Relazione Scientifica

Attività svolta dall'Ing. Davide Taibi durante il periodo di Short Term Mobility presso l'istituto di ricerca L3S di Hannover

Questa relazione riporta l'attività di ricerca svolta dall'Ing. Davide Taibi durante la partecipazione al programma di Short Term Mobility dal 12 Novembre al 03 Dicembre 2013 presso il centro di ricerca L3S della Leibniz Universität di Hannover in Germania.

Relazione Scientifica

Durante il periodo di visita svolto presso il centro di ricerca L3S di Hannover l'Ing. Davide Taibi ha collaborato con il Dr. Stefan Dietze e il Dr. Besnik Fetahu per la realizzazione di un approccio innovativo alla catalogazione delle risorse didattiche presenti nel Linked Open Data cloud. Tali attività si sono inserite all'interno delle attività previste dal progetto LinkedUp (http://linkedup-project.eu), progetto finanziato nell'ambito delle Azioni di Supporto del settimo programma quadro (FP7-ICT-2011-8).

La crescente disponibilità di dati nel Web, ed in particolare la diffusione degli approcci Linked Open Data (LOD) per la loro pubblicazione, offrono una risorsa senza precedenti per l'apprendimento online di tipo sia formale che informale. Benché la quantità di dataset nel LOD che offrono contenuti potenzialmente rilevanti per le attività didattiche sia numerosa, valutare se un dataset contiene risorse specifiche, particolarmente adatte per il raggiungimento di un obiettivo didattico è, attualmente, una sfida di ricerca ardua. Nel corso delle attività svolte durante il programma di STM si sono gettate le basi per lo sviluppo di un approccio volto a valutare la copertura di uno specifico argomento didattico da parte delle risorse contenute all'interno di un dataset, utilizzando le tecnologie del web semantico e in particolare del Linked Open Data. Di fatto, negli ultimi anni, le tecnologie e i principi del Linked Data hanno fornito le basi per la creazione di una rete di dati, in cui informazioni eterogenee provenienti da diverse sorgenti vengono interconnesse tra loro. Così come il successo del Web si è basato sulla capacità di avere documenti in formato di ipertesto collegati tra loro, il Linked Data basa il suo successo sulla possibilità di collegare tra loro diverse tipologie di dati. Le caratteristiche di eterogeneità dei dati trattati rendono necessaria l'applicazione delle tecnologie del web semantico che, con le sue ontologie, fornisce vocabolari comuni in grado di rendere la rete distribuita dei Linked Data sfruttabile per successive elaborazioni basate sui dati e sulle loro connessioni. I dataset presenti nel Linked Open Data sono stati pubblicati da istituzioni operanti in differenti domini, per cui tra essi si trovano dati di tipo geografico, dati pubblicati da enti governativi, pubblicazioni scientifiche, dati legati alla scienza della vita e così via.

Le attività svolte durante il periodo di STM, si sono focalizzate principalmente sui dataset legati al settore dell'apprendimento. In questa tipologia di dataset ricadono non soltanto i dataset i cui contenuti sono stati esplicitamente creati per scopi educativi, ma anche quei dataset contenenti informazioni non strettamente legati al mondo dell'educazione ma che possono essere sfruttati per fini didattici. Al primo insieme di dataset appartengono tutti i dati legati ai corsi accademici e alla struttura organizzativa, messi a disposizione da differenti università nel mondo. Di fatto negli ultimi anni il numero di università che espongono i loro dati in formato Linked Data per favorire una maggiore apertura verso l'esterno e consentire lo sviluppo di applicazioni basate sui loro dati, è cresciuto notevolmente (solo per citare alcuni esempi: l'Open University[1] di Milton Keynes,

[1] http://data.open.ac.uk/

l'università di Oxford[2], l'università di Münster[3]). Relativamente ai dataset non specificatamente creati per fini didattici, ma contenenti risorse che possono essere di interesse didattico, troviamo quei dataset messi a disposizione da musei, librerie o editori di articoli scientifici (come ad esempio: Europeana[4], Elsevier[5], Nature[6]). I dataset appartenenti a questa tipologia nascono con finalità diverse da quelle educative ma il loro impiego a fini didattici può essere rilevante. Pertanto la rilevanza didattica dei contenuti dei dataset presenti nel Linked Data cloud va al di là della mera classificazione attuata tramite le parole chiavi indicate da chi pubblica il dataset, e pertanto si rende necessario sviluppare dei metodi che focalizzano l'attenzione sulle caratteristiche delle risorse presenti in ciascun dataset.

Nell'ambito del progetto LinkedUp è stato svolto un primo lavoro di selezione dei dataset presenti nel Linked Data cloud utili a fini didattici. Tale insieme di dataset chiamata "Linked Education Cloud", disponibile al seguente indirizzo: http://data.linkededucation.org/linkedup/catalog/ è stato scelto come insieme di dataset iniziali per le attività svolte durante il periodo di STM. In particolare la seguente tabella riporta la lista dei dataset presi in esame:

Table 1: Dataset selezionati dal Linked Education Cloud

Learning Analytics and Knowledge (LAK) Dataset	http://datahub.io/it/dataset/lak-dataset
Oxford University	http://datahub.io/it/dataset/oxpoints
Yovisto - academic video search	http://datahub.io/it/dataset/yovisto
DBLP in RDF (L3S)	http://datahub.io/it/dataset/l3s-dblp
COLINDA - Conference Linked Data	http://datahub.io/it/dataset/colinda
education.data.gov.uk	http://datahub.io/it/dataset/education-data-gov-uk
TheSoz Thesaurus for the Social Sciences (GESIS)	http://datahub.io/it/dataset/gesis-thesoz
Morelab	http://datahub.io/it/dataset/morelab
Open Data from the Italian National Research Council	http://datahub.io/it/dataset/data-cnr-it
data.open.ac.uk, Linked Data from the Open University	http://datahub.io/it/dataset/data-open-ac-uk
Educational programs - SISVU	http://datahub.io/it/dataset/educationalprograms_sisvu
Linked Open Aalto Data Service	http://datahub.io/it/dataset/linked-open-aalto-data-service
ASN:US	http://datahub.io/it/dataset/asn-us
Linked Open Data of the University of Southampton	http://datahub.io/it/dataset/data-southampton-ac-uk
Open Courseware Consortium metadata	http://datahub.io/it/dataset/open-courseware-

[2] https://data.ox.ac.uk/
[3] http://lodum.de/
[4] http://www.europeana.eu/
[5] http://data.elsevier.com
[6] http://data.nature.com/

in RDF	consortium-metadata-in-rdf
Linked Open Data of the University of Bristol	http://datahub.io/it/dataset/university-of-bristol
UNISTAT-KIS in RDF (Key Information Set - UK Universities)	http://datahub.io/it/dataset/unistat-kis-in-rdf-key-information-set-uk-universities
LODUM	http://datahub.io/it/dataset/lodum
Italian public schools (LinkedOpenData.it)	http://datahub.io/it/dataset/italian-public-schools-linkedopendata-it
Nature Publishing Group - ALL	http://datahub.io/it/dataset/npg
http://datahub.io/it/dataset/rkb-explorer-kaunas	
SEEK-AT-WD ICT tools for education - Web-Share	http://datahub.io/it/dataset/seek-at-wd-ict-tools-for-education-web-share
Organic Edunet Linked Open Data","oxpoints":"OxPoints (University of Oxford)	http://datahub.io/it/dataset/organic-edunet
Aristotle University of Thessaloniki	http://datahub.io/it/dataset/aristotle-university
Universitat Pompeu Fabra - linked data	http://datahub.io/it/dataset/universitat-pompeu-fabra-linked-data
Publications of Charles University in Prague	http://datahub.io/it/dataset/publications-of-charles-university-in-prague
University of Huddersfield -- Circulation and Recommendation Data	http://datahub.io/it/dataset/hud-library-usagedata

Basandosi sulle risorse presenti nei dataset elencati in tabella 1, durante il periodo di STM sono state portate aventi le seguenti attività:

- Creazione del dataset-profile basato sulla tipologia delle risorse contenute in ciascun dataset
- Creazione di un modello di dati per la rappresentazione dei datset-profile
- Creazione di uno strumento per l'esplorazione dei profili creati

Ciascun punto viene di seguito descritto in dettaglio.

Creazione del dataset-profile basato sulla tipologia delle risorse contenute in ciascun dataset

In questa sezione viene descritto la metodologia adottata per la creazione del profilo dei dataset in esame. In primo luogo vengono selezionati, per ciascun dataset, le tipologie di risorse in esso presenti. Per esempio il dataset "Learning Analytics and Knowledge (LAK) Dataset", contenente gli articoli selezionati dalle principali conferenze sul Learning Analytics ed Educational Data Mining, ha al suo interno risorse appartenenti alle seguenti classi definite nelle rispettive ontologie (in questo caso le ontologie in questione sono: foaf[7], swrc[8], swc[9] e led[10]):

[7] http://xmlns.com/foaf/0.1/
[8] http://swrc.ontoware.org/ontology#

foaf:Person
foaf:Organization
swrc:Proceedings
swrc:InProceedings
swc:ConferenceEvent
led:Reference

Per ciascuna delle suddette tipologie di risorse vengono estratte dal dataset in maniera casuale un insieme di risorse. In questo modo si crea un campione rappresentativo di risorse del dataset, per ciascuna tipologia di risorsa. Successivamente, le risorse appartenenti ai campioni (per ogni tipologia) vengono elaborate mediante tecnologie di arricchimento dei dati, mediante l'uso del servizio DBpedia spotlight[11] al fine di associare ad ogni risorsa i concetti specifici dell'ontologia DBpedia. Ogni concetto di DBpedia ha una o più categorie associato ad esso. Ad esempio il concetto: <http://dbpedia.org/resource/Memory> ha associato le categorie Memory[12], Neuropsychological_assessment[13], Sources_of_knowledge[14], Mental_processes[15]. In questo modo si riesce a creare una corrispondenza tra risorse contenute in un dataset e le categorie espresse all'interno di DBpedia. Inoltre considerando che ogni categoria è legata alle categorie da cui essa deriva mediante la proprietà skos:broader, per ogni categoria associata ad un concetto è possibile estrarre un grafo di categorie che identificano le categorie di una risorsa a diversi livelli di granularità.

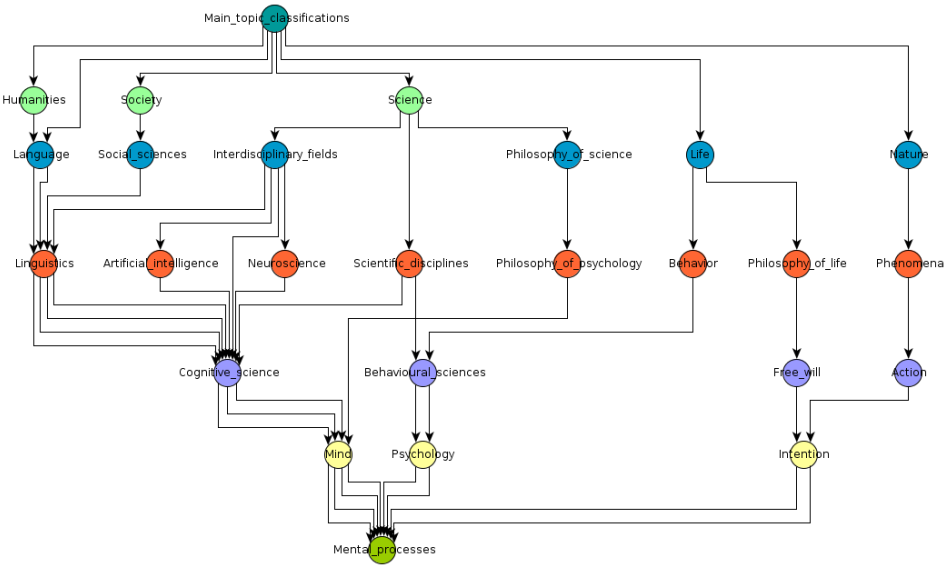

Figura 1: Grafo delle categorie legate alla categoria Mental_processes

[9] http://data.semanticweb.org/ns/swc/ontology#
[10] http://data.linkededucation.org/ns/linked-education.rdf#
[11] http://spotlight.dbpedia.org
[12] http://dbpedia.org/resource/Category:Memory (n.b. qui memory è una categoria)
[13] http://dbpedia.org/resource/Category:Neuropsychological_assessment
[14] http://dbpedia.org/resource/Category:Sources_of_knowledge
[15] http://dbpedia.org/resource/Category:Mental_processes

In figura 1 è mostrato il grafo delle categorie a cui la categoria Mental_processes è legata. Applicando le tecnologie sviluppate in (Fetahu B., Dietze S., Nunes B., Taibi D., Casanova M.A.: Generating structured Profiles of Linked Data Graphs. International Semantic Web Conference ISWC 2013: 113-116), è stato possibile determinare una misura del grado di associazione tra le risorse e le categorie. Questa misura stabilisce la forza dell'associazione tra le risorse di un dataset e le categorie. Pertanto per ciascun dataset è stato possibile determinare le principali categorie associate ad esso. A titolo di esempio, i risultati dell'applicazione di queste tecniche hanno consentito di affermare che per il dataset "TheSoz Thesaurus for the Social Sciences (GESIS)" le principali categorie ad esso associate sono:

db:Society
db:Social_systems
db:Structure
db:Knowledge
db:Classical_and_art_music_traditions
db:Classicism

mentre il dataset "Nature Publishing Group" è rappresentato, secondo le tecniche applicate, dalle seguenti categorie:

db:Bioinformatics
db:Biological_databases
db:Biorepositories
db:Biostatistics
db:Biotechnology
db:Computational_biology
db:Science
db:Scientific_buildings

Il profilo determinato, in questo modo, per ciascun dataset rappresenta l'associazione tra i dataset, le sua risorse e le corrispondenti categorie di DBpedia. L'ontologia "Describing RDF Resource Links with the Vocabulary of Links" (VoL), disponibile all'indirizzo http://data.linkededucation.org/vol/ è stata utilizzata per rappresentare questo tipo di legami.

Creazione di un modello per la rappresentazione dei profili dei dataset

I dati prodotti per il profilo di ciascun dataset sono stati resi accessibili attraverso lo Sparql endpoint disponibile al seguente indirizzo: http://meco.l3s.uni-hannover.de:8890/sparql utilizzando il grafo: "http://data-observatory.org/lod-profiles/linked-education-profile".

La seguente query SPARQL:

select distinct ?dataset ?category where {?dataset a void:Dataset . ?dataset dcterms:subject ?category}

che può essere eseguita direttamente dal seguente link:

http://meco.l3s.uni-hannover.de:8890/sparql?default-graph-uri=http%3A%2F%2Fdata-observatory.org%2Flod-profiles%2Flinked-education-profile&query=select+distinct+%3Fdataset+%3Fcategory+where+{%3Fdataset+a+void%3ADataset+.+%3Fdataset+dcterms%3Asubject+%3Fcategory}+&format=text%2Fhtml&timeout=0&debug=on

restituisce tutte le categorie associate ai dataset in esame. Considerato che ogni dataset contiene differenti tipologie di risorse, si possono anche ottenere le associazioni tra le tipologie di risorse e le categorie. Attualmente, il dataset contiene l'associazione tra risorse e categorie ma non l'associazione tra tipologie di risorse e categorie, per cui durante il periodo di visita si sono sviluppate delle procedure che a partire dalle risorse di ogni dataset, interrogando il corrispondente repository, si è individuato la tipologia di risorsa. In tal modo è stato possibile creare associazioni tra tipologie di risorse e categorie di DBpedia.

Le tipologie di risorse presenti in un dataset possono essere molteplici ma, considerando le relazioni esistenti tra le diverse tipologie di risorse, espresse ad esempio dalle proprietà *owl:equivalentClass*, *rdf:subClassOf*, è stato possibile ridurre le tipologie di risorse da prendere in esame per tutti i dataset considerati. In questa prima fase di analisi, l'attenzione si è focalizzata sulle seguenti tre tipologie di risorse: foaf:Document, foaf:Agent (nella quale ricadono con sottoclassi foaf:Person e foaf:Organization), aiiso:Course. Tutte le altre tipologie di risorse sono state al di sotto della categorie Other. In questo modo alla prima tipologia fanno parte le risorse didattiche, mentre in foaf:Agent ricadono tutte quelle risorse associate alle persone e alla organizzazioni, mentre aiiso:Course riguarda nella maggior parte dei casi, la descrizione dei corsi svolti dalle università.

Il risultato ottenuto al termine di questo passo è l'associazione, tra le diverse tipologie di risorse contenute in un dataset e le categorie. Questo risultato è un primo passo che consente di classificare le risorse presenti in un dataset (sia che essi rappresentino persone, organizzazioni, materiali didattici, ecc…) e consente anche di focalizzare l'attenzione solo verso quelle tipologia di risorsa di un dataset ritenuta più utile, per le finalità che si vogliono perseguire. L'analisi di questi dati è stata condotta attraverso uno strumento di esplorazione appositamente sviluppato durante il periodo di STM, e descritto nel prossimo paragrafo.

Creazione di uno strumento per l'esplorazione dei profili creati.

L'esplorazione dei dati contenuti nei profili è stata supportata dall'applicazione "Dataset Profile Explorer", sviluppata durante il periodo di STM e disponibile al seguente indirizzo:

http://meco.l3s.uni-hannover.de:9080/observatory/led-explorer/

La figura 2 mostra una schermata dell'applicazione.

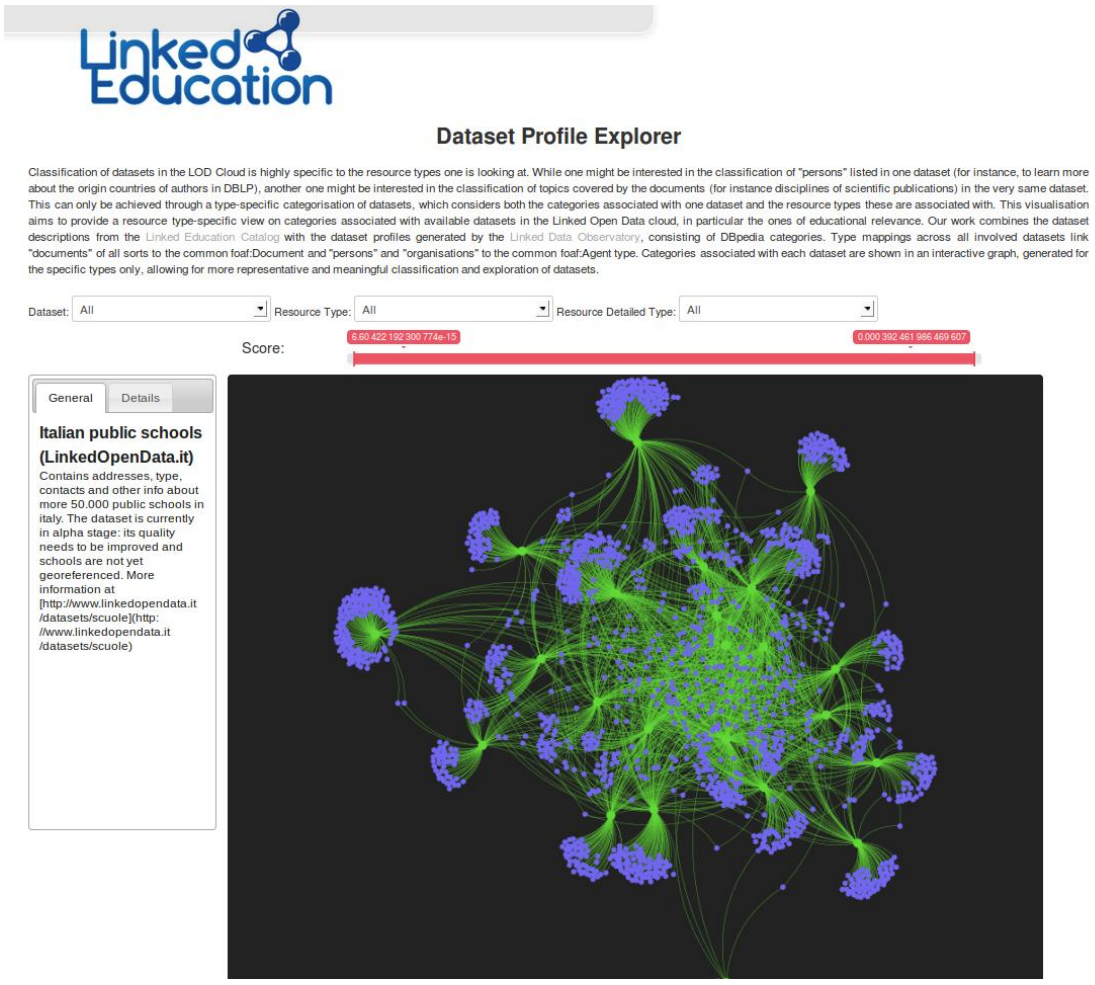

Figura 2 : L'applicazione "dataset profile explorer"

In particolare, l'applicazione per l'esplorazione dei profili, visualizza la rete tra i dataset e le categorie associate, consentendo di filtrare i risultati secondo differenti parametri relativi: al dataset, alla tipologia di risorse, alla sotto-tipologia (secondo le relazione di equivalenza e sottoclasse, tra le tipologie), al punteggio dei legami, dataset-categoria.

Ad esempio è possibile focalizzare l'attenzione sulla sottorete relativa ad un solo dataset come mostrato in figura 3.

La figura 3 mostra la parte di rete relativa solo al dataset LODUM contenente le risorse dell'università di Munster.

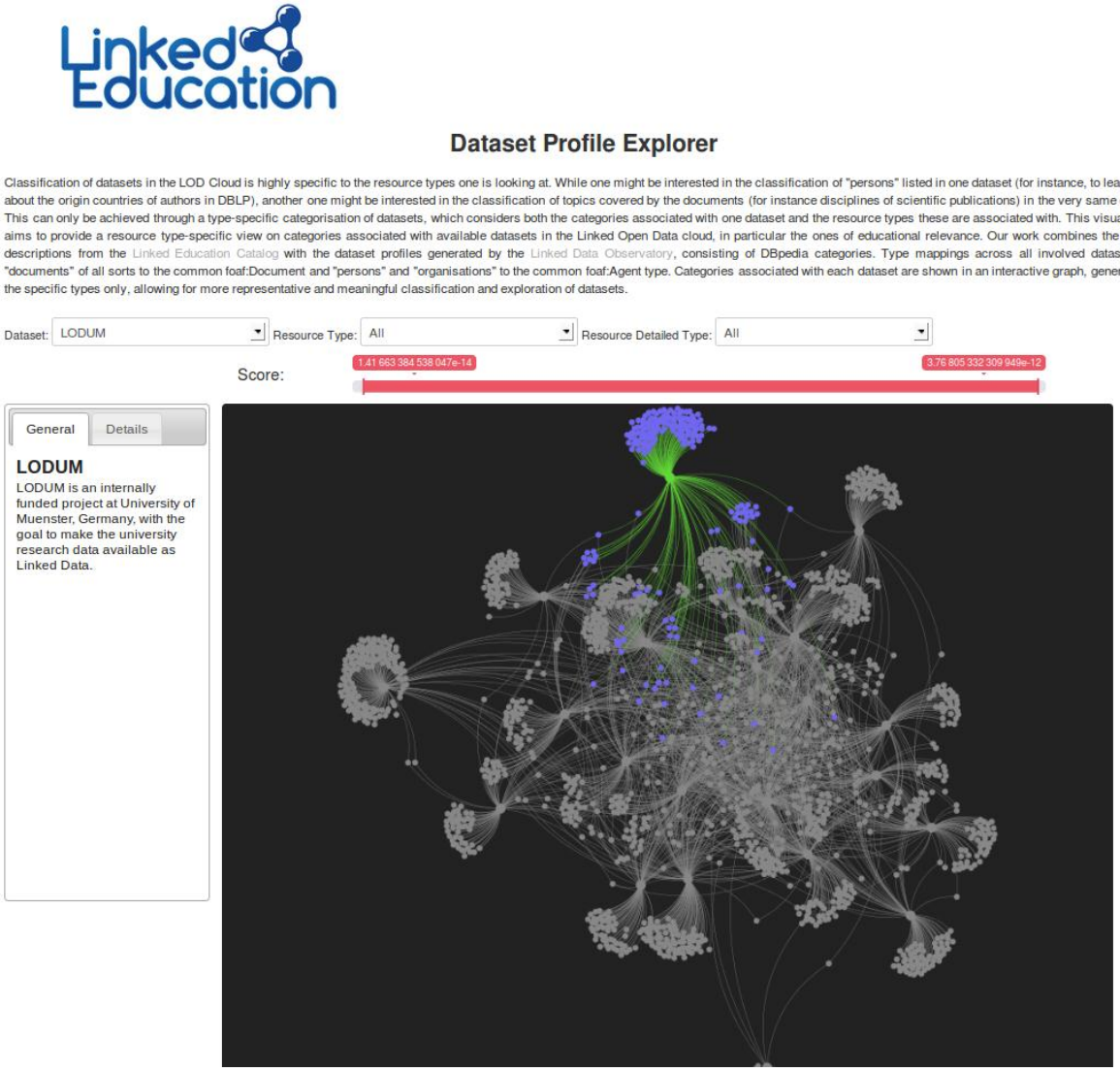

Figura 3 : Grafo delle categorie associate al dataset LODUM

Cliccando su un nodo rappresentante un dataset, si ottengono sia delle informazioni generali relative alla descrizione sul contenuto del dataset ma anche delle informazioni di dettaglio relative alle dieci categorie associate al dataset con punteggio più alto. La seguente figura mostra le 10 categorie con punteggio più alto associate al dataset Gesis.

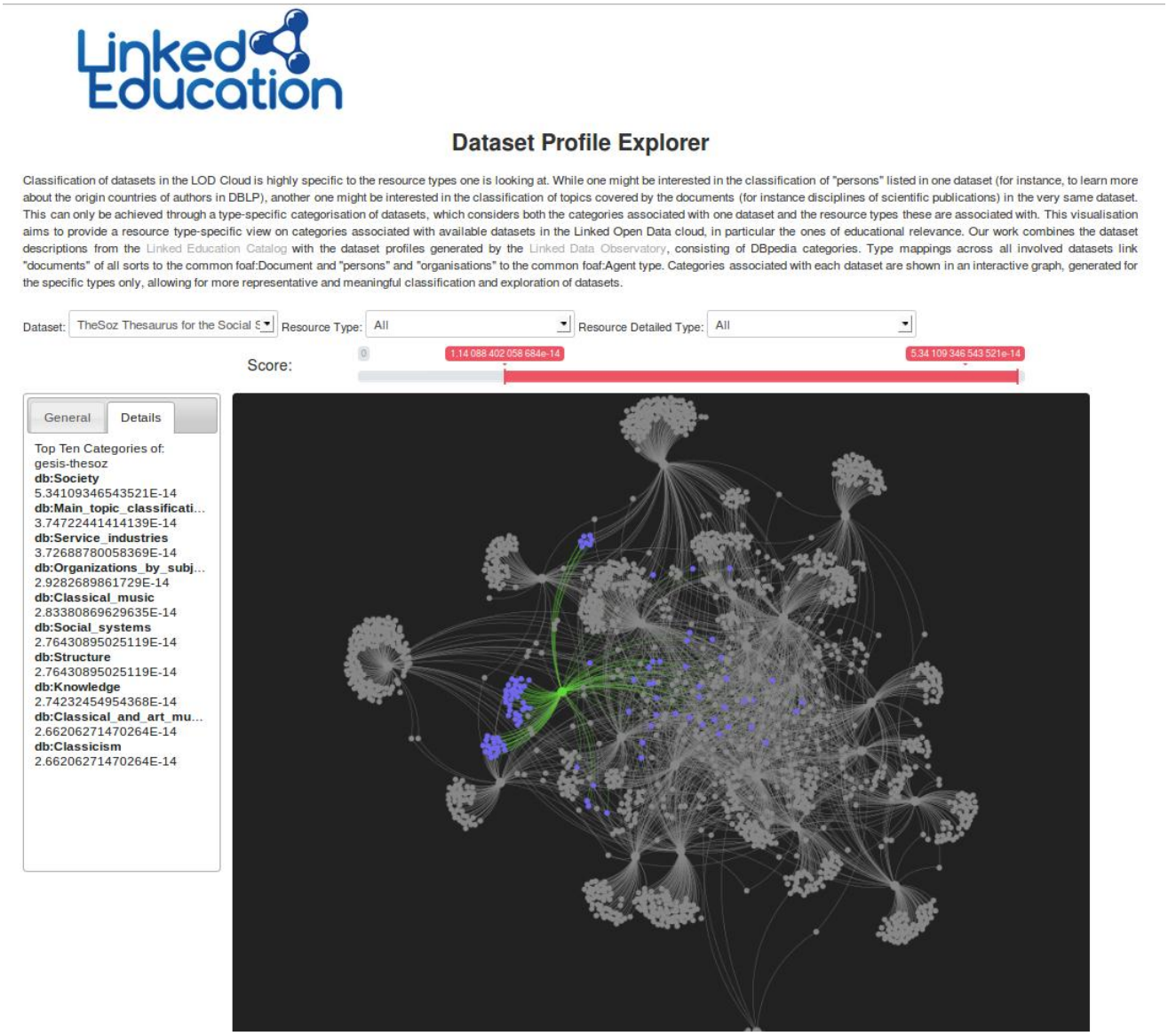

Figura 4: Le dieci categorie più rilevanti per il dataset Gesis

Altri filtri che possono essere applicati riguardano la tipologia di risorse (nelle 3 tipologie definite come principali in questa fase del progetto: foaf:Document, foaf:Agent, aiiso:Course), oppure stabilita la tipologia di risorsa è anche possibile focalizzare l'attenzione su una specifica sotto-tipologia di risorsa ad essa collegata (ad esempio si possono filtrare tutti i bibo:Book di uno specifico dataset).

In figura 5, ad esempio, viene evidenziata solo la parte della rete relativa ai dataset contenenti le risorse di tipo foaf:Document.

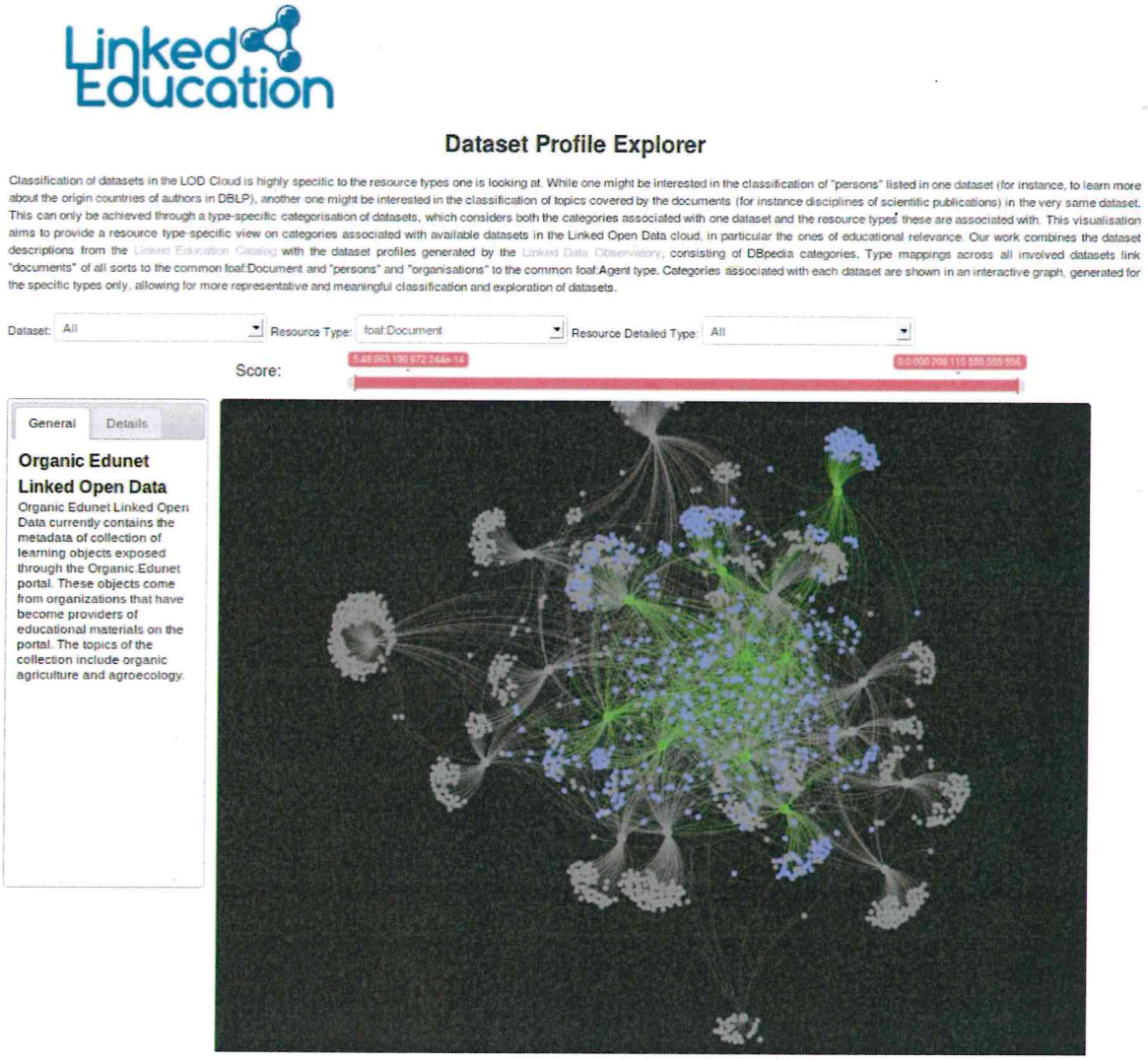

Figura 5: Visualizzazione dei Dataset contenenti risorse del tipo foaf:Document

Allo stesso modo è possibile applicare e combinare tra loro tutte le diverse tipologie di filtro.

Conclusioni

I dataset presenti nel Linked Open Data costituiscono una fonte preziosa di informazioni collegate tra loro semanticamente. Durante le attività svolte nel periodo di Short Term Mobility presso il centro di ricerca L3S della Leibniz Universität di Hannover, sono stati analizzati i dataset del "Linked Education Cloud" prodotto all'interno del progetto LinkedUp, al fine di proporre un approccio per la classificazione di tali dataset basato sulle categorie associate alle tipologie di risorse presenti nei dataset. Una applicazione per l'esplorazione dei risultati è stata sviluppata e resa disponibile online. Infine, si sta valutando la possibilità di presentare questo lavoro e le sue evoluzioni in conferenze specifiche di settore.

Palermo, 07/01/2014

Firma