF con le regole illustrate nella sezione precedente. Un esempio di citazione ben formata è il seguente:

```
<ref id="R32">
        <label>32</label>
        <element-citation publication-type="journal">
                <person-group person-group-type="author">
                        <name>
                                <surname>Iriarte</surname>
                                <given-names>FB</given-names>
                        </name>
                        <name>
                                <surname>Balogh</surname>
                                <given-names>B</given-names>
                        </name>
                        <name>
                                <surname>Momol</surname>
                                <given-names>MT</given-names>
                        </name>
                        <name>
                                <surname>Smith</surname>
                                <given-names>LM</given-names>
                        </name>
                        <name>
                                <surname>Wilson</surname>
                                <given-names>M</given-names>
                        </name>
                        <name>
                                <surname>Jones</surname>
                                <given-names>JB</given-names>
                        </name>
                </person-group>
                <article-title>Factors affecting survival of
bacteriophage on tomato leaf surfaces</article-title>
                <source>Appl Environ Microbiol</source>
```

```
                <year>2007</year>
                <volume>73</volume>
                <fpage>1704</fpage>
                <lpage>11</lpage>
                <pub-id pub-id-type="doi">10.1128/AEM.02118-06</pub-id>
                <pub-id pub-id-type="pmid">17259361</pub-id>
        </element-citation>
</ref>
```

Esso contiene tutte le informazioni utili per identificare correttamente la citazione come ad esempio i nomi degli autori, il titolo, il journal, l'anno, il volume e le pagine di inizio e fine, nonché il PubMedId e il DOI.
Di seguito invece riportiamo alcuni esempi di citazioni che necessitano di un'analisi più approfondita. Questo non vuol dire che i dati ivi contenuti vengano persi ma vengono salvati in una stringa di testo e annotati con la proprietà `dcterms:BibliographicCitation`[18].
Esempio 1:

```
<ref id="b4">
        <label>4</label>
        <element-citation publication-type="other">
                <collab>American Medical Association</collab>
                <comment>Report 2 of the Council on Scientific
 Affairs (A-04). Impact of drug formularies and therapeutic
 interchange on health outcomes. 2004. Available at
                <ext-link ext-link-type="uri" xlink:href="
http://www.ama-assn.org/ama/no-index/about-ama/13675.shtml">
http://www.ama-assn.org/ama/no-index/about-ama/13675.shtml
                </ext-link> (last accessed 29 April 2009)
        </comment>
        </element-citation>
</ref>
```

Esempio 2:

```
<ref id="B29">
        <mixed-citation publication-type="other">
                <collab>National Center for Biotechnology
Information Entrez Gene database</collab>
                <ext-link ext-link-type="uri" xlink:href=
"http://www.ncbi.nlm.nih.gov/gene">
```

```
http://www.ncbi.nlm.nih.gov/gene</ext-link>
        </mixed-citation>
</ref>
```

Esempio 3:

```
<ref id="B5">
        <citation citation-type="book">
                <person-group person-group-type="author">
                        <collab>IOM/NAS</collab>
                </person-group>
                <source>Modeling Community Containment for Pandemic
 Influenza</source>
                <year>2006</year>
                <publisher-name> Institute of Medicine of the
National Academies, The National Academies Press,
Washington, DC 20001</publisher-name>
                                        </citation>
</ref>
```

Esempio 4:

```
<ref id="bib1">
        <label>1.</label>
        <citation citation-type="journal">Chen, L. B.
        <year>1988</year>. Mitochondrial membrane potential in
 living cells. <source>Annu. Rev. Cell Biol.</source><volume>4
</volume>:<fpage>155</fpage>&#x2013;181.
<pub-id pub-id-type="pmid">3058159</pub-id></citation>
</ref>
```

Esempio 5:

```
<ref id="R39">
        <label>39</label>
        <mixed-citation publication-type="thesis">Balogh B.
Characterization and use of bacteriophages associated with citrus
bacterial pathogens for disease control. PhD thesis 2006. Univ. FL:
Gainesville.</mixed-citation>
</ref>
```

Esempio 6:

```
<ref id="R38">
        <label>38</label>
        <mixed-citation publication-type="confproc">Svircev AM,
Lehman SM, Kim W, Barszcz E, Schneider KE, Castle AJ. Control
of the fire blight pathogen with bacteriophages. In: Zeller W,
Ullrich C, Seeheim/Darmstadt eds. Proceedings of the 1st
International Symposium on Biological Control of Bacterial
Plant Diseases. Land- Forstwirtsch Germany: Mitt Biol Bundesanst,
2006; 408:259-61.</mixed-citation>
</ref>
```

Capitolo 3

Il prototipo CiNeX (Citation Network eXtractor)

Scopo del programma realizzato è quello di rendere fruibili sottoforma di linked data le informazioni contenute nei file XML descritti in JATS. Per fare questo il software prende in input uno o più file in XML e fornisce in output un file Turtle con la traduzione in RDF. In questo capitolo verrà illustrato il prototipo CiNeX (Citation Network eXtractor) sia per quanto concerne la logica applicativa sia la parte implementativa. Inoltre verranno approfonditi alcuni strumenti utilizzati per lo sviluppo, Jena e Fuseki, e verrà brevemente introdotto il linguaggio SPARQL per l'interrogazione di dataset in RDF.

3.1 La logica

Il software si occupa, in due fasi differenti, di due problemi. Il primo, riguarda l'individuazione dei tag XML all'interno dei documenti di input e il loro trasferimento su una struttura dati locale, poi, a partire da questa struttura, il tool identifica tramite le regole definite nel capitolo 2 le parti del documento e le traduce in RDF.
Per quanto riguarda il primo problema bisogna individuare 3 macro aree:

- `<journal-meta>`

- `<article-meta>`

- `<ref-list>`

Nella prima sono contenute tutte le informazioni relative al journal di riferimento dell'articolo, nella seconda i metadati dell'articolo stesso e nella terza la lista dei riferimenti bibliografici. Individuate queste aree si scende nel dettaglio con tutti i tag relativi e si popola la struttura dati locale.
La traduzione RDF deve avvenire necessariamente utilizzando delle specifiche regole. In questa fase sono stati riscontrati i seguenti problemi:

- *Nomi abbreviati e nomi per esteso*: solitamente nel tag `<article-meta>` i nomi degli autori vengono messi per esteso mentre nei riferimenti bibliografici i nomi vengono abbreviati con le iniziali. Ciò ha comportato un problema di individuazione univoca degli stessi perché non c'era omogeneità nell'abbreviazione (ad esempio venivano annotati sia il primo che il secondo nome, a volte con punti, a volte senza). Il problema è stato risolto in parte con un apposito algoritmo che ricerca l'esistenza dell'autore prima sul cognome e poi sulle iniziali del primo nome. Se vi è corrispondenza, controlla la lunghezza della stringa del nome più corto, e se questa è uguale a 1, sostituisce il nome con quello più lungo.

- *Omonimia*: non abbiamo trovato, per ora, una soluzione al problema legato all'omonimia tra autori

- *Errori di battitura / scrittura di Journal e Articoli*: un problema piuttosto comune è stato quello di dover far fronte ad articoli identici ma con nomi differenti. In questo caso basta controllare il DOI e verificare la corrispondenza oppure, nel caso di alcuni Journal, l'unica differenza era un punto alla fine del nome, quindi abbiamo deciso di eliminarlo manualmente nel tool.

3.2 Implementazione

Il software è stato realizzato con linguaggio Java, con Java Runtime Enviroment 7 e con IDE Netbeans 7.4[1].
Per il parsing dei documenti in XML/JATS è stata utilizzata la libreria SAX Parser messa a disposizione da Java mentre per le operazioni relative alla creazione dell'output in formato RDF si è scelto il framework Jena.
La soluzione adottata consta di due livelli. In un primo tempo viene effettuato il parse dei documenti XML inserendo i valori in una struttura dati locale mentre in un secondo momento da questa struttura venutasi a creare si modella il grafo RDF con Jena. Come mostrato nella figura 3.1 il processo è il seguente: il file XML di input viene processato dal tool che in output fornisce un file in RDF Turtle il quale verrà successivamente caricato su uno strumento del Web Semantico, Fuseki, che permette l'interrogazione del dataset creato tramite query SPARQL.

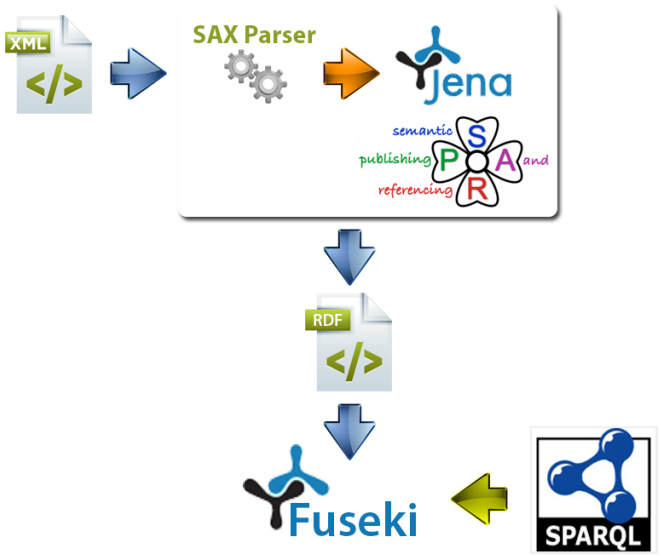

Figura 3.1: Schema delle fasi di utilizzo del tool

[1]https://netbeans.org/

La struttura dati utilizzata è abbastanza semplice e prevede l'utilizzo di alcune entità base:

- Paper: il documento che contiene i riferimenti bibliografici a cui è associato un journal specifico;
- Journal: la rivista che contiene i Paper dove è specificato un publisher;
- Reference: i riferimenti bibliografici contenuti nei Paper;
- Author: l'autore di un Paper o di una Reference.

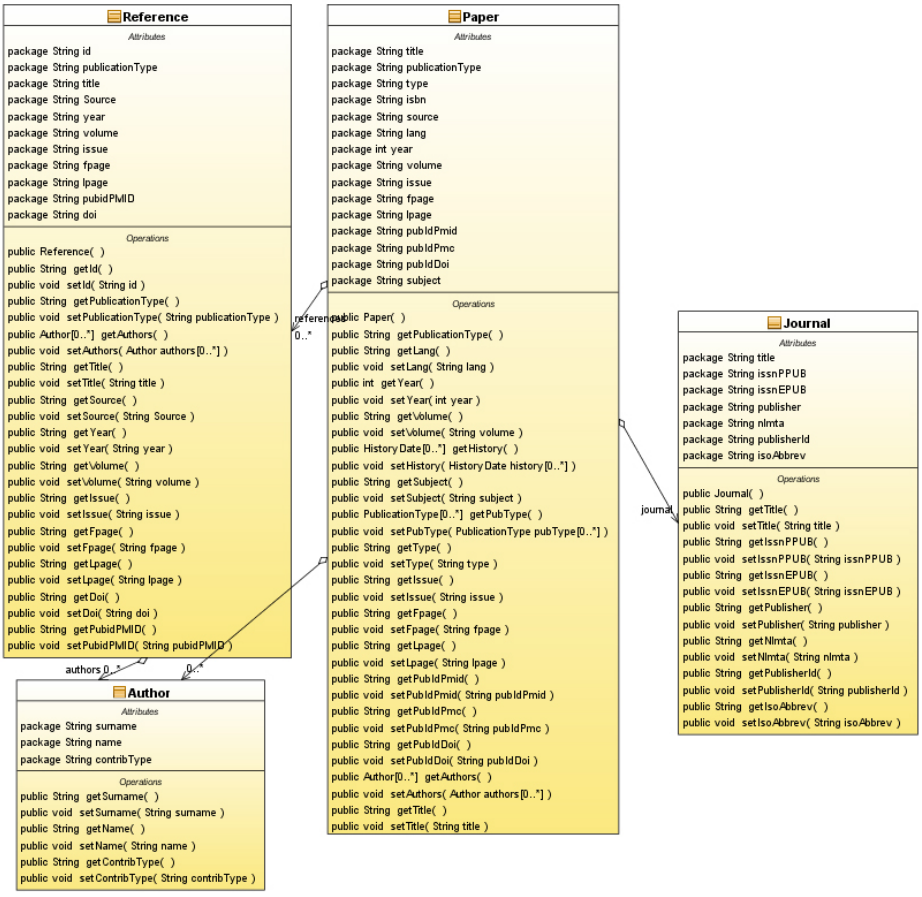

Figura 3.2: Diagramma delle Classi essenziale del software prodotto

3.2.1 Jena

Jena[2] è un framework open source per il Semantic Web che fornisce delle API per effettuare operazioni su grafi RDF. I grafi sono rappresentati come *model* i quali si vengono a creare da dati su file, database, URL o una combinazione di questi. Un modello può essere inoltre interrogato attraverso SPARQL e aggiornato attraverso SPARUL.
Jena fornisce supporto per OWL e possiede al suo interno diversi reasoner tra cui il Pellet, inoltre supporta la serializzazioni di grafi RDF in database relazionali, RDF/XML, Turtle, Notation 3. I tre concetti principali in Jena sono:

- *grafo*, la vista matematica di relazioni orientate tra nodi in una struttura connessa
- `Model`, una API Java per gli sviluppatori
- `Graph`, una libreria più semplice che ha lo scopo di estendere alcune funzionalità di Jena

Le informazioni RDF sono contenute in un grafo di nodi connessi di cui esistono due tipologie: URI e letterali. Questo denota rispettivamente, una risorsa sulla quale possiamo fare delle asserzioni e valori concreti che si presentano in queste. Un esempio è il seguente

```
:detroia foaf:name "Alessandro"
```

In questo caso `:detroia` è una risorsa che denota una persona e `Alessandro` è il valore della proprietà della risorsa. In Jena l'interfaccia `Resource` rappresenta le risorse mentre l'interfaccia `Literal` i letterali.
La classe usata per rappresentare le triple è `Statement` dove solo le risorse possono essere il soggetto di una tripla RDF mentre l'oggetto può essere sia una risorsa sia un letterale. Esistono tre metodi per l'estrazione di elementi da uno `Statement`:

[2] http://jena.apache.org/

- `getSubject()` che ritorna una `Resource`
- `getObject()` che ritorna un `RDFNode`
- `getPredicate()` che ritorna una `Property`

In Jena la classe `Property` è una sottoclasse di `Resource` che a sua volta una sottoclasse di `RDFNode`.

3.2.2 Fuseki

Fuseki è un'interfaccia HTTP ai dati RDF. Supporta SPARQL per l'aggiornamento e l'interrogazione dei dati ed è un sottoprogetto di Jena sviluppato in forma di servlet ma può comunque girare in maniera stand-alone. Nel nostro caso si è scelto di utilizzare la versione 1.0.1 stand-alone su Windows scaricandola dal sito ufficiale[3].
Per far girare Fuseki su una macchina Windows basta utilizzare il comando

```
java -jar fuseki-server.jar [parametri]
```

Dove in `[parametri]` si possono scegliere le varie configurazioni offerte dall'applicazione:

- `--mem` crea un dataset vuoto
- `--file=FILE` crea un dataset vuoto e poi carica `FILE`
- `--loc=DIR` crea un database TDB[4] esistente
- `--desc=assemblerFile` costruisce un dataset basato su una descrizione in assembler
- `--config=ConfigFile` costruisce uno o più endpoint basati su un file di configurazione

[3]http://jena.apache.org/download/index.cgi

[4]TDB è un componente di Jena utilizzato per l'immagazzinamento e l'interrogazione RDF

Il comando utilizzato per far partire Fuseki con un dataset vuoto è il seguente:

```
java -jar fuseki-server.jar --update --mem /ds
```

Fuseki gira sulla porta 3030 e vi si accede digitando l'URL `http://localhost:3030` ottenendo la schermata della figura 3.3.

Creato il dataset vuoto andremo a riempirlo con i dati generati accedendo

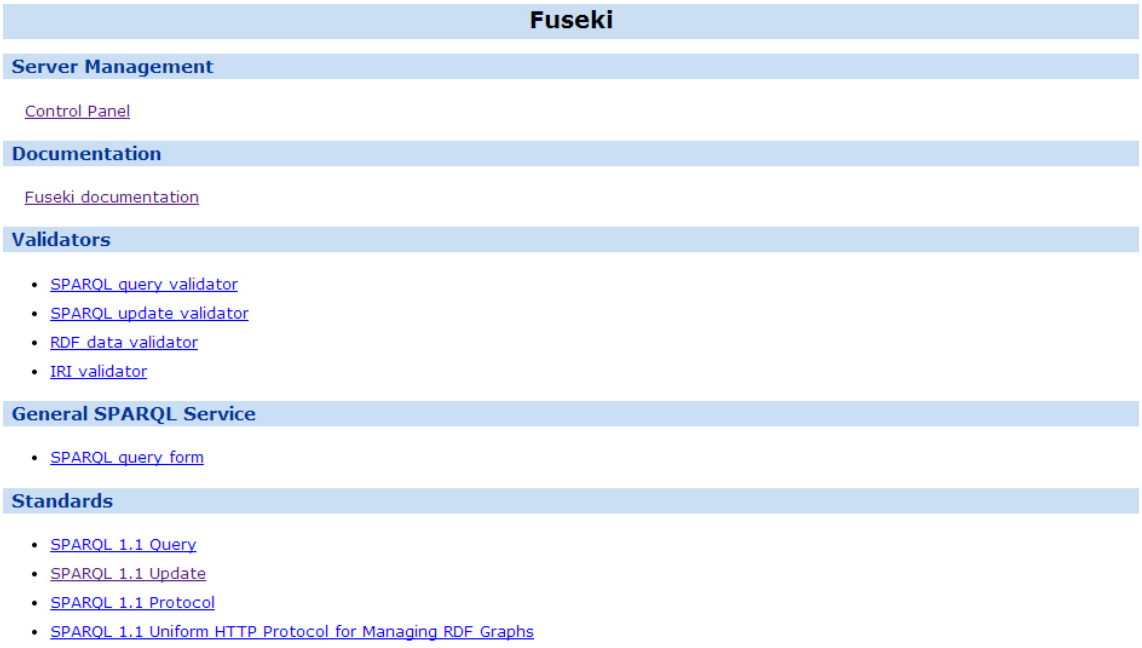

Figura 3.3: Schermata iniziale di Fuseki

al Control Panel. Qui potremo scegliere il dataset su cui andare ad effettuare le nostre operazioni (figura 3.4).

Fuseki mette a disposizione tre operazioni: Query[5], Update[6] e File upload grazie alle quali potremo rispettivamente effettuare delle interrogazioni sul dataset, modificarlo o caricare dei file in turtle.

[5]http://www.w3.org/TR/sparql11-query/

[6]http://www.w3.org/TR/sparql11-update/

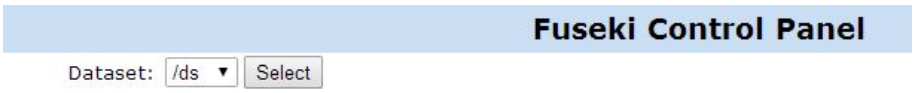

Figura 3.4: Schermata di scelta del dataset

Figura 3.5: Schermata per la manipolazione del dataset

3.3 SPARQL

L'acronimo SPARQL sta per *SPARQL Protocol and RDF Query Language*[7] ed è un linguaggio per l'interrogazione dell'RDF, è stato accolto come ultimo tassello per l'edificazione del Web Semantico ed è un W3C Recommendation dal 15 Gennaio 2008. Permette di effettuare query utilizzando la sintassi Turtle, un'estensione di N-Triples, alternativa estremamente sintetica e intuitiva rispetto al tradizionale RDF/XML.

[7]http://www.w3.org/TR/rdf-sparql-query/

Le query SPARQL si basano sul meccanismo del *pattern matching* e in particolare su un costrutto, il *triple pattern*, che ricalca il modello a triple di RDF fornendo un modello flessibile per la ricerca delle corrispondenze.

```
?titolo ex:autore ?autore
```

In questo triple pattern ci sono due variabili contrassegnate dal carattere `?`. Le variabili fungono da incognite mentre `ex:autore` funge da costante. Le triple RDF che trovano riscontro nel modello associeranno i propri termini alle variabili corrispondenti.
Ora prendiamo ad esempio le seguenti triple RDF:

```
@prefix ex: <http://example.org/ex/>
@prefix: <http://example.org/esempio/>
:TheRisingTied ex:autore "Mike Shinoda".
:Redemption ex:autore "Mike Shinoda".
:Meteora ex:autore "Linkin Park".
:Meteora ex:anno 2003.
```

Le asserzioni sono espresse nella sequenza soggetto-predicato-oggetto e delimitate da un punto. `@prefix` introduce i prefissi e i namespace mentre i due punti senza prefisso definiscono il namespace di default. Gli URI sono inclusi tra parentesi uncinate e i letterali di tipo stringa sono contrassegnati da virgolette.
Un esempio di query SPARQL è la seguente:

```
PREFIX ex: <http://example.org/ex/>
SELECT ?titolo ?autore ?anno
FROM  <http://example.com/listacd.ttl>
WHERE {?titolo cd:autore ?autore .
    ?titolo cd:anno ?anno .
}
```

La sintassi è molto simile a SQL. `PREFIX` dichiara prefissi e namespace, `SELECT` definisce le variabili di ricerca da prendere in considerazione nel risultato, `FROM` specifica il set di dati su cui verrà effettuata la query (nel caso

specifico all'indirizzo `http://example.com/listacd.ttl`), la clausola `WHERE` definisce i triple pattern che faranno da criterio di selezione.
Applicando la query all'esempio di triple precedente avremo come risultato: La query precedente ha escluso le triple che possiedono solo due termini

titolo	*autore*	*anno*
Meteora	Linkin Park	2003

Tabella 3.1: Risultato della query 1

richiesti (invece che tre). Si può riformulare la query in maniera da comprendere l'eventuale assenza di alcuni termini inserendo il costrutto `OPTIONAL`.

```
PREFIX ex: <http://example.org/ex/>
SELECT ?titolo ?autore ?anno
FROM  <http://example.com/listacd.ttl>
WHERE {?titolo cd:autore ?autore .
   OPTIONAL { ?titolo cd:anno ?anno . }
}
```

Il secondo pattern è dichiarato come opzionale quindi le variabili prive di valore compariranno vuote. Il risultato della query è il seguente: Un ulte-

titolo	*autore*	*anno*
TheRisingTied	Mike Shinoda	
Redemption	Mike Shinoda	
Meteora	Linkin Park	2003

Tabella 3.2: Risultato della query 2

riore modo per assicurare un certo grado di elasticità è inserire la parola chiave `UNION` che funziona come un `OR` logico, quindi cattura sia le triple che soddisfano il primo triple pattern sia il secondo. Un esempio di query è il seguente:

```
PREFIX ex: <http://example.org/ex/>
SELECT ?titolo ?autore ?anno
FROM  <http://example.com/listacd.ttl>
WHERE {
{?titolo cd:autore ?autore }
     UNION
{ ?titolo cd:anno ?anno }
}
```

SPARQL mette a disposizione un costrutto per porre delle restrizioni sui valori delle variabili utilizzando il comando `FILTER`:

```
PREFIX ex: <http://example.org/ex/>
SELECT ?titolo ?autore ?anno
FROM  <http://example.com/listacd.ttl>
WHERE {?titolo cd:anno ?anno .
   FILTER { ?anno > 2001 }
}
```

Nell'esempio verranno visualizzati solo le triple che avranno nella variabile una valore superiore a `2001`. Nel costrutto `FILTER` è possibile associarlo al comando `regex` per il pattern matching con espressioni regolari.

Capitolo 4

Evaluation

In questo capitolo si presenteranno i risultati ottenuti sul prototipo. Si passerà da una valutazione qualitativa confrontando i dati con dei golden standard utilizzati nella Semantic Publishing Challenge 2014 a una valutazione quantitativa cercando di analizzare il numero di riferimenti bibliografici che il software riesce a rilevare su un particolare dataset.

4.1 Semantic Publishing Challenge

La realizzazione di CiNeX è scaturita seguendo le linee guida del Task 2 della *ESWC-14 Challenge*[1], primo di una serie di eventi all'European Semantic Web Conference[2] che ha come scopo la produzione di dati per il Semantic Publishing. La caratterizzazione dell'evento del 2014 ha avuto al centro della discussione l'estrazione di informazioni e il loro utilizzo per valutare la qualità della produzione scientifica. Come visto nel capitolo 1 esistono dataset in linked data, come DBLP, ma coprono solo informazioni bibliografiche di base che non permettono di valutarne la qualità. In sostanza lo scopo della challenge è quello di ricercare le applicazioni più innovative e di impatto nel contesto emergente.

[1] http://challenges.2014.eswc-conferences.org/index.php/SemPub
[2] http://2014.eswc-conferences.org/

Tra i Task approntati il lavoro esposto in questa tesi si è soffermato sulla seconda attività: estrazione e caratterizzazione delle citazioni. Ai partecipanti è richiesto di processare un insieme di articoli scientifici in XML e creare una rete citazionale. I Dataset di input sono codificati in JATS e selezionati dai PubMedCentral Open Access Subset[3], dai Pensfot Biodiversity Data Journal[4] e dall'archivio Zookeys[5]. L'elaborazione è stata effettuata dapprima su un Training Dataset di 150 file e in seguito su un Evaluation dataset contenente 400 articoli.
Il fine ultimo del task è quello di rispondere a 10 query mentre il lavoro di tesi qui esposto si è limitato a fornire risposte alle prime quattro con relativi fattori di valutazione della qualità delle risposte.
Le 10 query sono:

1. Identificare tutti gli articoli citati dall'articolo X
2. Identificare tutte le riviste citate dall'articolo X
3. Identificare tutti gli autori citati dall'autore il cui cognome è X
4. Identificare tutti gli articoli citati dall'articolo X e scritti dallo stesso autore (o alcuni di questi)
5. Identificare tutti gli articoli citati più volte dal paper X
6. Identificare tutti gli articoli citati più volte nello stesso paragrafo dall'articolo X
7. Identificare la sovvenzione (o più di una) che ha supportato la ricerca presentata nell'articolo X, insieme all'ente che l'ha finanziata
8. Identificare la sezione Literature review delll'articolo X
9. Identificare tutti gli articoli di cui l'articolo X dichiara di utilizzare metodologie e teorie

[3]http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[4]http://biodiversitydatajournal.com/
[5]http://www.pensoft.net/journals/zookeys/

10. Identificare tutti gli articoli di cui l'articolo X dichiara di fornire un'estensione dei risultati

4.2 Le query

Di seguito vengono elencate le query richieste dalla Challenge con relativo codice SPARQL. Esse permettono di andare ad interrogare il dataset creato utilizzando l'ontologia SPAR. Per tutte le query valgono i prefissi:

```
PREFIX : <http://www.eswc14-sp-challenge.com/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX co: <http://purl.org/co/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX frapo: <http://purl.org/cerif/frapo/>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX pro: <http://purl.org/spar/pro/>
PREFIX biro: <http://purl.org/spar/biro/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
```

Q2.1a, Identify all papers cited by the paper whose DOI is '10.3897/zookeys.194.3308'

```
SELECT DISTINCT ?cited ?doi ?pmid ?title
WHERE { ?a prism:doi "10.3897/zookeys.194.3308".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
?c dcterms:title ?title
OPTIONAL {?c prism:doi ?doi . }
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
```

Q2.1b, Identify all papers cited by the paper whose PMID is '18298831'

```
SELECT DISTINCT ?cited ?doi ?pmid ?title
WHERE { ?a fabio:hasPubMedId "18298831".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
?c dcterms:title ?title
OPTIONAL {?c prism:doi ?doi . }
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
```

Q2.1c, Identify all papers cited by the paper whose DOI is '10.1093/aje/kwn259'

```
SELECT DISTINCT ?cited ?doi ?pmid ?title
WHERE { ?a prism:doi "10.1093/aje/kwn259".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
?c dcterms:title ?title
OPTIONAL {?c prism:doi ?doi . }
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
```

Q2.1d, Identify all papers cited by the paper whose DOI is '10.1155/2011/307152'

```
SELECT DISTINCT ?cited ?doi ?pmid ?title
WHERE { ?a prism:doi "10.1155/2011/307152".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
?c dcterms:title ?title
OPTIONAL {?c prism:doi ?doi . }
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
```

Q2.1e, Identify all papers cited by the paper whose TITLE is 'Extracellular Charge Adsorption Influences Intracellular Electrochemical Homeostasis in Amphibian Skeletal Muscle'

```
SELECT DISTINCT ?cited ?doi ?pmid ?title
WHERE { ?a dcterms:title "Extracellular Charge Adsorption Influences
Intracellular Electrochemical Homeostasis in Amphibian Skeletal Muscle".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
?c dcterms:title ?title
OPTIONAL {?c prism:doi ?doi . }
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
```

Q2.2a, Identify all journal papers cited by the paper whose DOI is '10.1186/1751-0759-7-4'

```
SELECT DISTINCT ?cited ?doi ?pmid ?papertitle ?journaltitle ?journalvolume
?journalissue
WHERE {
{ ?a prism:doi "10.1186/1751-0759-7-4".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?issue .
?issue prism:issueIdentifier ?journalissue .
?issue frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a prism:doi "10.1186/1751-0759-7-4".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
```

```
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a prism:doi "10.1186/1751-0759-7-4".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
}
```

Q2.2b, Identify all journal papers cited by the paper whose TITLE is 'Cluster randomized trials of prescription medicines or prescribing policy: public and general practitioner opinions in Scotland'

```
SELECT DISTINCT ?cited ?doi ?pmid ?papertitle ?journaltitle ?journalvolume
?journalissue
WHERE {
{ ?a dcterms:title "Cluster randomized trials of prescription medicines or
prescribing policy: public and general practitioner opinions in Scotland".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?issue .
?issue prism:issueIdentifier ?journalissue .
?issue frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a dcterms:title "Cluster randomized trials of prescription medicines or
```

```
prescribing policy: public and general practitioner opinions in Scotland".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a dcterms:title "Cluster randomized trials of prescription medicines or
prescribing policy: public and general practitioner opinions in Scotland".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
}
```

Q2.2c, Identify all journal papers cited by the paper whose TITLE is 'Size dependent heat generation of magnetite nanoparticles under AC magnetic field for cancer therapy'

```
SELECT DISTINCT ?cited ?doi ?pmid ?papertitle ?journaltitle ?journalvolume
?journalissue
WHERE {
{ ?a dcterms:title "Size dependent heat generation of magnetite nanoparticles
under AC magnetic field for cancer therapy".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?issue .
```

```
?issue prism:issueIdentifier ?journalissue .
?issue frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a dcterms:title "Size dependent heat generation of magnetite nanoparticles
under AC magnetic field for cancer therapy".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a dcterms:title "Size dependent heat generation of magnetite nanoparticles
under AC magnetic field for cancer therapy".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
}
```

Q2.2d, Identify all journal papers cited by the paper whose DOI is '10.1155/2011/307152'

```
SELECT DISTINCT ?cited ?doi ?pmid ?papertitle ?journaltitle ?journalvolume
?journalissue
```

```
WHERE {
{ ?a prism:doi "10.1155/2011/307152".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?issue .
?issue prism:issueIdentifier ?journalissue .
?issue frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a prism:doi "10.1155/2011/307152".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a prism:doi "10.1155/2011/307152".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
}
```

Q2.2e, Identify all journal papers cited by the paper whose PMID is '22028970'

```
SELECT DISTINCT ?cited ?doi ?pmid ?papertitle ?journaltitle ?journalvolume
?journalissue
WHERE {
{?a fabio:hasPubMedId "22028970".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?issue .
?issue prism:issueIdentifier ?journalissue .
?issue frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a fabio:hasPubMedId "22028970".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?volume .
?volume prism:volume ?journalvolume .
?volume frbr:partOf ?journal .
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
UNION
{ ?a fabio:hasPubMedId "22028970".
?b frbr:realization ?a .
?b cito:cites ?cited .
?cited frbr:realization ?c .
OPTIONAL {?c dcterms:title ?papertitle .}
OPTIONAL { ?c frbr:partOf ?journal .
```

```
?journal dcterms:title ?journaltitle . }
OPTIONAL {?c prism:doi ?doi .}
OPTIONAL {?c fabio:hasPubMedId ?pmid .}}
}
```

Q2.3a, Identify all authors cited by the author whose surname is 'Avigan'

```
SELECT DISTINCT ?author ?surname ?name
WHERE { ?authorPaper foaf:familyName "Avigan" .
?work frbr:creator ?authorPaper .
?work cito:cites ?reference .
?refAuthor pro:relatesToDocument ?reference .
?author pro:holdsRoleInTime ?refAuthor .
?author foaf:givenName ?name .
?author foaf:familyName ?surname .
}
```

Q2.3b, Identify all authors cited by the author whose surname is 'Pogue'

```
SELECT DISTINCT ?author ?surname ?name
WHERE { ?authorPaper foaf:familyName "Pogue" .
?work frbr:creator ?authorPaper .
?work cito:cites ?reference .
?refAuthor pro:relatesToDocument ?reference .
?author pro:holdsRoleInTime ?refAuthor .
?author foaf:givenName ?name .
?author foaf:familyName ?surname .
}
```

Q2.3c, Identify all authors cited by the author whose surname is 'Hostens'

```
SELECT DISTINCT ?author ?surname ?name
WHERE { ?authorPaper foaf:familyName "Hostens" .
?work frbr:creator ?authorPaper .
?work cito:cites ?reference .
?refAuthor pro:relatesToDocument ?reference .
?author pro:holdsRoleInTime ?refAuthor .
```

```
?author foaf:givenName ?name .
?author foaf:familyName ?surname .
}
```

Q2.3d, Identify all authors cited by the author whose surname is 'Ramadass'

```
SELECT DISTINCT ?author ?surname ?name
WHERE { ?authorPaper foaf:familyName "Ramadass" .
?work frbr:creator ?authorPaper .
?work cito:cites ?reference .
?refAuthor pro:relatesToDocument ?reference .
?author pro:holdsRoleInTime ?refAuthor .
?author foaf:givenName ?name .
?author foaf:familyName ?surname .
}
```

Q2.3e, Identify all authors cited by the author whose surname is 'Campagne'

```
SELECT DISTINCT ?author ?surname ?name
WHERE { ?authorPaper foaf:familyName "Campagne" .
?work frbr:creator ?authorPaper .
?work cito:cites ?reference .
?refAuthor pro:relatesToDocument ?reference .
?author pro:holdsRoleInTime ?refAuthor .
?author foaf:givenName ?name .
?author foaf:familyName ?surname .
}
```

Q2.4a, Identify all papers cited by the paper whose DOI is '10.3897/zookeys.118.1165' and written by the same authors (or some of them)

```
SELECT DISTINCT ?referenceWork ?doi ?pmid ?referenceTitle
WHERE { ?paperExpr prism:doi "10.3897/zookeys.118.1165".
?paperWork frbr:realization ?paperExpr .
?author pro:relatesToDocument ?paperWork .
```

```
?paperWork cito:cites ?referenceWork .
?author pro:relatesToDocument ?referenceWork .
?referenceWork frbr:realization ?referenceExpr .
?referenceExpr dcterms:title ?referenceTitle .
OPTIONAL { ?referenceExpr prism:doi ?doi . }
OPTIONAL { ?referenceExpr fabio:hasPubMedId ?pmid .}
}
```

Q2.4b, Identify all papers cited by the paper whose PMID is '19634910' and written by the same authors (or some of them)

```
SELECT DISTINCT ?referenceWork ?doi ?pmid ?referenceTitle
WHERE { ?paperExpr fabio:hasPubMedId "19634910".
?paperWork frbr:realization ?paperExpr .
?author pro:relatesToDocument ?paperWork .
?paperWork cito:cites ?referenceWork .
?author pro:relatesToDocument ?referenceWork .
?referenceWork frbr:realization ?referenceExpr .
?referenceExpr dcterms:title ?referenceTitle .
OPTIONAL { ?referenceExpr prism:doi ?doi . }
OPTIONAL { ?referenceExpr fabio:hasPubMedId ?pmid .}
}
```

Q2.4c, Identify all papers cited by the paper whose DOI is '10.1186/1471-2261-12-91' and written by the same authors (or some of them)

```
SELECT DISTINCT ?referenceWork ?doi ?pmid ?referenceTitle
WHERE { ?paperExpr prism:doi "10.1186/1471-2261-12-915".
?paperWork frbr:realization ?paperExpr .
?author pro:relatesToDocument ?paperWork .
?paperWork cito:cites ?referenceWork .
?author pro:relatesToDocument ?referenceWork .
?referenceWork frbr:realization ?referenceExpr .
?referenceExpr dcterms:title ?referenceTitle .
OPTIONAL { ?referenceExpr prism:doi ?doi . }
OPTIONAL { ?referenceExpr fabio:hasPubMedId ?pmid .}
}
```

Q2.4d, Identify all papers cited by the paper whose TITLE is 'A pilot randomized controlled trial to improve geriatric frailty' and written by the same authors (or some of them)

```
SELECT DISTINCT ?referenceWork ?doi ?pmid ?referenceTitle
WHERE { ?paperExpr dcterms:title "A pilot randomized controlled trial to
improve geriatric frailty".
?paperWork frbr:realization ?paperExpr .
?author pro:relatesToDocument ?paperWork .
?paperWork cito:cites ?referenceWork .
?author pro:relatesToDocument ?referenceWork .
?referenceWork frbr:realization ?referenceExpr .
?referenceExpr dcterms:title ?referenceTitle .
OPTIONAL { ?referenceExpr prism:doi ?doi . }
OPTIONAL { ?referenceExpr fabio:hasPubMedId ?pmid .}
}
```

Q2.4e, Identify all papers cited by the paper whose PMID is '16022735' and written by the same authors (or some of them)

```
SELECT DISTINCT ?referenceWork ?doi ?pmid ?referenceTitle
WHERE { ?paperExpr fabio:hasPubMedId "16022735".
?paperWork frbr:realization ?paperExpr .
?author pro:relatesToDocument ?paperWork .
?paperWork cito:cites ?referenceWork .
?author pro:relatesToDocument ?referenceWork .
?referenceWork frbr:realization ?referenceExpr .
?referenceExpr dcterms:title ?referenceTitle .
OPTIONAL { ?referenceExpr prism:doi ?doi . }
OPTIONAL { ?referenceExpr fabio:hasPubMedId ?pmid .}
}
```

4.3 Risultati

In questa sezione verranno illustrati i risultati della valutazione relativa agli output delle query. Questi ultimi sono stati comparati con un golden

standard e misurati in base alla precisione e alla recall. La recall è calcolata a partire dalla formula

$$\frac{TP}{TP + FN}$$

dove TP è il numero dei True Positives ossia i risultati attesi che il software ottiene e FN è il numero di False Negatives, ossia il numero di risultati attesi che il tool non riesce ad ottenere. Mentre la Precision è calcolata con la formula

$$\frac{TP}{TP + FP}$$

dove FP è il numero dei Falsi Positivi quindi i risultati duplicati o che non fanno parte dei TP.
Sia il golden standard che l'output in fase di verifica devono essere dei file CSV con una struttura comune. La valutazione è effettuata da un software in PHP che mette a confronto ogni coppia dei corrispondenti file in CSV.
Sono supportate due tipi di valutazioni:

- *strict*: solo le corrispondenze esatte sono considerate corrette.
- *loose*: sono considerate corrette anche le corrispondenze parziali.

La Precision media risulta essere 0.624 mentre la media della Recall 0.684. Le query con Precision più bassa sono quelle del gruppo *Q2.2** ed è dovuta ad un'elevata quantità di Falsi Positivi, probabilmente sarà necessario affinare le query di questa categoria considerata la mole di risultati duplicati. Tre query risultano con Precision e Recall uguale a zero a causa del mancato riconoscimento da parte del tool delle citazioni per quei particolari documenti. Si è calcolato inoltre il numero di riferimenti bibliografici non supportati dal tool sul totale dei 16626 con una percentuale di errore dell'8% (quindi con 1445 citazioni non conformi). Per quest'ultimo tipo di riferimento si è deciso di salvare l'intera stringa contenente le informazioni sull'articolo e in una prossima release si potrebbe pensare di estrarre tutti i dati attraverso delle espressioni regolari.
Il tempo per il parse dei 400 documenti è stato di 87 secondi su un Intel Core i7 Q740 a 1.73 GHz mentre il file di output generato ha un peso di 41,5 MB.

Query	Precision	Recall	TP	FN	FP
$Q2.1a$	1	1	8	0	0
$Q2.1b$	1	1	24	0	0
$Q2.1c$	1	1	28	0	0
$Q2.1d$	0.973	0.986	73	1	2
$Q2.1e$	0	0	0	18	0
$Q2.2a$	0.26	1	13	0	37
$Q2.2b$	0.409	0.9	9	1	13
$Q2.2c$	0.414	0.75	12	4	17
$Q2.2d$	0.327	0.986	71	1	146
$Q2.2e$	0.328	0.95	19	1	39
$Q2.3a$	0.951	0.983	58	1	3
$Q2.3b$	1	0.478	22	24	0
$Q2.3c$	0.909	0.866	290	45	29
$Q2.3d$	0	0	0	69	0
$Q2.3e$	0.946	0.614	35	22	2
$Q2.4a$	1	1	4	0	0
$Q2.4b$	0.944	0.944	17	1	1
$Q2.4c$	0	0	0	3	0
$Q2.4d$	1	0.2	1	4	0
$Q2.4e$	1	1	1	0	0

Tabella 4.1: Tabella relativa alla verifica dei risultati delle query della Challenge

4.4 Test su dataset PMCentral

Si è deciso di testare il software su un altro dataset da journal scelti sempre da PubMedCentral[6]. I risultati del test sono mostrati nelle tabelle 4.3, 4.4 e 4.5. Sono stati analizzati circa 6500 documenti JATS per un totale di 200263

[6]L'archivio FTP è raggiungibile all'indirizzo http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

Documenti analizzati	600
Rif. bib. ottenuti	200263
Rif. bib. non supportati	36840 (18%)
Tempo di esecuzione	1196 secondi

Tabella 4.2: Tabella riassuntiva dei test su dataset PMC

riferimenti bibliografici(Ref.). 36840 sono invece le reference non supportate dal software(Ref. NS) (circa il 18% del totale). Il tempo di esecuzione è stato di 1196 secondi. In 20 casi la percentuale di riferimenti non supportati ha superato il 50% di cui 17 con soglia superiore al 90% mentre 61 sono i casi in cui la percentuale di Reference individuate e supportate è superiore al 90%. Una parte corposa di tale percentuale risiede nei journal *Acta_Crystallogr* i quali non descrivono in maniera dettagliata le reference.
Il grafico 4.1 mostra quali journal hanno una maggiore incidenza di riferimenti bibliografici non supportati.

Nome	Ref.	Ref. NS	% NS	Tempo	N. Files
AAPS_J	2535	284	11	31	52
AAPS_PharmSciTech	642	123	19	6	15
Abdom_Imaging	413	17	4	6	20
Acad_Emerg_Med	129	17	13	4	4
Acc_Chem_Res	1074	16	1	8	22
Accid_Anal_Prev	12	6	50	1	1
Account_Res	143	55	38	3	3
ACS_Appl_Mater_Interfaces	520	440	84	5	12
ACS_Catal	87	80	91	3	3
ACS_Chem_Biol	3324	16	0	17	74
ACS_Chem_Neurosci	672	28	4	7	13
ACS_Comb_Sci	35	0	0	2	2
ACS_Macro_Lett	94	91	96	2	2
ACS_Med_Chem_Lett	704	21	2	8	29
ACS_Nano	3558	24	0	15	62
ACS_Sustain_Chem_Eng	69	0	0	2	2
ACS_Synth_Biol	484	2	0	6	10
Acta_Anaesthesiol_Scand	98	1	1	3	4
Acta_Biomater	494	28	5	5	9
Acta_Biotheor	224	42	18	5	8
Acta_Crystallogr_A	572	572	100	6	19
Acta_Crystallogr_B	354	354	100	4	11
Acta_Crystallogr_B	430	430	100	4	7
Acta_Crystallogr_C	56	56	100	2	5
Acta_Crystallogr_D	11202	11202	100	44	350
Acta_Crystallogr_F	305	305	100	4	11
Acta_Crystallogr_Sect_F	4145	4145	100	24	242
Acta_Diabetol	867	74	8	7	23
Acta_Ethol	332	20	6	4	7
Acta_Histochem	32	0	0	2	1
Acta_Histochem_Cytochem	5202	25	0	27	208
Acta_Inform_Med	3382	345	10	22	168
Acta_Mater	321	320	99	4	10
Acta_Med_Scand	0	0	0	4	1
Acta_Myol	2339	40	1	18	146
Acta_Naturae	11105	10575	95	37	216
Acta_Neurochir_(Wien)	1852	76	4	15	70
Acta_Neurol_Belg	148	9	6	4	3
Acta_Neuropathol	8111	211	2	31	162
Acta_Neuropathol_Commun	5386	4	0	24	108
Acta_Obstet_Gynecol_Scand	191	18	9	4	8

Tabella 4.3: Tabella relativa ai test su dataset PMC (1)

Nome	Ref.	Ref. NS	% NS	Tempo	N. Files
Acta_Odontol_Scand	174	6	3	4	6
Acta_Oncol	362	3	0	6	13
Acta_Ophthalmol	203	6	2	4	9
Acta_Orthop	16612	193	1	93	658
Acta_Ortop_Bras	2880	219	7	21	135
Acta_Otolaryngol	367	1	0	6	21
Acta_Otorhinolaryngol_Ital	5956	244	4	32	240
Acta_Paediatr	1053	46	4	9	53
Acta_Pharmacol_Sin	2113	11	0	9	26
Acta_Physiol_(Oxf)	608	4	0	7	18
Acta_Psychiatr_Scand	267	17	6	5	9
Acta_Psychol_(Amst)	331	43	12	5	5
Acta_Radiol	55	1	1	2	2
Acta_Radiol_Short_Rep	842	6	0	11	67
Acta_Theriol_(Warsz)	1053	178	16	8	25
Acta_Trop	182	21	11	4	5
Acta_Vet_Scand	10463	122	1	60	688
Acupunct_Med	570	21	3	6	19
Acute_Card_Care	5	0	0	2	1
Addict_Behav	97	20	20	3	2
Addict_Biol	197	9	4	4	5
Addict_Health	2209	309	13	15	97
Addict_Res_Theory	114	13	11	3	3
Addict_Sci_Clin_Pract	3758	180	4	32	297
Addiction	1525	282	18	10	48
Adipocyte	4064	33	0	18	101
Adm_Policy_Ment_Health	1011	135	13	7	21
Adolesc_Health_Med_Ther	3571	226	6	13	53
Adv_Appl_Bioinform_Chem	468	14	2	6	13
Adv_Appl_Bioinforma_Chem	1545	53	3	13	35
Adv_Bioinformatics	3496	187	5	19	101
Adv_Biol_Regul	46	0	0	2	1
Adv_Biomed_Res	7326	38	0	36	293
Adv_Biosci_Biotechnol	216	1	0	4	4
Adv_Breast_Cancer_Res	24	1	4	2	1
Adv_Cogn_Psychol	6717	419	6	27	121
Adv_Eng_Mater	62	62	100	2	1
Adv_Eng_Softw	9	6	66	2	1
Adv_Enzyme_Regul	84	1	1	3	2
Adv_Funct_Mater	76	76	100	2	2

Tabella 4.4: Tabella relativa ai test su dataset PMC (2)

Nome	Ref.	Ref. NS	% NS	Tempo	N. Files
Adv_Health_Sci	1879	188	10	12	55
Adv_Healthc_Mater	3	3	100	1	1
Adv_Hematol	10419	120	1	42	190
Adv_Mater	603	603	100	5	18
Adv_Math_(N_Y)	188	54	28	6	6
Adv_Med_Educ_Pract	3062	532	17	18	101
Adv_Nutr	770	114	14	6	17
Adv_Orthop	3967	81	2	34	120
Adv_Pharmacol_Sci	5142	169	3	21	101
Adv_Phys	621	620	99	4	1
Adv_Physiother	67	5	7	3	3
Adv_Prev_Med	2709	310	11	14	67
Adv_Sch_Ment_Health_Promot	44	13	29	2	1
Adv_Space_Res	12	2	16	2	1
Adv_Synth_Catal	269	267	99	3	6
Adv_Ther	1270	166	13	10	42
Adv_Urol	8383	135	1	37	319
Adv_Virol	7801	54	0	23	102
Aerobiologia_(Bologna)	700	125	17	6	14
Totale	200263	36840	18	1196	6490

Tabella 4.5: Tabella relativa ai test su dataset PMC (3)

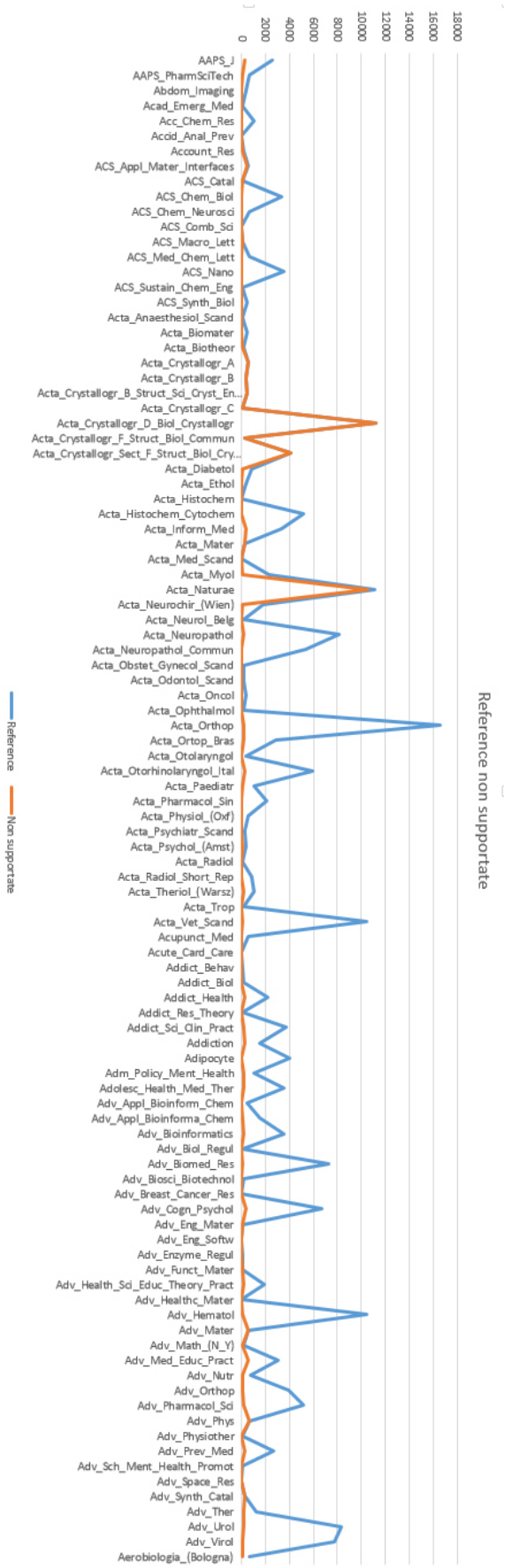

Figura 4.1: Grafico riassuntivo delle reference non supportate

Conclusioni

In questo lavoro si è cercato di sviluppare un software che traducesse in linguaggio RDF, quindi in una rete citazionale, dei riferimenti bibliografici descritti in linguaggio XML/JATS utilizzando i vocabolari e le ontologie per il Semantic Publishing.
Si è effettuata una rassegna delle ontologie utilizzate in questo ambito e analizzato alcuni dataset in Linked Data contenenti pubblicazioni scientifiche. Poi, utilizzando l'ontologia SPAR, si sono stilate alcune regole di traduzione dal formato JATS al linguaggio RDF per la creazione della rete citazionale e l'annotazione delle informazioni relative ai documenti presi in esame. Queste regole sono confluite nel prototipo CiNeX che si occupa di generare il dataset RDF il quale, caricato su Fuseki, viene interrogato tramite query SPARQL.
I risultati ottenuti sembrano positivi nonostante vi siano diversi punti su cui lavorare. In primo luogo si potrebbe centralizzare la gestione dei Journal effettuando una ricerca su qualche database in maniera da eliminare i duplicati dovuti alle molteplici forme di identificazione degli stessi. Inoltre, per quanto riguarda gli autori, l'abbreviazione del nome spesso non coincideva con il nome scritto per esteso e anche in questo caso si sono riscontrati dei problemi di identificazione; stessa cosa dicasi per l'omonimia che in un futuro bisognerà considerare in qualche modo. Inoltre, come descritto nel capitolo 3, c'è da sciogliere il nodo dei riferimenti non supportati studiando una soluzione ad hoc come ad esempio l'utilizzo delle espressioni regolari per l'identificazione delle parti di informazione utili visto che, da una prima analisi, sembra esserci una certa ricorrenza di pattern simili. L'idea è quella di integrare servizi

come Mendeley[7] o CrossRef[8] per rintracciare i riferimenti incompleti o non strutturati .

Infine la base di dati locale è predisposta per l'elaborazione e la traduzione in linguaggio RDF dei paragrafi con annesse citazioni nel testo dell'articolo, cosa che permetterebbe di rispondere alle successive query della Challenge.

[7] http://www.mendeley.com/

[8] http://www.crossref.org/

Bibliografia

[1] D. Shotton. Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing, Volume 22, Number 2, April 2009, pp. 85-94

[2] J. Carroll, G. Klyne. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 10 Febbraio 2004. World Wide Web Consortium. http://www.w3.org/TR/rdf-concepts/ (Ultima visita 22 Maggio 2014)

[3] B. Motik, P. F. Patel-Schneider, B. Parsia. OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax. W3C Recommendation, 27 Ottobre 2009. World Wide Web Consortium. http://www.w3.org/TR/owl2-syntax/ (Ultima visita 22 Maggio 2014)

[4] S. Peroni. Semantic Publishing: issues, solutions and new trends in scholarly publishing within the Semantic Web era. Tesi di dottorato, Università di Bologna, 2012.

[5] S. Pettifer, P. McDermott, J. Marsh, D. Thorne, A. Villeger, T. K. Attwood. Ceci n'est pas un hamburger: modelling and representing the scholarly article. Learned Publishing, Vol. 24 Number 3, July 2011, pp. 207-220

[6] S. Peroni, D.A. Lapeyre, D. Shotton. From Markup to Linked Data: Mapping NISO JATS v1.0 to RDF using the SPAR (Semantic Publishing and Referencing) Ontologies, Journal Article Tag Suite Conferen-

ce (JATS-Con) Proceedings 2012 [Internet], Bethesda (MD): National Center for Biotechnology Information (US), 2012.

[7] A. de Waard. From Proteins to Fairytales: Directions in Semantic Publishing. Intelligent Systems, IEEE, Vol. 25, Issue 2, March-April 2010, pp. 83-88

[8] Dublin Core Metadata Initiative (2010). Dublin Core Metadata Element Set, Version 1.1. DCMI Reccomendation. http//dublincore.org/documents/dces/ (Ultima visita 19 Luglio 2014)

[9] Dublin Core Metadata Initiative (2010). DCMI Metadata Terms. DCMI Recommendation. http://dublincore.org/documents/dcmi-terms/ (Ultima visita 19 Luglio 2014)

[10] International Digital Enterprise Alliance (2009). Publishing Requirements for Industry Standard Metadata Specification. Alexandria, VA, USA: IDEAlliance. http://www.prismstandard.org (Ultima visita 22 Maggio 2014)

[11] D'Arcus, B., Giasson, F. (2009). Bibliographic Ontology Specification. Specification Document, 4 November 2009. http://bibliontology.com/specification (Ultima visita 19 Luglio 2014)

[12] International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records (2009). Functional Requirements for Bibliographic Records Final Report. International Federation of Library Associations and Institutions. http://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf (Ultima visita 19 Luglio 2014)

[13] P. Ciccarese, E. Wu, J. Kinoshita, G. Wong, M. Ocana, A. Ruttenberg, T. Clark. The SWAN Biomedical Discourse Ontology. Journal of Biomedical Informatics, Vol. 41 5, 2008, pp. 739-751

[14] A. Miles, S. Bechhofer, (2009). SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. World Wide Web Consortium. http://www.w3.org/TR/skos-reference/ (Ultima visita 19 Luglio 2014)

[15] D. Shotton, S. Peroni. Introducing the Semantic Publishing and Referencing (SPAR) Ontologies. Ottobre 2010, http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/ (Ultima visita 19 Luglio 2014)

[16] L. Jael Garcia Castro, C. McLaughlin, A. Garcia. Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data. Journal of Biomedical Semantics 2013, 4(Suppl 1):S5

[17] D. Shotton. Open citations. Nature 502, 2013, pp. 295-297 http://www.nature.com/news/publishing-open-citations-1.13937 (Ultima visita 15 Agosto 2014)

[18] A. Di Iorio, A. Nuzzolese, S. Peroni, D. Shotton, F. Vitali. Describing bibliographic references in RDF. Proceedings of 4th Workshop on Semantic Publishing (SePublica 2013), pp. 63-74

[19] Y. Sure, S. Bloedhom, P. Hase, J. Hartmann, D. Oberle. The SWRC Ontology - Semantic Web for Reasearch Communities. Progress in Artificial Intelligence, Lecture Notes in Computer Science, Volume 3808, 2005, pp. 218-231

[20] J. Hendler, T. Berners-Lee, E. Miller. Integrating Applications on the Semantic Web. Journal of the Institute of Electrical Engineers of Japan, 2002, Vol 122(10) pp. 676-680

[21] C. Bizer, T. Heath, T. Berners-Lee. Linked Data - the story so far. International Journal on Semantic Web and Information Systems, 2009 5, (3), pp. 1-22

[22] S. Peroni, T. Gray, A. Dutton, D. Shotton. Setting our bibliographic references free: towards open citation data. In pubblicazione per il Journal of Documentation, 71(2), 2015. http://speroni.web.cs.unibo.it/publications/peroni-in-press-setting-bibliographic-references.pdf (Ultima visita 15 Agosto 2014)

Ringraziamenti

Vorrei ringraziare in primis il prof. Di Iorio per la disponibilità, il tempo e l'attenzione concessami, poi vorrei menzionare il Dr. Silvio Peroni per i preziosi consigli su SPAR e le ontologie utilizzate per questo scritto. Infine un ringraziamento va al mio collega Matteo Fanciulli per tutte le giornate spese con me a studiare e al supporto che mi ha dato nel redigere la tesi.