

Università degli Studi di Modena e Reggio Emilia

Facoltà di Ingegneria 'Enzo Ferrari' – Sede di Modena

Corso di Laurea Magistrale (D. M. 270/04) in
Ingegneria Informatica

Linked Open Data: pubblicazione, arricchimento semantico e linking di dataset pubblici attraverso il sistema MOMIS

Relatore:

Prof.ssa Sonia Bergamaschi

Correlatore:

Dott. Ing. Serena Sorrentino

Candidato:

Elisa Fusari

Sommario

Introduzione	4
Capitolo 1	
1. Linked Data	7
1.1. Sintassi RDF	11
1.2. Applicazioni con Linked Data	12
Capitolo 2	
2. Open Data	14
2.1. Open Data in Italia	20
2.2. Open Linked Data	23
2.2.1. LOD Project	26
2.2.2. LATC, LOD Around-The-Clock	33
2.2.3. LOD in Italia	36
Capitolo 3	
3. Strumenti per la pubblicazione di Linked Data	40
3.1. Modelling	40
3.1.1. Neologism	45
3.2. Pubblicazione	49
3.2.1. D2R Server	50
3.3. Linking	54
3.3.1. Silk Workbench	58
Capitolo 4	
4. Integrazione semantica nei LOD: stato dell'arte	66
4.1. BLOOMS	67
4.2. AGROPub	70
4.3. Stratosphere	73
4.4. WebSmatch	76
Capitolo 5	
5. Caso d'uso: dal dataset Precarietà ai LOD	78
5.1. Il dataset sulle "Precarietà Giovanili" del Comune di Modena	79
5.2. Modellazione e conversione dei dati in RDF	80
5.3. Annotazione semantica e linking a WordNet	83

5.3.1. WordNet	83
5.3.2. Creazione dei link con Protégé	84
5.4. Pubblicazione e SPARQL endpoint	87
5.5. MOMIS	96
5.5.1. MOMIS nella creazione dei Linked Open Data	98
Conclusioni e Future Works	100
Bibliografia	103
Appendice	108
Appendice A: Neologism	108
Appendice B: D2R Server	111
Appendice C: Silk Workbench	117
Appendice D: Questionario	124
Appendice E: Modello Relazionale	127

Introduzione

Con il diffondersi delle politiche di Open Government, partite dagli Stati Uniti e dal Regno Unito, sta sempre più prendendo piede, anche in Italia, l'idea di rendere pubblici tutti i dati relativi al Governo e alle Pubbliche Amministrazioni, offrendo un maggiore trasparenza ai cittadini e la possibilità di interagire con essi. Gli Open Data, i dati aperti, ovvero i dati resi disponibili in rete, a tutti gli utenti, senza la necessità di licenze e senza restrizioni, sono diventati così numerosi all'interno del Web, che ha preso piede l'idea di poterli standardizzare, organizzare e collegare gli uni con gli altri.

Il progetto Linked Open Data (LOD), si è proprio occupato di questo. Una volta redatti da Tim Berners Lee, fondatore del Semantic Web, i principi base per la pubblicazione dei dati, si è passati alla creazione di link tra di essi, creando un'unica nuvola di dati, completamente interconnessi ed appartenenti a domini differenti.

Si è così costituito il cosiddetto Web of Data, in cui i dati risultano organizzati e rintracciabili, e in cui si ha la possibilità di interrogare diverse sorgenti dati contemporaneamente, appartenenti a contesti differenti, in modo da vederne le possibili relazioni ed aggiungere valore all'informazione.

Ciò è stato possibile grazie alla creazione, da parte del W3C (World Wide Web Consortium), dello standard RDF, linguaggio che permette la pubblicazione dei dati in triple formate da *soggetto-predicato-oggetto*, in cui ciascun componente può essere un URI relativo alla risorsa che si vuole pubblicare.

I dati, sono per la maggior parte collegati a livello di istanza, con link che legano tra loro gli elementi relativi allo stesso oggetto reale, mentre le relazioni semantiche, a livello di schema, sono quasi del tutto assenti rendendo il Web of Data, carente a livello

semantico: pur essendo stati sviluppati tantissimi tool che facilitano gli utenti nel processo di pubblicazione dei dati, pochi sono quelli che arricchiscono e collegano i dati a livello di schema.

In questa tesi si è studiato e sperimentato il processo di creazione e pubblicazione di Linked Open Data partendo da un caso di studio concreto, i dati forniti dal Comune di Modena, relativi al progetto sulle Precarietà Giovanili, coordinato dall'Assessore Fabio Poggi.

Ciò ha comportato la rielaborazione dei dati in modo strutturato e la creazione del relativo database; l'analisi e il test di diversi strumenti per la conversione dei dati in RDF, per l'arricchimento semantico, per la loro pubblicazione e, infine, per il linking con gli altri dataset presenti nel Linked Open Data Cloud.

In particolare, si è approfondita la possibilità di creare link tra i diversi dataset, per mezzo dell'annotazione, operazione che permette di assegnare a ciascun termine il proprio significato, rendendolo più comprensibile, e facilitando l'integrazione dei dati ad esso relativi.

Si è inoltre analizzato il framework MOMIS (Mediator envirOnment for Multiple Information Sources), strumento per l'integrazione di sorgenti di dati eterogenee sviluppato da DBGroup, e dal suo spinoff Datariver, del Dipartimento di Ingegneria 'Enzo Ferrari' dell'Università di Modena, in modo da inserire anch'esso all'interno del processo di creazione dei LOD, e fornendo le basi per possibili sviluppi futuri del software.

Il contenuto della tesi è così strutturato:

- Capitolo 1: *Linked Data*; il capitolo offre una panoramica sui Linked Data, sul linguaggio RDF e sulle applicazioni che li utilizzano;
- Capitolo 2: *Open Data*; il capitolo presenta lo sviluppo della mentalità dei 'dati aperti', dall'Open Government e alle politiche che l'hanno portato avanti concretamente; si ha inoltre, una sintesi sullo stato degli Open Data nel mondo, e in Italia, e viene approfondita l'evoluzione dei Linked Open Data, le sue caratteristiche ed l'analisi dei progetti ad essi relativi;
- Capitolo 3: *Strumenti per la pubblicazione dei LOD*; il capitolo propone l'analisi delle fasi del processo di creazione dei Linked Open Data e il testing di diversi tool utilizzati;

- Capitolo 4: *Integrazione semantica nei LOD: stato dell'arte*; il capitolo offre una panoramica del problema di mancanza di semantica nei LOD e presenta un quadro dei progetti e degli strumenti che si sono già interessati alla tematica;
- Capitolo 5: *Caso d'uso: dal dataset Precarietà ai LOD*; il capitolo propone l'applicazione del processo di pubblicazione dei Linked Open Data al caso concreto utilizzando il dataset 'Precarietà Giovanili' fornitoci dal Comune di Modena; quindi approfondisce le fasi di arricchimento semantico e linking al dataset di WordNet, presente nei LOD, per mezzo del thesaurus multilingua MultiWordNet; ed infine, analizza il framework MOMIS, proponendo spunti per il suo possibile sviluppo ed utilizzo nell'ambito dei Linked Open Data.

Capitolo 1

1. Linked Data

Per Linked Data [10] si intende comunemente, il metodo di pubblicazione di dati strutturati interconnessi, in modo da creare link tra sorgenti diverse e aumentare così la loro possibilità di utilizzo; tecnicamente essi, fanno riferimento ai dati pubblicati sul Web in modo “machine-readable”, ovvero automaticamente leggibili dai computer, i cui contenuti sono esplicitamente definiti e che possono essere collegati ed interrogati insieme a dataset esterni.

I Linked Data basano il loro funzionamento sullo standard “Resource Description Framework (RDF)”, il quale si sviluppa a sua volta sulle due tecnologie fondamentali del Web, “HyperText Transfer Protocol (HTTP)”, come meccanismo di accesso universale, e “Uniform Resource Identifier (URI)”, come meccanismo globale di identificazione univoca.

Tim Berners-Lee, direttore del World Wide Web Consortium (W3C), coniò il termine ‘linked data’ nel 2006, in una Design Issue sul Semantic Web, e definì una serie di ‘regole’ per pubblicarli: seguendo queste indicazioni i dati vengono, così, inseriti in uno spazio globale, il cosiddetto “Web of Data”. Le regole prendono il nome di “Linked Data Principles” [10] e si compongono di quattro punti:

1. *Use URIs as names for things.*
2. *Use HTTP URIs so that people can look up those names.*
3. *When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).*

4. *Include links to other URI, so that they can discover more things.*

Ovvero, usare gli URI come nomi per le cose, sia per le entità stesse, oggetti e concetti astratti, che per i documenti che le descrivono, facendo attenzione a differenziare i due elementi; usare HTTP URI in modo che le persone possano effettuare ricerche di questi nomi, ad esempio all'interno di descrizioni dell'oggetto considerato, semplicemente de-referenziando le URI attraverso il protocollo HTTP; fornire informazioni sull'oggetto, quando la sua URI è de-referenziata, utilizzando standard come RDF e SPARQL, ottenendo così un unico modello per la pubblicazione dei dati strutturati nel Web; ed infine, inserire in queste descrizioni, altre URI, in modo da ampliare la possibilità di ricerca di informazioni relative all'elemento cercato.

Quindi, come detto prima, le entità sono identificate da URI che utilizzano lo schema *http://*, in modo da poter essere cercate e trovate de-referenziando l'URI con il protocollo HTTP; in questo modo, il protocollo HTTP diventa un meccanismo semplice ma universale per il recupero sia delle risorse serializzabili in flussi di byte, sia delle descrizioni delle risorse che non possono essere inviate attraverso la rete in questo modo. Ad esempio, de-referenziando l'URI *http://dbpedia.org/resource/Berlin*, e chiedendo il contenuto *application/rdf+xml*, dopo il reindirizzamento, si ottiene la descrizione associata, equivalente alla descrizione RDF di *http://dbpedia.org/resource/Berlin* contenente l'informazione sulla risorsa *http://dbpedia.org/data/Berlin* (il processo di de-referenziazione degli URI, sarà approfondito nel Paragrafo 3.1).

URI e HTTP sono, poi integrate da una tecnologia fondamentale per il Web of Data, il linguaggio "RDF (Resource Description Framework)"[15]: mentre l'HTML fornisce un mezzo per strutturare i documenti e i link sul Web, l'RDF fornisce un generico, modello di dati graph-based, con cui strutturare i dati e i link che descrivono le cose nel mondo.

Il modello dati RDF rappresenta le informazioni come grafi aventi nodi e archi orientati; esso è stato progettato per la rappresentazione di informazioni integrate, provenienti da più fonti, ed ha lo scopo di essere usato come lingua franca, in grado di mediare tra i vari modelli di dati utilizzati sul Web.

In RDF, la descrizione di una risorsa presenta la struttura di una *tripla*, le cui tre parti rispecchiano la struttura di una semplice frase e sono chiamate *soggetto*, *predicato* e *oggetto* (vedi Figura 1).

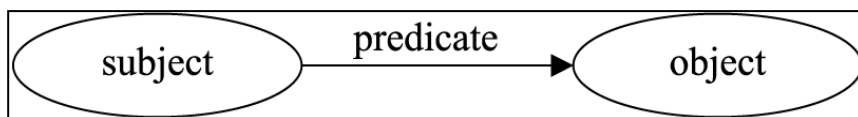


Figura 1: struttura di una tripla in RDF.

Il soggetto di una tripla è l'URI che identifica la risorsa descritta; l'oggetto può essere un semplice valore letterale, come una stringa o un numero, o l'URI di un'altra risorsa in qualche modo legata al soggetto; il predicato, nel mezzo, indica il tipo di relazione esistente tra soggetto e oggetto. Il predicato è anch'esso identificato da un URI e può essere selezionato all'interno dei vocabolari messi a disposizione per la costruzione dei Linked Data relativi ad un determinato dominio.

Si possono quindi distinguere due tipi principali di triple:

- *Literal Triple*, sono le triple che hanno come oggetto un elemento letterale come una stringa, un numero o una data; esse sono usate soprattutto per descrivere le proprietà delle risorse;
- *RDF Link*, sono le triple costituite da tre riferimenti URI e che descrivono la relazione tra due risorse (soggetto e oggetto); questi link possono essere interni, se gli URI di soggetto e oggetto si trovano nello stesso namespace, mentre sono detti esterni, se gli URI appartengono a namespace diversi; questi ultimi sono fondamentali per collegare le varie fonti isolate in un unico spazio di dati globale.

Gli RDF Link possono poi essere suddivisi in base al tipo di predicato usato; le categorie principali sono tre:

- *Relationship Link*: se legano fonti di dati diverse, ad esempio consentono alle persone di dare informazioni sul luogo dove vivono, sulle loro attività e così via;
- *Identity Link*: se collegano *URI-alias*, ovvero diverse fonti di dati che descrivono lo stesso oggetto, consentendo agli utenti di recuperare descrizioni differenti e diversi punti di vista sulla stessa entità;
- *Vocabulary Link*: se collegano i dati alle definizioni dei termini usati, o se collegano queste definizioni alle definizioni dei termini correlati nei diversi vocabolari; ciò permette di rendere i dati auto-descrittivi e consente, alle applicazioni che usano i Linked Data, di comprenderli ed integrarli meglio.

I termini utilizzati per descrivere i Linked Data sono presi da appositi vocabolari, i quali seguono due tendenze: da una parte si cerca di evitare la troppa eterogeneità cercando di

riutilizzare i termini più usati, facendo riferimento quindi ai vocabolari che contengono le classi e le proprietà più importanti, come RDF-Schema (RDFS)¹, Simple Knowledge Organization System (SKOS)² o Web Ontology Language (OWL)³, che forniscono una base per la descrizione delle entità e delle varie relazioni; d'altra parte si cerca di rendere i vocabolari più auto-descrittivi possibili, in modo che una volta pubblicato un vocabolario specifico, di dominio, ci si possa poi relazionare a quelli più conosciuti collegando i vari termini alle definizioni già presenti. Questo è possibile in quanto anche i vocabolari stessi sono espressi in RDF, e si usano i predicati come: `owl:equivalentClass`, `owl:equivalentProperty`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, `skos:broadMatch` e `skos:narrowMatch` per definire i diversi *mapping* tra di essi.

I principali vantaggi nell'utilizzare il modello di dati RDF in un contesto Linked Data sono i seguenti [8]:

- Usando HTTP URI come identificatori univoci globali per gli elementi dei dati, nonché per i termini di vocabolario, il modello di dati RDF è intrinsecamente destinato ad essere utilizzato su scala mondiale, permettendo a chiunque di fare riferimento a qualcosa;
- Gli utenti possono cercare qualsiasi URI in un grafo RDF sul Web per recuperare informazioni aggiuntive; ogni tripla RDF è parte del Web of Data e può essere utilizzata come punto di partenza per esplorare questo spazio dati globale;
- Il modello di dati consente di impostare i collegamenti RDF tra i dati provenienti da fonti diverse;
- Informazioni provenienti da fonti diverse possono essere facilmente combinate attraverso la fusione delle due serie di triple in un unico grafo;
- RDF consente di rappresentare informazioni, anche se espresse utilizzando schemi diversi, in un unico grafo; ciò significa che è possibile combinare le definizioni di vocabolari differenti per rappresentare i dati.
- In combinazione con i linguaggi come RDF-Schema e OWL, il modello di dati consente la rappresentazione dei dati che si desiderano, ovvero sia dati strutturati che semi-strutturati.

¹ <http://www.w3.org/TR/rdf-schema/>

² <http://www.w3.org/TR/skos-reference/>

³ <http://www.w3.org/TR/2004/REC-owl-features-20040210/>

1.1. Sintassi RDF

È importante ricordare, inoltre, come RDF non sia un formato di dati, ma un modello per descrivere le risorse in forme di triple; esistono poi diverse sintassi RDF [8] che possono essere utilizzate per pubblicare i Linked Data:

- *RDF/XML*, è standardizzato dal W3C ed è ampiamente utilizzato per pubblicare i dati collegati sul Web; un suo elemento a sfavore è che presenta una sintassi difficile da leggere e scrivere per gli umani e, quindi, per quegli utilizzi che necessitano la visualizzazione dei dati non automatizzata si deve prendere in considerazione l'uso di altre forme;
- *RDFa*, è un formato di serializzazione che incorpora triple RDF in documenti HTML, intrecciando i due elementi all'interno del "HTML Document Object Model (DOM)"; ciò significa che il contenuto esistente all'interno della pagina può essere contrassegnato con RDFa modificando il codice HTML, in modo da esporre i dati strutturati sul Web. RDFa è popolare in contesti in cui i pubblicatori di dati sono in grado di modificare i modelli HTML, ma hanno poco controllo sull'infrastruttura di pubblicazione; quando si utilizza RDFa è, inoltre, importante mantenere una distinzione inequivocabile tra gli oggetti reali descritti dai dati ed il documento in HTML+RDFa che incarna queste descrizioni, ad esempio utilizzando l'attributo RDFa *about=* per gli URI rappresentanti i primi;
- *Turtle*, è il formato di testo per la serializzazione di dati RDF che, grazie al supporto per i prefissi dei namespace e altre abbreviazioni, viene scelto quando è necessaria la lettura o la scrittura manuale di triple RDF;
- *N-Triples*, è un sottoinsieme di Turtle, che fa a meno di caratteristiche come i prefissi degli spazi dei nomi e delle abbreviazioni; il risultato è una sintassi con un sacco di ridondanza, in quanto tutti gli URI devono essere specificati per intero in ogni tripla: di conseguenza, un file N-triples può essere piuttosto grande rispetto ad altri formati, ma diventa vantaggioso, in quanto risulta analizzabile per righe, ideale per il caricamento di file di grandi dimensioni che non rientrano in memoria principale, e molto comprimibile, riducendo così il traffico di rete quando si scambiano file; ciò rende questa sintassi lo standard de facto per lo scambio di grandi insiemi di Linked Data, ad esempio ai fini di backup o mirroring;

- *JSON*, si riferisce agli sforzi per fornire una sintassi “JavaScript Object Notation (JSON)” per RDF; la più diffusa è la specifica *Talis*⁴. Mettere a disposizione una sintassi JSON per RDF è molto importante, in quanto un numero crescente di linguaggi di programmazione fornisce supporto nativo per JSON, compresi JavaScript e PHP, e ciò renderebbe i Linked Data accessibili agli sviluppatori Web, senza la necessità di installare librerie software aggiuntive per l’analisi e la manipolazione di dati RDF.

1.2. Applicazioni con Linked Data

Poiché la disponibilità dei Linked Data è un fenomeno relativamente recente, le applicazioni che li utilizzano sono ancora in via di sviluppo ma danno già un’idea dei modelli architettonici emergenti e di quello che sarà possibile eseguire in un prossimo futuro.

Le applicazioni per i Linked Data possono essere classificate in due categorie: applicazioni *generiche* e applicazioni *di dominio specifico*.

Le applicazioni generiche possono processare dati appartenenti a qualsiasi argomento, dalle pubblicazioni di un autore alle formule della biologia, e si dividono in due tipologie base: i *Linked Data Browser* e i *Linked Data Search Engines*.

I primi consentono agli utenti di navigare tra le fonti di dati seguendo i link RDF; essi scoprono nuove sorgenti dati automaticamente seguendo i predicati `owl:sameAs`, che collegano i dati riguardanti uno stesso soggetto, quindi prelevano i dati di interesse e li uniscono, presentando all’utente i risultati ottenuti relativi ad uno stesso oggetto. Esempi di tali browser includono **Tabulator**⁵, che dispone di diverse modalità per l’analisi dei dati risultanti come linee temporali o mappe, **Marbles**⁶, che tiene traccia della *provenance* dei dati mostrando le diverse fonti, oppure altri progetti come **Fenfire**⁷ o **VisiNav**⁸, che permette ricerche con query molto articolate.

I secondi invece, sono motori di ricerca che scansano il Web seguendo i link a livello dei dati; sono quindi in grado di aggregare dati da migliaia di sorgenti e forniscono la possibilità di interrogarli eseguendo delle query. Esempi di questo tipo di applicazioni

⁴ http://n2.talis.com/wiki/RDF_JSON_Specification

⁵ <http://www.w3.org/2005/ajar/tab>

⁶ <http://marbles.sourceforge.net/>

⁷ <http://fenfire.org/>

⁸ <http://sw.deri.org/2009/01/visinav/>

sono **Falcons**⁹ e **Semantic Web Search Engine (SWSE)**¹⁰, che permettono la ricerca dei dati agli utenti attraverso l'utilizzo di parole chiave, oppure come **Sindice**¹¹, **Swoogle**¹² e **Watson**¹³, che sono invece orientati alla ricerca automatizzata, quella necessaria alle applicazioni costruite sui Linked Data, e che utilizzano API per trovare i documenti RDF che fanno riferimento ad una certa URI o che contengono certe parole. Anche Yahoo e Google hanno iniziato ad utilizzare i Linked Data all'interno delle loro applicazioni: Yahoo, fornisce la scansione e l'accesso ai Linked Data tramite le sue API **BOSS**¹⁴, mentre Google, utilizza la ricerca di dati RDF, progettando di utilizzarne i risultati per migliorare i suoi **Rich Snippet**¹⁵, ovvero i risultati delle ricerche su prodotti, recensioni e persone.

Le applicazioni Linked Data per domini particolari, invece, agiscono su argomenti specifici e coprono le esigenze delle rispettive community. Esempi di queste applicazioni sono le innumerevoli *app* che sono state create in seguito alla pubblicazione sul Web dei dati relativi ai governi e alle pubbliche amministrazioni, (di ciò si tratterà in modo più approfondito nei capitoli successivi). Altri esempi includono **DBpedia Mobile**¹⁶, un'applicazione smart-phone di aiuto ai turisti che esplorano una città, che raccoglie tutte le informazioni relative agli elementi di quella città grazie ai dati di DBpedia e alla posizione satellitare; **Revyu**¹⁷, un sito Web di valutazioni che aumenta la consistenza delle revisioni con informazioni provenienti dal Web of Data, e **Talis Aspire**¹⁸, uno strumento in PHP per la gestione delle risorse dei corsi universitari che viene utilizzato da 40.000 studenti dell'Università di Plymouth e dell'Università del Sussex, esso assegna ad ogni dato un HTTP URI memorizzandolo secondo il modello RDF.

⁹ <http://iws.seu.edu.cn/services/falcons/documentsearch/>

¹⁰ <http://www.swse.org/>

¹¹ <http://sindice.com/>

¹² <http://swoogle.umbc.edu/>

¹³ <http://kmi-web05.open.ac.uk/Overview.html>

¹⁴ <http://developer.yahoo.com/search/boss/>

¹⁵ <http://googlewebmastercentral.blogspot.com/2009/10/help-us-make-web-better-update-on-rich.html>

¹⁶ <http://wiki.dbpedia.org/DBpediaMobile>

¹⁷ <http://revyu.com/>

¹⁸ <http://www.w3.org/2001/sw/sweo/public/UseCases/Talis/>

Capitolo 2

2. Open Data

Uno dei termini che è sempre più legato al discorso dei Linked Data, è *Open Data* [1], ma a cosa si riferisce esattamente? Per Open Data si intende l'idea per la quale alcuni dati debbano essere completamente a disposizione del pubblico, sia per l'utilizzo che per essere ulteriormente diffusi, senza restrizioni come copyright, licenze o altri meccanismi di controllo.

Lo scopo del movimento Open Data è simile a quello portato avanti dagli altri movimenti Open come l'Open Source, Open Content, Open Access, che enfatizzano l'importanza del contributo di ciascun individuo nel progresso tecnologico e scientifico; una tipica espressione di questa necessità ci viene dal commento di John Wilbanks, Direttore Esecutivo di Science Commons¹⁹:

“Numerous scientists have pointed out the irony that right at the historical moment when we have the technologies to permit worldwide availability and distributed process of scientific data, broadening collaboration and accelerating the pace and depth of discovery[...], we are busy locking up that data and preventing the use of correspondingly advanced technologies on knowledge”.

Gli Open Data fanno, spesso, riferimento a materiale non testuale come mappe, formule matematiche o scientifiche, dati medici o biologici, composti chimici, i quali non hanno un valore commerciale presi singolarmente, ma che possono acquistarlo se aggregati in modo utile ed efficace. È importante quindi che questi dati siano completamente *free*,

¹⁹ <http://creativecommons.org/science>

ovvero non vi siano limitazioni nella consultazione ma soprattutto che non servano ulteriori permessi per il loro riutilizzo.

Le maggiori sorgenti di Open Data sono la comunità scientifica, con l'Open Science Data [7], e le pubbliche amministrazioni, con l'Open Government [2].

Il concetto di Open Science Data pone l'attenzione sulla pubblicazione delle osservazioni e dei risultati delle attività scientifiche, in modo da poterli analizzare e riutilizzare; questa idea è stata istituzionalmente stabilita con la formazione del sistema World Data Center alla fine degli anni '60, in cui l'odierno International Council for Science stabilì diversi world data center per minimizzare il rischio di perdita di dati e massimizzarne la possibilità di accesso; nel 2004, i Ministeri delle Scienze appartenenti all'OECD (Organisation for Economic Co-operation and Development), che comprende la maggior parte delle nazioni più sviluppate, firmarono un accordo in cui si stabilì che tutti gli archivi creati con fondi pubblici sarebbero stati anche disponibili e utilizzabili dal pubblico. Esempi di siti contenenti Open Data scientifici sono:

- *data.uni-muenster.de*, dell'Università di Muenster, Germania, lanciato nel 2011;
- *linkedscience.org/data*, i cui dati sono codificati come Linked Data, (che analizzeremo in seguito), lanciato anch'esso nel 2011.

L'Open Government, invece, è un movimento che si sta sviluppando in questi anni e che prevede la pubblicazione dei dati dei vari governi nazionali. Questa condotta è stata promossa dal presidente degli Stati Uniti, Barack Obama, il quale, nel 2009, ha emanato una direttiva sull'Open Government nella quale si legge: *“Fin dove possibile e sottostando alle sole restrizioni valide, le agenzie devono pubblicare le informazioni on-line utilizzando un formato aperto (open) che possa cioè essere recuperato, soggetto ad azioni di download, indicizzato e ricercato attraverso le applicazioni di ricerca web più comunemente utilizzate. Per formato open si intende un formato indipendente rispetto alla piattaforma, leggibile dall'elaboratore e reso disponibile al pubblico senza che sia impedito il riuso dell'informazione veicolata”*.

Come conseguenza pratica ad essa è stato messo in esercizio il sito *www.data.gov*, il portale americano creato con lo scopo di diventare l'unico punto d'accesso e di riferimento per tutte le informazioni pubbliche prodotte dal governo.

Il servizio permette alle aziende, ai singoli e alle altre istituzioni di prelevare direttamente i dati grezzi oppure di estrarli tramite la mediazione di strumenti,

tipicamente applicazioni o siti web; una sezione particolare è riservata ai geodata pubblici, che contengono informazioni geografiche.

L'Open Government ha l'obiettivo primario di diminuire la distanza fra i cittadini e l'apparato pubblico mettendo al primo posto il tema della trasparenza; come secondo effetto, la stessa amministrazione scommette su un consistente risparmio di fondi pubblici e, soprattutto su un nuovo slancio economico indotto dalle opportunità di utilizzo dei dati da parte delle aziende; infine diventa evidente come si favorisca in questo modo lo sviluppo e la diffusione di applicazioni leggere (widget, mashup, ecc.) attraverso l'uso intelligente dell'open source, dei social media e del cloud computing.

La Finlandia e il Regno Unito (UK) hanno dato seguito alla strada intrapresa dall'amministrazione Obama e, successivamente, molti altri stati hanno fatto lo stesso, Italia inclusa; qui di seguito vi è un elenco [1] dei maggiori siti governativi contenenti Open Data attualmente attivi:

- *data.gov*, sito dell'open government degli Stati Uniti, Maggio 2009;
- *data.gov.uk*, sito dell'open government del Regno Unito, Settembre 2009;
- *data.norge.no*, sito dell'open government della Norvegia, Aprile 2010;
- *opengovdata.ru*, OpenGovData Russia Catalog, lanciato nel 2010;
- *data.gov.au*, sito dell'open government dell'Australia, Marzo 2011;
- *data.gc.ca*, sito dell'open government del Canada, Marzo 2011;
- *opendata.go.ke*, sito dell'open government del Kenia, Luglio 2011;
- *data.overheid.nl*, sito dell'open government dell'Olanda;
- *data.govt.nz*, iniziativa della Nuova Zelanda di pubblicare i dati del Governo sotto una licenza del tipo Creative Commons;
- *datos.gob.cl*, sito dell'open government del Cile, Settembre 2011;
- *data.gov.it*, sito dell'open government dell'Italia, Ottobre 2011;
- *datos.gob.es*, sito dell'open government della Spagna, Ottobre 2011;
- *datos.gub.uy*, sito dell'open government dell'Uruguay, Novembre 2011;
- *data.gouv.fr*, sito dell'open government della Francia, Dicembre 2011;
- *dados.gov.br* (beta), sito dell'open government del Brasile, la versione Beta è del Dicembre 2011, mentre il sito completo è previsto per Aprile 2012;
- *www.opendata.ee*, sito dell'open government dell'Estonia;
- *dados.gov.pt*, sito dell'open government del Portogallo.

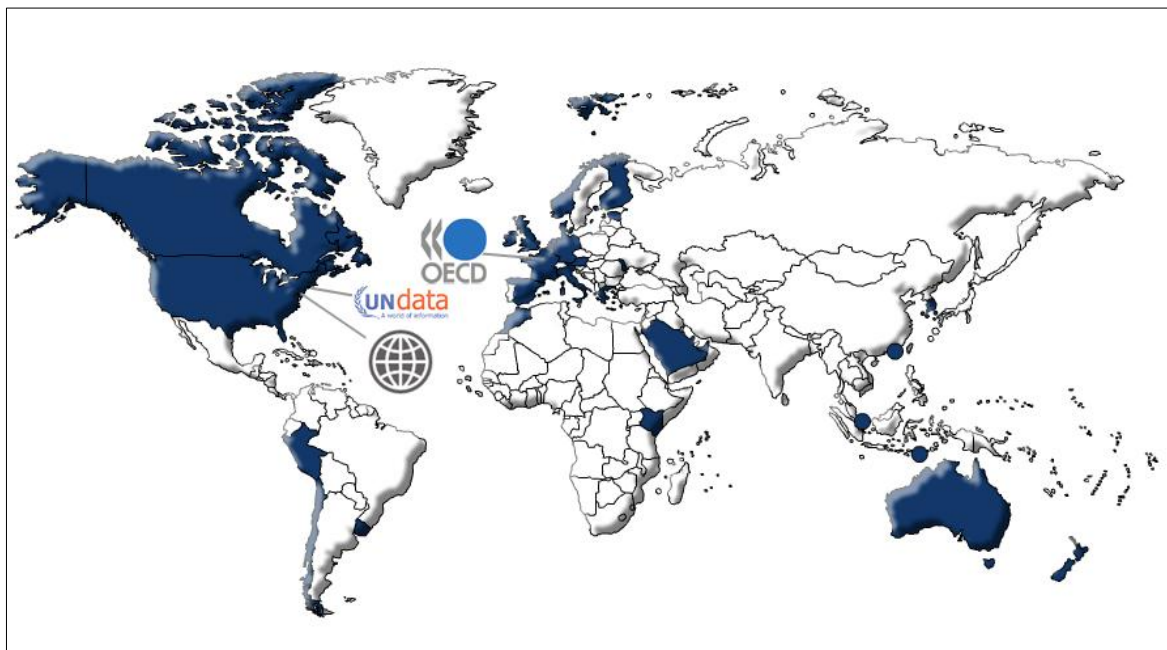


Figura 2: la distribuzione dei siti di open government nel mondo.

La pubblicazione e l'utilizzo di open data negli Stati Uniti, iniziata nel 2009, ha permesso in questi anni la raccolta di migliaia di dataset e la realizzazione di innumerevoli applicazioni; dal sito del Governo americano [3], *data.gov*, si possono infatti vedere le seguenti cifre:

- 390.831 sono i dataset raccolti, sia di dati in formato tabellare sia di dati in formato geospaziale;
- 1.231 sono le applicazioni create dal Governo americano;
- 236 sono, invece, le applicazioni sviluppate con gli open data messi a disposizione, dai cittadini;
- 85 sono le applicazioni sviluppate per i dispositivi mobili.

I dati disponibili sono scaricabili nei formati CVS o XLS, per quelli di tipo testuale o tabellare, nei formati KML o SHP (Shapefile), invece, per i cosiddetti *geodata*, ovvero quei dati contenenti mappe o altre informazioni geo-referenziate. Essi sono poi suddivisi per categoria a seconda dell'argomento che trattano; un quadro riassuntivo è rappresentato in Figura 3.

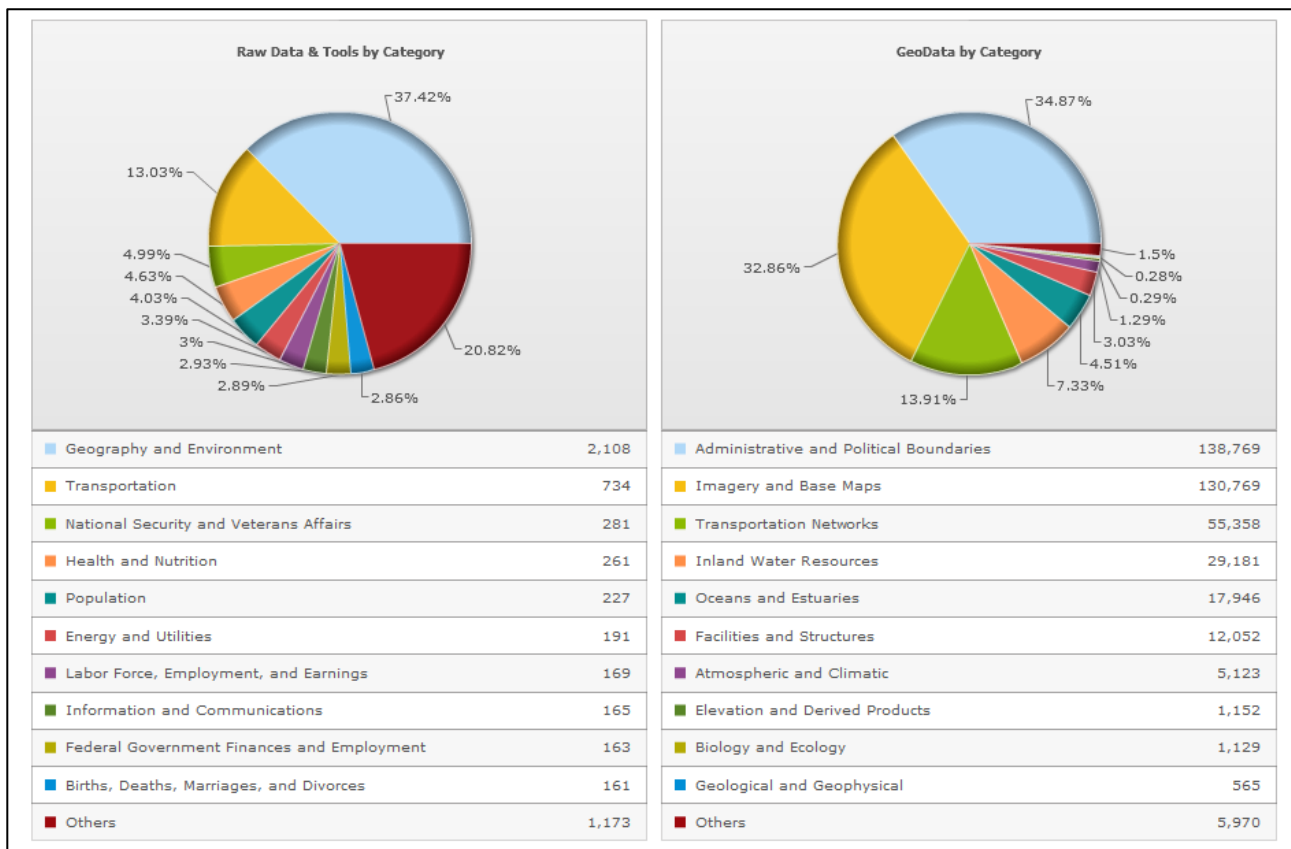


Figura 3: dataset del governo americano suddivisi per categoria.

Il sito *data.gov* ha inoltre eseguito un sondaggio per vedere quali altri dataset potrebbero essere utili ai cittadini e quindi resi disponibili al pubblico come open data: su 900 dataset suggeriti dal lancio del sito, si è calcolato che il 16% dei dati è già pubblicato sul sito, il 26% lo sarà prossimamente e il 36% lo potrà essere più avanti, e che solo il 22% dei dati proposti non potrà essere pubblicato per motivi di sicurezza, privacy o di vincoli tecnologici; la Figura 4 riassume queste percentuali.

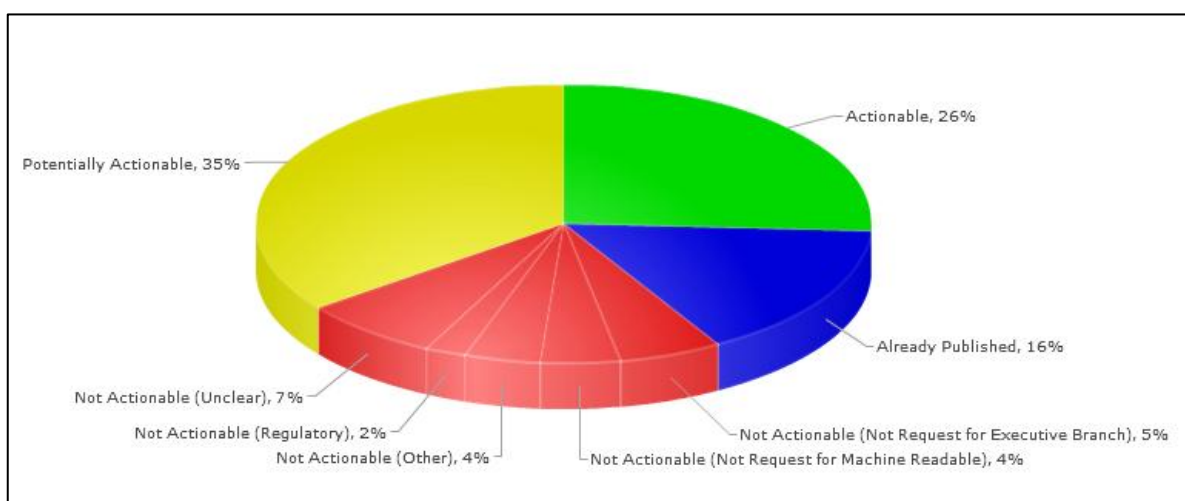


Figura 4: disponibilità dei dataset suggeriti su data.gov.

Un esempio a livello europeo, può essere il portale della città di Saragozza, Spagna, in particolare la sezione riguardante i dati ‘aperti’, <http://www.zaragoza.es/ciudad/risp/>. Esso presenta un catalogo di ben 106 dataset, riguardanti diverse tematiche della città, dal turismo al lavoro, dalla salute all’ambiente, di cui una buona percentuale è già in formato di Linked Data e quindi facilmente riutilizzabile per diverse applicazioni. La Figura 5 mostra la suddivisione dei vari dataset nei diversi formati: si può infatti notare come i dati in RDF siano più del 20% (sezioni in giallo) e che vi è anche una componente sostanziosa di dati in JSON, più dell’11% (in rosso).

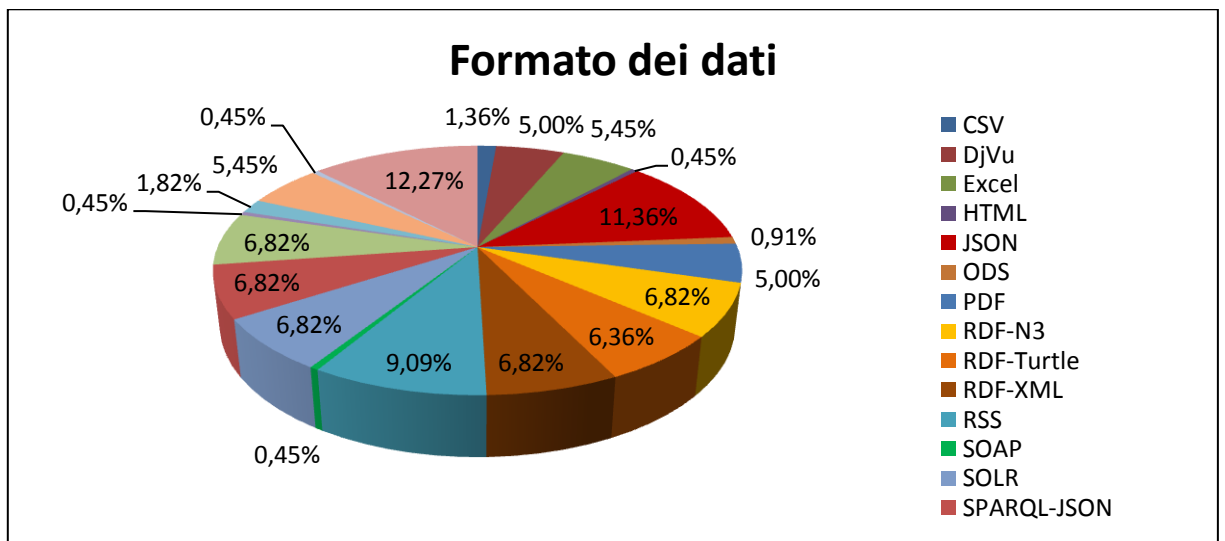


Figura 5: percentuali dei formati di dati nel sito <http://www.zaragoza.es/ciudad/risp/>.

Nel portale sono inoltre presenti diverse applicazioni e *widget* creati con l’utilizzo degli Open Data messi a disposizione, e vi è, inoltre, una sezione aperta all’inserimento continuo di nuove proposte ed *app* realizzate dagli utenti. Le applicazioni spaziano in tutte le diverse tematiche, dalle informazioni turistiche della città, a quelle relative alle manifestazioni religiose, o alle notizie sul traffico (l’elenco completo è disponibile al sito <http://www.zaragoza.es/ciudad/risp/aplicaciones.htm>).

E l’Italia? Per quanto riguarda l’Italia la situazione è meno avanzata, in quanto l’uso di Open Data è ancora un concetto recente, ma che sta avendo sempre più importanza e possibilità di sviluppo; il paragrafo successivo mostra una visione più approfondita dello scenario degli open data italiani.

2.1. Open Data in Italia

Il primo *data store* italiano [4] è stato quello della Regione Piemonte, dati.piemonte.it, all'interno del quale sono catalogati dati aperti riconducibili ai vari enti regionali (comuni, province,...). Dopo circa un anno dalla nascita del *data store* piemontese anche la regione Emilia-Romagna, ad ottobre 2011, pubblica online il suo catalogo di dataset (dati.emilia-romagna.it). Con il lancio del portale dati.gov.it, avvenuto il 18 ottobre 2011, si è aperta una nuova stagione per l'innovazione e la trasparenza nella Pubblica Amministrazione, una strada verso l'Open Data italiano.

Esso conta, attualmente, di ben 211 dataset di 32 diverse amministrazioni; i dati sono disponibili in diversi formati, come mostrato in Figura 6, e coprono diversi argomenti legati al territorio italiano, riassunti in Figura 7; utilizzando questi dati, sono state sviluppate, inoltre, innumerevoli applicazioni, soprattutto per dispositivi mobili, come iPhone o iPad, sia da parte della Pubblica Amministrazione che da parte di enti regionali, provinciali o comunali.

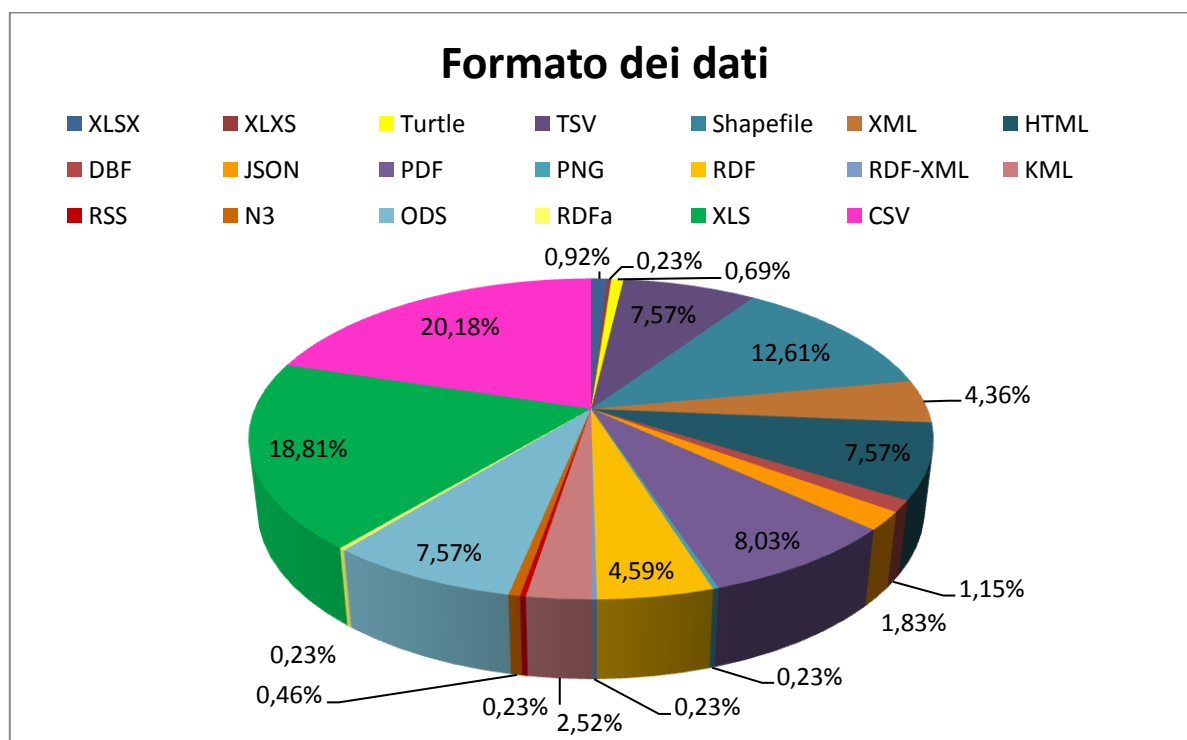


Figura 6: percentuali dei formati di dati nel sito dati.gov.it

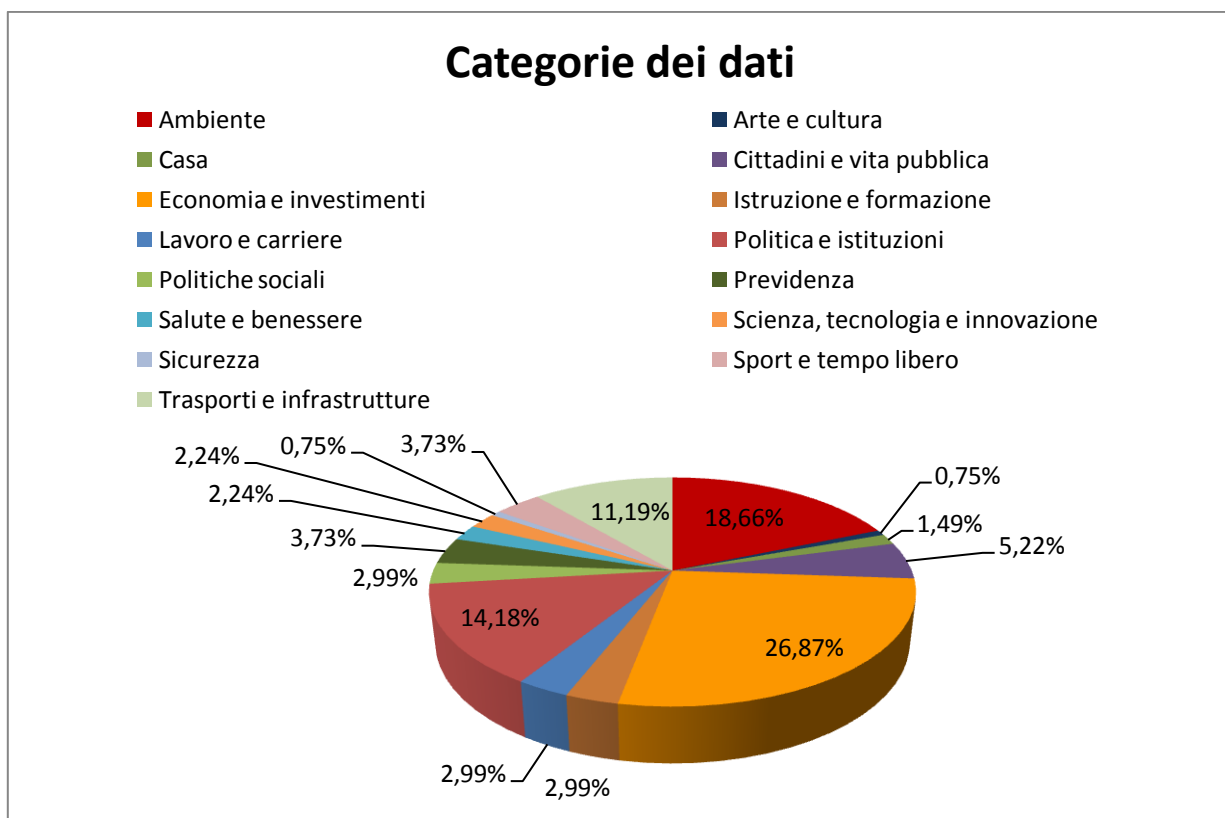


Figura 7: categorie dei dati nel sito *dati.gov.it*

Tra i portali regionali, il sito del Piemonte, *dati.piemonte.it*, è quello più avanzato in materia di Open Data. Come accennato precedentemente è il primo ad essere stato lanciato, e raccoglie ben 346 voci di dati scaricabili, soprattutto nei formati CVS, XLS e Shapefile. Mette inoltre a disposizione diverse applicazioni create integrando le informazioni fornite dai differenti dati: un esempio è ICONVIS²⁰, una *Rich Internet Application* che permette di rappresentare e navigare i dati in forma di ontologie, realizzata con il supporto del Politecnico di Torino.

Sono diversi i siti italiani che mettono a disposizione delle basi di dati di tipo ‘open’ ed un portale utile alla loro raccolta è *www.spaghettiopendata.org* [6]; esso è un sito gestito da cittadini italiani interessati al rilascio di dati pubblici, che può essere utilizzato dagli utenti come punto di raccolta dei link ai dataset aperti e degli strumenti utilizzati per organizzarli e distribuirli. Esso raccoglie i link di 50 basi di dati aperte che mettono a disposizione i propri dati prevalentemente nei formati CSV, XLS e TXT; raccoglie inoltre un elenco di ben 34 strumenti che possono essere utilizzati per elaborare le varie informazioni.

²⁰ <http://iconvis.polito.it/iconvis/>

Gli strumenti risultano essere di diverse tipologie sia per quanto riguarda l'aspetto tecnologico, suddividendosi in applicazioni desktop, servizi Web, framework server o client, sia per quanto riguarda quello che permettono di fare con i dati: vi sono strumenti per il refining dei dati come **Google Refine**²¹, **Data Wrangler**²² o **Document Cloud**²³; strumenti per la rappresentazione grafica dei dataset come **Gephi**²⁴, **Protovis**²⁵, **HiDE**²⁶ o **ManyEyes**²⁷, o per una visualizzazione dei dati geografici su mappe interattive come **MapServer**²⁸ e **Map4RDF**²⁹; vi sono inoltre applicazioni per la costruzione di linee del tempo dinamiche con i dati sequenziali come **Simile Timeplot**³⁰ o **TimeFlow**³¹, ed infine strumenti per il mash-up dei dati come **DERI Pipes**³², **Yahoo Pipes**³³ o **Managing News**³⁴, e la costruzione di pagine Web personalizzate come **Tableau Public**³⁵ o **Socrata**³⁶.

Successivamente alle regioni Piemonte ed Emilia-Romagna, sono stati attivati i portali regionali di Open Government delle regioni Lombardia e Veneto, rispettivamente dati.lombardia.it e dati.veneto.it, da cui possono essere scaricati i dati relativi a queste amministrazioni.

Oltre ai siti precedentemente citati, ve ne sono molti altri che possono essere usati come fonti di open data; qui di seguito proponiamo una lista [6] dei principali:

- *dbpedia.org*, mette a disposizione dati strutturati estrappolati da Wikipedia;
- *it.ckan.net*, raccoglie ben 253 diversi archivi di open data italiani;
- *dati.istat.it*, è la raccolta di dataset originali dell'Istat;

²¹ <http://code.google.com/p/google-refine/>

²² <http://vis.stanford.edu/wrangler/>

²³ <http://www.documentcloud.org/home>

²⁴ <http://gephi.org/>

²⁵ <http://mbostock.github.com/protovis/>

²⁶ <http://www.gicentre.org/hide/>

²⁷ <http://www-958.ibm.com/software/data/cognos/manyeyes/>

²⁸ <http://mapserver.org/index.html>

²⁹ <http://mccarthy.dia.fi.upm.es/map4rdf/>

³⁰ <http://www.simile-widgets.org/timeplot/>

³¹ <https://github.com/FlowingMedia/TimeFlow/wiki/>

³² <http://pipes.deri.org/>

³³ <http://pipes.yahoo.com/pipes/>

³⁴ <http://managingnews.com/>

³⁵ <http://www.tableausoftware.com/public/>

³⁶ <http://www.socrata.com/>

- *biennaledemocrazia.it/dataset*, è un esempio di dataset cittadini, legati alla città di Torino in questo caso, che dà un'idea di cosa dovremmo avere disponibile per ogni comune;
- *scraperwiki.com*, è una community in cui ci si aiuta nel recuperare open data quando questi non vengono messi direttamente a disposizione delle persone;
- *evpsi.org*, è un progetto torinese per il riutilizzo degli open data;
- *linkedopendata.it*, è centrato sulla connessione degli open data fra di loro.

L'ultimo punto anticipa l'impiego che sta diventando sempre più frequente degli open data, ovvero la loro connessione e di conseguenza, il loro utilizzo come Linked Open Data; questo argomento sarà approfondito maggiormente nel paragrafo successivo.

2.2. Open Linked Data

Come visto precedentemente spesso i 'dati aperti' non permettono il loro completo riutilizzo perché pubblicati sul Web in diversi formati: i dati strutturati sono spesso resi disponibili con i forms o integrati nelle pagine HTML, altri dati sono invece pubblicati come tabelle CSV, fogli di calcolo Excel, o in molti altri formati per domini specifici, altri ancora, invece, hanno iniziato a consentire l'accesso diretto alle loro banche dati via Web API specializzate.

Quale diventa, allora la motivazione per l'adozione degli Open Linked Data, ovvero la pubblicazione dei dati sottoforma di dati collegati e collegabili tra loro? Principalmente perché i Linked Data forniscono un paradigma di pubblicazione più generico, più flessibile e che facilita gli utenti nella scoperta e nell'integrazione di dati provenienti da un gran numero di fonti diverse.

In particolare, i vantaggi dei Linked Data [8] possono essere così riassunti, in quanto forniscono:

- un modello unificante per i dati; i Linked Data usano come unico formato di dati il linguaggio RDF, il quale è stato appositamente progettato per la condivisione globale dei dati, al contrario, gli altri metodi per la pubblicazione di dati sul Web, basandosi su una varietà di modelli, devono poi compensare questa eterogeneità risultante nel processo di integrazione;
- un meccanismo standardizzato di accesso ai dati; i Linked Data sono legati all'utilizzo del protocollo http, il quale consente l'accesso alle fonti di dati

- tramite browser generici e la scansione di esse con l'uso di motori di ricerca; le Web API, invece, sono rese accessibili solo tramite interfacce proprietarie;
- una discovery dei dati di tipo hyperlink-based; i Linked Data usano URI come identificatori globali per le diverse entità e consentono quindi la creazione di collegamenti ipertestuali tra entità in diverse fonti di dati, creando così un unico spazio globale di dati e permettendo alle applicazioni di scoprire nuove fonti di dati a run-time; al contrario, le Web API usate come depositi di dati in formati proprietari, restano contenitori di dati isolati;
- dati auto-descrittivi; i Linked Data facilitano l'integrazione dei dati provenienti da fonti diverse perché basati su vocabolari condivisi, le cui definizioni sono recuperabili e i cui termini sono collegati fra loro da link.

Tim Berners-Lee ha definito una classifica a 5 stelle [9], che descrive il livello di riusabilità dei dati pubblicati, in modo che i vari editori possano dare una valutazione, utile ai fine del riutilizzo dei dati, al proprio dataset in base ai seguenti criteri:

- *1 Stella = On the Web and Open License* ; i dati sono disponibili sul web (qualsiasi formato), ma con una licenza aperta;
- *2 Stelle = Machine-readable data*; i dati sono disponibili come dati strutturati e machine-readable (ad esempio, con Microsoft Excel, piuttosto di una scannerizzazione di una tabella);
- *3 Stelle = Non-proprietary format*; i dati sono disponibili come nel punto (2) ma in un formato non proprietario (ad esempio, CSV al posto di Excel);
- *4 Stelle = RDF Standards*; i dati sono disponibili come nei punti precedenti, ma con in più l'uso di standard aperti del W3C per identificare le entità (RDF e SPARQL), in modo da permettere agli utenti la creazione di link;
- *5 Stelle = Linked RDF*; i dati sono disponibili come in tutti i punti precedenti, ma con in più la possibilità di creare link verso dati di altre persone in modo da crearne un contesto.



Questa classifica rappresenta più che altro la progressiva transizione che si dovrebbe eseguire sui dati in modo da trasformarli in Linked Data a tutti gli effetti.

Un numero sempre maggiore di individui e organizzazioni stanno adottando i Linked Data come modo per pubblicare i loro dati, ovvero non li pongono solamente sul Web, ma li collegano tra loro in modo da creare un spazio di dati sempre più vasto e globale, il cosiddetto *Web of Data* [10].

Il Web dei dati, può essere visto come un grafico gigante e globale costituito da miliardi di dichiarazioni RDF, provenienti da fonti diverse, che coprono tutti i tipi di argomenti, quali la geografia, le persone, la società, i libri, le pubblicazioni scientifiche, i film, la musica, i programmi televisivi e radiofonici, i geni, le proteine, i farmaci e le sperimentazioni cliniche, i dati statistici, i risultati dei censimenti, le comunità online e le recensioni.

Il Web of Data, può essere visto come un ulteriore livello strettamente intrecciato a quello del Web classico, formato dai normali documenti, e che ha molte delle stesse proprietà:

- Il Web of Data è generico e può contenere qualsiasi tipo di dato;
- Chiunque può pubblicare dati nel Web of Data;
- Il Web of Data è in grado di rappresentare le informazioni discordanti e contraddittorie su un soggetto;
- Le entità sono collegate da link RDF, creando un grafico di dati globale che collega le diverse fonti e che consente di scoprirne di nuove; ciò significa che le applicazioni non devono essere costruite su un insieme fisso sorgenti di dati, ma possono scoprire nuove fonti in fase di esecuzione, proprio seguendo i link RDF;
- Coloro che pubblicano i dati non sono vincolati nella scelta dei vocabolari con cui rappresentarli;
- I dati sono auto-descrittivi; se un'applicazione che utilizza Linked Data incontra dati descritti con un vocabolario sconosciuto, può automaticamente dereferenziare gli URI presenti nelle terne RDF che ne identificano i termini, per trovarne la loro definizione;
- L'uso standardizzato di HTTP come meccanismo di accesso dati e di RDF come modello dati, semplifica l'accesso ai dati rispetto alle Web API, che si basano, invece, su modelli di dati e di interfacce di accesso, eterogenei.

2.2.1. LOD Project

Il concetto di Web of Data nasce nella comunità di ricerca sul Semantic Web e in particolare ha origine con il progetto Linking Open Data (LOD), attività della W3C fondata nel Gennaio 2007. L'obiettivo fondamentale del progetto, che ha generato una vivace e crescente comunità di Linked Data, è stato quello di, in qualche modo, ricavare il Web of Data, attraverso l'individuazione dei dataset attualmente disponibili con licenze aperte, la loro conversione in RDF, secondo i principi dei Linked Data, e la conseguente pubblicazione sul Web.

Il progetto è sempre stato aperto a tutti coloro che pubblicano i dati secondo i principi dei Linked Data, e questo è stato un fattore di successo, soprattutto nell'avvio automatico del progetto.

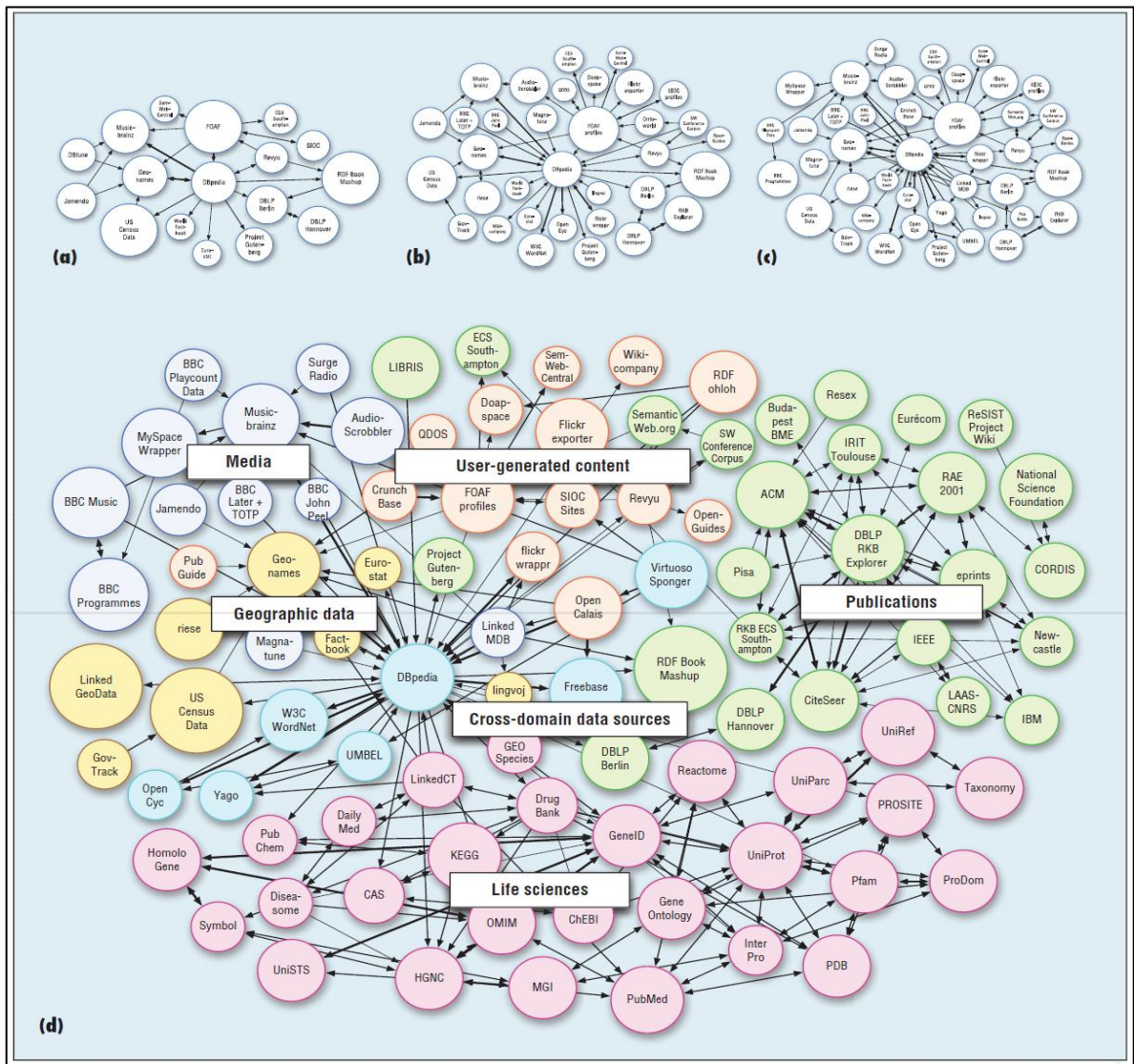


Figura 8: Crescita della Linking Open Data Cloud

La Figura 8 dimostra la crescita del numero di Linked Data e come sono collegati sul Web, dall'inizio del progetto nel Luglio 2007 (a), nell'Aprile 2008 (b), nel Settembre 2008 (c) e nel Luglio 2009(d): ogni nodo del diagramma rappresenta un insieme di dati distinto, pubblicato come Linked Data; gli archi indicano l'esistenza di collegamenti tra gli elementi dei due insiemi di dati; allo spessore degli archi corrisponde il numero di collegamenti, mentre la direzione degli archi specifica la struttura del link.

La Figura 9, mostra lo stato della Linking Open Data Cloud aggiornata a Settembre 2011: i colori classificano i dataset a seconda dei diversi domini evidenziando le diversità tra i dati; il grafico è disponibile sul sito <http://lod-cloud.net>, in cui viene periodicamente aggiornato in base al LOD Cloud Data Catalog³⁷, catalogo dei dataset open-license gestito dalla CKAN (Comprehensive Knowledge Archive Network).

Dal sito <http://lod-cloud.net/state/>, è inoltre possibile osservare le statistiche riassuntive calcolate in base al catalogo del CKAN; dalla Tabella 1, si ottiene un'idea generale sui tipi di dati e sulla quantità di dataset e triple RDF presenti ad oggi nel LOD Project: i dataset sono attualmente classificati in Geografia, Politica, Media, Pubblicazioni, Scienza, User-Generated Content (UGC) e Cross-Domain [11], e dalla tabella si ha una panoramica del numero di dati per ciascuno di questi domini.

Domain	Number of datasets	Triples	%	(Out-)Links	%
Media	25	1,841,852,061	5.82 %	50,440,705	10.01 %
Geographic	31	6,145,532,484	19.43 %	35,812,328	7.11 %
Government	49	13,315,009,400	42.09 %	19,343,519	3.84 %
Publications	87	2,950,720,693	9.33 %	139,925,218	27.76 %
Cross-domain	41	4,184,635,715	13.23 %	63,183,065	12.54 %
Life sciences	41	3,036,336,004	9.60 %	191,844,090	38.06 %
User-generated content	20	134,127,413	0.42 %	3,449,143	0.68 %
	295	31,634,213,770		503,998,829	

Tabella 1: numero dei dataset, delle triple e dei link RDF divisi per dominio.

³⁷ <http://www.ckan.net/group/lodcloud>

- **Media**

Un importante fonte di Linked Data nel settore dei media è la British Broadcasting Corporation (BBC)³⁸: i siti dei programmi e della musica della BBC, forniscono dati relativi a episodi di programmi TV e radio. I dati vengono interconnessi con MusicBrainz, un database musicale open-license, e con DBpedia, una versione linked-data di Wikipedia, consentendo alle applicazioni di recuperare e combinare i dati.

Altre compagnie di media che hanno annunciato l'intenzione di pubblicare i dati in forma di Linked Data sono il New York Times³⁹, CNET e Thomson Reuters, il quale ha inoltre sviluppato OpenCalais⁴⁰, un servizio di annotazione per i testi delle notizie con gli URI della Linked Data Open Cloud, in modo da creare riferimenti a luoghi, aziende e persone.

- **Pubblicazioni**

La US Library of the Congress e la German National Library of Economics pubblicano le intestazioni dei soggetti delle loro tassonomie come Linked Data. Altri Linked Data sulle pubblicazioni scientifiche sono disponibili dal L3S Research Center⁴¹, che ospita una versione linked-data della bibliografia DBLP. Il progetto ReSIST pubblica e collega fra di loro database bibliografici tra cui l'IEEE Digital Library, CiteSeer e vari archivi istituzionali.

Il wrapper delle API di Amazon e Google Base, RDF Book Mashup⁴², invece, fornisce dati riguardanti i libri, e, infine, l'Open Archives Initiative ha basato il suo nuovo standard, Object Reuse and Exchange (OAI-ORE)⁴³, sui principi dei Linked Data, dando la possibilità di accelerare ulteriormente la creazione di collegamenti tra dati relativi alle pubblicazioni.

³⁸ <http://www.bbc.co.uk/>

³⁹ <http://data.nytimes.com/>

⁴⁰ <http://www.opencalais.com/>

⁴¹ <http://dblp.l3s.de/>

⁴² <http://www4.wiwiw.fu-berlin.de/bizer/bookmashup/>

⁴³ http://version1.europeana.eu/c/document_library/get_file?uuid=9783319c-9049-436c-bdf9-25f72e85e34c&groupId=10602

- **Scienza**

Uno dei principali fornitori di Linked Data relativi alla scienza, in particolare alla Biologia, è il progetto Bio2RDF⁴⁴, il quale collega più di 30 dataset biologici ampiamente utilizzati, tra cui UniProt (Universal Protein Resource), KEGG (Kyoto Encyclopedia of Genes and Genomes), CAS (Chemical Abstracts Service), PubMed e Gene Ontology: complessivamente, il set di dati di Bio2RDF comprende più di due miliardi di triple RDF.

Inoltre, all'interno del W3C Linking Open Drug Data⁴⁵, le società farmaceutiche Eli Lilly, AstraZeneca e Johnson & Johnson, stanno cooperando per collegare tra loro i dati 'open' relativi ai farmaci e alle sperimentazioni cliniche per facilitarne la ricerca.

- **Geografia**

Il maggior fornitore di dati geografici è Geonames⁴⁶, un database geografico open-license, che pubblica e collega i dati di circa otto milioni di località. Un altro è il progetto LinkedGeoData⁴⁷, che pubblica una versione linked-data di OpenStreetMap, fornendo informazioni sulle caratteristiche spaziali di più di 350 milioni di elementi: le località in Geonames e LinkedGeoData sono interconnessi utilizzando elementi corrispondenti corrispondenti in DBpedia.

L'Ufficio di British Ordnance Survey⁴⁸ ha iniziato, inoltre, a pubblicare informazioni topologiche sulle aree amministrative del Regno Unito, come Linked Data, e sono anche disponibili versioni linked-data dei dataset di Eurostat⁴⁹, World Factbook⁵⁰ e US Census⁵¹.

- **UGC**

Una quantità crescente di metadati derivati dai contenuti generati dagli utenti sta diventando disponibile come Linked Data: esempi sono il wrapper di Flickr⁵² per il servizio di photosharing, il sistema di gestione dei contenuti di Drupal⁵³ e le bacheche

⁴⁴ <http://bio2rdf.org/>

⁴⁵ <http://esw.w3.org/HCLSIG/LODD>

⁴⁶ <http://www.geonames.org/>

⁴⁷ <http://linkedgeodata.org/About>

⁴⁸ <http://data.ordnancesurvey.co.uk/>

⁴⁹ <http://ckan.net/package?q=eurostat&groups=lodcloud>

⁵⁰ <http://www4.wiwiw.fu-berlin.de/factbook/>

⁵¹ <http://www.rdfabout.com/demo/census/>

⁵² <http://www4.wiwiw.fu-berlin.de/flickrwrapp/>

⁵³ <http://drupal.org/>

di phpBB⁵⁴. Zemanta⁵⁵ fornisce strumenti per l'arricchimento semiautomatico dei post del blog, e Faviki⁵⁶ propone un servizio per annotare il contenuto del Web con gli URI dei Linked Data, collegando il Web classico con il Web of Data.

- **Cross-domain**

Le fonti dei dati dette 'cross-domain', sono quelle che forniscono informazioni che coprono più domini; esse sono cruciali per il collegamento dei dati in un unico spazio globale, evitando così la creazione di isole di dati separate le une dalle altre.

Un esempio di tale sorgente è DBpedia⁵⁷, che pubblica i dati estratti dalle 'InfoBoxes', comunemente visibili sul lato destro degli articoli di Wikipedia: essa copre una gamma molto varia di argomenti e ha un elevato grado di sovrapposizione concettuale con altri insiemi di dati, fattore che ha reso DBpedia uno dei i nodi centrali nella Linking Open Data Cloud, condizionando la maggior parte dei fornitori di dati nell'impostare i collegamenti dalle proprie sorgenti dati fino ad essa.

Una seconda fonte importante di dati di tipo 'cross-domain' è Freebase⁵⁸, un database open-license che gli utenti possono modificare con modalità simile a quella utilizzata per Wikipedia.

Ulteriori ontologie 'cross-domain' che mettono a disposizione i Linked Data sono WordNet⁵⁹, OpenCyc⁶⁰, YAGO⁶¹ e UMBEL⁶²: esse si interconnettono con DBpedia, in modo da permettere alle applicazioni il mashup dei dati proveniente da tutte queste fonti.

- **E-Government**

I governi stanno cominciando a capire il potenziale dei Linked Data per le applicazioni del settore pubblico: come accennato in precedenza, sia le amministrazioni statunitensi che quelle britanniche hanno intrapreso le iniziative *data.gov* e *data.gov.uk* pubblicando volumi significativi di dati in RDF. L'amministrazione Obama ha puntato più sulla conversione di grandi volumi di dati, mentre nel Regno Unito ci si è concentrati sulla

⁵⁴ <http://www.phpbb.com/>

⁵⁵ <http://www.zemanta.com/>

⁵⁶ <http://www.faviki.com/>

⁵⁷ <http://dbpedia.org/>

⁵⁸ <http://www.freebase.com>

⁵⁹ <http://wordnet.princeton.edu/>

⁶⁰ <http://sw.opencyc.org/>

⁶¹ <http://www.mpi-inf.mpg.de/yago-naga/yago/>

⁶² <http://umbel.org/>

creazione di un'infrastruttura a livello dati, per la pubblicazione dei Linked Data, formata da URI stabili ai quali è più facile collegare una quantità crescente di i dati.

Un'iniziativa molto interessante è stata portata avanti dallo UK Civil Service⁶³, che ha iniziato a marcare i posti di lavoro vacanti utilizzando con la sintassi RDFa, fornendo informazioni sui posti disponibili attraverso form strutturati, in modo da facilitare la distribuzione delle informazioni per i portali di lavoro.

Dal sito *data.gov.uk* si può, infine, accedere ad una lista di dataset pubblicati come Linked Data e alle risorse che possono facilitarne l'utilizzo: essi sono divisi per settori e riguardano principalmente il lavoro del governo, le strutture organizzative, i parlamentari e i servizi come i trasporti pubblici; inoltre a ciascuno di essi è associato uno o più endpoint SPARQL, che permettono l'esplorazione dei dataset.

Anche in Italia ci si sta muovendo verso la conversione dei semplici 'dati aperti' in Linked Open Data, grazie al lavoro della Pubblica Amministrazione sia a livello nazionale che regionale; il paragrafo successivo mostra una panoramica più completa sui Linked Data italiani.

⁶³ <http://www.civilservice.gov.uk/>

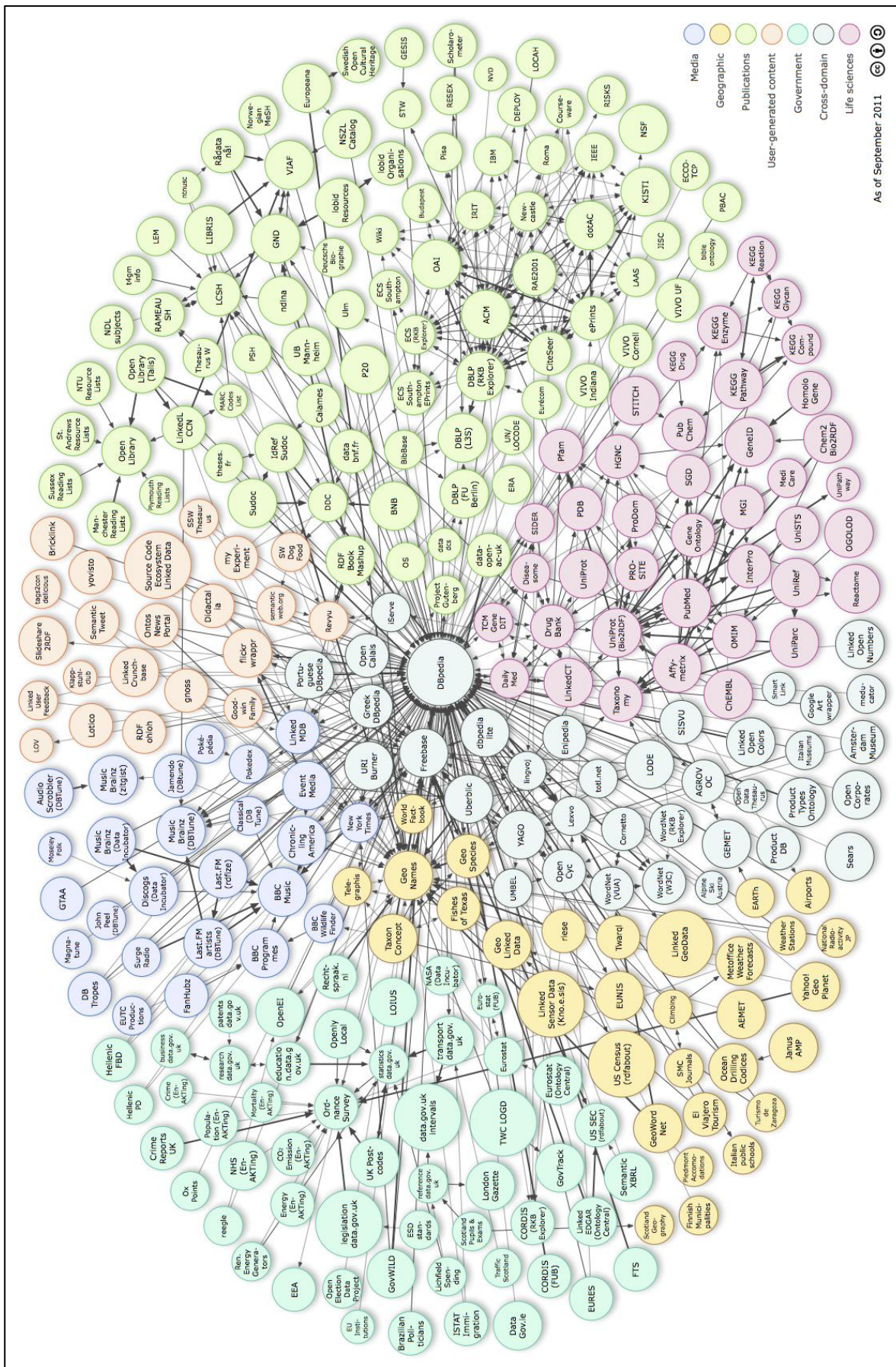


Figura 9: Linking Open Data Cloud, di Richard Cyganiak e Anja Jentzsch, Sett. 2011.

2.2.2. LATC, LOD Around-The-Clock

Con la crescita esplosiva della nuvola di Linked Open Data, la quantità e la qualità dei legami tra le fonti rimane però rada e priva di una ben documentata libreria software per l'accesso e la pubblicazione dei dati durante il loro ciclo di vita. Il progetto, o meglio la Specific Support Action LATC, Linked Open Data Around-The-Clock [16], si pone come obiettivo, nel contesto del programma FP7 ICT (Seventh Framework Programme for Information and Communication Technologies), di sostenere le persone e le organizzazioni nella pubblicazione e nell'utilizzo dei Linked Open Data.

L'iniziativa europea ha avuto inizio nel Settembre 2010; essa è coordinata dalla Digital Enterprise Research Institute (DERI)⁶⁴, Irlanda, e vanta partner molto importanti come la Vrije Universiteit Amsterdam (VUA)⁶⁵, Olanda, la Freie Universität Berlin (FUB)⁶⁶ e l'Institute for Applied Informatics e.V. (InfAI)⁶⁷, Germania, e Talis Information Ltd (TALIS)⁶⁸, del Regno Unito.

Con LATC, si affrontano, quindi i problemi relativi ai Linked Open Data proponendo diversi strumenti di supporto: per aumentare il numero, la qualità e l'accuratezza dei legami tra dataset, LATC mette a disposizione una piattaforma per l'interconnessione 24/7, sempre attiva; per supportare le parti interessate nella pubblicazione e nel consumo dei Linked Data, LATC pubblica e mantiene una libreria chiamata, *Publication & Consumption Tools Library*, composta da diversi strumenti, tutorial e descrizioni relativi all'utilizzo dei dati; per fornire un approfondito banco di prova per le applicazioni, infine, LATC pubblica in formato di Linked Data, dati prodotti dalla Commissione Europea, dal Parlamento Europeo e dalle altre istituzioni europee.

L'infrastruttura del progetto [17] si basa su due parti distinte, una parte interna, accessibile solo ai membri del consorzio LATC o alle persone affiliate al progetto, il cosiddetto *LATC BaseCamp*⁶⁹, e una parte pubblica, il sito Web, <http://latc-project.eu>, a cui tutti possono accedere e contribuire con possibili feedback.

La prima si basa appunto su BaseCamp, un software web-based per la gestione dei progetti, accessibile via Web browser e da dispositivi mobili, che permette di pianificare

⁶⁴ <http://www.deri.ie/>

⁶⁵ <http://www.few.vu.nl/en/index.asp>

⁶⁶ <http://www.wiwiss.fu-berlin.de/>

⁶⁷ <http://infai.org/en/Welcome>

⁶⁸ <http://www.talis.com/platform/>

⁶⁹ <https://lidrc.basecamphq.com/projects/4313391>

compiti, step, obiettivi e condividere documenti; esso offre inoltre una varietà di meccanismi di notifica ed è ben noto per essere di semplice utilizzo.

La seconda invece, è appunto composta dal sito web, implementato con il Content Management System (CMS) Drupal 7⁷⁰, in modo da pubblicare le varie informazioni in modo leggibile sia dalle persone che dalle macchine; esso si compone di quattro sezioni: *24/7 Platform*, la piattaforma di interconnessione, che permette di generare link RDF tra i dataset presenti nella LOD Cloud; *Tools*, la libreria di strumenti open-source, a supporto degli utenti che utilizzano i Linked Data, per tutte le fasi del processo di pubblicazione e consumo dei dati; *Datasets*, la lista dei dataset delle istituzioni europee utilizzabili per testare le varie applicazioni, pubblicati sottoforma di LOD Cloud Europea; ed infine, *Tutorial & Support*, sezione contenente i documenti riguardanti le istruzioni e le migliori pratiche in materia di pubblicazione ed utilizzo dei Linked Data.

Il programma LATC, è inoltre coinvolto in numerosi progetti internazionali [18] riguardanti i Linked Data per garantire un maggiore sviluppo della tematica, tra cui:

- *European Environment Agency*⁷¹: è l'agenzia dell'ambiente dell'Unione Europea (AEA), il cui compito consiste nel fornire informazioni valide e indipendenti sull'ambiente.
- *BlogForever*⁷², progetto FP7: ha l'obiettivo di sviluppare robuste strutture per la gestione, la conservazione e la diffusione dei weblog; strutture che siano in grado di catturare la natura dinamica e la continua evoluzione dei weblog, la loro rete e struttura sociale, e che favoriscano lo scambio di concetti e idee tra weblogger, in modo da arrivare alla creazione di un modello unico e generico di weblog.
- *LOD2*⁷³ (Creating Knowledge out of Interlinked Data), progetto FP7: è un progetto di integrazione su larga scala, co-finanziato dalla Commissione Europea, che coinvolge ben 15 partner provenienti da diversi paesi europei e non; esso si prefigge di continuare il lavoro, ed il successo, del progetto LOD arrivando a migliorare la coerenza e la qualità dei dati pubblicati sul Web, chiudendo così il divario tra la gestione dei dati relazionali e RDF; il progetto affronta quindi diverse sfide, portando avanti lo sviluppo di strumenti e

⁷⁰ <http://www.drupal.it>

⁷¹ <http://www.eea.europa.eu/>

⁷² <http://blogforever.eu/>

⁷³ <http://lod2.eu/>

metodologie enterprise-ready per la gestione e la pubblicazione di grandi quantità di informazioni strutturate sul Web, di una rete di testbed di alta qualità, multi-dominio e multi-linguaggio, che usa fonti come Wikipedia e OpenStreetMap, di algoritmi di *machine-learning* per l'interconnessione e la fusione automatica dei dati nel Web, di norme e standard che assicurino la provenienza, la sicurezza e l'affidabilità dei dati, nonché per valutare la qualità delle informazioni, ed infine, di strumenti per la ricerca, la navigazione e l'*authoring* dei Linked Data.

- *PlanetData*⁷⁴, progetto FP7: esso mira a creare una comunità europea di ricercatori in grado di supportare le organizzazioni nell'espone i loro dati in modi nuovi e utili, rendendo efficaci ed efficienti le enormi quantità di dati che sono pubblicati online in modo continuo.
- *SEALS*⁷⁵ (Semantic Evaluation at Large Scale), progetto FP7: esso sta sviluppando una infrastruttura di riferimento, la piattaforma SEALS, per facilitare la valutazione formale delle tecnologie semantiche.
- *SpitFIRE*⁷⁶ (Semantic-Service Provisioning for the Internet of Things using Future Internet Research by Experimentation), progetto FP7: questo progetto si pone all'interno della corrente che cerca di unificare i due rami principali del computing, il general-purpose e l'embedded, con lo sviluppo di applicazioni che individuino i concetti unificati, i metodi e le infrastrutture software necessarie.
- *DataLift project*⁷⁷, Francia: è un progetto di ricerca sperimentale finanziato dall'Agenzia di ricerca nazionale francese che ha come obiettivo lo sviluppo di una piattaforma per la pubblicazione e l'interconnessione dei dataset al Web of Data, trasformando dati grezzi in Linked Data.
- *Semic.EU*⁷⁸: progetto, migrato sulla piattaforma Joinup⁷⁹, che attraverso il programma di Interoperability Solutions for Public Administrations (ISA) offre una nuova serie di servizi per aiutare i professionisti di e-government a condividere la loro esperienza.

⁷⁴ <http://www.planet-data.eu/>

⁷⁵ <http://www.seals-project.eu/>

⁷⁶ <http://spitfire-project.eu/>

⁷⁷ <http://datalift.org/>

⁷⁸ <http://www.semic.eu/>

⁷⁹ <http://joinup.ec.europa.eu/>

2.2.3. LOD in Italia

Il movimento Open Data, soprattutto a livello governativo, come discusso prima, è in grande crescita nel mondo e sta muovendo i primi decisivi passi anche in Italia, anche grazie all'iniziativa *www.dati.gov.it*: l'obiettivo base è che le Pubbliche Amministrazioni 'liberino' i dati in loro possesso e li rendano disponibili ai cittadini e alle aziende, assicurando ai primi il diritto alla trasparenza ed offrendo ai secondi la possibilità di creare nuova ricchezza dal mashup di questi dati, ad esempio con la creazione di nuove applicazioni che migliorino l'efficienza di tutto l'apparato produttivo.

Per rendere i dati 'aperti' ci sono molti modi, e come visto in precedenza, la pubblicazione di essi nei formati tabellari standard (ad esempio Excel o CSV), è solo il primo passo: la nuova frontiera deve diventare il Semantic Web e, in particolare, i Linked Open Data.

I dati possono essere integrati nelle pagine HTML utilizzando anche, i cosiddetti *Microformat* (o *Microdata*)⁸⁰, che permettono di pubblicare dati strutturati rappresentanti entità specifiche; ma per essi, è necessario seguire modalità precise di pubblicazione, per rendere i dati automaticamente estraibili dalle applicazioni, ed inoltre, descrivono le entità con solo pochi attributi, senza poter stabilire relazioni tra di esse, risultando poco adatto alla condivisione di dati sul Web.

Utilizzando l'RDF, invece, i dati diventano reperibili tramite interrogazioni di motori semantici, dall'esterno e in real-time, senza alcuna intermediazione di formato o processando i dati in un secondo tempo, e le interrogazioni avvengono in maniera standard, facendo riferimento a vocabolari pubblici ed univoci (ontologie), riducendo la possibilità di errori o di fraintendimenti.

Le informazioni risultano trasparenti all'utente finale e allo stesso tempo analizzabili dalle macchine, e le pagine web, incorporando dati semantici, vanno così a comporre il grafo globale del Web, la LOD Cloud, diventando piccoli elementi di un unico database globale ed interrogabile.

LinkedPA [12] è uno dei progetti che si inserisce in questo nuovo filone del Linked Open Data portato avanti dalla Pubblica Amministrazione italiana. Esso offre una piattaforma open source per la Pubblica Amministrazione che consente di:

⁸⁰ <http://microformats.org/>

- inserire nel backend del sistema tutti gli atti amministrativi che vengono automaticamente mappati secondo l'ontologia della PA;
- gli atti amministrativi sono poi archiviati in un triple store e resi disponibili all'esterno tramite un endpoint SPARQL aperto;
- pubblicare i contenuti web incorporando automaticamente le informazioni semantiche codificate con lo standard RDFa;
- fare riferimento a un'ontologia creata dal basso per le esigenze di base della PA.

LinkedPA ambisce a essere un'iniziativa aperta che vuole rappresentare un punto di aggregazione e proposta per il decollo dell'Open Data in Italia: per questo motivo rende già disponibile a tutti, tramite l'apposita sezione nel sito www.ontologiapa.it/pa, sia l'ontologia della PA che le informazioni relative ad essa.

L'ontologia della PA può essere vista come un punto di partenza, inserito nella vasta nuvola delle ontologie pubbliche, con la prospettiva di crescere ed essere modificata per adattarsi alle diverse esigenze dei cittadini e del mondo produttivo: essa presenta materiale in forma di Linked Data, come si può vedere da Figura 10, riguardante bandi, concorsi, giunte, e sulle persone legate alle diverse delibere.

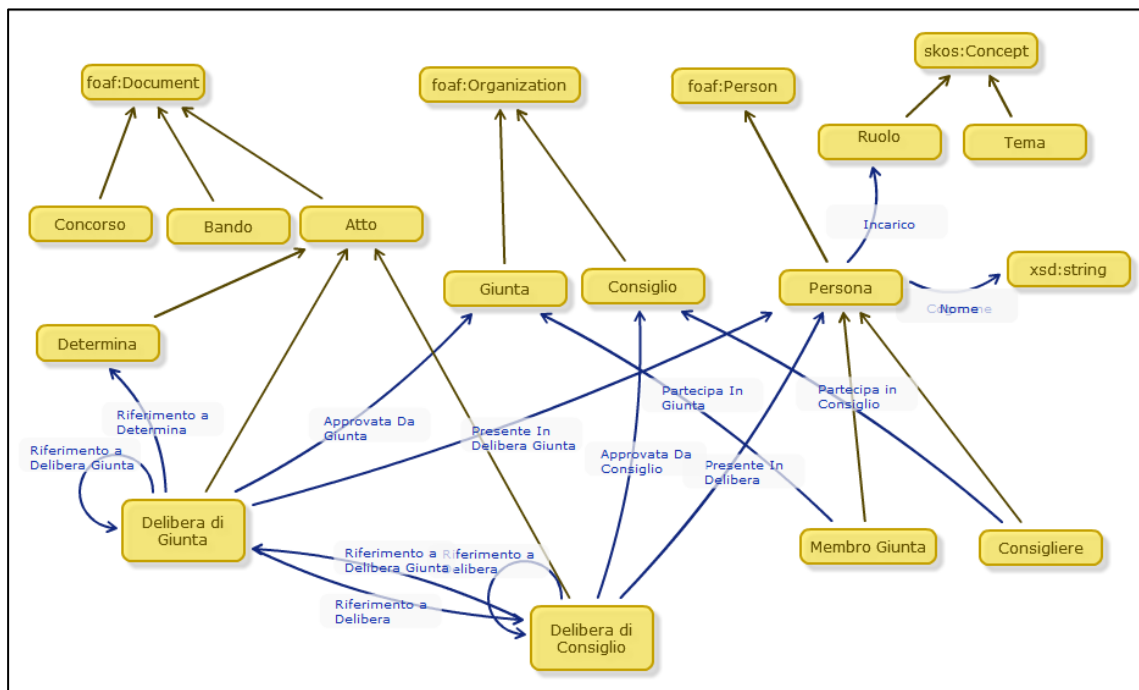


Figura 10: diagramma riassuntivo di OntologiaPA.

Uno dei primi esempi di connessione diretta tra la produzione degli atti amministrativi e la loro messa a disposizione in un'ottica Open Data senza intermediari, è avvenuta con la creazione del portale dell'amministrazione provinciale di Carbonia Iglesias, in Sardegna. Dal settembre 2011, è infatti on-line il portale semantico della provincia, www.provincia.carboniaiglesias.it, sviluppato con Drupal 7⁸¹, in cui sono stati mappati secondo l'OntologiaPA, precedentemente descritta, tutti gli atti amministrativi emanati dall'ente, permettendo così l'accesso in formato 'aperto' alle varie delibere.

L'iniziativa provinciale, come lo sviluppo del progetto LinkedPA, è stato affidato ad un team composto da tre aziende, Ablativ⁸², Linkalab⁸³ e Istos⁸⁴, che si sono occupate della catalogazione e dell'estrazione delle informazioni, basandosi sugli standard della W3C (RDF e SPARQL). La sezione contenente questi dati è <http://www.provincia.carboniaiglesias.it/albo-pretorio/>, mentre le interrogazioni possono essere eseguite dal punto di accesso SPARQL, <http://www.provincia.carboniaiglesias.it/sparql/>, che permette di eseguire query personalizzate e di scegliere il formato con cui visualizzare i vari risultati.

Questo primo esempio ha dimostrato come si sia stati in grado di conferire un valore aggiunto considerevole alle informazioni prodotte e pubblicate dall'ente, e si auspica venga visto come esempio da seguire.

Un altro fornitore di Linked Data è il portale della Camera dei Deputati [13]: esso mette a disposizione numerosi dataset contenenti le informazioni sulle persone, i mandati, le leggi e i riferimenti storici riguardanti la Camera, in diversi formati come RDF, n3, Turtle e JSON; dispone inoltre di una sezione apposita del sito contenente l'endpoint SPARQL specifico per la loro interrogazione.

Mostra infine alcune delle possibili applicazioni web costruite utilizzando i dati della Camera insieme ad altri elementi, come ad esempio i riferimenti spazio-temporali ad essi relativi; vi sono infatti, l'Atlante Istituzionale⁸⁵, mappa interattiva che mostra i luoghi di nascita dei presidenti dal Regno d'Italia da oggi, la Mappa delle Elezioni del

⁸¹ <http://www.drupal.it>

⁸² <http://www.ablattiv.org/>

⁸³ <http://www.linkalab.it/>

⁸⁴ <http://www.istos.it/>

⁸⁵ <http://dati.camera.it/apps/atlane-istituzionale/presidentiHome.html>

Regno d'Italia⁸⁶, applicazione che mostra la distribuzione geografica dei risultati delle elezioni fino alla prima Guerra Mondiale, ed il Portale Storico della Camera dei Deputati⁸⁷, che unisce i dati della camera agli archivi multimediali contenenti foto e video, costruendo così, diversi percorsi tematici.

Infine a livello regionale, è menzionabile il portale della regione Piemonte, *dati.piemonte.it*, il quale è stato sia il primo a trattare l'argomento dei dati 'aperti', come visto precedentemente, sia il primo a mettere a disposizione dataset in RDF, utilizzabili quindi come Linked Data. Le Tabella 2, mostra l'elenco degli identificativi URI degli oggetti al momento disponibili nella sezione apposita del portale regionale e una prima versione sperimentale del dizionario di Dati Piemonte relativo ai dati pubblicati in RDF; essi rappresentano l'inizio di un percorso di pubblicazione di una quantità via via crescente di informazioni secondo lo standard Linked Data.

Dataset	HTML	RDF
scuole del Piemonte	<i>scuole-piemonte.html</i>	<i>scuole-piemonte.rdf</i>
comuni del Piemonte	<i>comuni-piemonte.html</i>	<i>comuni-piemonte.rdf</i>
province del Piemonte	<i>province-piemonte.html</i>	<i>province-piemonte.rdf</i>
Dizionario	HTML	OWL
dati Piemonte ontology	<i>rponto.html</i>	<i>rponto.owl</i>

Tabella 2: Linked Data presenti sul sito <http://www.dati.piemonte.it/linked-data.html>

⁸⁶ <http://dati.camera.it/apps/elezioni/>

⁸⁷ <http://storia.camera.it/>

Capitolo 3

3. Strumenti per la pubblicazione di Linked Data

I dati sono presenti nel Web (e non) in diversi formati, e per renderli consultabili come Linked Data, sono necessarie diverse fasi di elaborazione.

Solitamente si ha prima una fase di *publication*, la quale segue uno specifico procedimento, dopodiché si ha la fase di utilizzo, o *consumption*: la pubblicazione dei Linked Data consiste nel mettere a disposizione sul web un determinato insieme di dati, il quale deve essere prima convertito dal formato originale al formato RDF, e deve quindi essere controllato attraverso operazioni di consolidamento, in modo da garantirne la conformità rispetto ai principi dei Linked Data (Paragrafo 2.2); il successivo consumo dei dati, invece, consiste nell'accedere ai dati, o ad un loro sottoinsieme, ed aggregarli o integrarli, al fine di realizzare un obiettivo specifico.

La prima fase del processo di divulgazione dei Linked Data, si suddivide principalmente in modellazione, pubblicazione e collegamento: nei prossimi paragrafi saranno analizzate in modo più approfondito.

3.1. Modelling

Per prima cosa è necessaria una fase di *modelling* [19], modellazione. Essa consiste nell'individuazione delle *risorse* da pubblicare, ovvero per prima cosa è necessario identificare quali sono gli elementi di interesse nel nostro dominio, aventi le proprietà e le relazioni che vogliamo descrivere. Le risorse si distinguono in *risorsa informativa* e *risorsa non-informativa*: questa distinzione è importante nel contesto dei Linked Data in

quanto tutte le informazioni presenti nel Web (documenti, immagini, ecc) sono da considerarsi risorse contenenti informazioni, ma spesso molti degli elementi che vogliamo condividere utilizzando i Linked Data sono dei cosiddetti ‘real-world object’ ovvero elementi che esistono anche al di fuori del Web, come persone, luoghi, prodotti e così via. Ogni risorsa sarà poi identificata con un HTTP URI, come visto nel Capitolo 1, il quale permetterà sia di identificarle in modo semplice ed univoco, sia di accedere ad esse.

Ciascuna risorsa avrà inoltre un particolare tipo di rappresentazione, ovvero il formato del proprio stream di byte (HTML, RDF/XML, JPEG, ecc): le risorse possono avere associati diversi tipi di rappresentazione. Questo risulta fondamentale nel processo detto di *URI Dereferencing*, ovvero il processo di ricerca dell’URI nel Web al fine di trovare più informazioni sulla risorsa referenziata. I due tipi di risorse si comportano diversamente durante questo procedimento: dereferenziando l’URI di una risorsa informativa, si ottiene, dal server contenente l’URI, una nuova rappresentazione della risorsa e del suo stato che viene restituita al client nella risposta HTTP avente codice 200 OK; le risorse non-informative, invece, non possono essere dereferenziate direttamente, ma si usa la cosiddetta redirezione, *303 redirect*, ovvero il server restituisce al client l’URI di una risorsa informativa che descrive la prima, utilizzando la risposta HTTP avente codice 303 See Other, e permette così al client di ottenere da quest’ultima URI, in un secondo momento, una rappresentazione della prima risorsa.

Un’alternativa alle *URI 303 redirect* sono le cosiddette *Hash URI* [8], le quali sono anch’esse utilizzate per identificare degli oggetti reali, ed evitano il doppio procedimento di richiesta per la dereferenziazione. Esse hanno una particolare struttura: l’URI contiene il simbolo cancelletto (#) che separa l’indirizzo base da una parte speciale detta *fragment identifier*: in questo modo il client sa già che l’URI contenente quel simbolo non è direttamente dereferenziabile, tronca l’indirizzo eliminando la parte dopo il cancelletto ed invia la richiesta al server che gli restituisce automaticamente, e istantaneamente, il documento che descrive quella risorsa. Gli Hash URI completi rappresentano un ottimo metodo per identificare in modo univoco ‘real-world object’ e concetti astratti senza creare ambiguità.

Nella richiesta HTTP, è possibile specificare in che formato si vuole ottenere la rappresentazione della risorsa, questa fase è detta *negoziazione* del contenuto:

nell'header della richiesta si indica il formato preferito, come RDF o HTML, che verrà poi restituito dal server con la redirectione; solitamente, infatti, per ogni risorsa non-informativa si memorizzano tre URI, il primo identifica il 'real-world object', il secondo fa riferimento ad una sua descrizione in RDF/XML e il terzo fa riferimento a quella in HTML. La Figura 11 mostra il processo di deferenza URI .

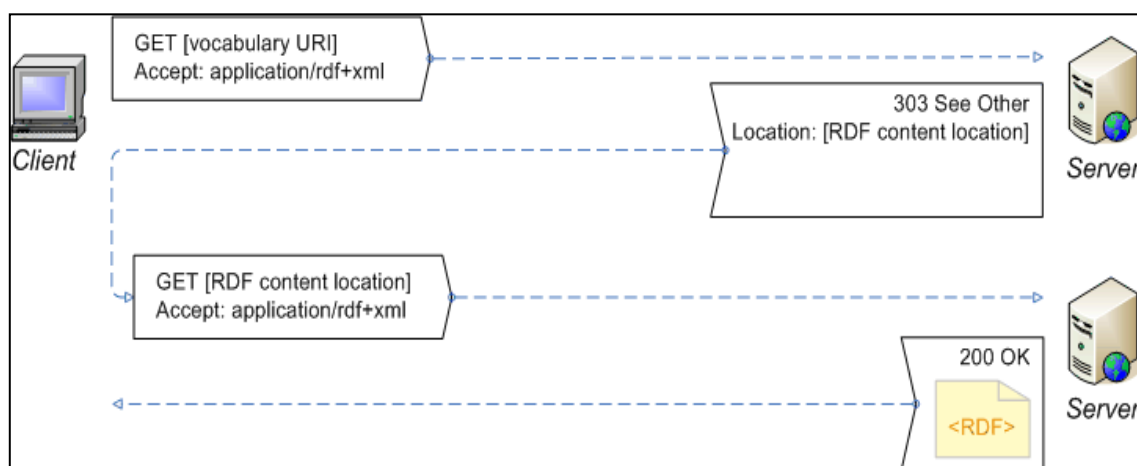


Figura 11: URI Dereferencing.

Ciò dimostra quanto sia importante la selezione degli URI per le varie risorse. Nella modellazione quindi, è necessario spendere molta attenzione nella loro scelta, in modo che gli altri pubblicatori possano facilmente utilizzarli per collegarvi i propri dati e che allo stesso tempo abbiano le caratteristiche tecniche adatte alla deferenza. Le regole per l'utilizzo delle URI possono essere così riassunte:

- Usare HTTP URI per tutto, essendo lo schema `http://` ampiamente supportato dai vari strumenti ed infrastrutture;
- Definire gli URI nel proprio namespace, e non in quello di altri, in modo da assicurare la loro deferenza;
- È meglio utilizzare nomi corti, semplici e mnemonici; inoltre, se le risorse hanno identificatori univoci significativi nel proprio dominio, è bene inserirli negli URI;
- Cercare di rendere i propri URI stabili e persistenti, in quanto modificarle in un secondo momento, rovinerebbe i link già stabiliti;
- La creazione di un URI rimane comunque limitata dal proprio ambiente tecnico, ad esempio dal nome del proprio server, può essere quindi utile aggiungere alcune regole di *URI rewriting* alla configurazione del proprio web server;

- Ad una singola non-information resource è bene associare tre URI distinti, uno per l'oggetto, uno per la sua rappresentazione HTML e uno per la sua rappresentazione RDF.

Dopo avere definito i vari URI, per la modellazione è fondamentale esprimere le relazioni tra di essi e, per fare ciò, si utilizzano dei vocabolari contenenti i concetti utili alla descrizione delle risorse e dei vari collegamenti che possono essere stabiliti tra di essi.

È conveniente utilizzare vocabolari a larga diffusione, in modo da favorire le applicazioni client nell'elaborazione dei dati. I vocabolari sono generalmente creati per adattarsi ad uno specifico settore da modellare; i più utilizzati sono:

- Friend-of-a-Friend (FOAF)⁸⁸, per descrivere persone;
- Dublin Core (DC)⁸⁹, che definisce attributi per metadata generici;
- Semantically-Interlinked Online Communities (SIOC)⁹⁰, vocabolario per rappresentare le online community;
- Description of a Project (DOAP)⁹¹, vocabolario utilizzato per descrivere progetti;
- Simple Knowledge Organization System (SKOS)⁹², per la rappresentazione di tassonomie, domini di conoscenza con strutture semplici;
- Music Ontology⁹³, fornisce termini per la descrizione di artisti, album e tracce;
- Review Vocabulary⁹⁴, vocabolario per la rappresentazione di recensioni;
- Creative Commons (CC)⁹⁵, vocabolario per la descrizione dei termini delle licenze.

Una lista più completa di vocabolari è mantenuta dalla community del progetto W3C SWEO Linking Open Data⁹⁶, nel quale sono presenti le varie sorgenti di dati da cui prelevare le URI riguardanti luoghi geografici, aree di ricerca, artisti, libri, CD o argomenti generali, in modo avere URI già deferenziabili, il cui concetto è ritrovabile

⁸⁸ <http://xmlns.com/foaf/0.1/>

⁸⁹ <http://dublincore.org/documents/dcmes-xml/>

⁹⁰ <http://sioc-project.org/>

⁹¹ <http://usefulinc.com/doap/>

⁹² <http://www.w3.org/2004/02/skos/>

⁹³ <http://musicontology.com/>

⁹⁴ <http://purl.org/stuff/rev>

⁹⁵ <http://creativecommons.org/ns>

⁹⁶ <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

nel Web, e che sono già collegate ad altre sorgenti, in modo da permettere una maggiore crescita della rete dei dati .

È pratica comune mescolare termini presi da vocabolari diversi, inoltre si raccomanda l'utilizzo di proprietà come *rdfs:label* e *foaf:depiction*, essendo termini ben supportati dalle applicazioni client.

Se però non si riescono a trovare vocabolari esistenti che coprono tutte le classi e le proprietà necessarie, allora è indispensabile definirne uno nuovo.

Per definire un nuovo vocabolario è consigliato avere una buona dimestichezza con il linguaggio RDF Schema, e si possono seguire le linee guida presenti nel RDF Vocabulary Description Language 1.0: RDF Schema⁹⁷; in sintesi è utile:

- Non definire il vocabolario dal nulla, ma completare quelli esistenti aggiungendo i termini mancanti;
- Fornire informazioni, sia per le macchine che per gli umani, non dimenticandosi di utilizzare le proprietà *rdfs:comment*, per aggiungere una descrizione ai termini inventati, o *rdfs:label*, per etichettarli correttamente;
- Rendere de-referenziabili gli URI dei termini, in modo da permettere la ricerca delle definizioni dei termini nel Web;
- Utilizzare termini di altre persone, o mapparli ad essi, in modo da accrescere il livello di *linking* tra i dati del Web, utilizzando ad esempio le proprietà *rdfs:subClassOf* e *rdfs:subPropertyOf*;
- Rendere esplicite tutte le informazioni importanti, in modo facilitare il lavoro alle macchine;
- Lasciare flessibilità e spazio di crescita al vocabolario, evitando l'utilizzo di modelli con troppi vincoli per facilitarne poi anche il riutilizzo da parte di altri utenti.

Un altro metodo per la modellazione dei dati, che supporti anche la creazione di nuovi vocabolari, è quello di avvalersi di strumenti che facilitino e automatizzino queste operazioni. Uno di questi strumenti è Neologism⁹⁸, una piattaforma per la redazione e la pubblicazione di vocabolari per il Web of Data, con una particolare attenzione per la compatibilità con i Linked Data Principles.

⁹⁷ <http://www.w3.org/TR/rdf-schema/>

⁹⁸ <http://neologism.deri.ie/>

3.1.1. Neologism

Neologism [20] permette la creazione di nuovi vocabolari, semplificando e standardizzando la creazione di classi e proprietà: esso è un semplice sistema web-based per l'editing e il publishing di vocabolari in RDF Schema. Il suo scopo principale è quello di ridurre drasticamente il tempo necessario alla creazione, pubblicazione o modifica dei vocabolari del Semantic Web.

È scritto in PHP e si basa su Drupal 6, piattaforma open-source per il management dei contenuti di milioni di siti Web. Neologism [22] è supporto per la creazione di sottoclassi, sottoproprietà, domini, range, proprietà dirette ed inverse, partendo dai principali vocabolari come FOAF o OWL; permette inoltre l'output sia in RDF/XML che in N3, la personalizzazione di diagrammi di classi, l'import di dati da file o dal Web e il mapping verso vocabolari esterni.

Per testare il funzionamento di questo strumento si è utilizzato un semplice database relazionale realizzato con MySQL, il cui schema è rappresentato in Figura 12.

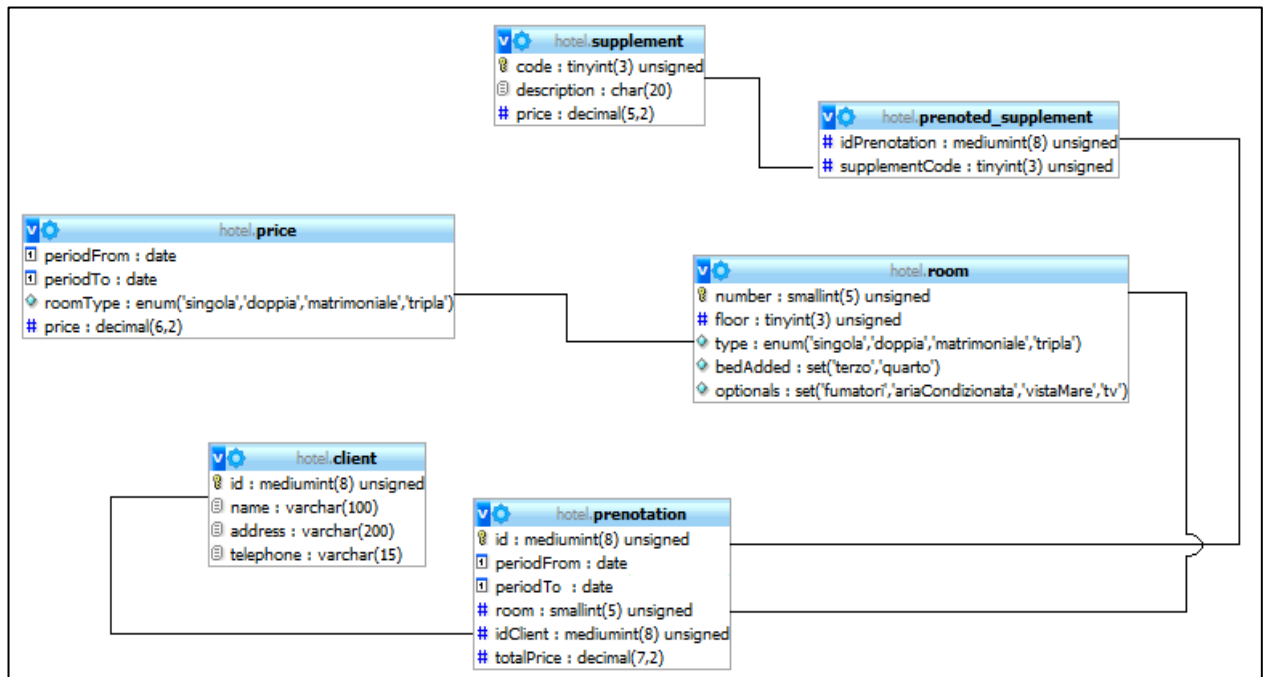


Figura 12: schema relazionale del database.

Per prima cosa si è creato un nuovo vocabolario [21] apposito per questo database con i comandi *Create content > Vocabulary*, assegnando ad esso un identificativo, un titolo, un namespace URI, l'autore ed una descrizione. Quindi si è passati alla definizione delle varie termini e proprietà del vocabolario, utilizzando per le classi il comando *Create*

new class, e facendo riferimento alle principali classi presenti nel database, nel nostro caso si sono utilizzati gli elementi Client, Room, Prenotation e Supplement, e per le proprietà il comando *Add New Property*, definendo sia relazioni tra classi che relazioni con dei dati, come numeri, stringhe o date.

Per creare una classe si sono definiti l'URI, l'etichetta, la descrizione e la super-classe a cui relazionarla, in modo da integrare i nuovi termini con quelli più utilizzati; lo stesso è stato necessario per le proprietà, definendo inoltre il dominio di partenza della proprietà ed il range dei valori possibili. Neologism permette inoltre di visualizzare pagine riassuntive di tutte le classi e proprietà create e di ottenerne un diagramma complessivo.

Il risultato ottenuto vede quindi, una pagina riassuntiva riferita al vocabolario (Figura 13), una pagina riguardante le classi (Figura 14) ed una relativa a tutte le proprietà (Figura 15). Infine è possibile salvare il file del vocabolario modellato nei formati Turtle o RDF/XML; il codice ottenuto in Turtle è nell'Appendice A.

The screenshot shows the Neologism web interface for a vocabulary named 'vocabolario hotel'. The page has a blue header with the URL 'prova.it' and a navigation bar with 'Home', 'View', 'Edit', 'Add new class', and 'Add new property'. The main content area includes:

- Author:** elisa
- Last update:** 02 July 2012
- Namespace URI:** http://localhost/neologism/db_hotel#
- Abstract:** vocabolario relativo ad un database di alberghi
- All terms at a glance:** Classes: client | prenotation | room | supplement; Properties: belong | belong_prenotation | prenotation_of | price | room_type
- Classes:** foaf:Person, db_hotel:client, owl:Thing, db_hotel:supplement, db_hotel:room, db_hotel:prenotation
- Properties:** dc:relation, db_hotel:prenotation_of, db_hotel:belong, db_hotel:belong_prenotation, db_hotel:price, dc:type, db_hotel:room_type

Figura 13: informazioni relative al vocabolario modellato con Neologism.

Classes

Class: db_hotel:client
 client – person who is a customer of a shop

URI: http://localhost/neologism/db_hotel#client
 Used with: [db_hotel:belong](#)
 Superclass: [foaf:Person](#)

[Edit](#) [Create new subclass](#) [Create new property](#) [\[back to top\]](#)

Class: db_hotel:prenotation
 prenotation – reservation

URI: http://localhost/neologism/db_hotel#prenotation
 Properties include: [db_hotel:price](#), [db_hotel:belong](#), [db_hotel:prenotation_of](#)
 Used with: [db_hotel:belong_prenotation](#)
 Superclass: [owl:Thing](#)

[Edit](#) [Create new subclass](#) [Create new property](#) [\[back to top\]](#)

Class: db_hotel:room
 room – chamber

URI: http://localhost/neologism/db_hotel#room
 Properties include: [db_hotel:price](#), [db_hotel:room_type](#)
 Used with: [db_hotel:prenotation_of](#)
 Superclass: [owl:Thing](#)

[Edit](#) [Create new subclass](#) [Create new property](#) [\[back to top\]](#)

Class: db_hotel:supplement
 supplement – extra

URI: http://localhost/neologism/db_hotel#supplement
 Properties include: [db_hotel:price](#), [db_hotel:belong_prenotation](#)
 Superclass: [owl:Thing](#)

[Edit](#) [Create new subclass](#) [Create new property](#) [\[back to top\]](#)

Figura 14: classi modellate con Neologism.

Properties	
<p>Property: db_hotel:belong belong – its owner</p> <p>URI: http://localhost/neologism/db_hotel#belong Domain: db_hotel:prenotation Range: db_hotel:client Superproperty: dc:relation Subproperty: db_hotel:belong_prenotation</p> <p>Edit Create new subproperty Add inverse property [back to top]</p>	
<p>Property: db_hotel:belong_prenotation belong_prenotation – the prenotation wich owns the extra</p> <p>URI: http://localhost/neologism/db_hotel#belong_prenotation Domain: db_hotel:supplement Range: db_hotel:prenotation Superproperty: db_hotel:belong</p> <p>Edit Create new subproperty Add inverse property [back to top]</p>	
<p>Property: db_hotel:prenotation_of prenotation_of – the room which is booked</p> <p>URI: http://localhost/neologism/db_hotel#prenotation_of Domain: db_hotel:prenotation Range: db_hotel:room Superproperty: dc:relation</p> <p>Edit Create new subproperty Add inverse property [back to top]</p>	
<p>Property: db_hotel:price price – cost</p> <p>URI: http://localhost/neologism/db_hotel#price Domains: db_hotel:prenotation, db_hotel:supplement Range: rdfs:Literal</p> <p>Edit Create new subproperty Add inverse property [back to top]</p>	
<p>Property: db_hotel:room_type room_type – the kind of room</p> <p>URI: http://localhost/neologism/db_hotel#room_type Domain: db_hotel:room Range: rdfs:Literal Superproperty: dc:type</p> <p>Edit Create new subproperty Add inverse property [back to top]</p>	

Figura 15: proprietà modellate con Neologism.

3.2. Publication

A questo punto i dati devono essere convertiti in RDF per essere considerati Linked Data, e poi devono essere pubblicati nel Web, è necessaria quindi la fase di *publishing* [19].

Come pubblicare i propri dati dipende da diversi fattori, come le dimensioni del proprio dataset, in che formato corrente è disponibile e con che frequenza i suoi dati variano.

Una volta pubblicati è poi necessario assicurarsi dell'esistenza di link esterni che puntino al proprio dataset in modo da renderlo accessibile ed esplorabile agli altri utenti: si possono quindi collegare le risorse centrali del dataset al proprio profilo o facilitare agli altri il collegamento alla propria sorgente, con l'invio di richieste di link o con il suo inserimento nelle liste di dataset ufficiali presenti nel Web (Paragrafo 3.3).

Un primo metodo per la pubblicazione dei dati, è quello di creare dei file RDF statici e di caricarli in un server web: questo avviene creando manualmente i file RDF o generandoli con semplici applicazioni che hanno questo formato come file di output. Con questo approccio non si ha però il supporto alla redirezione, 303 Redirect, ma è necessario utilizzare gli Hash URI (vedi Paragrafo 3.1), che rendono più laboriosa la scelta dei vari indirizzi, in quanto non modificabili nel tempo. Ciò rende il metodo adatto a quei dataset la cui struttura e dimensione non varia, solitamente quest'ultima non supera le poche centinaia di kilobyte essendo i documenti RDF lenti e pesanti, sia per il caricamento che per il consumo della banda, o per quelle soluzioni provvisorie in cui la stabilità dei link non è così importante.

Se quindi il proprio dataset è più ampio, coinvolge più sorgenti dati e magari presenta gli elementi nei formati CSV, Microsoft Excel o BibTEX, è meglio utilizzare un altro approccio: ad esempio è possibile per prima cosa convertire i propri dati utilizzando degli strumenti di 'RDFizing', come ConverterToRdf⁹⁹ o RDFizers¹⁰⁰, dopodiché si può salvare il risultato in un *RDF store* appropriato (una lista di *RDF repositories* è al sito <http://esw.w3.org/topic/SemanticWebTools>) e crearvi un'interfaccia per l'accesso. Se lo store scelto non presenta un'interfaccia già implementata si può utilizzare un endpoint

⁹⁹ <http://esw.w3.org/topic/ConverterToRdf>

¹⁰⁰ <http://simile.mit.edu/RDFizers/>

SPARQL insieme all'applicazione web Pubby¹⁰¹, la quale converte il primo in un'interfaccia per i Linked Data.

Molte applicazioni web, come Google, Amazon, eBay e Yahoo, hanno iniziato a rendere disponibili i propri dati attraverso l'utilizzo di Web API, le quali non sono però, spesso, create per supportare il formato dei Linked Data. Un altro metodo diventa, quindi, quello di creare dei Linked Data Wrapper che estendano le funzionalità delle diverse Web API: essi assegnano alle risorse non-informative i cui dati sono forniti dalle API, un HTTP URI permettendo la loro deferenza, ciò non necessario per le risorse informative che sono dereferenziabili direttamente; quando uno di questi URI è dereferenziato con una richiesta del tipo `application/rdf+xml`, il wrapper riscrive la richiesta in modo comprensibile all'API, ed infine si occupa della trasformazione della risposta in formato RDF, prima di restituirla al client. In questo modo anche questi dati vengono resi rintracciabili ed accessibili con l'uso dei motori di ricerca e dei data browser.

Infine, un approccio molto interessante si ha per la pubblicazione di quei dati che si presentano sottoforma di database relazionale: esistono diversi tool che facilitano la loro pubblicazione, come **D2R Server**¹⁰², **OpenLink Virtuoso**¹⁰³ o **Triplify**¹⁰⁴.

L'attenzione è stata puntata sul primo dei tre, il quale è stato utilizzato per la pubblicazione online di diversi dataset molto popolari come Berlin DBLP Bibliography Server¹⁰⁵ o EuroStat Countries and Regions Server¹⁰⁶, ed è spesso usato nei progetti di W3C (World Wide Web Consortium) e LATC (LOD Around The Clock).

3.2.1. D2R Server

D2R Server [20] è uno strumento implementato in Java, distribuito sotto la Apache License V2.0, che mette a disposizione un'interfaccia HTML, un'interfaccia RDF o Linked Data, e un'interfaccia SPARQL per l'accesso ai dati pubblicati. L'architettura è rappresentata in Figura 16.

¹⁰¹ <http://www4.wiwiw.fu-berlin.de/pubby/>

¹⁰² <http://sites.wiwiw.fu-berlin.de/suhl/bizer/d2r-server/index.html>

¹⁰³ <http://virtuoso.openlinksw.com/wiki/main/Main/>

¹⁰⁴ <http://triplify.org/Overview>

¹⁰⁵ <http://www4.wiwiw.fu-berlin.de/dblp/>

¹⁰⁶ <http://www4.wiwiw.fu-berlin.de/eurostat/>

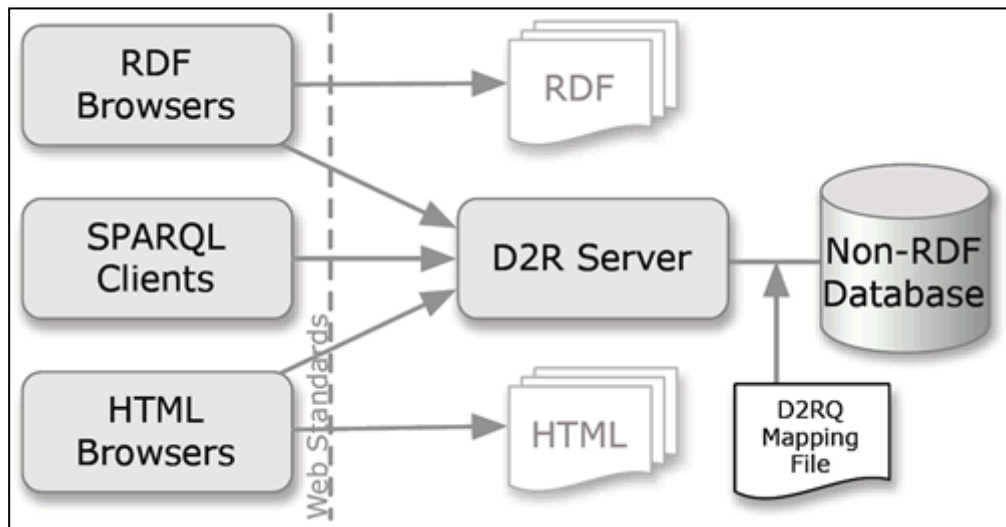


Figura 16: architettura di D2R Server.

Esso permette di pubblicare database relazionali nel Semantic Web e, ai browser RDF e HTML, di esplorare il contenuto del database, consentendo di interrogarlo attraverso il linguaggio SPARQL. Qui di seguito verrà descritta la procedura con la quale si è testato il funzionamento del tool.

Per prima cosa si è scaricato il package di D2R Server, e si è utilizzato il comando

```
d2r-server [-p port] [-b serverBaseURI]
           [--fast] [--verbose] [--debug]
           mapping-file-example.ttl
```

per avviarne il funzionamento, dal prompt dei comandi. Per analizzare lo strumento, si è poi utilizzato lo stesso database di hotel, precedentemente visto per Neologism (Paragrafo 3.1.1) e il vocabolario costruito con esso.

Si è, quindi, creato il `mapping-file.ttl` del database, un file nel linguaggio D2RQ¹⁰⁷, linguaggio dichiarativo che permette la trasformazione di schemi di database relazionali, nei linguaggi RDF e OWL, utilizzando i vari vocabolari a disposizione per il mapping. Utilizzando D2R Server, si può generare automaticamente questo file attraverso il comando

¹⁰⁷ <http://d2rq.org/d2rq-language>

```

generate-mapping [-u user] [-p password] [-d driver]
                 [-l script.sql]
                 [--[skip-] (schemas|tables|columns) list]
                 [--w3c] [-v] [-b baseURI] [-o outfile.ttl]
                 [--verbose] [--debug]
                 jdbcURL

```

sostituendo alle parentesi quadre, i parametri per la connessione con il database.

Il `generate-mapping`¹⁰⁸ è un'istruzione che crea automaticamente il file di mapping D2RQ analizzando lo schema del database esistente; esso propone un mapping di default, che associa ad ogni tabella una nuova classe RDFS, basata sul nome della tabella, e ad ogni colonna, una nuova proprietà, usando anche per esse il loro nome. Il file può essere utilizzato così come è o può essere modificato: nel nostro caso il file è stato modificato in modo da utilizzare le classi e le proprietà del vocabolario appositamente costruito con Neologism. Il testo del file `mapping_hotel.ttl` utilizzato, è presente nell'Appendice B.

Le prime linee indicano i vari vocabolari utilizzati, da notare la linea

```
@prefix db_hotel: <http://localhost/neologism/db_hotel#>
```

che indica il vocabolario precedentemente implementato.

Per ciascuna tabella si è creata la corrispondente classe ed, oltre alle normali proprietà di `rdfs:label` per le colonne, sono state definite le proprietà di `d2rq:join` in modo da permettere i collegamenti tra gli attributi delle diverse tabelle.

Una volta completato, si è messo in funzione D2R Server, passandovi il file come parametro, ed ottenendo così, all'indirizzo `http://localhost:2020`, la pagina iniziale per l'esplorazione del database (Figura 17). La porta 2020 indicata, è quella utilizzata nei test, ma il server è configurabile con qualsiasi porta.

È inoltre possibile ottenere la versione del database in RDF, cliccando sull'apposito simbolo, ed eseguire le query, utilizzando lo SPARQL endpoint presente all'indirizzo, nel nostro caso, `http://localhost:2020/sparql`.

¹⁰⁸ <http://d2rq.org/generate-mapping>

D2R Server pubblica il database Hotel
Running at <http://localhost:2020/>

Home | [client](#) [prenotation](#) [prenoted_supplement](#) [price](#) [room](#) [supplement](#)

This is a database published with D2R Server. It can be accessed using

1. your plain old web browser
2. Semantic Web browsers
3. SPARQL clients.

1. HTML View

You can use the navigation links at the top of this page to explore the database.

2. RDF View

You can also explore this database with **Semantic Web browsers** like [Disco](#) or [Marbles](#). To start browsing, open this entry point URL in your Semantic Web browser:

<http://localhost:2020/all>

3. SPARQL Endpoint

SPARQL clients can query the database at this SPARQL endpoint:

<http://localhost:2020/sparql>

The database can also be explored using [this AJAX-based SPARQL Explorer](#).

Generated by [D2R Server](#)

Figura 17: homepage della pubblicazione del database con D2R Server.

prenotation #1
Resource URI: <http://localhost:2020/resource/prenotation/1>

Home | [All prenotation](#)

Property	Value
db_hotel:belong	< http://localhost:2020/resource/client/1 >
dc:date	2006-06-01 (xsd:date)
dc:date	2006-06-15 (xsd:date)
rdfs:label	1 (xsd:unsignedInt)
rdfs:label	prenotation #1
db_hotel:prenotation_of	< http://localhost:2020/resource/room/101 >
db_hotel:price	825 (xsd:decimal)
rdf:type	db_hotel:prenotation

Generated by [D2R Server](#)

Figura 18: istanza all'interno della tabella 'prenotation'.

3.3. Linking

L'ultima fase è detta di *linking* [19], collegamento. Essa consiste, principalmente nella creazione di *RDF links*, in modo da permettere ai Linked Data browser e crawler di navigare tra le diverse sorgenti di dati.

È necessario quindi determinare quali proprietà RDF dovranno essere utilizzate come predicati: ad esempio le più utilizzate per la descrizione delle persone sono `foaf:knows`, `foaf:based_near` e `foaf:topic_interst`, utili a collegare i profili di diversi individui e i loro interessi.

Un altro link molto utilizzato è `owl:sameAs`, per creare i cosiddetti *URI alias*, ovvero quando URI differenti identificano lo stesso oggetto nella realtà; essi risultano fondamentali nel Web of Data, in quanto permettono di de-referenziare un URI in diverse risorse, esprimendo così più punti di vista ed opinioni diverse sullo stesso elemento.

Esistono diverse tecniche per creare gli RDF link tra sorgenti diverse, essi possono infatti essere creati manualmente o stabiliti in modo automatico con l'utilizzo di algoritmi appositi.

Per settare i link manualmente è necessario per prima cosa, conoscere i dataset a cui ci si vuole collegare; una lista dei vari dataset utilizzabili è mantenuta nella Linking Open Data Dataset list¹⁰⁹, i quali presentano spesso delle interfacce SPARQL utili alla ricerca degli URI corretti a cui collegarsi, altrimenti si possono usare i Linked Data browser, come **Tabulator**¹¹⁰ o **Disco**¹¹¹, in modo da esplorare i dataset e trovare gli URI necessari. Altri strumenti, possono essere i servizi come **Sindice**¹¹² o **Uriqr**¹¹³, i quali, data la risorsa a cui ci si vuole legare, ritornano gli URI più popolari ed utilizzati.

L'approccio manuale non è efficiente, ma soprattutto non è scalabile nel linking di grandi dataset; in questi casi è molto più utile utilizzare algoritmi che generino automaticamente i link tra le diverse sorgenti. Si ha però una mancanza di strumenti generici per l'auto-generazione dei RDF link, spesso si creano algoritmi ad hoc, specifici per ciascun dataset.

¹⁰⁹ <http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets>

¹¹⁰ <http://www.w3.org/2005/ajar/tab>

¹¹¹ <http://sites.wiwiw.fu-berlin.de/suhl/bizer/ng4j/disco/>

¹¹² <http://www.sindice.com>

¹¹³ <http://uriqr.com>

Gli algoritmi più utilizzati sono detti *deterministici* o *rule-based*, basati su regole, in quanto generano collegamenti in base al numero di corrispondenze tra gli identificatori individuali disponibili dei vari dataset: due elementi si dicono corrispondenti secondo una procedura di *record linkage* [24] deterministica, se tutti o alcuni dei loro identificatori, in numero maggiore di una certa soglia, sono identici.

Questo tipo di algoritmi è utilizzato per quei dataset aventi una buona qualità dei dati, soprattutto quando le entità sono individuate o da un identificatore comune, utilizzando i cosiddetti algoritmi *pattern-based*, o quando ci sono identificatori rappresentativi diversi, come nome, cognome, sesso nel caso delle persone, utilizzando quelli più complessi, detti *property-based*.

I primi sono utilizzati specialmente in quei domini in cui si ha già uno schema ben definito, con chiavi primarie precise e significative che possono essere utilizzate come identificatori precisi delle varie risorse: questi identificatori sono inseriti negli HTTP URI, in modo da creare un legame univoco tra indirizzo e risorsa, e possono essere usati come pattern da ricercare nelle URI delle diverse sorgenti, in modo da stabilire un legame di alias, `owl:sameAs`, tra le risorse che lo hanno in comune [19]. Un esempio è visibile nel linking tra le sorgenti **DBpedia** e **BookMashup**, basato sul codice ISBN dei vari libri (RDF Book Mashup¹¹⁴).

I secondi, invece, sono utilizzati nei casi in cui non esistano identificatori comuni tra i diversi dataset; essi sono più complessi in quanto devono confrontare più proprietà delle risorse per vedere la presenza delle possibili corrispondenze[19]. Un esempio è il linking tra le sorgenti **DBpedia** e **Geonames**, in cui si è usato un algoritmo basato sul confronto tra informazioni come latitudine, longitudine, nazione, popolazione e così via, in modo da identificare gli stessi luoghi presenti in entrambe i dataset (Link sto Geonames¹¹⁵).

Un'altra tipologia di algoritmi utilizzata, è quella detta *probabilistica* [24], o anche di *fuzzy matching*. Questi algoritmi adottano un approccio diverso al problema di *linking*, prendendo in considerazione una gamma più ampia di potenziali identificatori, calcolando i pesi per ogni identificatore, in base alla sua capacità stimata di identificare correttamente una corrispondenza o un *non-match*, una non-corrispondenza, e

¹¹⁴ <http://sites.wiwiw.fu-berlin.de/suhl/bizer/bookmashup/index.html>

¹¹⁵ <http://wiki.dbpedia.org/Downloads32>

utilizzando questi pesi per calcolare la probabilità che i due dati indicati si riferiscano alla stessa entità.

Un coppia di elementi si considera quindi una corrispondenza, se ha una probabilità al di sopra di una certa soglia definita, mentre le coppie con probabilità al di sotto un'altra soglia, sempre stabilita, sono considerate un non-match: le coppie che rientrano tra le due soglie, vengono poi valutate come 'possibili corrispondenze' e possono essere trattate di conseguenza, ad esempio, revisionate manualmente e collegate, o non collegate, a seconda delle esigenze. Questo processo risulta più comodo di quello deterministico in quanto non ci sono da stabilire regole a priori, e può essere implementato in modo da funzionare bene anche senza l'intervento umano, o almeno in quantità molto minore.

Il procedimento vede come prima cosa che si debbano stabilire le soglie di probabilità u e m : la prima indica la probabilità che due elementi non correlati siano corrispondenti per puro caso, ad esempio per il mese di nascita si ha $u=1/12 \approx 0,083$; la seconda indica, invece, la probabilità che due elementi effettivamente collegati, siano in corrispondenza o siano sufficientemente simili, ovvero si ha $m=1$ in caso di coincidenza perfetta. Quest'ultimo caso è molto raro, e spesso è necessaria una stima della probabilità, basata sulle conoscenze pregresse sugli insiemi di dati; si possono infatti eseguire manualmente una serie di link e non-link tra gli elementi, in modo da identificare in che modo eseguire i match e i non-match, e 'addestrare' l'algoritmo di record linkage, o iterarlo in modo da ottenere stime sempre più precise della probabilità.

Oltre ad utilizzare normali misure di uguaglianza tra le stringhe degli elementi si possono utilizzare anche, funzioni più complesse ma precise come la High Jaro-Winkler distance¹¹⁶ o la Low Levenshtein distance¹¹⁷, che danno un valore di probabilità m , per cui due elementi possono definirsi sufficientemente simili.

Determinare come impostare le soglie di match e non-match, comporta la valutazione dell'equilibrio tra la sensibilità dell'algoritmo, o *recall*, ovvero la percentuale di elementi che sono legati dall'algoritmo, e la sua predittività, o *precision*, ovvero la percentuale di record collegati correttamente dall'algoritmo, che veramente corrispondono. Esistono diversi metodi manuali e automatici, per stabilire le soglie in

¹¹⁶ http://en.wikipedia.org/wiki/Jaro-Winkler_distance

¹¹⁷ http://en.wikipedia.org/wiki/Levenshtein_distance

modo corretto, e spesso sono integrati in strumenti software che aiutano l'utente nella ricerca di valori più corretti (un esempio è al Paragrafo 3.2.1) .

Essendo questa metodologia spesso utilizzata in grandi insiemi di dati, il procedimento comporta una alta complessità computazionale; per ovviare a ciò si adotta spesso, una tecnica detta *blocking*, la quale permette di migliorare l'efficienza del processo: il *blocking*, permette infatti di limitare il confronto delle istanze, ai soli elementi in cui uno, o più identificatori discriminanti, corrispondono, ottenendo così l'aumento del valore di *precision*, a discapito della *recall*. Il problema principale dato da questa tecnica, è quello di tralasciare potenziali coppie corrispondenti, non avendo queste gli identificatori correlati; si possono quindi creare più passaggi di linking, con diversi tipi di blocking, in modo da raggruppare i dati in modo diversi e poi confrontarli tra loro, velocizzando il procedimento complessivo.

Queste tecniche sono spesso utilizzate nella ricerca medica e farmaceutica, ad esempio per associare i dati di virus sconosciuti a quelli già noti, permettendo con pochi dati noti, il collegamento completo di dataset anonimi [24].

Esistono diversi tool [35] che facilitano il linking tra le diverse sorgenti, **LIMES (Link Discovery framework for METric Spaces)**¹¹⁸, che utilizza approcci di linking basati sulle caratteristiche degli spazi metrici degli URI; **LDIF (Linked Data Integration Framework)**¹¹⁹, che trasforma Linked Data eterogenei in un'unica chiara rappresentazione mentre si tiene traccia della provenienza dei dati; **SERIMI**¹²⁰, strumento di generazione automatica di link che non richiede la conoscenza preventiva del dominio di dati o del loro schema; **SemMF**¹²¹, è un framework flessibile per il calcolo della similarità semantica tra gli oggetti che vengono rappresentati come grafici RDF; **MOAT (Meaning Of A Tag)**¹²², framework per il collegamento manuale dei tag con gli URI del Web semantico. Ma lo strumento più referenziato e utilizzato dai gruppi LOD (Linking Open Data) e LATC (LOD Around The Clock), è **Silk**¹²³, un framework per la scoperta di link all'interno del Web of Data, che utilizza diverse tecniche pattern-based e property-based che, combinate, permettono di scoprire le corrispondenze tra i vari URI.

¹¹⁸ <http://limes.sf.net/>

¹¹⁹ <http://www4.wiwiw.fu-berlin.de/bizer/ldif/>

¹²⁰ <https://github.com/samurarujo/SERIMI-RDF-Interlinking>

¹²¹ <http://sites.wiwiw.fu-berlin.de/suhl/radek/semmf/doc/index.html>

¹²² <http://moat-project.org/>

¹²³ <http://www4.wiwiw.fu-berlin.de/bizer/silk/>

Silk [20], è uno strumento che permette la scoperta di relazioni tra elementi appartenenti a diverse sorgenti di Linked Data, e può essere usato dai data publisher per stabilire link RDF tra i propri dati e quelli appartenenti ad altre sorgenti all'interno del Web.

3.3.1. Silk Workbench

Silk [25], il Link Discovery Framework per il Web of Data, è uno strumento utilizzato sia per la ricerca, nel Web, di elementi collegabili ai propri dataset, sia per l'effettiva creazione degli RDF link tra i dati.

Silk è disponibile in tre varianti, a seconda dei casi d'uso:

- **Silk Single Machine**¹²⁴, viene utilizzato per generare link RDF su una singola macchina; i dataset possono risiedere tutti sulla stessa macchina o su macchine remote, accessibili con il protocollo SPARQL; permette il multithreading ed il caching, e le sue prestazioni sono migliorabili con l'utilizzo degli algoritmi di blocking;
- **Silk MapReduce**¹²⁵, è usato per generare link RDF su cluster di più macchine; esso si basa su Hadoop¹²⁶ e può essere eseguito su Amazon, grazie alla sua scalabilità, permette di generare link per dataset molto ampi, distribuendo l'onere computazionale su più macchine;
- **Silk Server**¹²⁷ [26], può essere utilizzato come componente per *identity resolution*, processo che identifica quali elementi fanno riferimento ad un'unica entità, all'interno di applicazioni che utilizzano i Linked Data; esso fornisce una HTTP API per il matching delle entità provenienti da un flusso di dati RDF, tenendo traccia di quelle conosciute, e per questo può essere utilizzato, insieme a data crawler, per popolare cache duplicate-free.

Lo strumento preso in considerazione, è Silk Workbench¹²⁸, applicazione web che guida gli utenti nel processo di interlinkg di diverse sorgenti dati RDF. Esso, infatti, permette di gestire diversi dataset e diverse tipologie di link, offre un'interfaccia grafica che facilita la creazione e le modifiche alle specifiche dei link, ed esegue

¹²⁴ <http://www4.wiwiss.fu-berlin.de/bizer/silk/singlemachine/>

¹²⁵ <http://www4.wiwiss.fu-berlin.de/bizer/silk/mapreduce/>

¹²⁶ <http://hadoop.apache.org/>

¹²⁷ <http://www4.wiwiss.fu-berlin.de/bizer/silk/server/>

¹²⁸ <http://www4.wiwiss.fu-berlin.de/bizer/silk/workbench/>

automaticamente l'iterazione del processo di linking scelto, in modo da supportare l'utente nella valutazione dei link risultanti da una certa specifica.

Tutti questi moduli si basano sul Silk Link Discovery Engine¹²⁹, modulo che permette, data una specifica dei link da trovare/stabilire, di accedere ai dati, sia locali che remoti attraverso il linguaggio SPARQL, generando un flusso di elementi di dati. Quindi si ha la fase di blocking (opzionale), che permette l'elaborazione dei collegamenti per gruppi di dati; i dati, infatti, sono considerati a coppie, calcolando per ciascuna di esse un valore di somiglianza: se il meccanismo di blocking, non è abilitato si ottiene il prodotto cartesiano dei vari elementi, altrimenti si confrontano tra loro solo i dati appartenenti allo stesso cluster.

Il valore di somiglianza, tra 0 e 1, viene poi valutato a seconda delle condizioni di linking espresse precedentemente, e se soddisfatte, si genera un link preliminare tra i due elementi. Si ha poi una fase di filtraggio in cui si eliminano, per prima cosa, i link aventi confidenza minore della soglia definita dall'utente, e poi, i link che superano il numero di link massimi generabili da uno stesso elemento. Infine si scrivono i link ottenuti, in una destinazione specificata dall'utente.

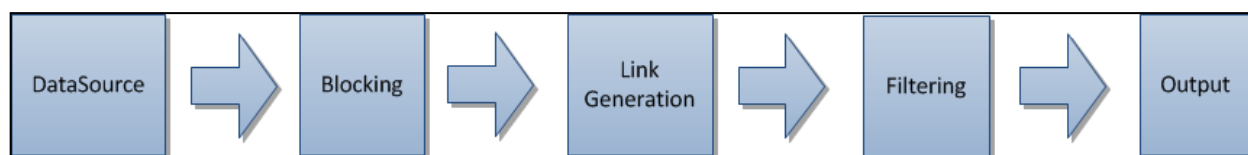


Figura 19: architettura del Silk Link Discovery Engine.

Le specifiche di linking devono essere espresse con il linguaggio dichiarativo Silk – Link Specification Language (Silk-LSL)¹³⁰, in cui gli sviluppatori possono specificare il tipo di RDF link da scoprire tra le fonti dati e le condizioni che devono essere soddisfatte perché due elementi possano essere interconnessi. Le condizioni riguardano la scelta delle metriche di similarità da utilizzare e anche gli altri collegamenti che un dato può già avere. Il Silk-SLS è espresso in XML ed ha come root tag <Silk>; il documento per il linking è solitamente diviso in quattro parti: la definizione dei prefissi dei vari vocabolari utilizzati, la definizione delle sorgenti dati e dei loro endpoint

¹²⁹ http://www.assembla.com/wiki/show/silk/Silk_Link_Discovery_Engine

¹³⁰ http://www.assembla.com/spaces/silk/wiki/Link_Specification_Language

SPARQL, le specifiche dei link ed, infine, la definizione dell'output (in Figura 20 si ha la struttura di un generico file Silk-LSL).

```
<?xml version="1.0" encoding="utf-8" ?>
<Silk>
  <Prefixes ... />
  ...
  <DataSources ... />
  ...
  [<Blocking ... />]
  ...
  <Interlinks ... />
  ...
  [<Outputs ... />]
  ...
</Silk>
```

Figura 20: struttura di un file Silk-LSL.

Per provare il funzionamento di Silk Workbench, si sono apportate delle piccole modifiche al database precedentemente utilizzato, aggiungendo alla tabella 'client' alcuni nomi di personaggi noti, in modo da vedere se effettivamente era possibile stabilire dei link tra il database 'hotel' ed il più generico ed ampio insieme di dati, DBpedia.

Dopo aver scaricato il package dell'applicazione, si è scelto di eseguire Silk Workbench da linea di comando,

```
java -jar workbench.war
```

dopodiché si è passati all'esecuzione delle diverse fasi che permettono la creazione dei link tra le diverse sorgenti (Figura 21).

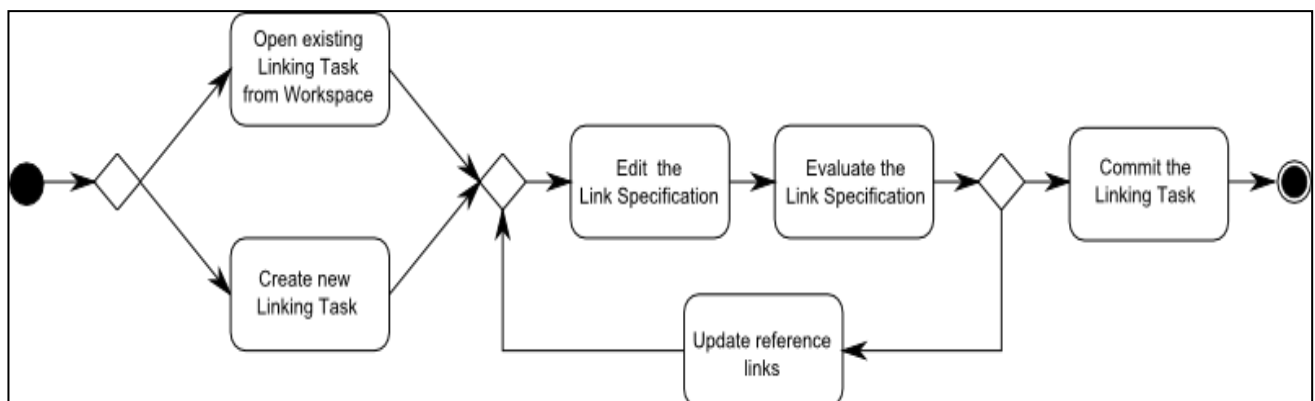


Figura 21: fasi per la costruzione di link con Silk Workbench.

Per prima cosa si sono definite le sorgenti di dati utilizzando gli endpoint SPARQL: per la versione italiana di DBpedia si è utilizzato l'URI `http://it.dbpedia.org/sparql`, mentre per il database si è utilizzato l'URI `http://localhost:2020/sparql`, creato con D2R Server (Paragrafo 3.3.1). Quindi si è definito il tipo di link da stabilire; nel nostro caso, la relazione `owl:sameAs`, restringendo il confronto alla tabella 'client' per il nostro database e alla classe 'Person' all'interno di DBpedia (Figura 22).

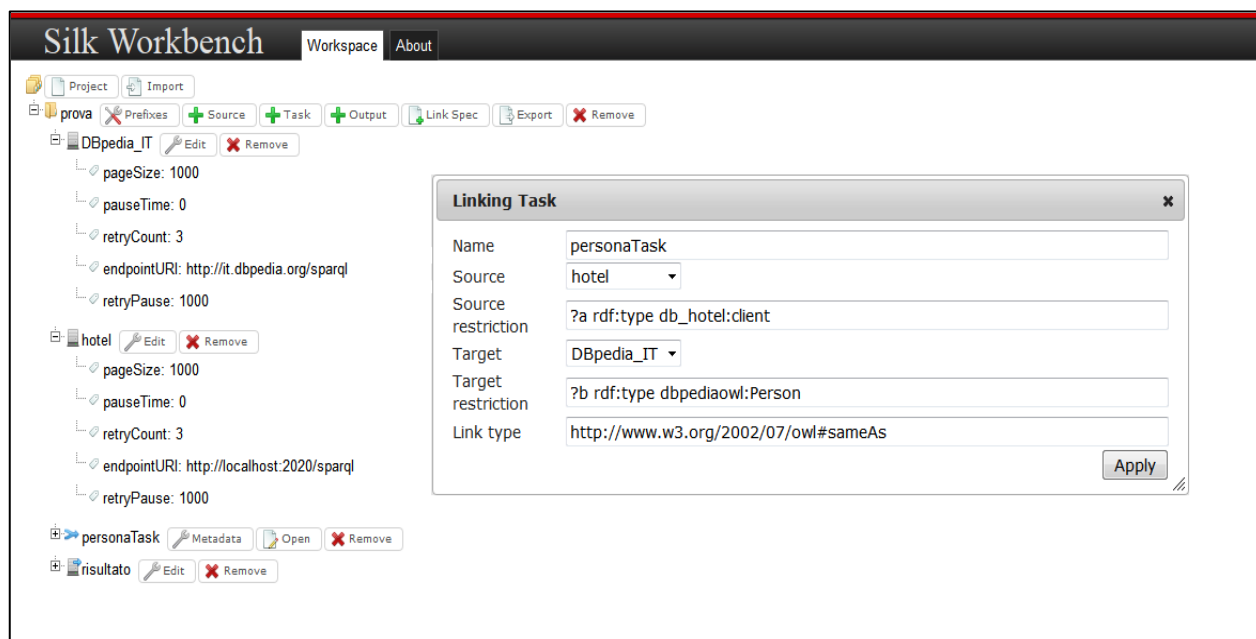


Figura 22: sorgenti e linking task in Silk Workbench.

Si è poi passati alla fase di dichiarazione delle specifiche del link: qui si utilizza un'interfaccia grafica, il Linkage Rule Editor¹³¹, che mette a disposizione diversi blocchi, ciascuno contenente una particolare operazione, che l'utente dispone in una sequenza, a piacere, in modo da ottenere il tipo di confronto e le soglie desiderate. Nel nostro esempio si è scelto di confrontare gli elementi `foaf:familyName` di 'client' e `foaf:name` di 'Person', di utilizzare due blocchi del tipo *Transformations:Lower case*, per la conversione delle stringhe in caratteri tutti minuscoli, e, per il confronto, il blocco *Comparators:Equality*, il quale restituisce 1 in caso di coincidenza perfetta delle stringhe e 0 in caso contrario (Figura 23). Per vedere inizialmente, quali coppie il framework selezionava, non sono state impostate soglie di nessun tipo.

¹³¹ http://www.assembla.com/spaces/silk/wiki/Linkage_Rule_Editor

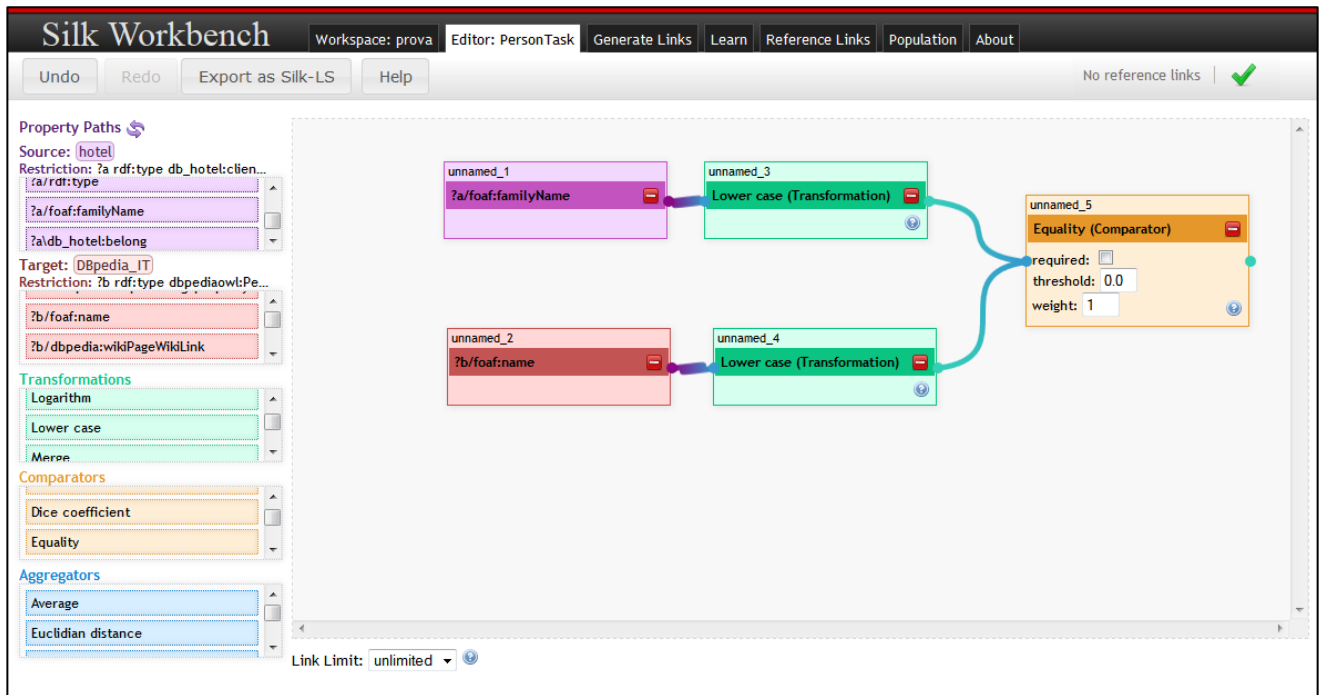


Figura 23: Linkage Rule Editor in Silk Workbench.

Quindi si è avviata la fase di generazione dei link, la quale ha trovato 4 coincidenze nei nomi della tabella; per ciascuna di essa si ha la possibilità di scegliere se:

- Confermare il link come corretto, inserendolo nella parte positiva dei Reference link;
- Se la correttezza del link non è decisa, e non saranno quindi inseriti nei Reference link;
- Confermare il link come incorretto, inserendolo nella parte negativa dei Reference link.

Nel nostro caso, si è confermata manualmente la correttezza della prima scelta, mentre per le altre tre corrispondenze si è dato valore interrogativo, in modo da decidere successivamente la correttezza delle corrispondenze (Figura 24).

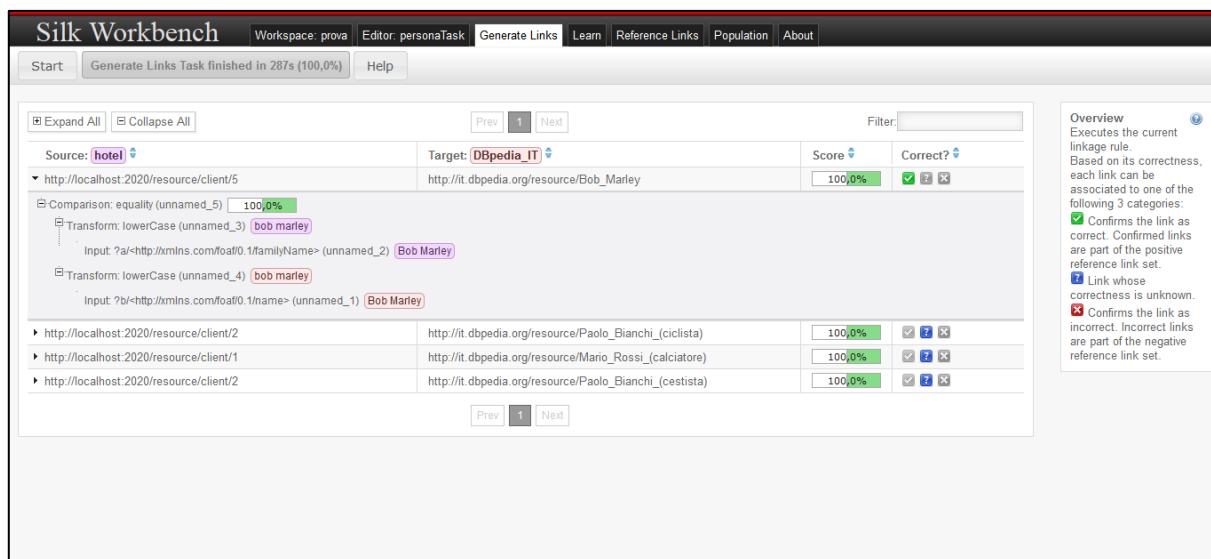


Figura 24: generazione dei link con Silk Workbench

Si è passati alla fase di gestione dei Reference link, i quali sono appunto dei link la cui correttezza deve essere confermata dall'utente, in modo da aumentare iterativamente l'esattezza della specifica di linking: nella parte negativa sono state inserite le tre corrispondenze precedentemente tralasciate, in quanto rappresentanti un caso di omonimia, ma senza la coincidenza dei soggetti; mentre nella parte positiva si è (automaticamente) inserita, la coppia precedentemente evidenziata come corretta. Quindi si è eseguita la fase di *learning*¹³²: Silk Workbench, supporta infatti il learning delle regole di linkage che può avvenire utilizzando i Reference link prima selezionati, o caricandoli da file esistenti. Nel nostro caso si sono utilizzati i link prima elaborati, quindi si è avviato l'apprendimento, il quale itera nuovamente la ricerca di link, utilizzando le nuove indicazioni, e si arresta una volta superato il numero massimo di iterazioni o con l'intervento manuale dall'utente. Esso dà come risultato, tutte le regole presenti nel workspace, le quali possono essere modificate in modo da inserirvi quella appena eseguita: nel nostro caso si è ottenuta una regola di linkage composta da un unico modulo *Comparators:Jaccard*, avente diversi valori di soglia e peso (Figura 25), la quale una volta ri-eseguita ha restituito automaticamente solo il link corretto sui precedenti quattro, avendo quindi i valori di *precision* e *recall* ottimali, pari a 1 (Figura 26).

¹³² http://www.assembla.com/spaces/silk/wiki/Learning_Linkage_Rules

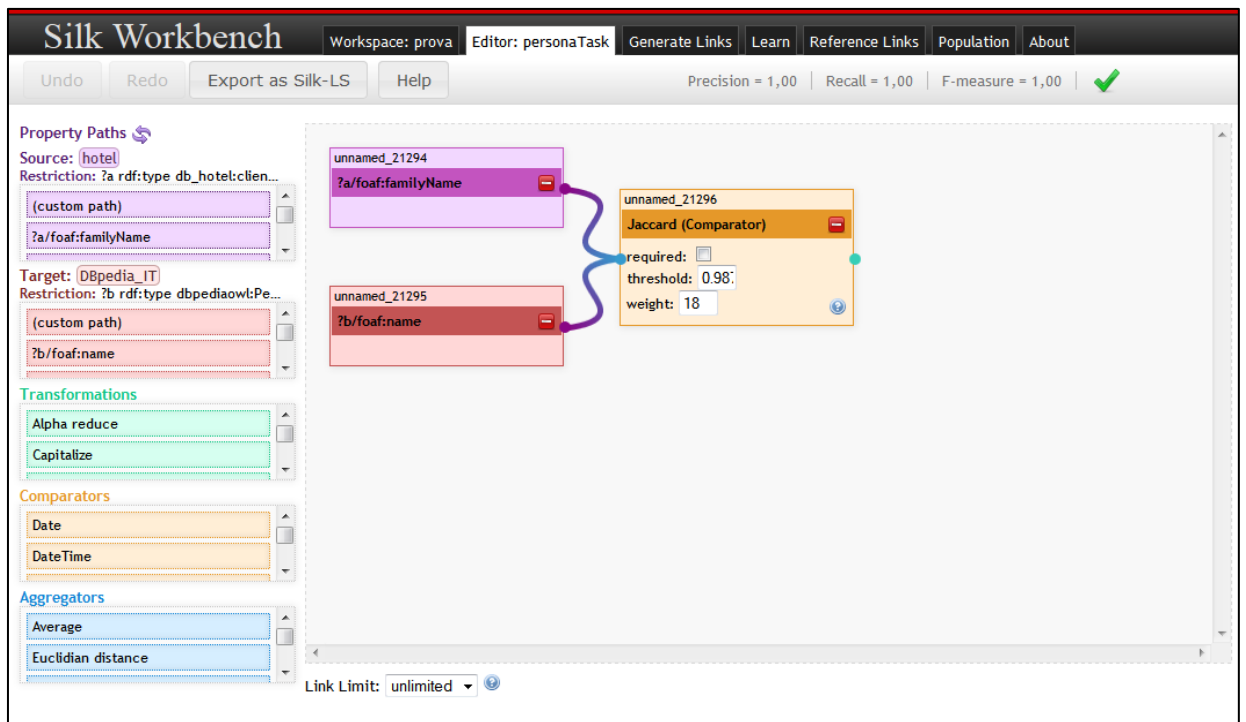


Figura 25: regola risultante dopo la fase di learning nel Linkage Rule Editor.

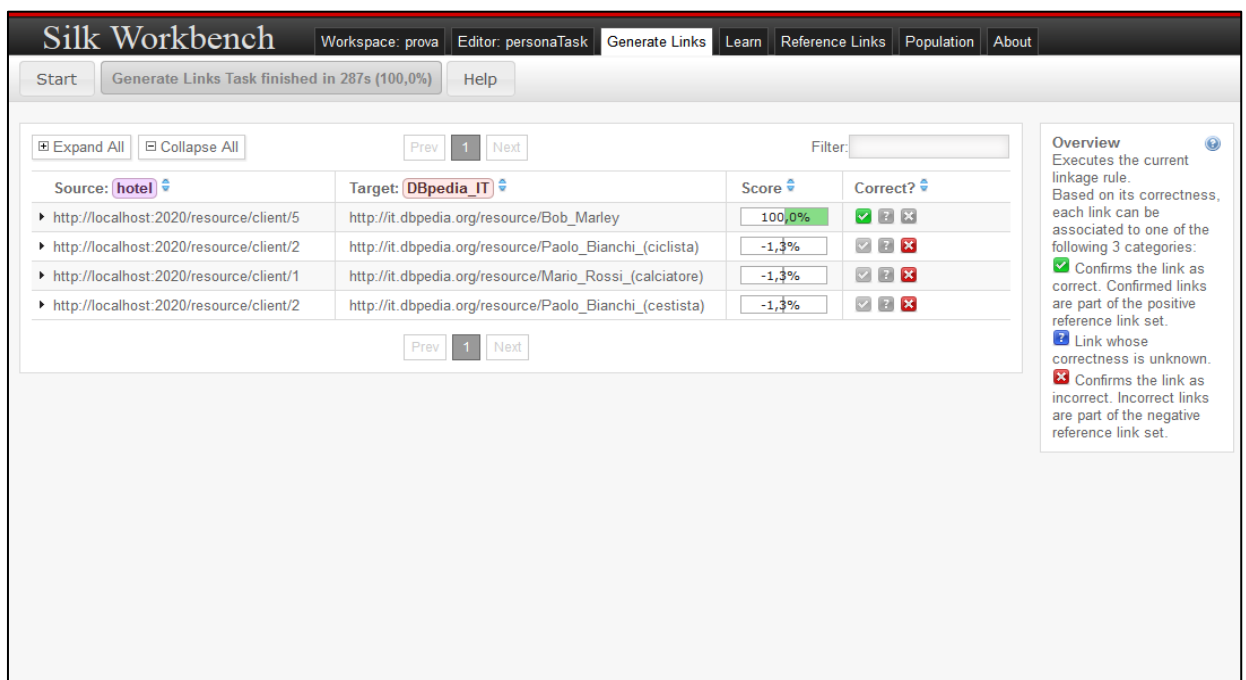


Figura 26: link risultanti con la regola finale, prima della sua esecuzione; si noti come siano già diverse le percentuali di correttezza.

Con Silk Workbench, è poi possibile ottenere il codice sia di tutto il progetto, che delle sole parti interessate, come il file in Silk-LSL che permette di personalizzare le

specifiche della regola di linkage o i link risultanti. Quest'ultimo è dato dalla tripla, nella sintassi N-Triples, nel file `output.nt`; il link dovrà poi essere inserito come istanza all'interno del database, o come tripla in un RDF store contenente i dati, in formato RDF, precedentemente pubblicati, in modo da renderlo attivo:

```
<http://localhost:2020/resource/client/5>  
<http://www.w3.org/2002/07/owl#  
sameAs> <http://it.dbpedia.org/resource/Bob_Marley> .
```

Per completezza, il codice elaborato è messo a disposizione nell'Appendice C.

Capitolo 4

4. Integrazione semantica nei LOD: stato dell'arte

Come visto in precedenza, i Linked Data Open (LOD) sono composti da diversi miliardi di triple RDF, un insieme che comprende la conoscenza di domini diversi e che è in costante crescita.

La creazione del Web of Data, è quindi da considerare una pietra fondamentale nella realizzazione del Semantic Web [27], ovvero nella trasformazione del World Wide Web, in un ambiente in cui tutte le informazioni pubblicate risultino associate a dati e metadati che ne specificano il contesto, rendendole adatte all'elaborazione automatica.

Tuttavia, in termini di usabilità, abbiamo visto come i LOD siano ancora in fase di sviluppo; diverse infatti sono le questioni che ancora devono essere studiate e risolte, e tra queste, una delle più importanti è ottenere il collegamento semantico dei dati.

I vari Linked Data, sono tra loro molto ben connessi a livello di istanza: utilizzando link di tipo `owl:sameAs`, si creano le varie corrispondenze tra i singoli dati.

D'altra parte si ha una grande carenza di integrazione a livello di schema dei vari dataset: con ciò si indica la mancanza di proprietà e di link del tipo `rdfs:subClassOf`, necessari alla costruzione di tassonomie, e quindi alla contestualizzazione, all'interno degli schemi, dei vari dati [28].

La creazione di link a livello di schema, risulta però fondamentale nei processi di integrazione ed allineamento di ontologie o di dataset di domini differenti; essi infatti

facilitano e velocizzano il riconoscimento delle corrispondenze tra i dati, e arricchiscono contestualizzando le informazioni disponibili.

Diversi sono i gruppi di ricerca che si sono orientati verso lo sviluppo di tool, framework e piattaforme per l'integrazioni dei Linked Data o di dataset all'interno dei LOD a livello semantico; in questo capitolo verranno descritti i principali strumenti realizzati.

4.1. BLOOMS

BLOOMS (Bootstrapping-based Linked Open Data Ontology Matching System) [28], è un sistema per l'allineamento di dataset a livello di schema, implementato da un team di esperti formato da Prateek Jain, Pascal Hitzler, Amit P. Sheth, Kunal Verma e Peter Z. Yeh, sui framework **Alignment API**¹³³, implementazione che permette l'allineamento e la condivisione di ontologie, e **Jena**¹³⁴, un framework Java per la costruzione di applicazione sul Semantic Web.

Esso è stato appositamente costruito per l'allineamento dei LOD a livello di schema, considerando come base dell'approccio di bootstrapping, la struttura costituita dalle categorie di DBpedia, elemento che occupa un ruolo centrale nella nuvola dei Linked Data, fondandosi sulle infobox di Wikipedia.

Per allineamento a livello di schema si intende la generazione di collegamenti tra le gerarchie di classi (tassonomie), ovvero la creazione di triple RDF aventi come predicato la proprietà `rdfs:subClassOf`.

Il metodo utilizzato da BLOOMS costruisce una 'foresta' per ogni categoria delle ontologie candidate al matching, ovvero crea un insieme di alberi per ciascuna classe, utilizzando una selezione delle sue superclassi, e poi confronta le due foreste in modo da decidere se e come le due classi possano essere allineate.

Più in dettaglio, una volta prese in input due ontologie contenenti informazioni a livello di schema, il procedimento di BLOOMS può essere suddiviso in quattro fasi;

¹³³ <http://alignapi.gforge.inria.fr/>

¹³⁴ <http://www.openjena.org/>

- *Pre-processing* delle ontologie; si utilizza un semplice algoritmo che trasforma i nomi delle classi in stringhe semplici, scomponendo le parole composte ed eliminando accenti, underscore e i caratteri di terminazione delle parole.
- *Costruzione della foresta relativa a ciascuna classe*; la stringa identificante la classe viene passata al servizio Web di Wikipedia, il quale ritornerà tutte le pagine di Wikipedia relative ad essa; queste verranno considerate come i vari *sense*, significati, della parola e per ciascuno di essi sarà costruito uno dei *tree*, alberi, utilizzando il *sense* come radice e come nodi figli, le sue sottocategorie presenti in Wikipedia.

Si può notare come per BLOOMS si sia scelto di utilizzare Wikipedia per la disambiguazione dei significati delle varie classi delle ontologie, invece di thesauri o vocabolari utilizzati in altri sistemi di integrazione, ad esempio come WordNet, in modo da rimanere più legati all'insieme dei Linked Open Data.

- *Confronto delle foreste di due classi*; per decidere in che modo collegare le due classi è quindi necessario confrontare tutti gli alberi di una classe della prima ontologia, con tutti gli alberi che costituiscono la foresta della classe nella seconda ontologia; per fare ciò si utilizza una funzione che una volta confrontati tutti i nodi, calcola un valore di somiglianza degli alberi, compreso tra 0 e 1, in base al numero di nodi totale e a quanti di essi sono risultati coincidenti; quindi si confrontano i valori ottenuti con un soglia x per permettere la creazione del link; se gli alberi risultano identici si usa il link `owl:equivalentClass`, altrimenti si confrontano i valori risultanti dalla stessa coppia esaminata nei due diversi ordini e il link diventa `A rdfs:subClassOf B`, se $f(A,B) > f(B,A)$, oppure `B rdfs:subClassOf A`, se $f(A,B) < f(B,A)$.
- *Post-processing*; una volta stabiliti i link tra le ontologie si utilizza l'Alignment API per cercare le informazioni sullo schema di ciascuna ontologia ed infine si invoca Jena, uno strumento di reasoning, che permette di dedurre altre relazioni implicite tra le due ontologie, sulla base di quelle trovate dall'API e di quelle costruite con BLOOMS.

BLOOMS è stato valutato sia come strumento per l'allineamento di ontologie, confrontandolo con altri sistemi dell'Ontology Alignment Evaluation Initiative¹³⁵ (OAEI), sia come strumento per l'integrazione a livello di schema nei LOD, comparandolo ad altri sistemi di ontology matching e utilizzando i dataset dei Linked Open Data.

In questa seconda valutazione, sono stati scelti diversi LOD dataset di domini sia generici che specifici visibili nella Tabella 3, e si sono eseguiti i vari allineamenti, confrontando poi i risultati sia tra loro che con quelli ottenuti dagli altri sistemi, raggiungendo, come si può vedere dalla Tabella 4, i migliori risultati sia in termini di *recall* che di *precision*.

Schema	LOD datasets	D	# C	Linked datasets
DBpedia	DBpedia	4	204	Geonames, US Census, Freebase
Geonames	Geonames, Geospecies	2	11	DBpedia, Jamendo, FOAF Profiles
Music Ontology	Jamendo, Music Brainz, DBTunes	4	136	GovTrack, DBpedia, Geonames
BBC Program	BBC Programs, BBC Music	4	100	BBC Music, BBC Playcount Data
FOAF Profiles	FOAF, Music Brainz	3	16	Crunch Base, QDOS, SIOC Sites
SIOC	DBpedia, LinkedMDB	2	14	Virtuoso Sponger, FOAF Profiles, SemanticWeb.org
AKT Reference Ontology	ACM, DBLP	5	17	Pisa, IEEE, eprints
Semantic Web Conference Ontology	SW Conference Corpus	5	177	SemanticWeb.org, Revyu

Tabella 3: dataset utilizzati per la valutazione di BLOOMS.

¹³⁵ <http://oei.ontologymatching.org/>

Linked Open Data Schema Ontology Alignment													
Test	Alignment API OMViaUO		RiMoM		S-Match		AROMA		BLOOMS				
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec			
M,B	0.4	0	1	0	err	err	0.04	0.28	0	0	0.63	0.78	
M,D	0	0	0	0	err	err	0.08	0.30	0.45	0.01	0.39	0.62	
F,D	0	0	0	0	err	err	0.11	0.40	0.33	0.04	0.67	0.73	
G,D	0	0	0	0	err	err	0.23	1	0	0	0	0	
S,F	0	0	0	0	0.3	0.2	0.52	0.11	0.30	0.20	0.55	0.64	
W,A	0.12	0.05	0.16	0.03	err	err	0.06	0.4	0.38	0.03	0.42	0.59	
W,D	0	0	0	0	err	err	0.15	0.50	0.27	0.01	0.70	0.40	
Avg.	0.07	0.01	0.17	0	NA	NA	0.17	0.43	0.25	0.04	0.48	0.54	

Tabella 4: risultati della valutazione di BLOOMS.

Successivamente, è poi stato implementato BLOOMS+ [29], una versione aggiornata del sistema, contenente miglieorie sia nella metrica utilizzata per determinare quali classi allineare tra le due ontologie, sia nella considerazione di informazioni contestuali per supportare o rifiutare con più sicurezza un collegamento. Esso, infatti, utilizza un metodo per il calcolo della similarità delle classi, basato sulle gerarchie delle loro categorie, più sofisticato e allinea quelle con i risultati più alti, calcolati anche con l'utilizzo di informazioni sulla somiglianza contestuale, ottenendo risultati ancora più soddisfacenti.

4.2. AGROPub

Un altro sistema per la pubblicazione di risorse, è AGROPub [30], sviluppato per aiutare i fornitori di informazioni agro-ambientali nell'annotazione, pubblicazione ed integrazione dei loro dati nei Linked Open Data.

Il sistema comprende servizi e strumenti che permettono a chi possiede informazioni pubblicabili di annotare le proprie risorse con concetti appartenenti ad ontologie di dominio agricolo e ambientale, di generare e pubblicare descrizioni in RDF relative ad esse e di creare link tra le risorse e i LOD.

Il sistema si basa su **TaToo (Tagging Tool)**¹³⁶ framework che mette a disposizione diversi elementi per l'elaborazione di ontologie, tra cui diverse ontologie condivise di

¹³⁶ <http://www.tatoo-fp7.eu/tatooweb/about>

domini specifici, e utilizza un nuovo modello per la creazione di risorse di tipo ambientale, chiamato **MERM**¹³⁷ (**Minimal Environmental Resource Model**).

AGROPub, utilizza queste basi, per ottenere un'architettura estensibile e service-oriented, che aiuta sia i fornitori nell'integrazione semantica delle risorse all'interno dello spazio di dati globale dei LOD, sia i consumatori, nella visualizzazione delle informazioni e nel completamento dell'integrazione, interagendo con esse nel tempo attraverso uso, ricerca, annotazione, interlinking e valutazione.

AGROPub ha un'architettura client-server costruita su tre livelli, dati, servizi e presentazione; in quest'ultimo si ha l'interfaccia utente costituita dalla desktop application, AGROPub Manager, la quale permette di accedere ai vari tool messi a disposizione dal sistema; nel livello intermedio, invece, sono implementati i vari servizi del sistema, che possono essere così definiti:

- *Publishing and Annotation*; servizio che fornisce la possibilità di generare e pubblicare descrizioni RDF delle risorse agro-ambientali, facendo riferimento al modello MERM; questo le descrive combinando un insieme di informazioni di base, come le loro pagine Web, l'autore e la data di creazione, con un insieme di tag di annotazione basati su ontologie di dominio specifico, creando così collegamenti in forma di URI, tra queste ultime e le risorse; infine la descrizione viene memorizzata nell' AGROPub RDF repository, concludendo il processo di pubblicazione.

Si può quindi notare come con questo tipo di annotazione, ovvero tramite la selezione di tag predefiniti, sia i fornitori che, successivamente, i consumatori, possono creare una sorta di allineamento tra i dati e le ontologie di dominio pre-esistenti, facilitando la loro successiva integrazione nei LOD.

- *Semantic Linking*; servizio che permette la creazione di collegamenti tra le risorse agro-ambientali, indipendentemente dal fatto che le loro descrizioni in RDF risiedano nello stesso o in differenti repository; il servizio permette di stabilire automaticamente link generici identificando nei vari repository le risorse simili, utilizzando per la generazione dei collegamenti semantici le annotazioni *ontology-based*. Consente, inoltre, di stabilire manualmente i link *property-based*, più specifici come `skos:related`, `skos:closeMatch`,

¹³⁷ <http://www.tatoo-fp7.eu/tatooweb/about-tatoo-0>

skos:exactMatch, lasciando all'utente la possibilità di scegliere quali e come connettere le risorse.

- *Semantic Search*; servizio che implementa un semplice servizio di ricerca di tipo *concept-matching* tra le risorse agro-ambientali collegate; all'utente è data la possibilità di creare una query selezionando gli elementi dalle ontologie utilizzate in AGROPub, il servizio poi la trasformerà in una query SPARQL equivalente in modo da risultare eseguibile sui vari repository RDF disponibili.
- *Semantic Navigation*; servizio che consente agli utenti, per mezzo di un'interfaccia apposita, di navigare tra le risorse collegate, attraverso i vari link stabiliti; l'utente seleziona la prima risorsa, quindi il servizio propone i link che partono da essa e poi, scegliendo uno di essi, crea la query SPARQL necessaria alla navigazione.
- *Resource Evaluation*; servizio di valutazione delle risorse che consente agli utenti di esprimere il proprio parere in merito ai collegamenti tra le risorse agro-ambientali; gli utenti possono infatti valutare sia le risorse che le rispettive annotazioni.

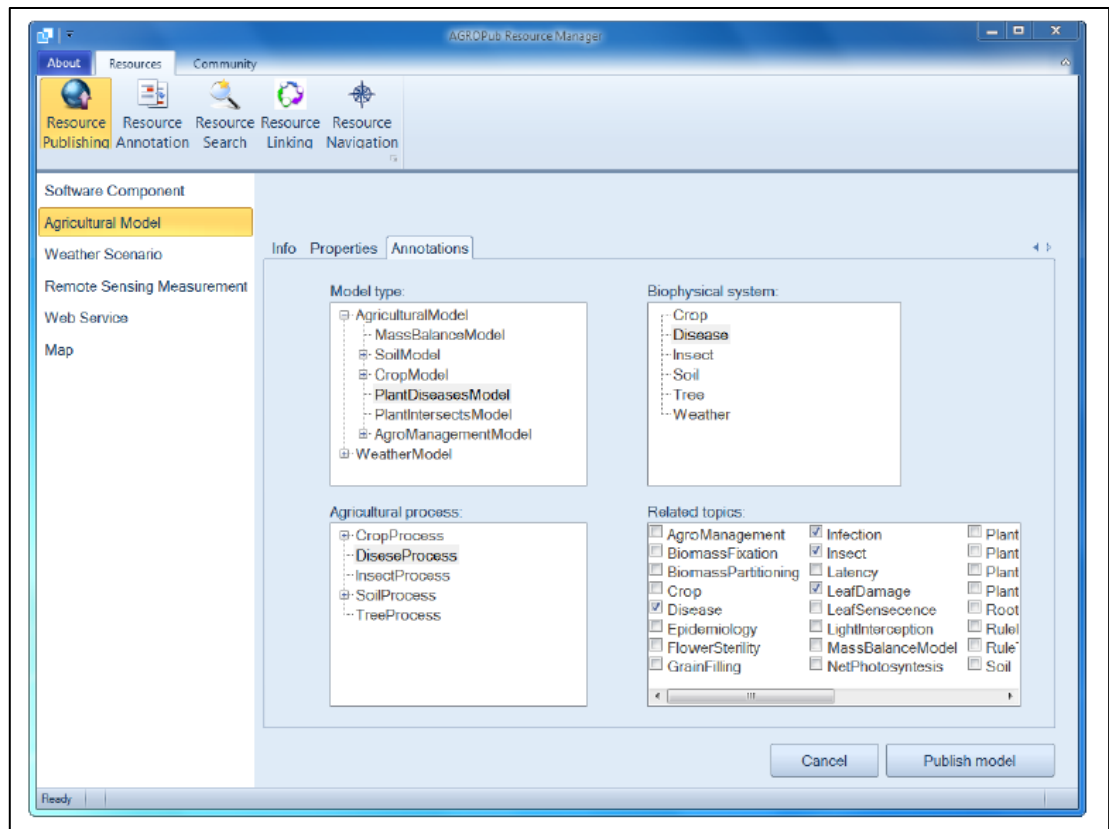


Figura 27: schermata di AGROPub Manager.

4.3. Stratosphere

Stratosphere¹³⁸ è un sistema open-source per il cluster e cloud computing, utilizzato per l'analisi dei Big Data. Esso comprende diversi insiemi di strumenti, utilizzati per complessi scopi nell'elaborazione dei dati, ed è stato esteso ed usato anche per l'integrazione di grandi dataset appartenenti ai Linked Open Data, in particolare per dimostrare la possibilità di unificare in un unico dataset i dati messi a disposizione con l'Open Government e i LOD più popolari, come DBpedia e Freebase [31].

Il progetto ha visto l'utilizzo di dati provenienti da diverse sorgenti, come siti e database, in diversi formati, relativi all'economia americana: si sono infatti usate informazioni riguardanti entrate ed uscite del governo americano, i dati anagrafici sui componenti del US Congress, e dati relativi all'argomento presenti in Freebase.

Ovviamente, per ottenere un insieme integrato di dati, è necessario integrare le fonti di dati dal punto di vista del formato, dello schema e della semantica, seguendo l'approccio classico di integrazione dei dati.

In primo luogo, i dataset vengono scaricati e convertiti nello stesso formato, in modo da eliminare l'eterogeneità tecnologica.

Quindi, le singole fonti di dati vengono esaminate con specifiche operazioni di *scrubbing*, atte ad eliminare gli errori che spesso vi occorrono; poi, per rimuovere l'eterogeneità legato allo schema tra le singole sorgenti, gli attributi si mappano nello schema integrato, mentre per eliminare l'eterogeneità semantica, i vari valori vengono convertiti in una forma comune, definita dal progettista.

Infine, i record di fonti differenti che rappresentavano la stessa entità del mondo reale, sono stati fusi insieme e si sono creati link tra di essi.

Stratosphere, a differenza dei normali DBMS, supporta dati sia text-based che semi-strutturati ed è quindi adatto per le operazioni di *cleansing*, pulizia dei dati, che spesso è necessaria nel trattamento di dati pubblici; inoltre, ha un'architettura particolare, composta da un motore di esecuzione, che permette l'elaborazione di massicci flussi di dati paralleli (Nephele), un modello di programmazione, PArallelization ConTract (PACT), implementato sul precedente, che permette lo svolgimento in parallelo di operazioni definite dall'utente, ed un framework estensibile, SOPREMO, che utilizza il

¹³⁸ <https://stratosphere.eu/>

linguaggio Json ed il suo modello di dati, e che contiene strumenti di alto livello per la realizzazione di operazioni di dominio specifico e di tipo relazionale; grazie ad essa, Stratosphere, permette di creare script specifici per l'integrazioni dei dati e per eseguire query in Jaql (Json Query Language), che interrogano più sorgenti contemporaneamente.

I dataset considerati sono stati integrati attraverso la realizzazione di script specifici; in Figura 28 si ha il riassunto delle operazioni eseguite. In particolare, per quanto riguarda il *record linkage* [24], si sono creati due nuovi operatori per facilitare la formulazione delle condizioni e dei parametri: `link record`, che esegue la risoluzione di base e ritorna il link tra i due record, e `cluster record`, che aggiunge i link anche a tutti gli elementi appartenenti agli stessi cluster dei record analizzati.

Dopodiché, una volta generati i link, questi vengono utilizzati per eseguire la *data fusion* [34], processo di integrazione di molteplici dati e conoscenze, rappresentanti lo stesso oggetto del mondo reale in un'unica, coerente e precisa rappresentazione, unendo i record collegati in un'unica entry; quest'ultime, prese insieme, formano poi quella visione complessiva e globale dei dati analizzati, accessibile per mezzo di query complesse, che si era stabilito di ottenere.

Operator	Syntax
Data Scrubbing	<code>scrub input with rules</code>
Entity Extraction	<code>extract from input into entities</code>
Record Linkage	<code>link/ cluster records input where similarity_condition partition on partition_key into target_expr</code>
Data Fusion	<code>fuse input with weights weight_map into result</code>
Dictionary Lookup	<code>replace input with dictionary default default_expr</code>

Figura 28: sommario delle operazioni di elaborazione dei dati con Stratosphere.

Oltre all'integrazione dei dati, per dare ai cittadini la possibilità di visionare ed interrogare in modo più semplice il dataset globale, si è implementato il portale **GovWild**¹³⁹ (Government Web Data Integration for Linked Data) [32], il quale aumenta la trasparenza delle transazioni monetarie governative, integrando le sorgenti di dati americane, europee e tedesche (Figura 29). Il dataset integrato è accessibile sia con un'interfaccia Web, attraverso interrogazioni SPARQL, che per il download, trasformando in modo automatico i file in Json nel formato di triple RDF, N-Triples, o in un'ontologia in OWL. Freebase, all'interno della Linked Open Data cloud, risulta essere la sorgente dati con più link ai dati di GovWild.

GovWILD

Barack Obama

Born: 04.08.1961
 Birth Place: Honolulu, Hawaii
 Nationality: [American](#)
 Political Party: Democratic
 Spouse(s): [Michelle Obama](#)
 Children: [Malia Ann Obama](#),
[Natasha Obama](#)
 Residence: The White House
 Chicago, Illinois
 Religion: Christianity

Relations:

Same as:

[The New York Times](#) [DBpedia](#) [Freebase](#)

The New York Times articles:

[Britain Reaffirms Support for Afghanistan Effort](#)
[Obama Pledges New Aid to Palestinians](#)
[A Modest Proposal: A King and Queen for America](#)

Data sheet based on sources:

[Freebase](#) | [US Congress](#) | [US Earmarks](#)

Figura 29: un'immagine del portale GovWild.

¹³⁹ www.govwild.org

4.4. WebSmatch

WebSmatch [33], è un ambiente flessibile per l'integrazione di dati Web, basato su un reale scenario di dati presi da Data Publica¹⁴⁰, compagnia francese che dà valore aggiunto ai propri dataset attraverso la creazione di dati integrati e la messa a disposizione di strumenti di visualizzazione. **WebSmatch** è uno strumento di supporto all'intero processo di importazione, perfezionamento ed integrazione delle fonti di dati, ed utilizza strumenti aggiuntivi per ottenerne una visualizzazione di alta qualità.

WebSmatch è una Rich Internet Application (RIA), ovvero può essere utilizzato, insieme a tutti i suoi componenti, senza installazione e come Web service.

Il primo passo per l'elaborazione dei dati vede la loro ricerca ed estrazione in modo da avere tutti i dati in un unico formato, DSPL (Data Set Publishing Language). I dati strutturati sono convertiti in modo automatico ma, anche se esistono formati strutturati per condividere e pubblicare i dati, la maggior parte dei dati pubblici disponibili sul Web sono fogli di calcolo Excel. Di conseguenza è necessario rilevare la differenza tra dati e metadati contenuta in essi; per fare ciò WebSmatch, si avvale di tecniche di visione computerizzate, per rilevare le complesse rappresentazioni tabulari presenti nei documenti, e di tecniche di *machine learning* per automatizzare i precedenti procedimenti.

La vera e propria integrazione delle sorgenti è ottenuta utilizzando tecniche di *schema matching*, in modo da unificare a livello di schema le strutture dati; spesso queste informazioni nel contesto degli open data, sono però assenti; WebSmatch utilizza **YAM++**¹⁴¹ (**Yet Another Matcher**), un potente strumento per lo schema matching e l'allineamento di ontologie, il quale combina 14 diverse tecniche per l'individuazione delle corrispondenze basate sulle similitudine tra stringhe, sulle informazioni presenti in dizionari e in thesauri come WordNet, o sulla correlazione tra istanze. In questo modo si vede la sovrapposizione tra le sorgenti, utile alla creazione di dati integrati, ed è inoltre possibile generare la loro corretta descrizione in DSPL, basandosi su concetti condivisi. In WebSmatch, infatti, è possibile eseguire una fase di *clustering* automatico basato sui collegamenti messi in luce dalla fase precedente di matching e sul calcolo delle distanze tra ciascuna coppia di documenti, che permette di visualizzare le somiglianze semantiche tra le sorgenti.

¹⁴⁰ <http://www.data-publica.com>

¹⁴¹ <http://nb.vse.cz/~svabo/oei2011.5/eval.html>

Infine, le informazioni sono strutturate ed esportate in un file DSPL, il quale viene poi caricato nel database di Data Publica. I dati sono accessibili con generiche API atte alla conversione in diversi formati (XML, CSV, JSON) o all'elaborazione con strumenti grafici, come la ricca collezione di tool di Google Data Explorer¹⁴².

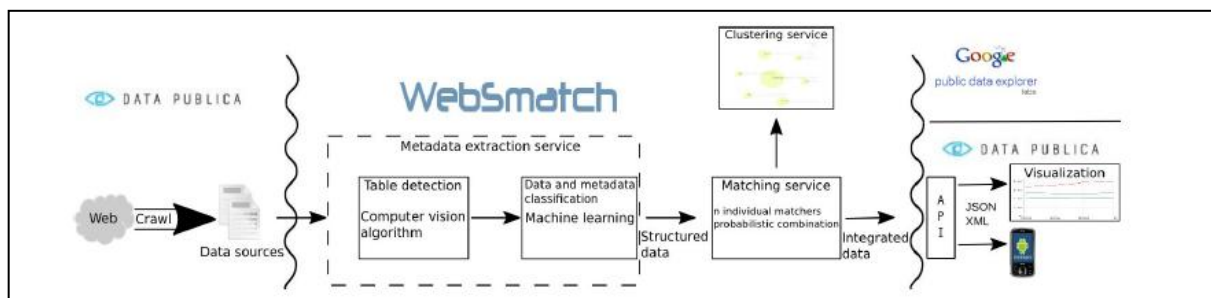


Figura 30: il processo di Data Integration di WebSmatch.

¹⁴² <http://www.google.com/publicdata/directory>

Capitolo 5

5. Caso d'uso: dal dataset Precarietà ai LOD

Nel Capitolo 3 è stato analizzato il procedimento generale di creazione dei Linked Open Data, i vari tool a disposizione e le eventuali carenze. In questo capitolo, andremo ad utilizzare tali tool su un caso d'uso reale: ci è stato fornito un dataset relativo al progetto sulle precarietà giovanili, dalla Pubblica Amministrazione del Comune di Modena, il quale verrà utilizzato per la conversione in Linked Open Data.

In particolare, i dati utilizzati nel caso di studio, sono stati presi dal Progetto di ricerca sulle Precarietà, progetto avviato dall'Assessore all'Innovazione e Comunicazione, ed al Patrimonio, Politiche giovanili e Cooperazione internazionale, Fabio Poggi, con la collaborazione del Dipartimento di Scienze del Linguaggio e della Cultura, dell'Università di Modena, coordinata dal Prof. Claudio Baraldi e dal Dott. Federico Farini. L'insieme di dati utilizzati, ci è poi stato fornito dal Dott. Sergio Ansaloni, responsabile di Stradanove, centro studi e documentazione sulla condizione giovanile per l'Assessorato alle Politiche Giovanili.

Il processo di conversione dei dati, in Linked Open Data, prevede per prima cosa, la trasformazione dei dati nel formato RDF; quindi, il loro arricchimento semantico, a livello di schema, attraverso l'annotazione; poi, l'utilizzo di quest'ultima per il linking semantico ad un altro dataset presente nella LOD Cloud, nel nostro caso WordNet¹⁴³; ed

¹⁴³ <http://www.cogsci.princeton.edu/wn>

infine, la loro pubblicazione e possibilità di interrogazione attraverso un endpoint SPARQL.

Le procedure utilizzate nel progetto, gli strumenti e i problemi incontrati, verranno descritti in modo più dettagliato nei paragrafi seguenti.

5.1. Il dataset sulle “Precarietà Giovanili” del Comune di Modena

L’Assessorato alle Politiche Giovanili, ha promosso la raccolta dati del progetto ‘Precarietà Giovanili’, al fine di analizzare la situazione di precarietà lavorativa ed emotiva che stanno attraversando i giovani, in particolare nella città di Modena, facendo emergere come non vi sia solo una semplice instabilità in ambito lavorativo, data dall’assenza di posti di lavoro, ma esista un senso di incertezza che si ripercuote in tutti gli aspetti della vita quotidiana: la formazione, gli affetti, la famiglia, e così via.

I dati utilizzati per la trasformazione in Linked Open Data, rappresentano solo una parte della ricerca del Comune di Modena, che prevede una raccolta dati più ampia.

Il progetto è composto, in primo luogo, dalla creazione di un questionario (il questionario è riportato in Appendice D), da distribuire e raccogliere in diversi luoghi di aggregazione della provincia modenese; quindi prevede, la realizzazione di diversi Focus Group, gruppi di analisi, incaricati di esaminare i risultati dei questionari nei diversi gruppi della città e di creare ulteriori questionari, o altri strumenti, per raccogliere dati aggiuntivi, sugli aspetti più significativi emersi in quei contesti; infine, con la collaborazione, dell’ente comunale locale, si arriverà all’integrazione dei dati, in modo da potervi ricavare un report reale sul profilo socio-culturale dei giovani, al giorno d’oggi.

Al momento, la fase di raccolta dati è ancora in corso, ma per lo studio sulla pubblicazione e linking dei dati nei LOD, è stato sufficiente l’utilizzo di quelli ricavati dalla compilazione dei primi questionari.

Il questionario, anonimo, è composto da 29 domande, che spaziano dalle caratteristiche personali dell’intervistato, come la sua provenienza e la sua situazione occupazionale, alle valutazioni degli ambienti sociali, famiglia, lavoro o scuola, coppia ed amici, in cui l’interessato si sente più o meno a suo agio, in cui ripone più fiducia e si sente più stabile, o in cui ha più incertezze e difficoltà, confrontando queste sensazioni con il proprio recente passato, e dando valutazioni su cosa si aspetta dal proprio futuro.

Il normale trattamento dei dati prevede l'inserimento manuale dei questionari all'interno di file **SPSS (Statistical Package for Social Science)**, un software di IBM, utilizzato per l'analisi statistica dei dati: il file è composto da due tabelle molto semplici, una contenente le etichette corrispondenti alle diverse domande del questionario, e una contenente i dati effettivi in formato numerico. Il software, poi, dispone di diversi strumenti che permettono l'analisi dei dati, ricavandone linee di tendenza e grafici.

5.2. Modellazione e conversione dei dati in RDF

Una volta ottenuto il dataset nel formato originale, il file `precarietà.spss`, si sono subito studiate le sue possibili modalità di conversione in database, in modo da ottenere i dati in un formato più indicato alla loro elaborazione.

Per prima cosa i dati sono stati esportati in formato Excel, ricavando un'unica tabella complessiva contenente circa un centinaio di record; dopodiché si è definito uno schema Entity-Relationship del questionario, in modo da rendere i dati strutturati. Per ottenere uno schema più chiaro e facilmente convertibile in un database, alcune delle domande sono state semplificate: ad esempio nelle domande 24-27 si è evitato di modellare il riferimento temporale "*negli ultimi sei mesi...*", il quale aggiunge un valore trascurabile all'informazione, che rimane lo stesso contestualizzata nel presente; inoltre non sono state modellate le domande 8-11, 28 e 29, in quanto supponevano un confronto tra le opinioni dell'intervistato in diversi tempi, "*Rispetto ad un anno fa, se pensi al tuo futuro cosa provi?*", informazione non rappresentabile in E/R.

Lo schema E/R vede la modellazione di tre gerarchie principali, `Persona`, `Futuro` e `Ambiente_sociale`, le quali sono state messe in relazione tra loro attraverso associazioni rappresentanti le domande del questionario: ogni associazione ha poi diversi attributi che danno la possibilità di specificare le risposte date da ciascun intervistato.

Prendiamo, ad esempio, la relazione `intervistato-futuro-ambiente_sociale`, che modella le domande 6 e 7 del questionario: essa permette di specificare per ogni intervistato, che cosa pensa del proprio futuro, prossimo e lontano, in ciascun ambiente sociale, specificando un valore in tutte le voci che compongono l'attributo complesso `RELAZIONI`; attraverso gli attributi binari dell'associazione, si possono, infatti,

assegnare gli interi 0 o 1, per indicare il primo o il secondo valore dell'attributo, ad esempio per l'attributo *Fiducia/Sfiducia*, 0 indicherà fiducia e 1 sfiducia. Lo schema E/R risultante è mostrato in Figura 31; esso è stato poi convertito nel modello relazionale (riportato in Appendice E), e successivamente in database.

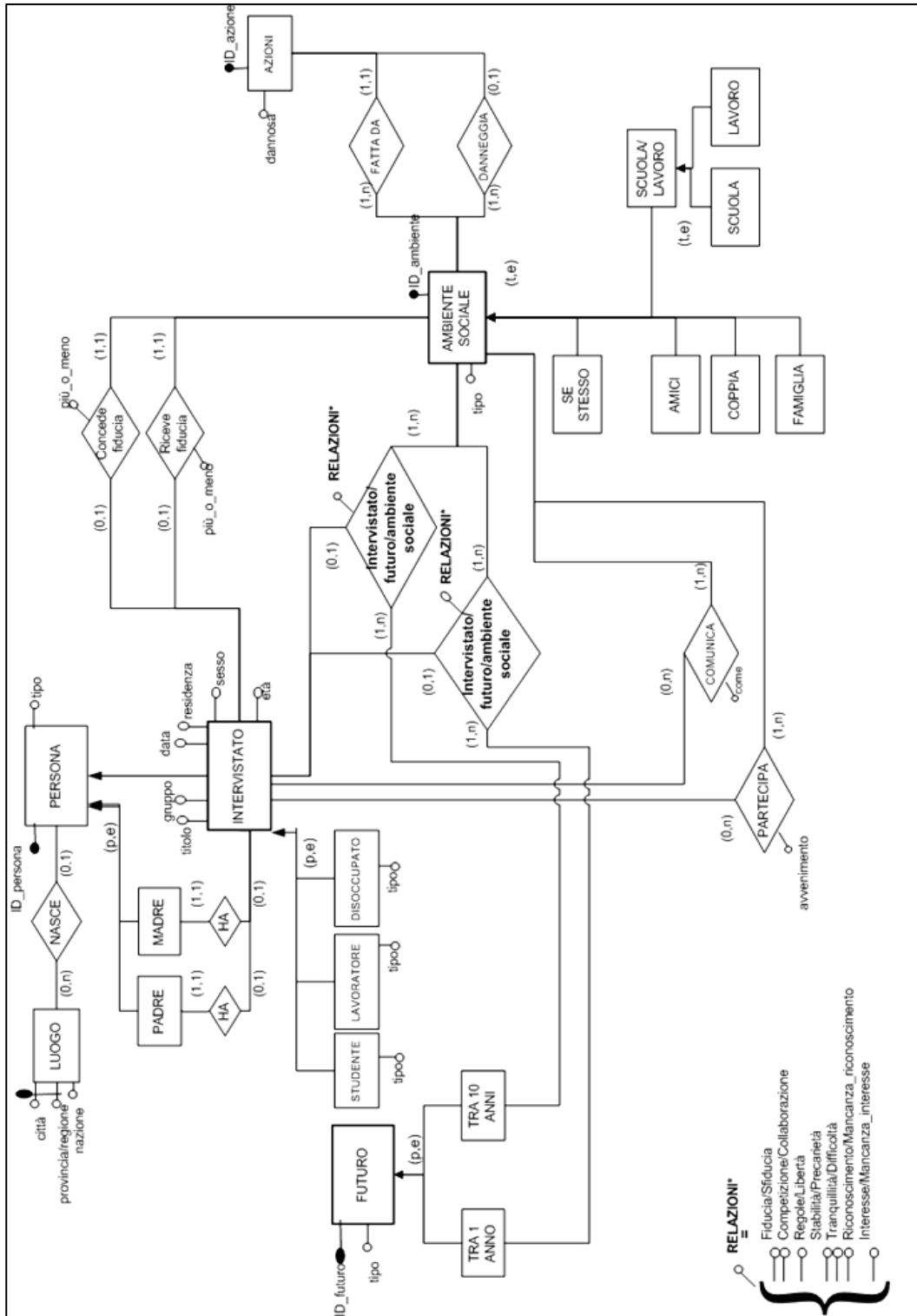


Figura 31: schema E/R ricavato dal questionario.

Il database 'precarieta' è stato creato in MySQL¹⁴⁴, DBMS open source molto diffuso, dopodiché è stato popolato importando i dati dal foglio Excel.

Quindi si sono iniziati a studiare i modi per effettuare la pubblicazione dei dati come Linked Open Data: partendo da un database si è subito pensato all'utilizzo del tool, precedentemente testato, D2R Server (Paragrafo 3.2.1), automatizzando così la trasformazione dei dati in RDF.

Si è creato il file di mapping, `mapping_precarieta.ttl`, in D2RQ, linguaggio che permette di mappare ciascuna tabella del database in una classe RDF, `rdfs:Class`, e ciascuna colonna in una sua proprietà, `rdfs:Property`. Rende inoltre esplorabili le istanze del database, creando dei mapping, del tipo `d2rq:join`, al posto delle Foreign Key presenti, consentendo il collegamento tra i dati di tabelle diverse. In tal modo, le tabelle contenenti solo chiavi esterne, non vengono automaticamente trasformate in classi, ma in proprietà, che collegano i dati presenti nelle tabelle a cui le chiavi esterne fanno riferimento. Un esempio nel database, è la tabella `Nasce`, che invece di diventare una classe, è stata trasformata in `rdfs:Property`, in modo da collegare l'identificatore della tabella `Persona` con quello della tabella `Luogo`, esprimendo così la località di nascita di ciascun individuo.

Inoltre, nel file di mapping, sono specificate le gerarchie presenti tra le classi: grazie al comando `d2rq:additionalClassDefinitionProperty` è, infatti, possibile aggiungere le proprietà `rdfs:subClassOf` e `rdfs:subPropertyOf`, arricchendo i dati RDF a livello di schema.

Attraverso il linguaggio D2RQ è possibile aggiungere qualsiasi relazione tra le classi e le proprietà del database, ma non sempre è possibile definirla correttamente; ad esempio, si possono aggiungere relazioni OWL o RDFS, che arricchiscono i dati a livello di schema, utilizzando le classi e le proprietà del database convertito come soggetti ed oggetti delle relazioni, ma non posso aggiungere relazioni che facciano riferimento ad elementi non appartenenti al mio dataset, né modificare manualmente la sintassi utilizzata.

Non essendo quindi più possibile proseguire il processo di realizzazione di Linked Open Data con lo stesso tool, si è cercato di ottenere una copia dei dati in RDF, in modo da

¹⁴⁴ <http://www.mysql.it/>

proseguire con altri strumenti. D2R Server mette a disposizione l'istruzione `dump-rdf`, in questo caso si è utilizzato il comando

```
dump-rdf -m mapping_precarieta.ttl -f RDF/XML  
-b http://localhost:2020/precarieta/ -o output.rdf
```

con il quale è stato possibile ottenere una copia di tutti i dati nel formato RDF desiderato; più precisamente, si è scelto il formato RDF/XML, in modo da proseguire nel progetto con Protégé, editor di ontologie che verrà analizzato più approfonditamente nel paragrafo successivo (Paragrafo 5.2).

5.3. Annotazione semantica e linking a WordNet

L'obiettivo del progetto, è quello di ottenere dei dati RDF, contenenti informazioni di tipo semantico, in modo da poterli collegare alla Linked Open Data Cloud, non attraverso semplici link tra istanze, ma a livello di schema. Si è pensato, quindi di annotare lo schema in modo da arricchirlo semanticamente e allo stesso tempo collegare l'intero dataset al LOD Cloud attraverso il mapping con i synset di WordNet (Paragrafo 5.3.1).

Per annotazione si intende, la fase dell'operazione di disambiguazione [36] che assegna a ciascun termine utilizzato il proprio significato: più precisamente si utilizzano dei vocabolari, o thesauri condivisi, dai quali, dato il lemma, o vocabolo, desiderato, si ottiene la sua glossa, o definizione, dando la possibilità di comprendere correttamente il significato della parola nel suo contesto.

Nel paragrafo verrà descritto in modo più approfondito il thesaurus WordNet e la metodologia utilizzata per l'annotazione e il linking del dataset Precarietà.

5.3.1. WordNet

Analizzata in precedenza la carenza di semantica dei LOD (Capitolo 4), ci si è soffermati principalmente, sulla ricerca di possibili metodi per la creazione di link tra i dataset, a livello di schema.

Si sono quindi scelte, come 'oggetti' degli RDF link, le risorse del dataset WordNet [36], database lessicale in lingua inglese sviluppato presso l'università di Princeton sotto la direzione del professore Gorge A.Miller, successivamente convertito in RDF e inserito nei LOD.

È stato scelto proprio WordNet, in quanto non è semplicemente un dizionario di termini inglesi, ma lega, a ciascuno di essi, numerose proprietà di tipo semantico: i termini, infatti, non sono disposti seguendo l'ordine alfabetico, ma per affinità di significato, andando a costituire i cosiddetti *synset*, gruppi di termini aventi tra loro relazioni di sinonimia, ai quali è associato un unico significato descritto tramite una *glossa*. WordNet suddivide i termini in quattro categorie sintattiche: nomi, verbi, aggettivi ed avverbi; un termine, ovviamente, può essere presente in più di una categoria sintattica, ed inoltre, può possedere più di un significato, ed essere, quindi, presente in più *synset*. Un'altro elemento rilevante, che contraddistingue WordNet da un semplice dizionario di vocaboli, è la presenza di relazioni fra i *synset*: essi possono essere collegati da diversi tipi di relazioni, come ad esempio, l'iperonimia e l'iponimia, che indicano legami di generalizzazione e specificazione tra i termini, tramite cui si è in grado di creare, all'interno dell'intera categoria sintattica, gerarchie di significato.

WordNet fa parte delle risorse pubblicate attraverso i LOD. La sua versione RDF include 2.727.068 triple rappresentanti i *synset* e le relazioni fra questi, che possono essere esplorate, grazie all'interfaccia web al sito <http://wordnet.rkbexplorer.com/> ed interrogate attraverso un endpoint SPARQL. Esso è inserita all'interno della LOD Cloud, e presenta innumerevoli link con altri dataset, come DBPedia e OpenCyc.

5.3.2. Creazione dei link con Protégé

Nella fase precedente di conversione dei dati con D2R Server, si è eseguito un primo livello di annotazione. Sono, infatti, già a disposizione, nel linguaggio di mapping D2RQ, le istanze della classe `owl:AnnotationProperty`, `rdfs:label` e `rdfs:comment`, le quali permettono di inserire un'etichetta e un commento alle classi e alle proprietà del database.

Nel nostro caso si sono inseriti, come *label*, i lemmi e, come *comment*, le glosse in italiano, prese da MultiWordNet¹⁴⁵, thesaurus multilingua che estende il più noto WordNet. MultiWordNet [37] è un database lessicale multilingua, sviluppato presso l'istituto FBK-irst di Trento, che si propone di allineare i diversi WordNet in diverse lingue (Italiano, Rumeno, Latino, Ebraico, Spagnolo e Portoghese), alla versione originale in lingua inglese, Princeton WordNet 1.6. Esso ha inoltre un'interfaccia web,

¹⁴⁵ <http://multiwordnet.fbk.eu/>

disponibile al sito http://multiwordnet.fbk.eu/online/multi_wordnet.php, con la quale è possibile verificarne il funzionamento.

MultiWordNet, è stato utilizzato perché contenente i termini in italiano e tutte le corrispondenze tra questi, e quelli presenti in WordNet, in inglese, con i quali vorremmo poi creare i link: in questo modo, è stato possibile il suo utilizzo come ponte tra l'annotazione in italiano e la corrispondente annotazione in inglese. Inizialmente i termini sono stati cercati grazie all'interfaccia on-line, ma poi si è passati ad una ricerca più approfondita all'interno del database, Italian WordNet, progetto distribuito sottoforma di dump di tabelle per un DBMS MySQL, contenente 38.654 *synset* e 45.089 *lemmi*.

Per poter inserire all'interno dell'annotazione, il collegamento ai lemmi inglesi di WordNet, è necessario creare una nuova istanza, personalizzata, della classe `owl:AnnotationProperty`; infatti si devono collegare istanze della classe `rdfs:Class`, le classi del database, con delle istanze della classe `rdfs:Resource`, i termini di WordNet, e per questo non è possibile sfruttare nessuna proprietà già presente negli standard RDF-Schema o OWL: si è quindi creata una proprietà specifica, che abbiamo chiamato `vocab:annotation` (il namespace `vocab`, fa riferimento al vocabolario che viene creato automaticamente da D2R in fase di conversione dei dati in RDF).

Non potendo creare la nuova property con D2R Server, si è passati all'utilizzo di un editor per ontologie in modo da arricchire correttamente i dati in RDF/XML a livello di schema: Protégé [38].

Protégé è una piattaforma open-source in Java, sviluppata dal Centro di Biomedical Informatics Research di Stanford, che fornisce strumenti per la modellazione di ontologie di dominio e la costruzione di applicazioni knowledge-based. Esso offre un supporto alla creazione, visualizzazione e manipolazione di ontologie in vari formati di rappresentazione; possono infatti, essere importate ed esportate in RDF, RDF-Schema, OWL e XML-Schema, ovvero ontologie per il Web semantico, che possono contenere descrizioni delle classi, proprietà e delle relative istanze.

Proprio la possibilità di definire le caratteristiche logiche delle classi come espressioni OWL, ci ha permesso di annotare correttamente i dati, inserendo la dicitura

```
<owl:AnnotationProperty rdf:about="&vocab;annotation"/>
```

per poi utilizzare `vocab:annotation`, come tag di annotazione personalizzato.

Nell'annotazione sono stati poi inseriti i link alla versione RDF di WordNet, i cui URI sono composti dal namespace di base, più la parola 'synset' e le caratteristiche del termine cercato, ovvero *http://wordnet.rkbexplorer.com/id/synset-parola-categoria_sintattica-numero_senso*. Questa costruzione degli URI, ci ha permesso di verificare la corrispondenza delle parole dall'italiano all'inglese attraverso il database di MultiWordNet, trovando i termini esattamente correlati.

Per esempio, la classe `luogo` con Protégé, risulta con l'annotazione, così definita:

```
<!-- http://localhost:2020/vocab/resource/luogo -->
<owl:Class rdf:about="&vocab;luogo">
  <rdfs:label xml:lang="it">
    luogo, posto, sito
  </rdfs:label>
  <vocab:annotation rdf:datatype="&xsd:anyURI">
    http://wordnet.rkbexplorer.com/id/synset-place-noun-2
  </vocab:annotation>
  <rdfs:comment xml:lang="it">
    qualsiasi area riservata ad un particolare scopo
  </rdfs:comment>
</owl:Class>
```

Tuttavia, pur definendo il tipo di dato `xsd:anyURI`, il link inserito viene visualizzato solo come informazione testuale, come commento, e non risulta un collegamento navigabile.

Abbiamo quindi inserito un nuovo namespace all'inizio del file RDF/XML, `xmlns:id=http://wordnet.rkbexplorer.com/id/`, e modificato, manualmente (non essendo possibile con Protégé), la dicitura delle annotazioni in `<vocab:annotation rdf:resource = "&id;synset-parola-categoria_sintattica-numero_senso"/>`; ad esempio la classe `luogo` diventa:

```

<!-- http://localhost:2020/vocab/resource/luogo -->
<owl:Class rdf:about="&vocab;luogo">
  <rdfs:label xml:lang="it">
    luogo, posto, sito
  </rdfs:label>
  <vocab:annotation rdf:resource="&id;synset-place-noun-2"/>
  <rdfs:comment xml:lang="it">
    qualsiasi area riservata ad un particolare scopo
  </rdfs:comment>
</owl:Class>

```

Una volta re-importata l'ontologia in Protégé, il tool riconosce i link come istanze della radice `owl:Thing`; si è quindi modificata anche la loro definizione, in modo da farle diventare istanze della classe `rdfs:Resource`, rispettando quindi la classificazione della versione RDF di WordNet. Ad esempio il link della classe `luogo`, è stato definito come:

```

<!-- http://wordnet.rkbexplorer.com/id/synset-place-noun-2 -->
<owl:Thing rdf:about="&id;synset-place-noun-2">
  <rdf:type rdf:resource="&rdfs;Resource"/>
</owl:Thing>

```

Il file RDF/XML così modificato, è stato poi utilizzato per la pubblicazione finale dei dati, attraverso il tool LMF (LinkedData Media Framework), e per la loro elaborazione, con il tool Fuseki, passaggi che verranno descritto in modo più dettagliato nel paragrafo successivo.

5.4. Pubblicazione e SPARQL endpoint

Per completare la trasformazione da semplici dati a Linked Open Data, è necessario renderli pubblici e accessibili agli altri utenti attraverso un endpoint SPARQL, punto da cui è possibile visualizzarli, esplorarli ed interrogarli per mezzo di query.

E' possibile inserire i propri dati in un triple store (http://www.w3.org/2001/sw/wiki/Category:Triple_Store), database appositamente costruiti per la memorizzazione e il recupero di dati in formato di triple RDF, e poi aggiungerci il collegamento ad uno SPARQL engine

(http://www.w3.org/2001/sw/wiki/Category:SPARQL_Endpoint), oppure utilizzare strumenti che hanno contemporaneamente, entrambe le funzionalità; per comodità si è scelta quest'ultima soluzione.

Il tool utilizzato è **LMF**¹⁴⁶ (LinkedData Media Framework) [39], una Java Web Application, che facilita gli utenti nella creazione dei Linked Open Data. Esso è open source, e si compone di un modulo centrale, *LMF Server*, e diversi moduli aggiuntivi, uniti dal primo, che vengono utilizzati a seconda delle necessità dell'utente; quelli già implementati sono:

- *LMF SPARQL*, che offre il supporto completo per il protocollo SPARQL 1.1, in modo da modificare l'archivio triple-based;
- *LMF LD Path*, che permette di interrogare facilmente la Linked Data Cloud, grazie al linguaggio LDPPath;
- *LMF Semantic Search*, che offre un servizio di ricerca semantica, basato su Apache SOLR, e che permette la configurazione di diversi indici di ricerca nella stessa istanza di LMF;
- *LMF Linked Data Cache*, che permette l'utilizzo trasparente della cache quando, usando i servizi di ricerca, di interrogazione e di query, si incontrano dati collegati alla LOD Cloud, consentendo di recuperare e memorizzare il contenuto della risorsa remota nel background locale;
- *LMF Reasoner*, che implementa un reasoner, rule-based, consentendo di elaborare regole sullo stile di Datalog, sulle triple RDF;
- *LMF Text Classification*, che fornisce servizi statistici per la classificazione del testo in categorie;
- *LMF Versioning*, che implementa, utilizzando degli snapshot dei metadati, le versioni per i loro aggiornamenti;
- *LMF Stanbol Integration*, che permette l'integrazione con Apache Stanbol per l'analisi e l'interconnessione dei contenuti;
- *LMF SKOS Editor*, che dà la possibilità di visualizzare e aggiornare direttamente il thesaurus SKOS, importato in LMF, utilizzando l'editor SKOSjs.

Grazie a questo tool si è importato il file RDF/XML, precedentemente creato, all'interno di un database, creato automaticamente grazie all'Embedded Database

¹⁴⁶ <http://code.google.com/p/lmf/>

Engine H2¹⁴⁷, ottenendo un archivio triple-based permanente composto da ben 30.392 triple. Quindi si è utilizzato il modulo LMF SPARQL, contenente il servizio Snorql¹⁴⁸, per esplorarlo e verificare la navigabilità dei link tra i dati del database Precarietà e WordNet.

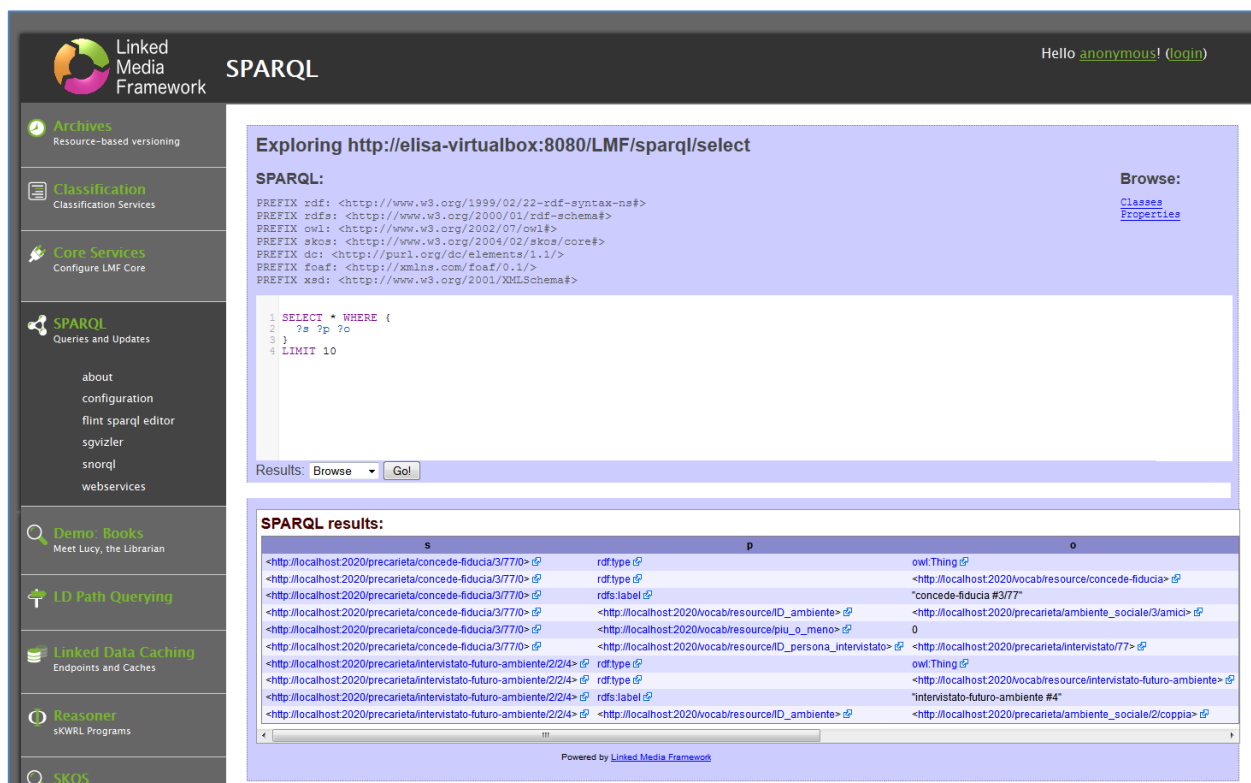


Figura 32: esempio di query in LMF SPARQL.

I link risultano perfettamente funzionanti, se cliccati portano al sito ospitante la versione RDF di WordNet, e se esplorati con il linguaggio SPARQL, si ottengono le informazioni delle relative pagine in RDF (Figura 33).

¹⁴⁷ <http://www.h2database.com/html/main.html>

¹⁴⁸ <http://data.semanticweb.org/snorql/>

Figura 33: esempio di esplorazioni dei link con LMF SPARQL.

Attraverso il linguaggio SPARQL è possibile eseguire qualsiasi tipo di query, in modo da analizzare approfonditamente la costruzione di ciascuna tripla, o ottenere informazioni di insieme sulla complessità dei dati, con la possibilità di ottenere informazioni anche statistiche sui dati pubblicati.

Per la precisione, con LMF, non è possibile utilizzare la sintassi COUNT o GROUP BY nelle query, e risulta quindi più complesso ottenere dati statistici, o visioni d'insieme sui dati complicati; questo risulta invece possibile con altri tool, come **Fuseki** [40], uno SPARQL Server, implementato da Apache-Jena¹⁴⁹, che permette di visualizzare i propri dati RDF, e di analizzarli con query aventi a disposizione un maggiore set di istruzioni grazie al query engine **ARQ**¹⁵⁰. Con quest'ultimo non è possibile però ottenere i dati in RDF sottoforma di link navigabili, ma solo come testo: a seconda, quindi, delle necessità dell'utente, si può scegliere il tool più appropriato.

¹⁴⁹ <http://jena.apache.org/index.html>

¹⁵⁰ <http://jena.apache.org/documentation/query/index.html>

Con Fuseki, si sono quindi eseguite alcune query con funzioni di raggruppamento, per riuscire ad avere una visione globale dei dati presenti nel dataset Precarietà, cercando di tornare ai risultati ottenibili con il software di elaborazione SPSS, inizialmente utilizzato per l'analisi dei risultati dei sondaggi.

Come esempio, prendiamo la ricostruzione delle domande 6 e 7 del questionario, le quali esprimono un giudizio con diverse voci, sull'idea del proprio futuro tra un anno e tra 10 anni, nei diversi ambienti sociali (famiglia, amici, coppia, scuola e lavoro): si è eseguita la query SPARQL dall'endpoint di Fuseki, selezionando le triple necessarie; si è usata l'istruzione *group by()*, per suddividere le risposte per ambiente; si è utilizzata la funzione *count()*, per quantificare le risposte degli intervistati in base al valore della risposta stessa; ed infine, si sono eseguite tante query quante sono le voci di giudizio delle domande, selezionando ogni volta quella appropriata, esportando poi ciascun risultato in formato CSV, in modo da poterlo graficare con gli strumenti di Microsoft Excel.

```
select count(*) ?ambiente ?valore
where {
  ?s rdf:type <http://localhost:2020/vocab/resource/intervistato-futuro-
    ambiente>.
  ?s <http://localhost:2020/vocab/resource/ID_futuro>
    <http://localhost:2020/precarieta/futuro/1>.
  ?s <http://localhost:2020/vocab/resource/ID_ambiente> ?ambiente.
  ?s <http://localhost:2020/vocab/resource/stabilita-precarieta> ?valore
}
group by ?ambiente ?valore
order by ?ambiente ?valore
```

Come possiamo notare, nella clausola *where* sono state inserite le condizioni per la selezione delle triple RDF corrette: la prima prende le triple relative all'originaria tabella *intervistato-futuro-ambiente*, in cui erano state modellate le domande 6 e 7; la seconda seleziona quelle relative al valore temporale 'tra 1 anno', aventi *ID_futuro=1* ('tra 10 anni' è *ID_futuro=2*); la terza, indica di selezionare anche il valore relativo all'ambiente sociale, *ID_ambiente*, che verrà utilizzato come variabile della funzione di raggruppamento; ed infine, la quarta specifica la voce di giudizio che si vuole analizzare, in questo caso l'idea di stabilità o precarietà. In questo modo si ottiene, per ogni ambiente sociale, il numero di intervistati che, tra un anno, pensano ci sarà stabilità e quelli che pensano ci sarà precarietà. I grafici successivi, danno un quadro complessivo dei risultati delle domande 6 e 7 del questionario (Figure 34-40).

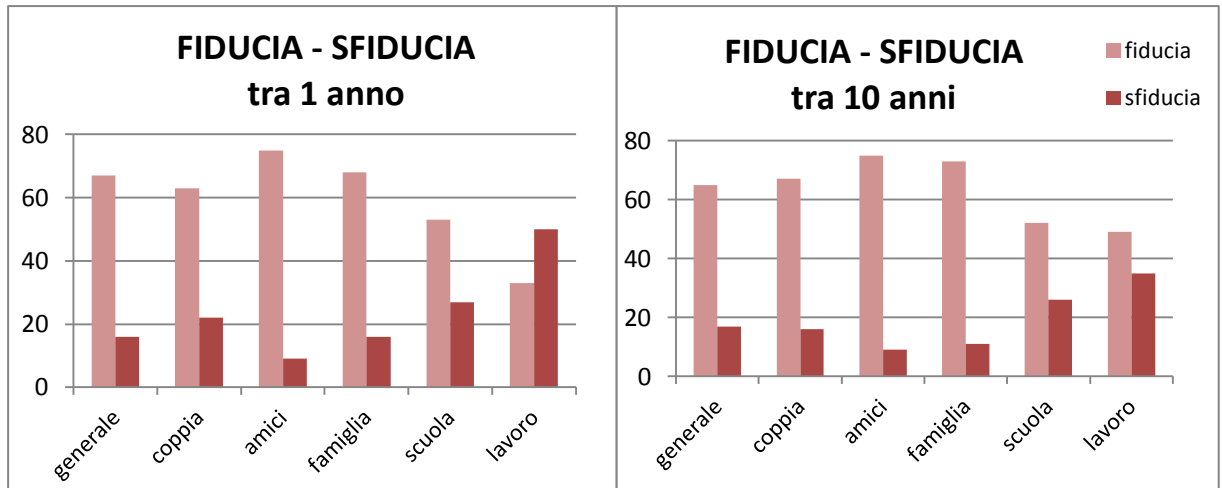


Figura 34: confronto sull'idea di fiducia e sfiducia.

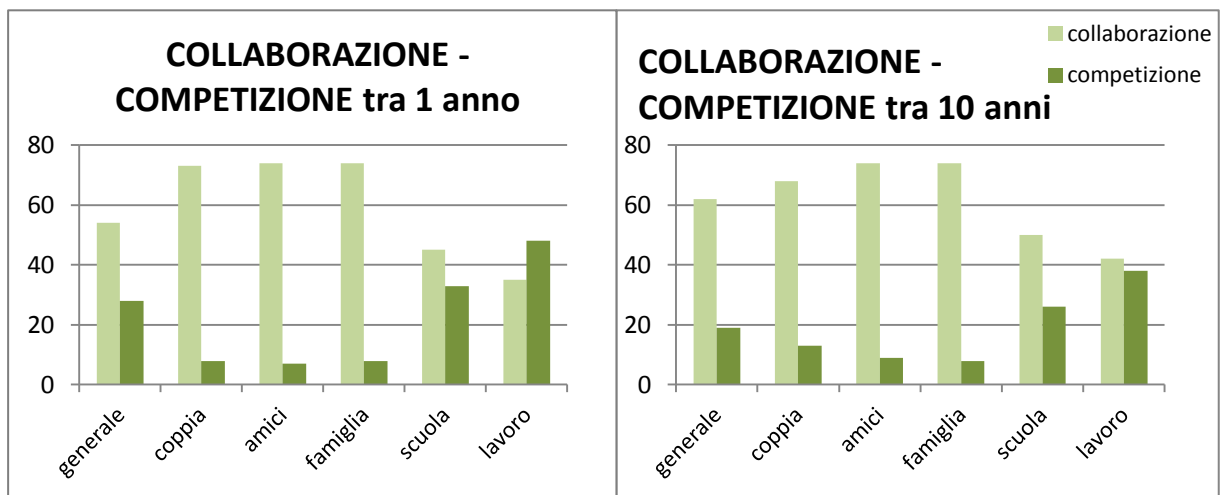


Figura 35: confronto sull'idea di collaborazione e competizione.

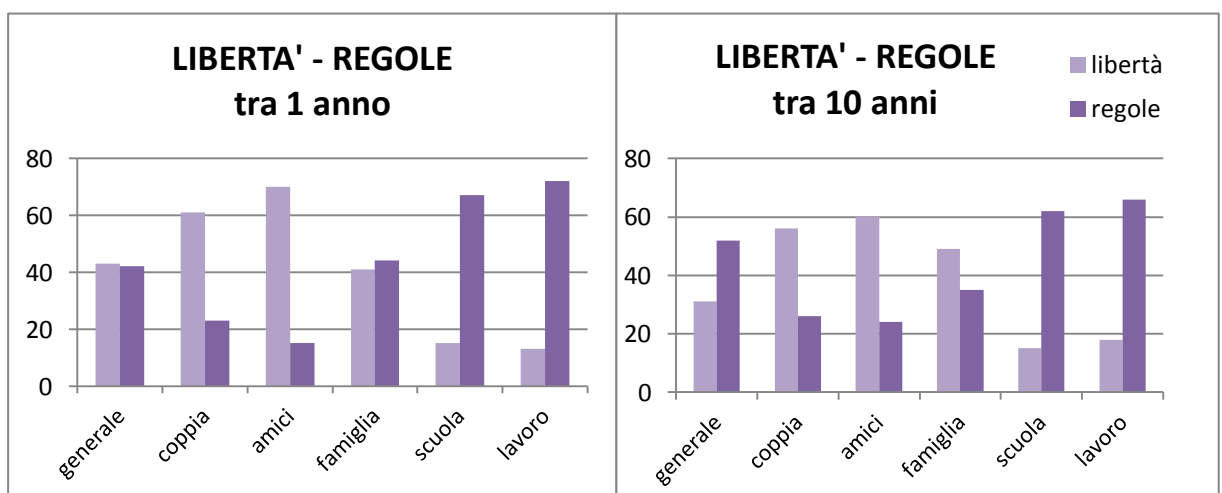


Figura 36: confronto sull'idea di libertà d'azione e presenza di regole

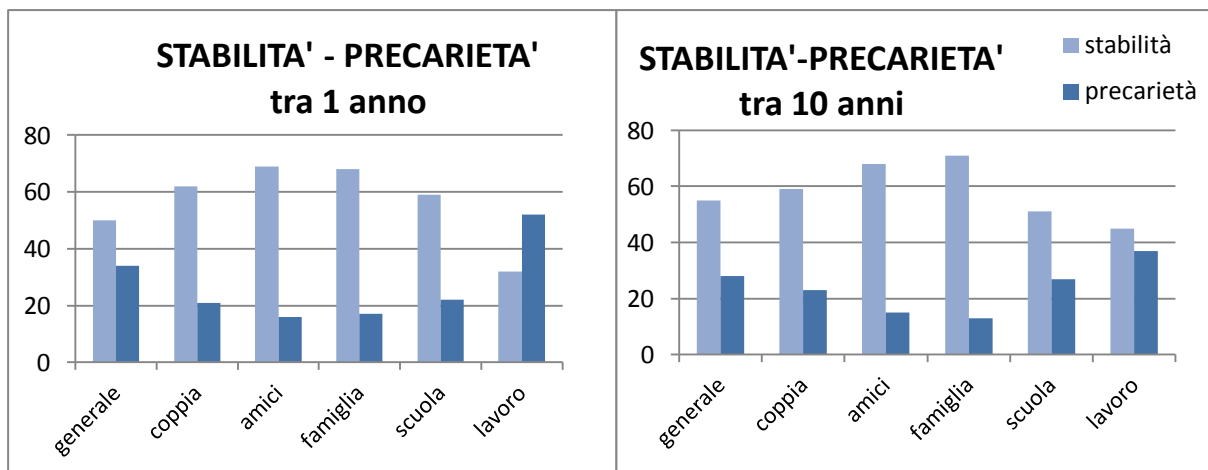


Figura 37: confronto sull'idea di stabilità e precarietà.

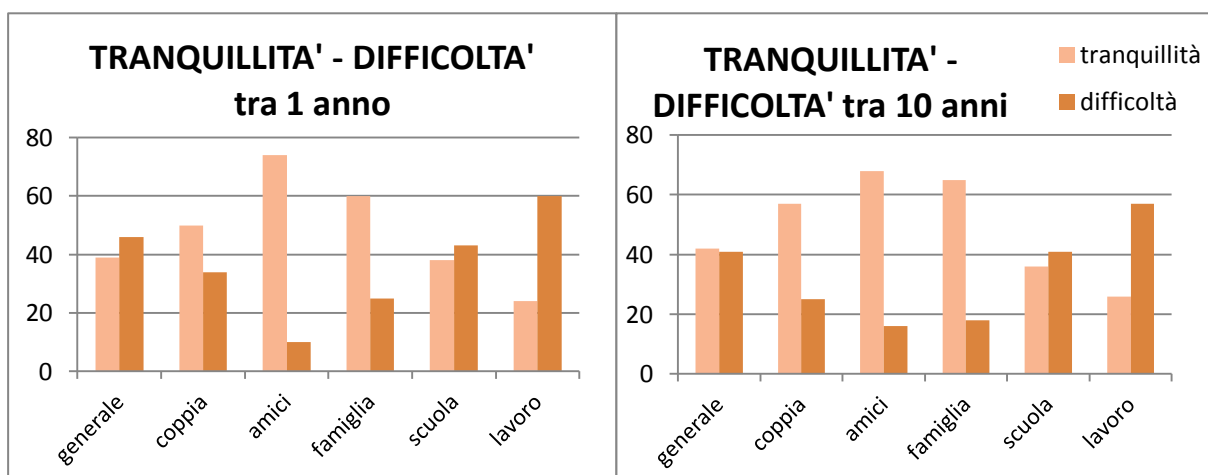


Figura 38: confronto sull'idea di tranquillità e difficoltà.

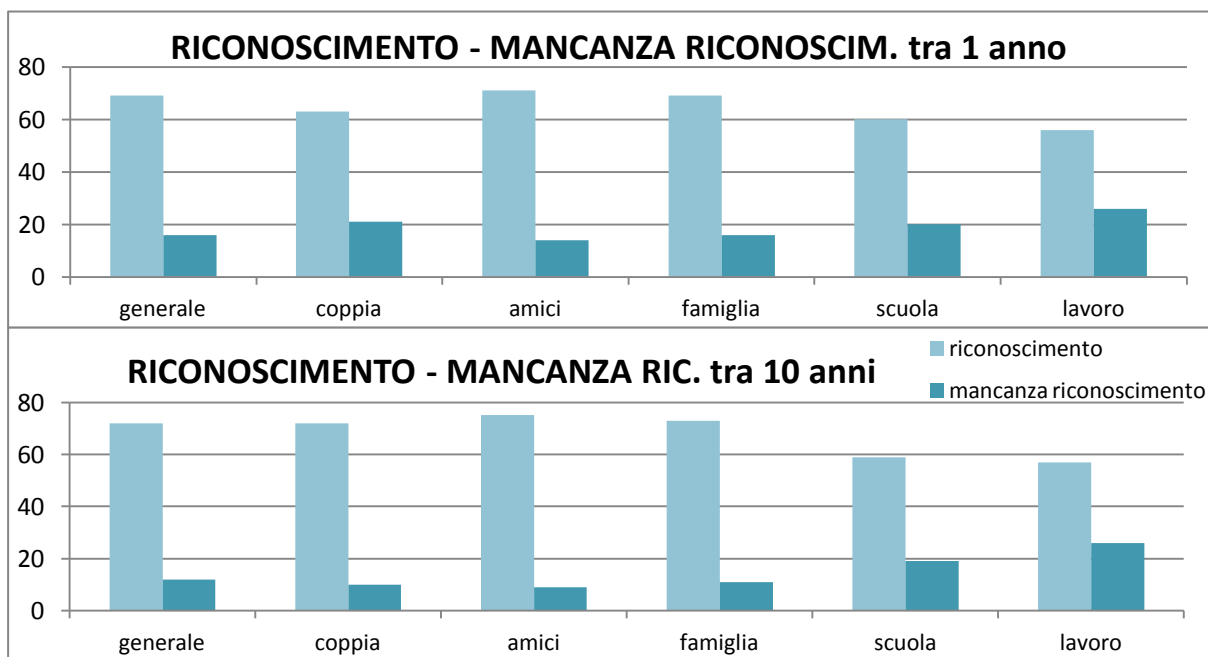


Figura 39: confronto sulla presenza o sulla mancanza di riconoscimento delle capacità.

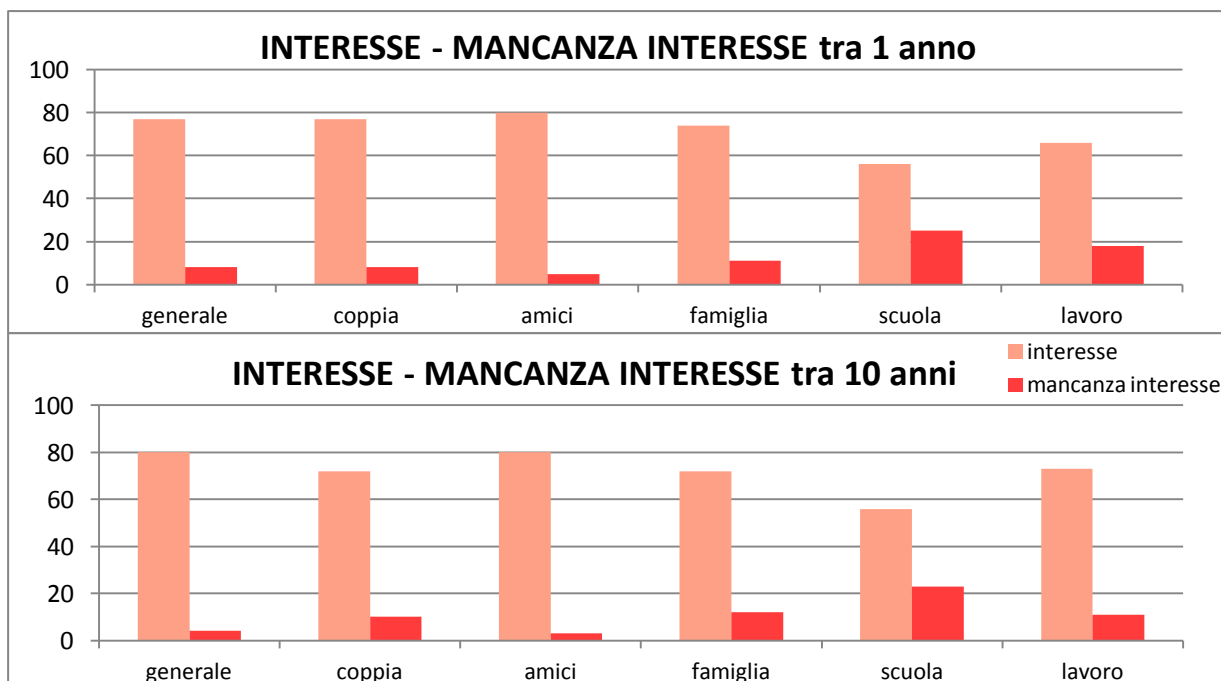


Figura 40: confronto sulla presenza o sulla mancanza di interesse.

Oltre ad avere una visione complessiva dei risultati del sondaggio, si possono eseguire query riguardanti le caratteristiche degli intervistati; ad esempio si può esaminare il loro stato di occupazione o la loro provenienza.

Eseguendo le seguenti query:

```

QUERY 1:
select count(*) ?occ ?tipo
where {
  { ?int <http://localhost:2020/vocab/resource/intervistato_tipo> ?occ.
    ?stud <http://localhost:2020/vocab/resource/
      ID_persona_intervistato_studente> ?int.
    ?stud <http://localhost:2020/vocab/resource/studente_tipo> ?tipo }
  union
  { ?int <http://localhost:2020/vocab/resource/intervistato_tipo> ?occ.
    ?lav <http://localhost:2020/vocab/resource/
      ID_persona_intervistato_lavoratore> ?int.
    ?lav <http://localhost:2020/vocab/resource/lavoratore_tipo> ?tipo }
  union
  { ?int <http://localhost:2020/vocab/resource/intervistato_tipo> ?occ.
    ?dis <http://localhost:2020/vocab/resource/
      ID_persona_intervistato_disoccupato> ?int.
    ?dis <http://localhost:2020/vocab/resource/disoccupato_tipo> ?tipo }
}
group by ?occ ?tipo

```

QUERY 2:

```

select count(*) ?luogo ?nazione
where { ?s <http://localhost:2020/vocab/resource/persona_tipo>
        'intervistato'.
        ?s <http://localhost:2020/vocab/resource/nasce> ?luogo.
        ?luogo <http://localhost:2020/vocab/resource/nazione>
        ?nazione.
}
group by ?luogo ?nazione
order by ?luogo

```

si possono ottenere rispettivamente, i risultati sottostanti (Figure 41-42).

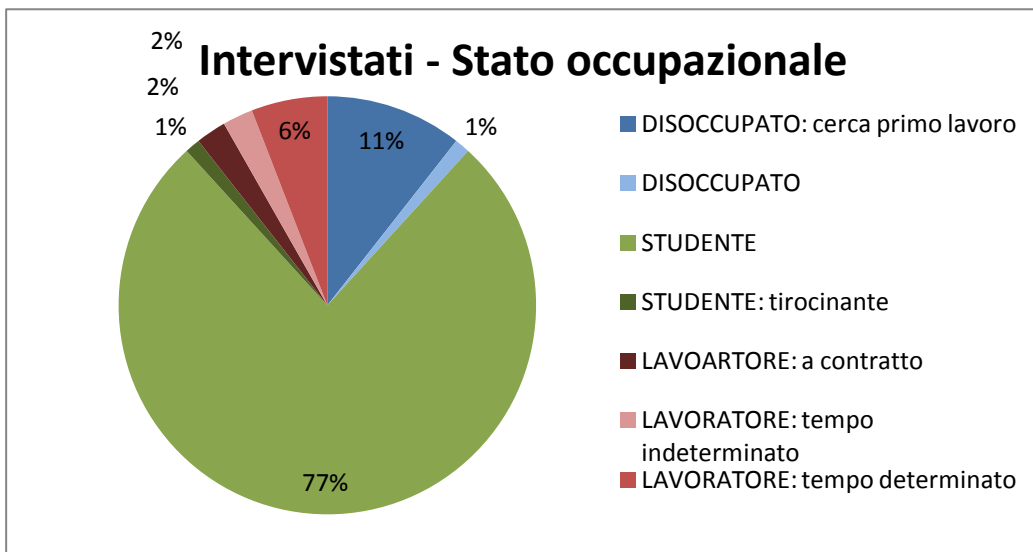


Figura 41: stato occupazionale degli intervistati dal questionario.

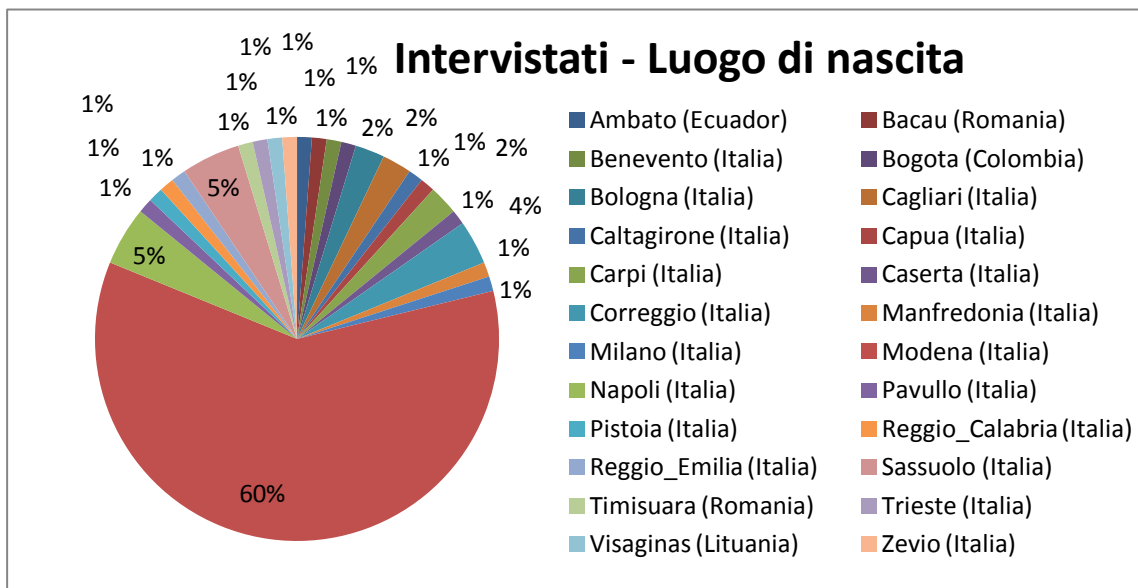


Figura 42: luogo di nascita degli intervistati dal questionario.

5.5. MOMIS

MOMIS ¹⁵¹ (Mediator envirOnment for Multiple Information Sources) [14] e' un framework per l'estrazione e l'integrazione di informazioni di sorgenti dati strutturate e semistrutturate, sviluppato inizialmente da DBGroup ¹⁵² e poi portato avanti da DataRiver ¹⁵³, un suo spin-off, che lo ha re-ingegnerizzato, esteso ed ottimizzato, migliorando il costo dei processi di integrazione, nonché reso open source per uso commerciale ed accademico.

Esso permette di scoprire nuove informazioni da dati esistenti e apparentemente non correlati, grazie all'utilizzo della semantica presente nelle sorgenti dati ed a tecniche di clustering e di inferenza delle logiche descrittive.

Il sistema MOMIS è basato sull'architettura convenzionale per i sistemi di Integrazione Intelligente dell'Informazione wrapper/mediator ed è sviluppato in Java, basando la GUI sul framework Eclipse RCP (Rich Client Platform).

MOMIS consente, quindi, di integrare sorgenti dati in formati diverso, in modo virtuale, all'interno di un unico schema globale, che può essere poi interrogato dall'utente utilizzando semplici query, che accedono alle sorgenti locali e fondono in modo intelligente i risultati parziali, in modo da ottenere un unico esito.

Esso permette queste operazioni grazie ai principali componenti che compongono la sua architettura (Figura 43): il linguaggio ODL_I³, il thesaurus WordNet, il Global Schema e il Query Manager.

¹⁵¹ http://www.datariver.it/index.php?option=com_content&view=article&id=16&Itemid=25&lang=it

¹⁵² <http://www.dbgroup.unimo.it/site2012/>

¹⁵³ http://www.datariver.it/index.php?option=com_content&view=article&id=18&Itemid=28&lang=it

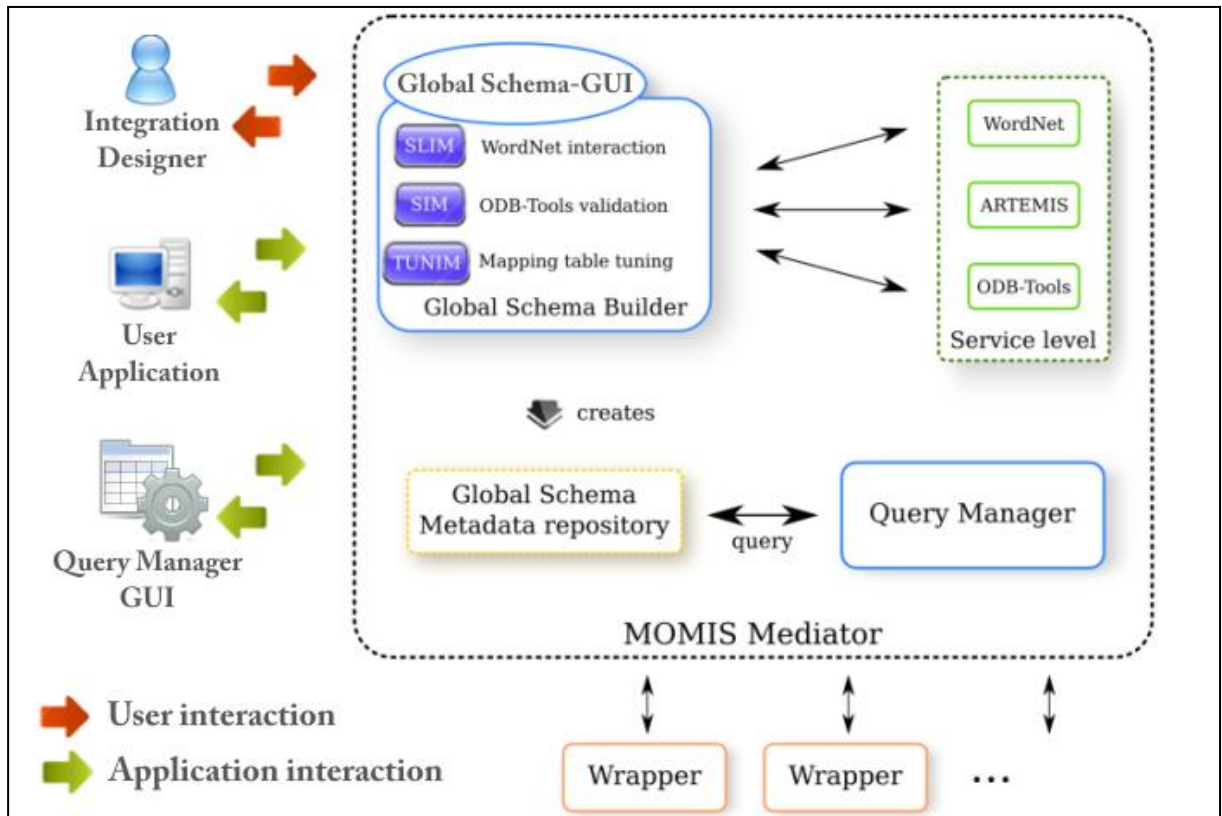


Figura 43: architettura di MOMIS.

Il linguaggio ODL_I^3 , è un linguaggio object-oriented, derivato dallo standard ODMG (Object Data Management Group), nel quale sono convertiti tutti gli schemi delle sorgenti che si vogliono integrare, in modo da averne le rappresentazioni in un linguaggio comune.

Grazie all'annotazione, si possono mappare gli elementi degli schemi ad un thesaurus comune, WordNet, ovvero si associano i significati presenti nel vocabolario ai nomi delle classi e degli attributi delle sorgenti; WordNet, infatti è un thesaurus in lingua inglese che raggruppa i termini, chiamati *lemmi*, in insiemi di sinonimi, i *synset*: esso fornisce, per ciascuno di essi, una breve definizione, detta *gloss*, e li collega attraverso una vasta rete di relazioni semantiche.

Basandosi sulle relazioni definite in WordNet, sulle descrizioni degli schemi in ODL_I^3 e sulla sua Description Logic, MOMIS deduce le relazioni tra gli elementi delle sorgenti, le quali possono essere di sinonimia (SYNonym), generalizzazione (Broader Term), specificazione (Narrower Term) e relatività (Related Term), andando poi a costituire il cosiddetto Global Schema (GS). Esso è una vista integrata delle sorgenti

coinvolte per le quali sono specificate regole di mapping e vincoli di integrità per la gestione delle eterogeneità (Mapping Table): grazie alle relazioni stabilite, si possono calcolare coefficienti di affinità, sia basati sul matching delle stringhe che su quello delle strutture, tra le classi locali ed individuare quelle semanticamente equivalenti, che vengono poi inserite nella stessa classe globale.

Quest'ultime sono, poi, interrogabili per mezzo del Query Manager, strumento che partendo dalla singola query formulata sullo schema globale, in base ai mapping del GS e utilizzando tecniche di unfolding, genera le query da inviare alle sorgenti locali coinvolte nell'interrogazione. Il Query Manager invia le corrispondenti sottoquery alle singole sorgenti tramite i wrapper, elabora i risultati parziali e restituisce all'utente/applicazione una risposta unificata, utilizzando anche particolari funzioni di risoluzione (Resolution Function) per eliminare gli eventuali conflitti tra i dati.

I vantaggi, quindi di MOMIS sono quello di avere una GUI che facilita il processo di integrazione, di utilizzare un insieme di tool che permettono al progettista di visualizzare in anteprima ogni fase del processo, di poter creare più viste unificate per esplorare il risultato globale dell'integrazione, di sfruttare una suite di strumenti che consente di annotare semanticamente le fonti ad un riferimento lessicale comune, di importare ed esportare le annotazioni e di estendere il riferimento lessicale con altri termini del dominio, ed, infine, di poter visualizzare, per ciascuna query globale eseguita, il suo piano di query, ovvero tutte le singole query che la compongono.

5.5.1. MOMIS nella creazione dei Linked Open Data

Proprio lo stesso metodo di annotazione di MOMIS, è stato usato per la creazione dei link tra il dataset Precarietà e WordNet. Si è infatti creata la corrispondenza tra i nomi delle tabelle e delle varie colonne, trasformate rispettivamente in classi e proprietà, con la pagina di WordNet contenente il lemma e la glossa corrispondenti, arricchendo così i dati a livello di schema e non di istanza.

I termini sono stati annotati manualmente in inglese, perché al momento, non è ancora possibile con MOMIS ottenere l'annotazione direttamente in italiano, pur essendo già state studiate alcune possibilità di integrazione del thesaurus MultiWordNet, al posto di WordNet. Si è comunque ricostruito il passaggio, trovando il termine italiano in MultiWordNet e, da esso, il corrispondente termine inglese di WordNet.

Abbiamo inoltre analizzato il framework MOMIS, per individuare le possibili modifiche ed aggiunte da implementare per rendere lo strumento in grado di lavorare direttamente con i Linked Open Data. Si è quindi concluso che, in MOMIS:

- in primo luogo è necessario creare un wrapper che permetta l'utilizzo dei dati RDF in modo da poter lavorare direttamente sui dataset dei Linked Open Data; al momento è in fatti possibile importare solo dati provenienti da diversi DBMS (MySQL, Postgres, ecc) o ontologie in linguaggio OWL, e sarebbe sicuramente utile estendere quest'ultimo in modo da rendere possibile l'integrazione dei dati RDF;
- in secondo luogo, è importante, ed è già in corso, l'integrazione del thesaurus MultiWordNet in italiano, dando la possibilità di annotare i dataset nel linguaggio più appropriato e creando così, nuove opportunità di collegamento automatico tra sorgenti dati differenti, ma di domini comuni;
- infine, sarebbe utile implementare la possibilità di esportare, non solo lo Schema Globale in cui vengono mappate tutte le diverse sorgenti, ma anche i rispettivi dati integrati in modo da ottenere dataset completi facilmente convertibili in Linked Open Data; al momento è infatti, possibile solo esportare lo Schema Globale, ed in linguaggio proprietario di MOMIS, l'ODL₁³, sarebbe quindi utile poterne esportare una versione in un linguaggio standard, come l'OWL o direttamente in RDF, facilitando la trasformazione dei dati in Linked Open Data.

Conclusioni e Future Work

All'interno di questa tesi si sono analizzati lo stato dell'arte degli Open Data, dei Linked Data e della fusione delle due correnti.

In particolare, si sono esaminate le caratteristiche dei dati RDF dei Linked Open Data e le eventuali carenze dei LOD già pubblicati, il procedimento per la loro pubblicazione e i tool necessari al supporto nelle varie fasi di trasformazione da semplici dati a LOD dataset.

Tali tool sono stati applicati e valutati su dati reali grazie all'utilizzo del dataset sulle 'Precarietà Giovanili' fornito dal Comune di Modena, trovando soluzioni alle piccole difficoltà tecniche incontrate, e delineando un percorso di conversione e pubblicazione del dataset nella Linked Open Data Cloud.

Al contrario dei dataset presi da esempi e tutorial, si è potuto vedere come esistano sempre piccole incongruenze nel passare dall'utilizzo di un tool ad un altro: ad esempio, per prima cosa si è dovuto lavorare con dati in un formato proprietario e con diverse difficoltà di conversione in un database; poi con D2R Server si è potuto convertire i dati in RDF, ma non aggiungere o modificare lo schema; quindi con Protégé, si sono convertiti i dati in ontologia, in cui si è potuta inserire l'annotazione, ma anche qui solo in forma di testo, di commento, mentre per la modifica di questo in link si è dovuto intervenire manualmente sul file; infine, i tool di pubblicazione e di interrogazione SPARQL, LMF e Fuseki, aventi funzioni implementate diverse, ma entrambe necessarie.

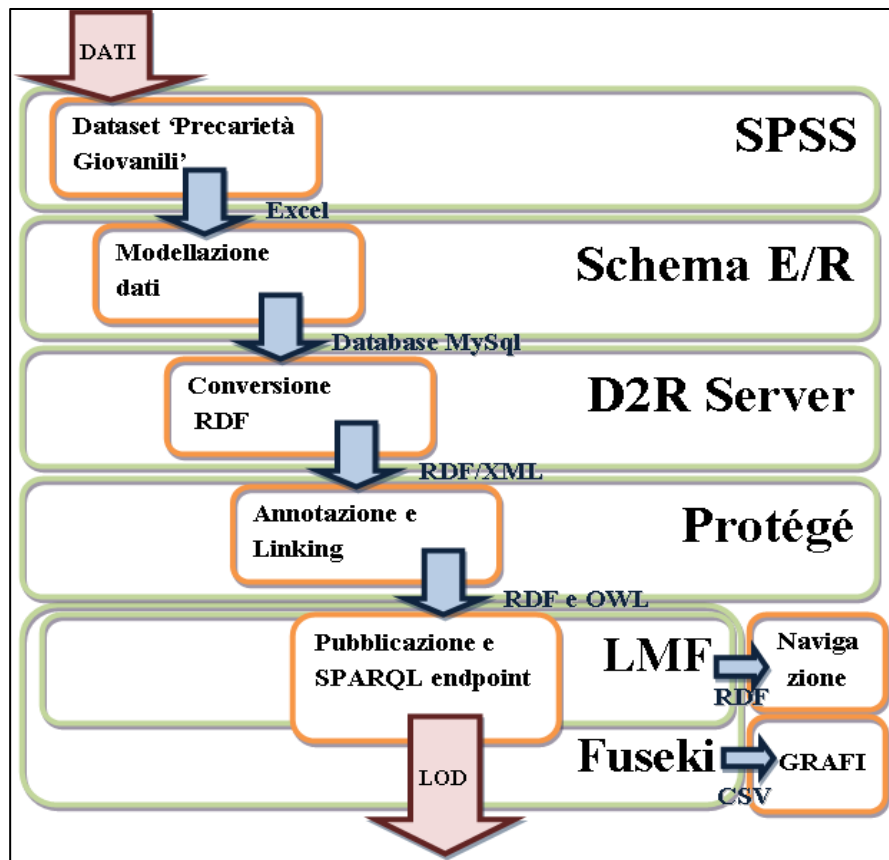


Figura 44: percorso per la trasformazione del dataset Precarietà in LOD.

Sarebbe sicuramente più funzionale avere un unico tool che integrasse sia la possibilità di navigare i link dei dati RDF, sia che permettesse l'esecuzione di query complesse, con operatori di raggruppamento e funzioni di conteggio. Inoltre, si ha carenza di strumenti che permettono la creazione di grafici statistici direttamente dai risultati delle query SPARQL: è infatti necessario esportarli in formato CSV e poi utilizzare strumenti, come Microsoft Office, per ottenerne delle rappresentazioni, quando sarebbe molto più indicato avere la possibilità di lavorare direttamente sul risultato delle query. Esistono molti altri strumenti utilizzabili, ma essendo questo ancora un argomento di ricerca, tali tool sono spesso disponibili in versioni ancora non aggiornate o in via di sviluppo, oppure non sono open-source ma in formato proprietario.

Inoltre, il processo seguito per la pubblicazione dei dati, ha visto l'installazione dei diversi strumenti su server locale, e sarà quindi interessante ripetere il percorso utilizzando un sito vero e proprio, collaborando con il Comune di Modena per mettere a disposizione dei cittadini tutti i dati relativi all'indagine sulle precarietà giovanili.

Infine, si è analizzato il framework MOMIS, sviluppato da DBGroup e Datariver del Dipartimento di Ingegneria 'Enzo Ferrari' dell'Università di Modena: si è preso spunto da esso nella scelta dell'utilizzo dell'annotazione a WordNet, come mezzo per il linking di dataset differenti; inoltre, si sono studiati i possibili cambiamenti da effettuare per poterlo utilizzare come strumento di integrazione ed esportazione di sorgenti dati RDF. Ad oggi, infatti, non esistono strumenti che integrino i LOD a livello di schema, ma solo strumenti per la scoperta di link a livello di istanza, e sarebbe quindi molto utile modificarlo al fine di inserire anch'esso fra gli strumenti per l'elaborazione dei Linked Open Data.

Bibliografia

- [1] Open Data, *Open Definition Home page*, site: <http://opendefinition.org/>
Conoscenza Aperta, site: <http://opendefinition.org/okd/italiano/>
- [2] Associazione Italiana per l'Open Government, site: <http://www.datagov.it/>
- [3] USA Open Government, site: <http://www.data.gov/>
- [4] Open Government Italiano, *I dati delle PA*, site: <http://www.dati.gov.it/>
- [5] Open Data in Italia, *Che dati pubblici abbiamo in giro?*,
site: <http://www.spaghettiopendata.org>
- [6] Blogosfere – Comunità Digitali, *Open Data Italia: Cosa sono, come sono e dove trovarli*, site: <http://comunitadigitali.blogosfere.it/2011/10/open-data-italia-dove-trovarli.html>
- [7] Open Science Data, *On the Full and Open Exchange of Scientific Data*, National Research Council, Washington DC, 1995; site: http://www.nap.edu/readingroom.php?book=exch&page=summary.html#sum_need
- [8] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a global dataspace*, Morgan & Claypool, 2011, site: <http://linkeddatabook.com/editions/1.0/>
- [9] Tim Berners Lee, *Linked Data – Design Issue*, site: <http://www.w3.org/DesignIssues/LinkedData.html>

- [10] Christian Bizer, Tom Heath, and Tim Berners-Lee, *Linked data – The Story So Far*, Int. J. Semantic Web Inf. Syst., 5(3):1–22, 2009, *site*: <http://dx.doi.org/10.4018/jswis.2009081901> DOI: 10.4018/jswis.2009081901
- [11] Bizer C., *The Emerging Web of Linked Data*, IEEE Intelligent Systems, pp. 87–92, September/October 2009, *site*: <http://lpis.csd.auth.gr/mtpx/sw/material/IEEE-IS/IS-24-5.pdf>
- [12] Linked PA, *Portale semantico per la Pubblica Amministrazione*, *site*: <http://www.ontologiapa.it>
- [13] Parlamento Italiano, *Camera dei Deputati*, *site*: <http://www.dati.camera.it>
- [14] Sonia Bergamaschi, Domenico Beneventano, Alberto Corni, Entela Kazazi, Mirko Orsini, Laura Po, Serena Sorrentino, *The Open Source release of the MOMIS Data Integration System*
- [15] W3C Recommendation, *Resource Description Framework (RDF): Concepts and Abstract Syntax*, 10 February 2004, *site*: <http://www.w3.org/TR/rdf-concepts/>
- [16] LATC (LOD Around The Clock), *Project Fact Sheet*, *site*: <http://latc-project.eu/sites/default/files/deliverables/latc-wp5-D512%20Project%20Fact%20Sheet.pdf>
- [17] LATC (LOD Around The Clock), *Project Infrastructure*, *site*: <http://latc-project.eu/sites/default/files/deliverables/latc-wp5-D511%20Project%20Infrastructure.pdf>
- [18] LATC (LOD Around The Clock), *Helping you in publish and consume Linked Data*, *site*: <http://latc-project.eu/about>
- [19] Tom Heath, Chris Bizer, Richard Cyganiak, *How to publish Linked Data*, 27th October 2008, Karlsruhe, Germany, *site*: <http://www4.wiwiwiss.fu-berlin.de/bizer/pub/linkeddatatutorial/>
- [20] LATC (LOD Around The Clock), *Data Publications and Consumption Tools*, *site*: <http://www4.wiwiwiss.fu-berlin.de/latc/toollibrary/>

- [21] LATC (LOD Around The Clock), *Data Publications and Consumption Tools* - Example, site: <http://www4.wiwiss.fu-berlin.de/latc/toollibrary/example.php>
- [22] Neologism, site: <http://neologism.der.i.e/>;
LATC site: <http://www4.wiwiss.fu-berlin.de/latc/toollibrary/tool.php?name=neologism>
- [23] D2R Server (Database to RDF), site: <http://d2rq.org/d2r-server/>;
LATC site: <http://www4.wiwiss.fu-berlin.de/latc/toollibrary/tool.php?name=d2rserver>
- [24] Tony Blakely, Clare Salmond, *Probabilistic record linkage and a method to calculate the positive predictive value*, Department of Public Health, Wellington School of Medicine, University of Otago, New Zealand, *International Journal of Epidemiology*, August 12, 2002.
- [25] Silk - Link Discovery Framework,
site: <http://www4.wiwiss.fu-berlin.de/bizer/silk/#manual>
- [26] Robert Isele, Anja Jentzsch, Christian Bizer, *Silk Server - Adding missing Links while consuming Linked Data*, 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010.
- [27] Semantic Web, "*W3C Semantic Web Activity*", World Wide Web Consortium (W3C), 7 November 2011, Retrieved 26 November 2011;
site: <http://www.w3.org/2001/sw/>
- [28] Prateek Jain, Pascal Hitzler, Amit Shet, Kunal Verma, Peter Yeh, *Ontology Alignment for Linked Open Data*, Proc 9th ISWC, Shanghai (CN), LNCS 6496:401-416, 2010.
- [29] Prateek Jain, Peter Z. Yeh, Kunal Verma, Reymonrod G. Vasquez, Mariana Damova, Pascal Hitzler, Amit P. Sheth *Contextual Ontology Alignment of LOD with an Upper Ontology: A Case Study with Proton*, Proceedings of ESWC 2011, Extended Semantic Web Conference, Heraklion, Greece, May 2011.

- [30] Saša Nešić, Andrea Emilio Rizzoli, Ioannis N. Athanasiadis, *Publishing and Linking Semantically Annotated Agro-Environmental resources to LOD with AGROPub*, In Conference Proceeding: Towards a Semantically Unified Environmental Information Space, Frameworks of e-Environment - 9th IFIP WG 5.11 International Symposium, ISESS 2011, Brno, Czech Republic, June 27-29, 2011.
- [31] Arvid Heise, Felix Naumann, *Integrating Open Government Data with Stratosphere for more Transparency*, J. Web Sem. (WS) 14:45-56 (2012).
- [32] Christoph Böhm, Felix Naumann, Markus Freitag, Stefan George, Norman Höfler, Martin Köppelmann, Claudia Lehmann, Andrina Mascher, Tobias Schmidt, *Linking Open Government Data: What Journalists Wish They Had Known*, In Proceedings the 6th International Conference on Semantic Systems, I-SEMANTICS 2010, Graz, Austria, September 1-3, 2010, 2010.
- [33] Remi Coletta, Emmanuel Castanier, Patrick Valduriez, Christian Frisch, DuyHoa Ngo, Zohra Bellahsene, *Public Data Integration with WebSmatch*, CoRR abs/1205.2555 (2012).
- [34] Lawrence A. Klein, *Sensor and data fusion: A tool for information assessment and decision making*, SPIE Press. p. 51, 2004.
- [35] W3C, *Equivalence Mining and Matching Frameworks*, SWEO Community Project: Linking Open Data on the Semantic Web;
 site: <http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining>
- [36] Tesi Laurea Specialistica VOD: Serena Sorrentino, *Metodi di Disambiguazione del Testo ed Estensioni di WordNet nel sistema MOMIS*, Introduzione e Capitolo 1, 2005/2006.
- [37] Tesi Laurea Specialistica VOD: Riccardo Saponi, *Integrazione multilingua di sorgenti dati PostgreSQL della Pubblica Amministrazione: accoppiamento dei sistemi MOMIS e SIAM*, Capitolo 3, 2008/2009.
- [38] Protégé, site: <http://protege.stanford.edu/>

[39] LMF- LinkedData Media Framework,

site: <http://code.google.com/p/lmf/wiki/DeveloperIntroduction>

[40] Fuseki: Serving RDF data over HTTP,

site: http://jena.apache.org/documentation/serving_data/index.html

Appendice

Appendice A: Neologism

Codice del vocabolario costruito con il tool Neologism; il linguaggio utilizzato è Turtle.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix vann: <http://purl.org/vocab/vann/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix db_hotel: <http://localhost/neologism/db_hotel#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
<http://localhost/neologism/db_hotel> a owl:Ontology;
    dc:title "vocabolario hotel";
    dc:description "vocabolario relativo ad un database di alberghi";
    dc:modified "2012-07-02"^^xsd:date;
    vann:preferredNamespaceUri "http://localhost/neologism/db_hotel#";
    vann:preferredNamespacePrefix "db_hotel";
    foaf:page <http://localhost/neologism/db_hotel.html>;
    dc:creator db_hotel:elisa .

db_hotel:elisa a foaf:Person;
    foaf:nick "elisa" .

db_hotel:client a rdfs:Class, owl:Class;
    rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
    rdfs:label "client";
    rdfs:comment "person who is a customer of a shop";
    rdfs:subClassOf foaf:Person .
```

```

db_hotel:prenotation a rdfs:Class, owl:Class;
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
  rdfs:label "prenotation";
  rdfs:comment "reservation";
  rdfs:subClassOf owl:Thing .

db_hotel:room a rdfs:Class, owl:Class;
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
  rdfs:label "room";
  rdfs:comment "chamber";
  rdfs:subClassOf owl:Thing .

db_hotel:supplement a rdfs:Class, owl:Class;
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
  rdfs:label "supplement";
  rdfs:comment "extra";
  rdfs:subClassOf owl:Thing .

db_hotel:price a rdf:Property;
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
  rdfs:label "price";
  rdfs:comment "cost";
  rdfs:domain
    db_hotel:prenotation,
    db_hotel:supplement ;
  rdfs:range rdfs:Literal .

db_hotel:room_type a rdf:Property, owl:DatatypeProperty;
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
  rdfs:label "room_type";
  rdfs:comment "the kind of room";
  rdfs:domain db_hotel:room;
  rdfs:range rdfs:Literal;
  rdfs:subPropertyOf dc:type .

db_hotel:belong a rdf:Property;
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
  rdfs:label "belong";
  rdfs:comment "its owner";
  rdfs:domain db_hotel:prenotation;
  rdfs:range db_hotel:client;
  rdfs:subPropertyOf dc:relation .

db_hotel:prenotation_of a rdf:Property;
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;
  rdfs:label "prenotation_of";
  rdfs:comment "the room which is booked";
  rdfs:domain db_hotel:prenotation;
  rdfs:range db_hotel:room;
  rdfs:subPropertyOf dc:relation .

```

```
db_hotel:belong_prenotation a rdf:Property;  
  rdfs:isDefinedBy <http://localhost/neologism/db_hotel>;  
  rdfs:label "belong_prenotation";  
  rdfs:comment "the prenotation wich owns the extra";  
  rdfs:domain db_hotel:supplement;  
  rdfs:range db_hotel:prenotation;  
  rdfs:subPropertyOf db_hotel:belong .
```

Appendice B: D2R Server

Codice del file di mapping utilizzato per testare il tool D2R Server.

```
@prefix map: <d2r-mappings/mapping_hotel.ttl#> .
@prefix db_hotel: <http://localhost/neologism/db_hotel#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix d2r: <http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-
server/config.rdf#> .
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#>
.
@prefix jdbc: <http://d2rq.org/terms/jdbc/> .

map:Server a d2r:Server;
  rdfs:label "D2R Server pubblica il database Hotel";
.

map:database a d2rq:Database;
  d2rq:jdbcDriver "com.mysql.jdbc.Driver";
  d2rq:jdbcDSN "jdbc:mysql://localhost/hotel";
  d2rq:username "elisa";
  d2rq:password "root";
  jdbc:autoReconnect "true";
  jdbc:zeroDateTimeBehavior "convertToNull";
.

# Table client
map:client a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern "client/@@client.id@";
  d2rq:class db_hotel:client;
  d2rq:classDefinitionLabel "client"@en;
.

map:client_label a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:client;
  d2rq:property rdfs:label;
  d2rq:pattern "client #@@client.id@";
.

map:client_id a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:client;
  d2rq:property rdfs:label;
  d2rq:propertyDefinitionLabel "client id"@en;
  d2rq:column "client.id";
  d2rq:datatype xsd:unsignedInt;
.

map:client_name a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:client;
  d2rq:property foaf:familyName;
  d2rq:propertyDefinitionLabel "client name"@en;
  d2rq:column "client.name";
.
```



```

map:client_address a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:client;
    d2rq:property rdfs:label;
    d2rq:propertyDefinitionLabel "client address"@en;
    d2rq:column "client.address";
.
map:client_telephone a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:client;
    d2rq:property rdfs:label;
    d2rq:propertyDefinitionLabel "client telephone"@en;
    d2rq:column "client.telephone";
.
# Table prenotation
map:prenotation a d2rq:ClassMap;
    d2rq:dataStorage map:database;
    d2rq:uriPattern "prenotation/@@prenotation.id@@";
    d2rq:class db_hotel:prenotation;
    d2rq:classDefinitionLabel "prenotation"@en;
.
map:prenotation__label a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:prenotation;
    d2rq:property rdfs:label;
    d2rq:pattern "prenotation #@@prenotation.id@@";
.
map:prenotation_id a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:prenotation;
    d2rq:property rdfs:label;
    d2rq:propertyDefinitionLabel "prenotation id"@en;
    d2rq:column "prenotation.id";
    d2rq:datatype xsd:unsignedInt;
.
map:prenotation_periodFrom a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:prenotation;
    d2rq:property dc:date;
    d2rq:propertyDefinitionLabel "prenotation periodFrom"@en;
    d2rq:column "prenotation.periodFrom";
    d2rq:datatype xsd:date;
.
map:prenotation_periodTo a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:prenotation;
    d2rq:property dc:date;
    d2rq:propertyDefinitionLabel "prenotation periodTo"@en;
    d2rq:column "prenotation.periodTo";
    d2rq:datatype xsd:date;
.
map:prenotation_room a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:prenotation;
    d2rq:property db_hotel:prenotation_of;
    d2rq:propertyDefinitionLabel "prenotation room"@en;
    d2rq:uriPattern "room/@@prenotation.room|urlify@@";
    d2rq:join "prenotation.room => room.number";
.

```

```

map:prenotation_idClient a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:prenotation;
  d2rq:property db_hotel:belong;
  d2rq:propertyDefinitionLabel "prenotation idClient"@en;
  d2rq:uriPattern "client/@@prenotation.idClient|urlify@";
  # d2rq:column "prenotation.idClient";
  # d2rq:datatype xsd:unsignedInt;
  d2rq:join "prenotation.idClient => client.id";
.
map:prenotation_totalPrice a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:prenotation;
  d2rq:property db_hotel:price;
  d2rq:propertyDefinitionLabel "prenotation totalPrice"@en;
  d2rq:column "prenotation.totalPrice";
  d2rq:datatype xsd:decimal;
.
# Table prenoted_supplement
map:prenoted_supplement a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern
  "prenoted_supplement/@@prenoted_supplement.supplementCode@";
  d2rq:class rdfs:Resource;
  d2rq:classDefinitionLabel "prenoted supplement"@en;
.
map:prenoted_supplement__label a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:prenoted_supplement;
  d2rq:property rdfs:label;
  d2rq:pattern "prenoted_supplement
  @@prenoted_supplement.supplementCode@";
.
map:prenoted_supplement_idPrenotation a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:prenoted_supplement;
  d2rq:property db_hotel:belong_prenotation;
  d2rq:uriPattern
  "prenotation/@@prenoted_supplement.idPrenotation|urlify@";
  d2rq:propertyDefinitionLabel "prenoted supplement
  idPrenotation"@en;
  # d2rq:column "prenoted_supplement.idPrenotation";
  # d2rq:datatype xsd:unsignedInt;
  d2rq:join "prenoted_supplement.idPrenotation =>
  prenotation.id";
.
map:prenoted_supplement_supplementCode a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:prenoted_supplement;
  d2rq:property rdfs:label;
  d2rq:propertyDefinitionLabel "prenoted supplement
  supplementCode"@en;
  d2rq:column "prenoted_supplement.supplementCode";
  d2rq:datatype xsd:unsignedByte;
.

```

```

# Table price
map:price a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern
    "price/@@price.periodFrom@@/@@price.periodTo@@/@@price.price@@/@
    @price.roomType|urlify@@";
  d2rq:class rdfs:Resource;
  d2rq:classDefinitionLabel "price"@en;
  .
map:price__label a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:price;
  d2rq:property rdfs:label;
  d2rq:pattern "price
  #@@price.periodFrom@@/@@price.periodTo@@/@@price.price@@/@@price
  .roomType@@";
  .
map:price_periodFrom a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:price;
  d2rq:property dc:date;
  d2rq:propertyDefinitionLabel "price periodFrom"@en;
  d2rq:column "price.periodFrom";
  d2rq:datatype xsd:date;
  .
map:price_periodTo a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:price;
  d2rq:property dc:date;
  d2rq:propertyDefinitionLabel "price periodTo"@en;
  d2rq:column "price.periodTo";
  d2rq:datatype xsd:date;
  .
map:price_roomType a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:price;
  d2rq:property dc:description;
  d2rq:propertyDefinitionLabel "price roomType"@en;
  d2rq:column "price.roomType";
  .
map:price_price a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:price;
  d2rq:property rdfs:label;
  d2rq:propertyDefinitionLabel "price"@en;
  d2rq:column "price.price";
  d2rq:datatype xsd:decimal;
  .

map:room_number a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:room;
  d2rq:property rdfs:label;
  d2rq:propertyDefinitionLabel "room number"@en;
  d2rq:column "room.number";
  d2rq:datatype xsd:unsignedShort;
  .

```

```

# Table room
map:room a d2rq:ClassMap;
    d2rq:dataStorage map:database;
    d2rq:uriPattern "room/@@room.number@@";
    d2rq:class db_hotel:room;
    d2rq:classDefinitionLabel "room"@en;
    .
map:room__label a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:room;
    d2rq:property rdfs:label;
    d2rq:pattern "room #@@room.number@@";

map:room_floor a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:room;
    d2rq:property rdfs:label;
    d2rq:propertyDefinitionLabel "room floor"@en;
    d2rq:column "room.floor";
    d2rq:datatype xsd:unsignedByte;
    .
map:room_type a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:room;
    d2rq:property db_hotel:room_type;
    d2rq:propertyDefinitionLabel "room type"@en;
    d2rq:column "room.type";
    .
map:room_bedAdded a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:room;
    d2rq:property rdfs:label;
    d2rq:propertyDefinitionLabel "room bedAdded"@en;
    d2rq:column "room.bedAdded";
    .
map:room_optionals a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:room;
    d2rq:property dc:description;
    d2rq:propertyDefinitionLabel "room optionals"@en;
    d2rq:column "room.optionals";
    .

# Table supplement
map:supplement a d2rq:ClassMap;
    d2rq:dataStorage map:database;
    d2rq:uriPattern "supplement/@@supplement.code@@";
    d2rq:class db_hotel:supplement;
    d2rq:classDefinitionLabel "supplement"@en;
    .
map:supplement__label a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:supplement;
    d2rq:property rdfs:label;
    d2rq:pattern "supplement #@@supplement.code@@";
    .

```

```
map:supplement_code a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:supplement;
  d2rq:property rdfs:label;
  d2rq:propertyDefinitionLabel "supplement code"@en;
  d2rq:column "supplement.code";
  d2rq:datatype xsd:unsignedByte;
.
map:supplement_description a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:supplement;
  d2rq:property dc:description;
  d2rq:propertyDefinitionLabel "supplement description"@en;
  d2rq:column "supplement.description";
.
map:supplement_price a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:supplement;
  d2rq:property db_hotel:price;
  d2rq:propertyDefinitionLabel "supplement price"@en;
  d2rq:column "supplement.price";
  d2rq:datatype xsd:decimal;
.
```

Appendice C: Silk Workbench

Codice del progetto utilizzato per testare il tool Silk Workbench.

```
<Project>
  <Config>
    <Prefixes>
      <Prefix namespace=http://www.w3.org/1999/02/22-rdf-syntax-ns#
id="rdf"></Prefix>
      <Prefix namespace="http://dbpedia.org/ontology/" id="dbpedia">
</Prefix>
      <Prefix namespace="http://www.w3.org/2002/07/owl#" id="owl">
</Prefix>
      <Prefix namespace="http://xmlns.com/foaf/0.1/" id="foaf"></Prefix>
      <Prefix namespace=http://www.w3.org/2000/01/rdf-schema# id="rdfs">
</Prefix>
      <Prefix namespace="http://dbpedia.org/ontology/" id="dbpediaowl">
</Prefix>
      <Prefix namespace=http://localhost/neologism/db_hotel#
id="db_hotel"></Prefix>
    </Prefixes>
  </Config>
  <SourceModule>
    <Tasks>
      <SourceTask>
        <DataSource type="sparqlEndpoint" id="DBpedia_IT">
          <Param name="pageSize" value="1000"></Param>
          <Param name="pauseTime" value="0"></Param>
          <Param name="retryCount" value="3"></Param>
          <Param name="endpointURI" value="http://it.dbpedia.org/sparql">
</Param>
          <Param name="retryPause" value="1000"></Param>
          <Param name="graph" value=""></Param>
          <Param name="login" value=""></Param>
          <Param name="entityList" value=""></Param>
          <Param name="password" value=""></Param>
          <Param name="additionalParameters" value=""></Param>
        </DataSource>
      </SourceTask>
      <SourceTask>
        <DataSource type="sparqlEndpoint" id="hotel">
          <Param name="pageSize" value="1000"></Param>
          <Param name="pauseTime" value="0"></Param>
          <Param name="retryCount" value="3"></Param>
          <Param name="endpointURI" value="http://localhost:2020/sparql">
</Param>
          <Param name="retryPause" value="1000"></Param>
          <Param name="graph" value=""></Param>
          <Param name="login" value=""></Param>
          <Param name="entityList" value=""></Param>
        </DataSource>
      </SourceTask>
    </Tasks>
  </SourceModule>
</Project>
```

```

        <Param name="password" value=""></Param>
        <Param name="additionalParameters" value=""></Param>
    </DataSource>
</SourceTask>
</Tasks>
</SourceModule>
<LinkingModule>
    <Tasks><LinkingTask>
        <LinkSpecification>
            <Interlink id="personaTask">
                <LinkType>owl:sameAs</LinkType>
                <SourceDataset dataSource="hotel" var="a">
                    <RestrictTo> ?a &lt;http://www.w3.org/1999/02/22-rdf-
syntax-ns#type&gt;
&lt;http://localhost/neologism/db_hotel#client&gt; .
                    </RestrictTo>
                </SourceDataset>
                <TargetDataset dataSource="DBpedia_IT" var="b">
                    <RestrictTo> ?b &lt;http://www.w3.org/1999/02/22-rdf-syntax-
ns#type&gt; &lt;http://dbpedia.org/ontology/Person&gt; .
                    </RestrictTo>
                </TargetDataset>
            </Interlink>
            <LinkageRule>
                <Compare weight="18" threshold="0.9875092758931159" required="false"
metric="jaccard" id="unnamed_21296">
                    <Input path="?a/foaf:familyName" id="unnamed_21294"></Input>
                    <Input path="?b/foaf:name" id="unnamed_21295"></Input>
                </Compare>
            </LinkageRule>
        </Filter></Filter>
        <Outputs> </Outputs>
    </Interlink>
</LinkSpecification>
<Alignment>
<rdf:RDF
    xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:align="http://knowledgeweb.semanticweb.org/heterogeneity/alignm
ent#"
    xmlns="http://knowledgeweb.semanticweb.org/heterogeneity/alignment#">
<Alignment>
    <map>
        <Cell>
            <entity1 rdf:resource="http://localhost:2020/resource/client/5">
            </entity1>
            <entity2 rdf:resource="http://it.dbpedia.org/resource/Bob_Marley">
            </entity2>
            <relation>=</relation>
            <measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0
            </measure></Cell></map>
    </map>
    <Cell>

```

```

    <entity1 rdf:resource="http://localhost:2020/resource/client/2">
  </entity1>
  <entity2
  rdf:resource="http://it.dbpedia.org/resource/Paolo_Bianchi_(ciclista)">
  </entity2>
    <relation>!=</relation>
    <measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0
  </measure></Cell></map>
<map>
  <Cell>
    <entity1 rdf:resource="http://localhost:2020/resource/client/1">
  </entity1>
    <entity2
  rdf:resource="http://it.dbpedia.org/resource/Mario_Rossi_(calciatore)">
  </entity2>
    <relation>!=</relation>
    <measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0
  </measure></Cell></map>
<map>
  <Cell>
    <entity1 rdf:resource="http://localhost:2020/resource/client/2">
  </entity1>
    <entity2
  rdf:resource="http://it.dbpedia.org/resource/Paolo_Bianchi_(cestista)">
  </entity2>
    <relation>!=</relation>
    <measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0
  </measure></Cell></map>
</Alignment>
</rdf:RDF>
</Alignment>
<Cache>
  <Cache>
  <EntityDescriptions>
    <Source>
    <EntityDescription>
      <Variable>a</Variable>
      <Restrictions>?a &lt;http://www.w3.org/1999/02/22-rdf-syntax-
  ns#type&gt; &lt;http://localhost/neologism/db_hotel#client&gt;
  .</Restrictions>
      <Paths>
        <Path>?a/&lt;http://xmlns.com/foaf/0.1/familyName&gt;</Path>
        <Path>?a/&lt;http://www.w3.org/2000/01/rdf-schema#label&gt;</Path>
        <Path>?a/&lt;http://localhost/neologism/db_hotel#from&gt;</Path>
        <Path>?a/&lt;http://www.w3.org/1999/02/22-rdf-syntax-
  ns#type&gt;</Path>
        <Path>?a&lt;http://localhost/neologism/db_hotel#belong&gt;</Path>
      </Paths>
    </EntityDescription>
  </Source>
  <Target>

```



```

<EntityDescription>
  <Variable>b</Variable>
  <Restrictions>?b &lt;http://www.w3.org/1999/02/22-rdf-syntax-ns#type&gt;
    &lt;http://dbpedia.org/ontology/Person&gt; .</Restrictions>
  <Paths>
    <Path>?b/&lt;http://xmlns.com/foaf/0.1/name&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/sexo&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/wikiPageWikiLink&gt;</Path>
    <Path>?b/&lt;http://www.w3.org/2000/01/rdf-schema#label&gt;</Path>
    <Path>?b/&lt;http://purl.org/dc/terms/subject&gt;</Path>
    <Path>?b/&lt;http://www.w3.org/1999/02/22-rdf-syntax-ns#type&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/nome&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/wikiPageUsesTemplate&gt;
      </Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/annoscita&gt;</Path>
    <Path>?b/&lt;http://www.w3.org/2002/07/owl#sameAs&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/luogonascita&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/nazionalità&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/attività&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/cognome&gt;</Path>
    <Path>?b/&lt;http://xmlns.com/foaf/0.1/page&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/thumbail&gt;</Path>
    <Path>?b/&lt;http://xmlns.com/foaf/0.1/depiction&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/birthYear&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/giornomesenascita&gt;</Path>
    <Path>?b/&lt;http://xmlns.com/foaf/0.1/surname&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/immagine&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/wikiPageExternalLink&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/annomorte&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/luogomorte&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/deathYear&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/giornomesemorte&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/abstract&gt;</Path>
    <Path>?b/&lt;http://www.w3.org/2000/01/rdf-schema#comment&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/birthPlace&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/deathPlace&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/postnazionalità&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/epoca&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/didascalia&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/titolo&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/postcognomevirgola&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/nomeOnorificenza&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/collegamentoOnorificenza&gt;
      </Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/forzaordinamento&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/luogo&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/periodo&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/precedente&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/carica&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/id&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/successivo&gt;</Path>

```

```

    <Path>?b/&lt;http://it.dbpedia.org/property/editore&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/anno&gt;</Path>
    <Path>?b/&lt;http://xmlns.com/foaf/0.1/homepage&gt;</Path>
    <Path>?b/&lt;http://dbpedia.org/ontology/stateOfOrigin&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/autore&gt;</Path>
    <Path>?b/&lt;http://it.dbpedia.org/property/url&gt;</Path>
  </Paths>
</EntityDescription></Target></EntityDescriptions>
<PositiveEntities>
  <Pair>
    <Source>
      <Entity uri="http://localhost:2020/resource/client/5">
        <Val>
          <e>Bob Marley</e>
        </Val></Entity> </Source>
    <Target>
      <Entity uri="http://it.dbpedia.org/resource/Bob_Marley">
        <Val>
          <e>Bob Marley</e>
        </Val> </Entity> </Target> </Pair>
  </PositiveEntities>
<NegativeEntities>
  <Pair>
    <Source>
      <Entity uri="http://localhost:2020/resource/client/2">
        <Val>
          <e>paolo bianchi</e>
        </Val> </Entity></Source>
    <Target>
      <Entity uri="http://it.dbpedia.org/resource/Paolo_Bianchi_(ciclista)">
        <Val>
          <e>Paolo Bianchi</e>
          <e>Paolo Alfredo Bianchi</e>
        </Val></Entity></Target>
  </Pair>
  <Pair>
    <Source>
      <Entity uri="http://localhost:2020/resource/client/1">
        <Val>
          <e>mario rossi</e>
        </Val></Entity></Source>
    <Target>
      <Entity uri="http://it.dbpedia.org/resource/Mario_Rossi_(calciatore)">
        <Val>
          <e>Mario Rossi</e>
        </Val></Entity></Target>
  </Pair>
  <Pair>
    <Source>
      <Entity uri="http://localhost:2020/resource/client/2">

```

```

    <Val>
      <e>paolo bianchi</e>
    </Val></Entity></Source>
  <Target>
    <Entity uri="http://it.dbpedia.org/resource/Paolo_Bianchi
      _(cestista)">
      <Val>
        <e>Paolo Bianchi</e>
      </Val></Entity></Target>
    </Pair>
  </NegativeEntities>
</Cache>
</Cache>
</LinkingTask>
</Tasks>
</LinkingModule>
</Project>

```

Codice delle specifiche della *linkage rule* ottenuta, nel linguaggio Silk-LSL.

```

<Silk>
  <Prefixes>
    <Prefix namespace=http://www.w3.org/1999/02/22-rdf-syntax-ns#
      id="rdf">
    </Prefix>
    <Prefix namespace="http://dbpedia.org/ontology/" id="dbpedia">
    </Prefix>
    <Prefix namespace="http://www.w3.org/2002/07/owl#" id="owl">
    </Prefix>
    <Prefix namespace="http://xmlns.com/foaf/0.1/" id="foaf"></Prefix>
    <Prefix namespace="http://www.w3.org/2000/01/rdf-schema#" id="rdfs">
    </Prefix>
    <Prefix namespace="http://dbpedia.org/ontology/" id="dbpediaowl">
    </Prefix>
    <Prefix namespace=http://localhost/neologism/db_hotel# id="db_hotel">
    </Prefix>
  </Prefixes>
  <DataSources>
    <DataSource type="sparqlEndpoint" id="hotel">
      <Param name="pageSize" value="1000"></Param>
      <Param name="pauseTime" value="0"></Param>
      <Param name="retryCount" value="3"></Param>
      <Param name="endpointURI" value="http://localhost:2020/sparql">
      </Param>
      <Param name="retryPause" value="1000"></Param>
      <Param name="graph" value=""></Param>
      <Param name="login" value=""></Param>
      <Param name="entityList" value=""></Param>
      <Param name="password" value=""></Param>
      <Param name="additionalParameters" value=""></Param>
    </DataSource>

```

```

</DataSources>
<Interlinks>
  <Interlink id="personaTask">
    <LinkType>owl:sameAs</LinkType>
    <SourceDataset dataSource="hotel" var="a">
      <RestrictTo> ?a &lt;http://www.w3.org/1999/02/22-rdf-syntax-
        ns#type&gt; &lt;http://localhost/neologism/db_hotel#client&gt; .
      </RestrictTo>
    </SourceDataset>
    <TargetDataset dataSource="DBpedia_IT" var="b">
      <RestrictTo> ?b &lt;http://www.w3.org/1999/02/22-rdf-syntax-
        ns#type&gt; &lt;http://dbpedia.org/ontology/Person&gt; .
      </RestrictTo>
    </TargetDataset>
    <LinkageRule>
      <Compare weight="18" threshold="0.9875092758931159" required="false"
        metric="jaccard" id="unnamed_21296">
        <Input path="?a/foaf:familyName" id="unnamed_21294"></Input>
        <Input path="?b/foaf:name" id="unnamed_21295"></Input>
      </Compare>
    </LinkageRule>
    <Filter></Filter>
    <Outputs> </Outputs>
  </Interlink>
</Interlinks>
</Silk>

```

Appendice D: Questionario

Il questionario utilizzato per la raccolta dati sulle precarietà, del Comune di Modena.

<p>1. Et� _____</p> <p>2. Sesso <input type="checkbox"/> 1. M <input type="checkbox"/> 2. F</p> <p>3. Puoi dirci dove sei nato? Citt�: _____ Nazione: _____</p> <p>4. Puoi dirci dove sono nati i tuoi genitori? Padre: _____ Citt�: _____ Nazione: _____ Madre: _____ Citt�: _____ Nazione: _____</p> <p>5. Attualmente sei...: <input type="checkbox"/> Studente (inclusi studenti lavoratori) <input type="checkbox"/> Lavoratore assunto a tempo indeterminato <input type="checkbox"/> Lavoratore assunto a tempo determinato <input type="checkbox"/> Lavoratore a contratto <input type="checkbox"/> Disoccupato <input type="checkbox"/> In cerca di prima occupazione</p>	<p>6. Che cosa prevale quando pensi al tuo futuro nel prossimo anno (un'opzione per coppia)?</p> <table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 60%;"></th> <th style="width: 10%; text-align: center;">In generale</th> <th style="width: 10%; text-align: center;">Scuola/lavoro</th> <th style="width: 10%; text-align: center;">Amici</th> <th style="width: 10%; text-align: center;">Coppia Famiglia</th> </tr> </thead> <tbody> <tr> <td>fiducia</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>sfiducia</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>collaborazione</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>competizione</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>presenza di regole</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>libert� di agire liberamente</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>stabilit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>precariet�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>difficolt�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>tranquillit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>riconoscimento delle mie capacit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>mancato riconoscimento delle mie capacit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>interesse</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>manca di interesse</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </tbody> </table>		In generale	Scuola/lavoro	Amici	Coppia Famiglia	fiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	sfiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	collaborazione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	competizione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	presenza di regole	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	libert� di agire liberamente	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	stabilit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	precariet�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficolt�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	tranquillit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	mancato riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	manca di interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																				
	In generale	Scuola/lavoro	Amici	Coppia Famiglia																																																																																												
fiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
sfiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
collaborazione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
competizione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
presenza di regole	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
libert� di agire liberamente	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
stabilit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
precariet�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
difficolt�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
tranquillit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
mancato riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
manca di interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
<p>7. Che cosa prevale quando pensi al tuo futuro nei prossimi dieci anni (un'opzione per coppia)?</p> <table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 60%;"></th> <th style="width: 10%; text-align: center;">In generale</th> <th style="width: 10%; text-align: center;">Scuola/lavoro</th> <th style="width: 10%; text-align: center;">Amici</th> <th style="width: 10%; text-align: center;">Coppia Famiglia</th> </tr> </thead> <tbody> <tr> <td>fiducia</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>sfiducia</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>collaborazione</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>competizione</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>presenza di regole</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>libert� di agire liberamente</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>stabilit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>precariet�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>difficolt�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>tranquillit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>riconoscimento delle mie capacit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>mancato riconoscimento delle mie capacit�</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>interesse</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>manca di interesse</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </tbody> </table>		In generale	Scuola/lavoro	Amici	Coppia Famiglia	fiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	sfiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	collaborazione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	competizione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	presenza di regole	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	libert� di agire liberamente	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	stabilit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	precariet�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficolt�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	tranquillit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	mancato riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	manca di interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<p>8. Rispetto ad un anno fa, se pensi al tuo futuro che cosa provi?</p> <table style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 60%;">pi� fiducia</td> <td style="width: 10%; text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>meno fiducia</td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </tbody> </table> <p>9. Rispetto ad un anno fa, se pensi al tuo futuro che cosa vedi?</p> <table style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 60%;">vedo pi� rischi</td> <td style="width: 10%; text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>vedo meno rischi</td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </tbody> </table> <p>10. Riguardo a che cosa sei pi� fiducioso per il futuro?</p> <table style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 60%;">scuola/lavoro</td> <td style="width: 10%; text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>amore</td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>amicizia</td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </tbody> </table> <p>11. Riguardo a che cosa sei MENO fiducioso per il futuro?</p> <table style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 60%;">scuola/lavoro</td> <td style="width: 10%; text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>amore</td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td>amicizia</td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </tbody> </table>	pi� fiducia	<input type="checkbox"/>	meno fiducia	<input type="checkbox"/>	vedo pi� rischi	<input type="checkbox"/>	vedo meno rischi	<input type="checkbox"/>	scuola/lavoro	<input type="checkbox"/>	amore	<input type="checkbox"/>	amicizia	<input type="checkbox"/>	scuola/lavoro	<input type="checkbox"/>	amore	<input type="checkbox"/>	amicizia	<input type="checkbox"/>
	In generale	Scuola/lavoro	Amici	Coppia Famiglia																																																																																												
fiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
sfiducia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
collaborazione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
competizione	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
presenza di regole	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
libert� di agire liberamente	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
stabilit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
precariet�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
difficolt�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
tranquillit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
mancato riconoscimento delle mie capacit�	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
manca di interesse	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																																																																												
pi� fiducia	<input type="checkbox"/>																																																																																															
meno fiducia	<input type="checkbox"/>																																																																																															
vedo pi� rischi	<input type="checkbox"/>																																																																																															
vedo meno rischi	<input type="checkbox"/>																																																																																															
scuola/lavoro	<input type="checkbox"/>																																																																																															
amore	<input type="checkbox"/>																																																																																															
amicizia	<input type="checkbox"/>																																																																																															
scuola/lavoro	<input type="checkbox"/>																																																																																															
amore	<input type="checkbox"/>																																																																																															
amicizia	<input type="checkbox"/>																																																																																															

12. (SE NON STAI STUDIANDO O LAVORANDO PASSA ALLA DOMANDA 13)

Quale delle seguenti frasi descrive meglio il modo in cui pensi alla tua esperienza scolastica (SE STUDI) o lavorativa (SE LAVORI)?

- comunico senza problemi con gli altri anche se non so come reagiranno
- per poter comunicare senza problemi con gli altri devo essere sicuro di come reagiranno
- di solito non sono disponibile a comunicare con gli altri

13. Quale delle seguenti frasi descrive meglio il modo in cui pensi ai tuoi rapporti con gli amici?

- comunico senza problemi con gli altri anche se non so come reagiranno
- per poter comunicare senza problemi con gli altri devo essere sicuro di come reagiranno
- di solito non sono disponibile a comunicare con gli altri

14. Quale delle seguenti frasi descrive meglio il modo in cui pensi alla tua famiglia?

- comunico senza problemi con gli altri anche se non so come reagiranno
- per poter comunicare senza problemi con gli altri devo essere sicuro di come reagiranno
- di solito non sono disponibile a comunicare con gli altri

15. Quale delle seguenti frasi descrive meglio il modo in cui pensi alla coppia (ANCHE SE NON SEI IN COPPIA)?

- comunico senza problemi con la/il partner anche se non so come reagiranno
- per poter comunicare senza problemi con la/il partner devo essere sicuro di come reagiranno
- di solito non sono disponibile a comunicare con la/il partner

16. In quale ambiente senti di avere più fiducia da chi sta con te?

- lavoro/scuola
- famiglia
- gruppo di amici
- coppia

17. In quale ambiente senti di avere MENO fiducia da chi sta con te?

- lavoro/scuola
- famiglia
- gruppo di amici
- coppia

18. In quale ambiente concedi più fiducia a chi sta con te?

- lavoro/scuola
- famiglia
- gruppo di amici
- coppia

19. In quale ambiente concedi MENO fiducia a chi sta con te?

- lavoro/scuola
- famiglia
- gruppo di amici
- coppia

20. Scegli, tra queste frasi, quella che corrisponde a ciò che pensi per i rapporti a scuola (SE STUDI) oppure sul lavoro (SE LAVORI) (SE NON STUDI O LAVORI PASSA ALLA DOMANDA 19)

- le mie azioni possono danneggiare gli altri
- le azioni degli altri possono danneggiarmi
- le mie azioni posso danneggiare gli altri e le azioni degli altri possono danneggiare me
- le mie azioni possono danneggiare solo me stesso
- le mie azioni possono danneggiare me stesso e anche gli altri
- nessuna azione, né mia, né degli altri, può fare danni

21. Scegli, tra queste frasi, quella che corrisponde a ciò che pensi per i rapporti con gli amici

- le mie azioni possono danneggiare gli altri
- le azioni degli altri possono danneggiarmi
- le mie azioni posso danneggiare gli altri e le azioni degli altri possono danneggiare me
- le mie azioni possono danneggiare solo me stesso
- le mie azioni possono danneggiare me stesso e anche gli altri
- nessuna azione, né mia, né degli altri, può fare danni

<p>22. Scegli, tra queste frasi, quella che corrisponde a ciò che pensi per i rapporti in famiglia</p> <p> <input type="checkbox"/> le mie azioni possono danneggiare gli altri <input type="checkbox"/> le azioni degli altri possono danneggiarmi <input type="checkbox"/> le mie azioni posso danneggiare gli altri e le azioni degli altri possono danneggiare me <input type="checkbox"/> le mie azioni possono danneggiare solo me stesso <input type="checkbox"/> le mie azioni possono danneggiare me stesso e anche gli altri <input type="checkbox"/> nessuna azione, né mia, né degli altri, può fare danni </p> <p>23. Scegli, tra queste frasi, quella che corrisponde a ciò che pensi per i rapporti di coppia</p> <p> <input type="checkbox"/> le mie azioni possono danneggiare gli altri <input type="checkbox"/> le azioni degli altri possono danneggiarmi <input type="checkbox"/> le mie azioni posso danneggiare gli altri e le azioni degli altri possono danneggiare me <input type="checkbox"/> le mie azioni possono danneggiare solo me stesso <input type="checkbox"/> le mie azioni possono danneggiare me stesso e anche gli altri <input type="checkbox"/> nessuna azione, né mia, né degli altri, può fare danni </p> <p>24. Negli ultimi 6 mesi, sono accadute queste cose con i compagni di classe/colleghi di lavoro? (se non studi o lavori vai alla domanda 25) (scegli tutte le risposte che credi giuste)</p> <p> <input type="checkbox"/> partecipare alle decisioni <input type="checkbox"/> aiutare a superare una brutta esperienza <input type="checkbox"/> collaborare nel fare cose insieme </p> <p>25. Negli ultimi 6 mesi, sono accadute queste cose con gli amici? (scegli tutte le risposte che credi giuste)</p> <p> <input type="checkbox"/> partecipare alle decisioni <input type="checkbox"/> aiutare a superare una brutta esperienza <input type="checkbox"/> collaborare nel fare cose insieme </p>	<p>26. Negli ultimi 6 mesi, sono accadute queste cose in famiglia? (scegli tutte le risposte che credi giuste)</p> <p> <input type="checkbox"/> partecipare alle decisioni <input type="checkbox"/> aiutare a superare una brutta esperienza <input type="checkbox"/> collaborare nel fare cose insieme </p> <p>27. (RISPONDI SE SEI STATO IMPEGNATO IN UN RAPPORTO DI COPPIA) Negli ultimi 6 mesi, sono accadute queste cose nel rapporto di coppia? (scegli tutte le risposte che credi giuste)</p> <p> <input type="checkbox"/> partecipare alle decisioni <input type="checkbox"/> aiutare a superare una brutta esperienza <input type="checkbox"/> collaborare nel fare cose insieme </p> <p>28. Rispetto al passato, ti capita di prendere decisioni</p> <p> <input type="checkbox"/> meno frequentemente <input type="checkbox"/> più frequentemente <input type="checkbox"/> nello stesso modo </p> <p>29. Rispetto al passato, prendere decisioni è per te</p> <p> <input type="checkbox"/> più facile <input type="checkbox"/> difficile allo stesso modo <input type="checkbox"/> più difficile </p>
---	--

Grazie per la preziosa collaborazione!

Appendice E: Modello Relazionale

Modello relazionale, utilizzato per l'implementazione del database Precarietà.

```
PERSONA (ID persona integer, tipo enum('padre','madre','intervistato'))

PADRE (ID persona padre integer, tipo enum('padre'))
  FK: ID_persona_padre REFERENCES PERSONA

MADRE (ID persona madre integer, tipo enum('madre'))
  FK: ID_persona_madre REFERENCES PERSONA

INTERVISTATO (ID persona intervistato, età integer, sesso integer[0,1],
  gruppo enum('progetto trasparente', 'servizio civile nazionale','città dei
  ragazzi', 'gruppo informale'), data date, tipo
  enum('studente','lavoratore', 'disoccupato'), residenza enum('Modena',
  'provincia di Modena'), titolo enum('diploma maturità', 'diploma
  professionale', 'licenza media', 'laurea triennale', 'laurea
  specialistica', 'nessuno'))
  FK: ID_persona_intervistato REFERENCES PERSONA

STUDENTE (ID persona intervistato studente integer, tipo
  enum('studente','tirocinante'))
  FK: ID_persona_intervistato_studente REFERENCES INTERVISTATO

LAVORATORE (ID persona intervistato lavoratore integer, tipo enum('a
  contratto','tempo determinato', 'tempo indeterminato'))
  FK: ID_persona_intervistato_lavoratore REFERENCES INTERVISTATO

DISOCCUPATO (ID persona intervistato disoccupato integer, tipo
  enum('disoccupato', 'primo lavoro'))
  FK: ID_persona_intervistato_disoccupato REFERENCES INTERVISTATO

HA_PADRE (ID persona intervistato, ID persona padre)
  FK: ID_persona_intervistato REFERENCES INTERVISTATO
  FK: ID_persona_padre REFERENCES PADRE

HA_MADRE (ID persona intervistato, ID persona madre)
  FK: ID_persona_intervistato REFERENCES INTERVISTATO
  FK: ID_persona_madre REFERENCES MADRE

LUOGO (città string, nazione string)

NASCE (ID persona, città, nazione)
  FK: ID_persona REFERENCES PERSONA
  FK: città,nazione REFERENCES LUOGO NOT NULL

FUTURO (ID futuro integer, tipo enum('tra 1 anno', 'tra 10 anni'))

AMBIENTE_SOCIALE (ID_ambiente integer, tipo enum('se stesso', 'coppia,
  'amici', 'famiglia', 'scuola/lavoro', 'generale'),
  ID_persona_intervistato)
  FK: ID_persona_intervistato REFERENCES INTERVISTATO

INTERVISTATO/FUTURO/AMBIENTE_SOCIALE (ID_persona_intervistato, ID_ambiente,
  ID_futuro, fiducia/sfiducia integer[0,1], collaborazione/competizione
  integer[0,1], libertà/regole integer[0,1], stabilità/precarietà
  integer[0,1], tranquillità/difficoltà integer[0,1],
  riconoscimento/mancanza_riconoscimento integer[0,1],
  interesse/mancanza_interesse integer[0,1])
  FK: ID_persona_intervistato REFERENCES INTERVISTATO
  FK: ID_futuro REFERENCES FUTURO
  FK: ID_ambiente REFERENCES AMBIENTE_SOCIALE
```



```
PARTECIPA (ID intervistato, ID ambiente, avvenimento enum('collaborazione',  
'supporto nel brutto momento', 'partecipazione decisioni')  
FK: ID_persona_intervistato REFERENCES INTERVISTATO  
FK: ID_ambiente REFERENCES AMBIENTE_SOCIALE  
  
CONCEDE_FIDUCIA (ID intervistato, ID ambiente, più_o_meno integer[0,1])  
FK: ID_persona_intervistato REFERENCES INTERVISTATO  
FK: ID_ambiente REFERENCES AMBIENTE_SOCIALE NOT NULL  
RICEVE_FIDUCIA (ID intervistato, ID ambiente, più_o_meno integer[0,1])  
FK: ID_persona_intervistato REFERENCES INTERVISTATO  
FK: ID_ambiente REFERENCES AMBIENTE_SOCIALE NOT NULL  
  
AZIONE (ID azione, ID fatta_da, ID danneggiato, dannosa integer[0,1])  
FK: ID_fatta_da REFERENCES AMBIENTE_SOCIALE NOT NULL  
FK: ID_danneggiato REFERENCES AMBIENTE_SOCIALE NOT NULL
```


Ringraziamenti

Desidero ringraziare la Prof. Sonia Bergamaschi e, in particolare, la Dott. Ing. Serena Sorrentino per tutto l'aiuto fornito durante la realizzazione di questa tesi.

Ringrazio la mia famiglia, per la pazienza, l'aiuto e l'incrollabile sostegno durante tutti questi lunghi anni di studio, e soprattutto per l'affetto che da sempre e in ogni istante mi dimostrano. Un particolare grazie a mia madre, Milena, che da sempre risulta essere la colonna portante della mia vita ed il mio incrollabile riferimento.

Ringrazio tutti gli amici (e Marco), che mi sostengono, accompagnano, e soprattutto sopportano, in tutti i momenti della mia vita con la speranza che continuino sempre a farlo.

Un sincero grazie a tutti quanti.