JLIS.it

Le tecnologie semantiche applicate ai linked data, esemplificate nel contesto del CNR

Aldo Gangemi

Evoluzione del web e dereferenziazione

Il web si sta evolvendo da una rete di documenti (chiamata a volte web 1.0) a una rete di entità (chiamata variamente web Semantico, web dei dati, web 3.0 con sottili differenze di senso), passando anche dalla possibilità da parte degli utenti di cambiarne i contenuti e di creare reti sociali attraverso semplici interazioni (web 2.0 o web sociale). Questa evoluzione sta avvenendo innanzi tutto grazie all'architettura del web che si è andata disegnando negli anni novanta,[1] che permette di "dereferenziare" e "linkare" contenuti (identificati da un indirizzo web) attraverso semplici protocolli di comunicazione (ex. HTTP). Per esempio, quello che succede quando scriviamo l'indirizzo (URI) http://www.cnr.it (l'indirizzo del portale web del Consiglio Nazionale delle Ricerche, CNR) in un browser (es. Firefox) e premiamo il tasto di invio è che il client del browser dereferenzia l'indirizzo comunicando con un server del CNR, che gli propone una pagina scritta nel linguaggio HTML, che il client visualizza per l'utente in modo appropriato. Altri indirizzi

[1] http://www.w3.org/TR/webarch.

Traduzione italiana a cura dell'Autore.

di pagina possono essere usati nella pagina visualizzata, creando quindi una rete di link ipertestuali, che permettono la navigazione (browsing). Questo è, all'osso, il web dei documenti. A volte la dereferenziazione è un po' più indiretta, come nel caso in cui l'indirizzo rappresenta una chiamata a un database: questo è ancora il web dei documenti, ma il sistemista ha creato un programma che esegue una query (interrogazione) al database, la cui risposta viene poi visualizzata nel linguaggio HTML, per esempio quando qualcuno cerca le informazioni relative alla propria posizione fiscale nel sito dell'Agenzia delle entrate. Il caso del web 2.0 è un esempio ancora più elaborato di questa dereferenziazione indiretta, che permette all'utente l'intervento diretto, non solo la navigazione: applicazioni diverse come chat, protocolli voce, email, sistemi di annotazione, algoritmi di analisi automatica dei link, log e di feedback utente etc., convergono in pagine HTML ricche, personalizzabili e dinamiche, come nel caso di Facebook.

Due problemi difficili: identità e interoperabilità

Il web dei documenti e anche il web sociale hanno però due problemi che esistono nei sistemi informativi da secoli: l'identità e l'interoperabilità. Il problema dell'identità è illustrato da questo esempio: Aldo Gangemi ha molte homepage diverse (una sul sito del suo istituto, l'ISTC-CNR, una sul wiki del suo laboratorio, l'STLab, uno sul sito di semanticweb.org etc.). Inoltre, è registrato su molti altri portali di servizi al cittadino (per esempio fiscali), di servizi per i membri di associazioni, conferenze etc., o commerciali. Poi, possiede molti account (utenze) di applicazioni e servizi web: Facebook, GMail, Yahoo, Flickr, iTunes, vari servizi di email etc. Ancora, Aldo

Gangemi è un dato in servizi pubblici e personali (dato provvisto di identificatori noti solo ai servizi in questione, che stabiliscono una sorta di "identità posizionale" in una tabella di database), come Google Scholar, DBLP etc. Infine, Aldo Gangemi è citato in altre pagine web: articoli, riferimenti bibliografici, report di eventi. Il problema è: come facciamo a sapere che Aldo Gangemi è l'entità (fisica o sociale che si voglia) a cui fanno capo le sue homepage, registrazioni, account, identificatori vari, citazioni? Il problema ovviamente non è limitato alle persone, ma anche a luoghi, organizzazioni, prodotti e servizi, eventi, leggi, concetti etc. Il problema dell'interoperabilità deriva in parte dal primo: se non sappiamo come "risolvere" l'identità di una cosa fra i diversi sistemi che ne "parlano", è molto difficile aggregare le informazioni relative a quella cosa. Inoltre, le relazioni fra quella cosa e altre cose possono essere molto simili nei diversi sistemi: per esempio, la relazione fra Aldo Gangemi e i messaggi di posta a lui diretti (o fra Aldo Gangemi e altri utenti di posta elettronica) è molto simile, sia che queste relazioni si manifestino su GMail, sia su Facebook, Yahoo Mail o Apple Mail: ma quei sistemi assegneranno identità diverse alle stesse persone. Come aggravante, ogni sistema si basa su infrastrutture eterogenee: diversi linguaggi, formati, protocolli etc., e il tutto rende molto difficile l'integrazione fra dati di diversi sistemi.

Le soluzioni tradizionali

Negli ultimi anni si è creato una specie di cartello fra sistemi commerciali come Facebook, Google etc. per scambiarsi i dati delle loro reti sociali, ma le soluzioni sono completamente ad hoc e riguardano solo le interazioni commercialmente utili fra quei sistemi.
Sul fronte dei database, quando occorre integrarne i contenuti, si procede a complicati procedimenti di schema integration, di identity

resolution, di data warehousing etc. Ogni procedimento è effettuato ad hoc su una coppia di database. Soluzioni parziali all'integrazione vengono poi da tecniche diverse di data mining e di trattamento del linguaggio naturale: in questi casi, si usano approcci statistici per riconoscere entità (named entity recognition), per stabilire la somiglianza fra dati diversi, per scoprire relazioni indirette fra dati etc. L'annotazione dei documenti è un approccio che risale almeno all'inizio del novecento: un documento o una sua parte (paragrafo, termine) sono "annotati" con una categoria presa da qualche knowledge organization system: thesauri, schemi di classificazione, nomenclature, vocabolari controllati, che la maggior parte delle discipline scientifiche, archivistiche, commerciali etc. hanno prodotto nel corso di decenni. Casi macroscopici di questi sforzi esistono per esempio in ambito medico (SnoMed, ICD, MeSH), museale (Getty), agricolo (Agrovoc). In tempi recenti, le procedure di annotazione sono assistite sia mediante supporti all'attività manuale, sia mediante algoritmi di annotazione automatica (ex. text classification), ovviamente con precisione variabile.

Il web dei dati

Nel 2006, Tim Berners-Lee ha proposto i linked data, un metodo elegante ed efficace[2] per semplificare e omogeneizzare le soluzioni ai problemi di identità e interoperabilità. Il metodo mira a creare un web dei dati (o delle entità o delle cose) e si colloca nell'ambito delle tecnologie per il web semantico, di cui si parla in maggiore dettaglio nella prossima sezione. Il metodo consiste sostanzialmente di quattro principi e di molte buone pratiche per la sua applicazione. I quattro principi sono:

[2] http://www.w3.org/DesignIssues/LinkedData.html.

1. usare indirizzi web (URI) come nomi per le cose;
2. usare URI utili al protocollo HTTP in modo che sia possibile cercare e dereferenziare quei nomi;
3. quando qualcuno cerca una URI, fornire informazione utile;
4. includere link ad altre URI, così chi cerca può scoprire nuovi collegamenti.

Fra le buone pratiche, è utile ricordare almeno quelle che hanno contribuito di più all'esplosione dei Linking Open Data, visualizzati nella loro evoluzione per mezzo della cosiddetta LOD Cloud:[3]

- usare licenze aperte per ottenere dati altamente riusabili;
- usare formati non-proprietari (ex. CSV invece di Excel);
- usare standard aperti del W3C, come RDF,[4] SPARQL,[5] Web Ontology Language (OWL),[6] per identificare le cose, relazionarle, interrogarle, ragionarci etc.

Queste pratiche si armonizzano con le raccomandazioni fornite dal movimento degli Open Data e sono attualmente adottate in molti campi diversi, tra cui i dati delle Pubbliche Amministrazioni[7] e usati per applicazioni di integrazione e arricchimento di dati, per esempio nella ricerca di esperti.[8]

La LOD Cloud contiene linked data di molti domini diversi, con particolare importanza per i dati biomedici, culturali, multimediali, bibliografici, geografici etc. Un esempio della potenzialità dei linked

[3] http://linkeddata.org.
[4] http://www.w3.org/RDF.
[5] http://www.w3.org/TR/rdf-sparql-query.
[6] http://www.w3.org/2004/OWL.
[7] http://data.gov; http://data.gov.uk; http://dati.gov.it.
[8] http://data.cnr.it.

data è fornito nella figura 1, che mostra un grafo emerso da un'applicazione (RelFinder[9]) che costruisce incrementalmente le relazioni fra due cose qualsiasi, a patto che esse abbiano una identità sul web dei dati. Nella figura 1 l'esempio parte dalle entità:

```
<http://dbpedia.org/wiki/Neo-positivism>
<http://dbpedia.org/wiki/Francis_Bacon>
```

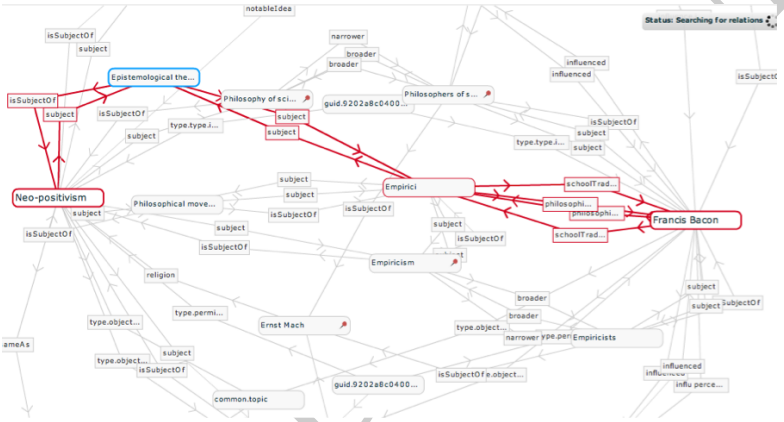

Figura 1: Le relazioni emergenti fra due entità attraverso il grafo dei linking open data.

Il web semantico

Gli standard aperti del W3C: RDF,[10] SPARQL[11] e OWL,[12] permettono di rappresentare, interrogare e ragionare su gran parte delle strutture

[9] http://www.visualdataweb.org/relfinder.php.
[10] http://www.w3.org/RDF.
[11] http://www.w3.org/TR/rdf-sparql-query.
[12] http://www.w3.org/2004/OWL.

di dati tradizionali in modo elegante ed omogeneo. RDF si basa su una struttura di dati ricorsiva, chiamata tripla, formata da un Soggetto, un Predicato e un Oggetto, analogamente alla struttura grammaticale più astratta delle lingue occidentali, la SVO (Soggetto-Verbo-Oggetto).

Listing 1: Alcuni esempi di triple RDF

```
<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://www.w3.org/2000/01/rdf-schema#label>
  ''Aldo Gangemi''

<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  <http://www.cnr.it/ontology/cnr/personale.owl#
      UnitaDiPersonaleInterno>

<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://purl.org/dc/terms/subject>
  <http://dbpedia.org/resource/Category:Semantic_Web>
```

Le triple RDF possono essere interrogate con il linguaggio SPARQL.

Listing 2: Query sulle triple mostrate nel listato 1

```
SELECT ?l
WHERE {
?x <http://purl.org/dc/terms/subject>
<http://dbpedia.org/resource/Category:Semantic_Web>.
?x <http://www.w3.org/2000/01/rdf-schema#label> ?l}
```

La query mostrata nel listato 2 nella pagina precedente fornisce questa risposta:

Listing 3: Risposta alla query mostrata nel listato 2 nella pagina precedente

```
1
''Aldo Gangemi''
```

Ogni tripla contiene Soggetti e Oggetti che devono avere un tipo che è a sua volta una Classe, per esempio:

```
<http://www.cnr.it/ontology/cnr/personale.owl#
    UnitaDiPersonaleInterno>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <http://www.w3.org/2002/07/owl#Class>
```

Ogni tripla contiene un Predicato (o Property), che insieme alle classi, forma il vocabolario (chiamato anche schema o ontologia) usato da un dataset. Nei casi migliori, il vocabolario è scritto in OWL,[13] un linguaggio che permette l'uso di ragionatori automatici per derivare inferenze logiche dalla struttura dei dati, per esempio un ragionatore automatico genera le inverse delle triple, le triple simmetriche, le triple che valgono transitivamente (laddove una regola transitiva sia stata introdotta nel vocabolario), l'eredità delle caratteristiche di una classe (laddove queste caratteristiche siano state definite nel vocabolario) etc.
La capacità espressiva di OWL e di SPARQL sul web dei dati permettono di porre domande complesse a sorgenti di conoscenza eterogenee, per esempio nel dominio del Diritto Romano (le domande non sono formalizzate nell'esempio: i termini sottolineati indicano classi, i **termini** in grassetto proprietà del vocabolario, i *termini* in corsivo cose specifiche o valori):

[13] http://www.w3.org/2004/OWL.

-quali fonti **contengono** massime **riguardo alla** *stipulatio*, **citano** *Ulpiano* e **includono** commentari **prodotti negli** *ultimi 10 anni*?

-quali casi **apparsi in** sistemi di *Common Law* **contengono** interpretazioni **relative a** contratti **analoghi alla** *stipulatio*?

Per migliorare la qualità dei vocabolari e le capacità inferenziali, può essere necessario definire anche altri assiomi di tipo più sofisticato (ex. quale tipo di entità è analogo a quale altro, cosa può essere in citato in un certo contesto etc.). I vocabolari richiedono quindi una certa accuratezza nella loro realizzazione, che può essere raggiunta usando un approccio orientato ai requisiti degli utenti, riusando vocabolari standard oppure pattern ontologici (ontology design patterns[14]) che siano noti per risolvere i problemi di modellazione che emergono dai requisiti utente.

Le applicazioni semantiche

La disponibilità di una grande quantità di dati aperti ha fornito la motivazione principale per sviluppare applicazioni di nuova generazione, che facciano tesoro delle soluzioni tradizionali e di quelle più recenti, centrandosi sul paradigma semantico: l'uso del significato dei dati come schema di organizzazione a tutti i livelli.
Il ciclo di produzione di una applicazione semantica, nel caso più tipico, è il seguente:

1. reingegnerizzazione dei dati esistenti e produzione dei dataset di base in RDF (dati) e OWL (schemi, vocabolari o ontologie);

2. linking fra entità presenti nei dati e produzione di nuove triple;

[14] http://www.ontologydesignpatterns.org.

3. estrazione di nuove entità e triple mediante tecniche statistiche di analisi del testo e dei dati;

4. ragionamento sulla struttura logica derivata dai passi precedenti e produzione di eventuali nuove triple;

5. pubblicazione dei dataset su piattaforme appropriate per la loro interrogazione via SPARQL;

6. presentazione dei dati arricchiti in modo fruibile da utenti del web: testuale, grafico, rich snippet etc.

Questo ciclo di produzione riflette una multipla interpretazione del termine "semantica": nel caso dei passi 2 e 3 ci si riferisce soprattutto alla semantica linguistica implicita nei testi analizzati; le tecnologie relative sono quelle di analisi del testo e dei dati e ha come fine il riconoscimento di collegamenti fra entità, di nomi propri, termini, relazioni, fatti, argomenti etc. Nel caso dei passi 1, 4 e 5, ci si riferisce alla semantica logica (o formale) dei dati e della loro ontologia; le tecnologie relative sono quelle del web semantico presentate nella sezione precedente. Nel passo 6, ci si riferisce alla semantica dell'interazione con l'utente.
Le tecnologie orientate alla semantica linguistica permettono di riconoscere entità nei testi e di risolverne l'identità rispetto a dataset noti. Appena l'identità è risolta, è possibile arricchire il dataset con le relazioni note fra quell'entità e altre entità. Per esempio, dato il seguente testo trascritto da una seduta del Parlamento UE:

> The sensitivities of Northern Ireland are too important for any ill-informed bandwagoning on the International Fund for Ireland. Raytheon has been welcomed to Derry by no less than Nobel Peace Prize winners, John Hume - one of our own colleagues, and David Trimble. Raytheon will be funded by the Industrial Development Board in

> Northern Ireland. Not one euro nor one Irish pound from the International Fund for Ireland is going to Raytheon.

È possibile usare un "named entity recognizer"[15] per riconoscere nomi propri (come Northern Ireland, International Fund for Ireland, Derry, John Hume etc.), la cui identità può essere risolta automaticamente da un "entity resolver"[16] come le entità:

```
<http://dbpedia.org/resource/Northern_Ireland>
<http://dbpedia.org/resource/John_Hume>
<http://dbpedia.org/resource/Derry>
<http://dbpedia.org/resource/David_Trimble>
```

Una volta identificate le entità, possiamo fare un'interrogazione al LOD per trovare altre entità collegate, ex.

```
<http://dbpedia.org/wiki/Mark_Durkan>
<http://dbpedia.org/wiki/David_Cameron>
```

A questo punto, se cerchiamo le quattro entità di partenza e le due nuove, possiamo far emergere dalle triple LOD un grafo complesso, che può visualizzato per esempio mediante l'applicazione RelFinder.[17] Le seguenti sono alcune delle triple trovate in questo modo:

```
<http://dbpedia.org/wiki/Mark_Durkan> <http://dbpedia.org/
    ontology/placeOfBirth> <http://dbpedia.org/resource/Derry>

<http://dbpedia.org/resource/Derry> <http://dbpedia.org/
    ontology/country>
<http://dbpedia.org/resource/Northern_Ireland>
```

[15] http://www.alchemyapi.com/api/demo.html.
[16] http://wit.istc.cnr.it/stlab-tools/wikifier.
[17] http://www.visualdataweb.org/relfinder.php.

```
<http://dbpedia.org/wiki/Mark_Durkan> <http://dbpedia.org/
    ontology/predecessor> <http://dbpedia.org/resource/
    John_Hume>
<http://dbpedia.org/resource/Northern_Ireland> <http://dbpedia.
    org/ontology/leaderName>
<http://dbpedia.org/resource/Derry>
```

La figura 2 riassume questo tipo di procedimento: la conoscenza linguistica estratta può essere usata per arricchire (e dare accesso a) la conoscenza logico-formale, mentre quest'ultima, oltre a generare le conseguenze implicite nella sua struttura (inferenze deduttive), serve anche da *background knowledge* agli algoritmi che estraggono nuova conoscenza linguistica (come quelli usati nel compito di risoluzione di entità).

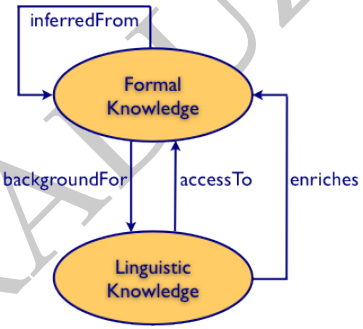

Figura 2: L'ibridazione di tecnologie linguistiche e logiche.

Le interpretazioni linguistica e formale e l'integrazione delle tecnologie relative permettono una metodologia ibrida che arricchisce e potenzia la strutturazione e l'interrogazione della conoscenza. Un progresso recente per tale ibridazione è fornito dal compito di *deep*

knowledge extraction, come è implementato nell'applicazione FRED.[18] FRED esegue un'analisi profonda delle frasi di un testo in linguaggio natural, produce strutture formalmente corrette in RDF e OWL e arricchisce I risultati con la risoluzione di entità. La figura 3 mostra un frammento del grafo RDF prodotto da FRED a partire dalla frase di esempio del Parlamento UE.

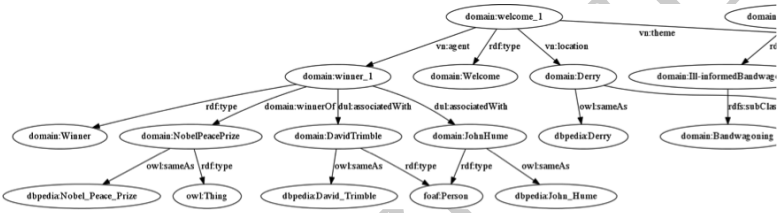

Figura 3: Frammento del grafo RDF prodotto da FRED a partire dalla frase di esempio del Parlamento UE.

Un caso di studio: il Consiglio Nazionale delle Ricerche (CNR)

Un esempio concreto di dati costruiti seguendo questo ciclo di produzione è http://data.cnr.it, un insieme di dataset che contengono i dati della ricerca del CNR, arricchiti mediante estrazione automatica dell'informazione, categorizzazione automatica dei profili relativi alle persone e ai progetti del CNR, e materializzazione automatica delle inferenze prodotte a partire dai dati e dai loro schemi, progettati nella forma di ontologie modulari (figura 8 a pagina 17), che sono a loro volta allineate con i vocabolari di riferimento di Linked Open Data; una parte della tassonomia principale risultante dalle

[18] http://wit.istc.cnr.it/stlab-tools/fred.

ontologie modulari è mostrata in figura 9 a pagina 17. I dati sono anche collegati dove possibile a entità dotate di un'identità pubblica, per esempio entità dei dataset di DBpedia,[19] GeoNames,[20] etc.
L'accesso ai dati avviene in modalità diverse a seconda di chi li deve "consumare". Applicazioni di sistemi informativi useranno il cosiddetto SPARQL endpoint per eseguire query. Invece gli utenti umani hanno a disposizione molti modi per cercare ed esplorare i dati. Un'applicazione sviluppata per il "consumo" dei dati di data.cnr.it è il Semantic Scout,[21] una web application che assiste l'expert finding sulle competenze scientifiche presenti nel CNR, come sono rappresentate nei dataset di data.cnr.it. La figura 4 nella pagina successiva mostra una ricerca a parole-chiave effettuata con il Semantic Scout per trovare chi si occupa di diritto romano nel CNR, la figura 5 a fronte mostra la navigazione gerarchica nella rete sociale semantica costruita a partire dalla conoscenza relativa a una delle persone trovate; la figura 6 a pagina 16 mostra l'interfaccia di "ricerca esplorativa" sui dati in uno spazio sferico e la figura 7 a pagina 16 mostra la possibilità di esportare i risultati dell'esplorazione personalizzati sulla base delle scelte dell'utente durante l'esplorazione.

Conclusioni

Le tecnologie semantiche forniscono una soluzione semplice ai problemi dell'identità e dell'interoperabilità, sfruttando l'accessibilità diretta dei dati, gli standard web, la precisione logica degli schemi e la facilità di ibridazione fra tecniche orientate all'esplicitazione della semantica linguistica e tecniche orientate alla gestione della semantica logico-formale.

[19] http://dbpedia.org.
[20] http://www.geonames.org.
[21] http://bit.ly/semanticscout.

Figura 4: Ricerca con il Semantic Scout.

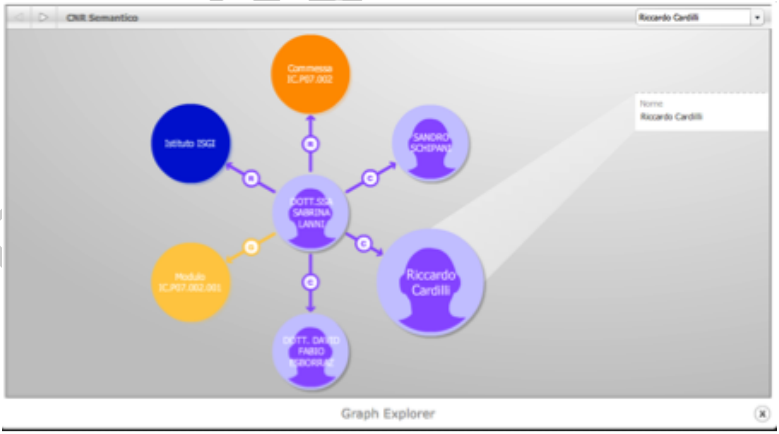

Figura 5: Navigazione semantica con il Semantic Scout.

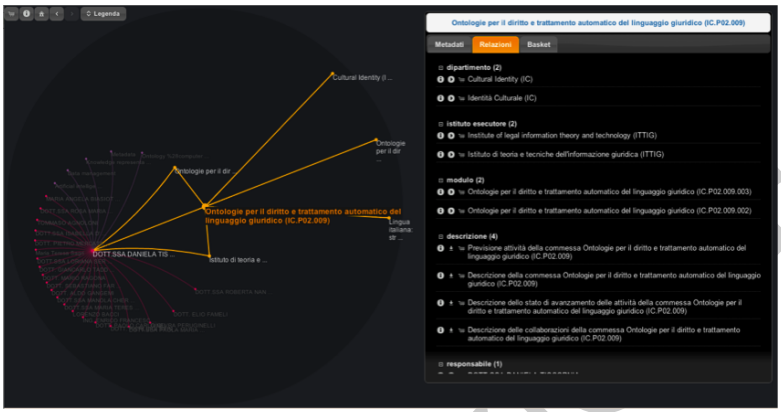

Figura 6: Interfaccia di ”ricerca esplorativa” sui dati in uno spazio sferico.

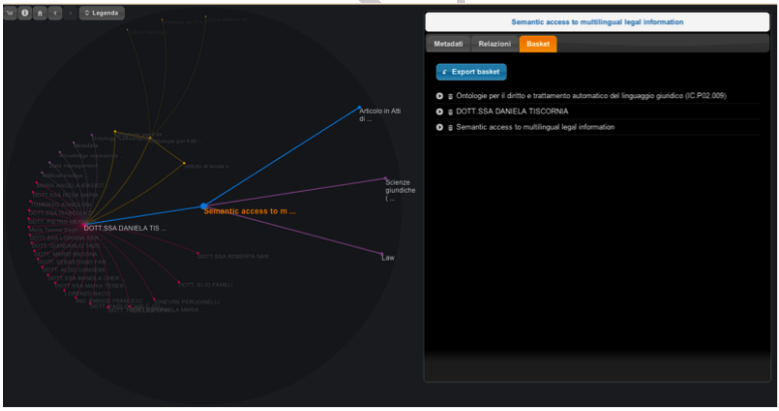

Figura 7: Esportazione dei risultati dell’esplorazione a partire dalle scelte dell’utente.

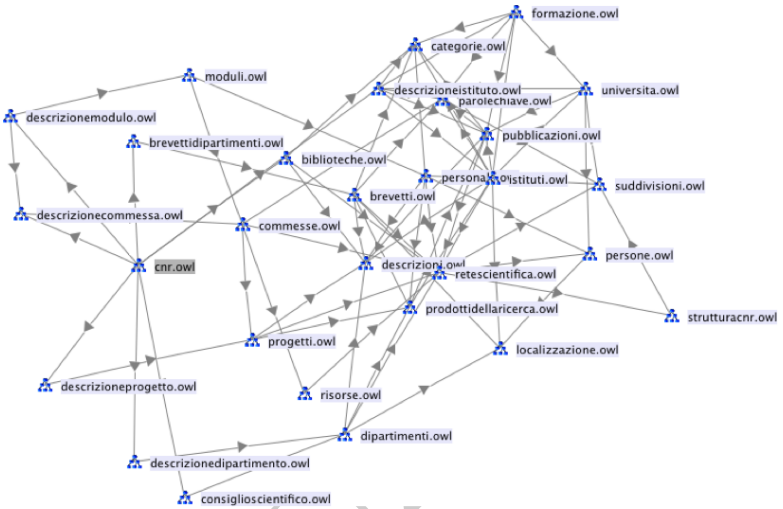

Figura 8: La rete di ontologie modulari del CNR.

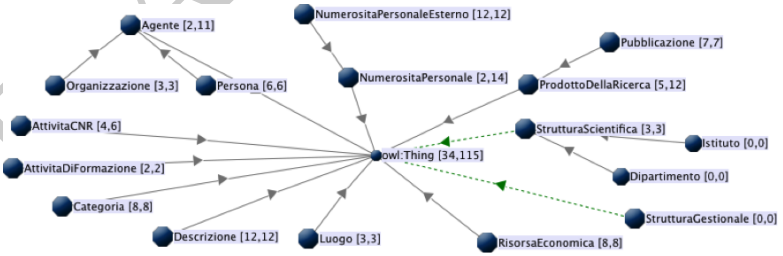

Figura 9: La tassonomia centrale dell'ontologia CNR.

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

Gangemi, A. "Semantic technologies and linked data, with a case study at theConsiglio Nazionale delle Ricerche (CNR)". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5457. DOI: 10.4403/jlis.it-5457. Web.