Alma Mater Studiorum
Università di Bologna Dipartimento di Filologia Classica e Italianistica

Ivan Heibi,
ivan.heibi2@unibo.it

Relazione conclusiva per l'assegno di ricerca: "Extending and Visualising the OpenCitations Corpus"

L'assegno di ricerca e i suoi obiettivi

L'OpenCitations Corpus (OCC, http://opencitations.net) è una collezione aperta di dati citazionali, messi a disposizione nel Pubblico Dominio, che rende disponibili, in RDF, informazioni sulle citazioni presenti negli articoli accademici. L'OpenCitations Enhancement Project finanziato dalla Alfred P. Sloan Foundation di New York vuole rendere l'OCC ancora più utile per la comunità accademica estendendo il volume dei dati citazionali presenti nel Corpus e migliorando le sue interfacce utente. A questo proposito, l'OpenCitations Enhancement Project ha finanziato l'assegno di ricerca in oggetto.
L'obiettivo principale del progetto da svolgere è quello di sviluppare nuove interfacce di visualizzazione e nuovi servizi di interrogazione sui dati citazionali presenti all'interno del corpus.

Lo svolgimento

La durata dell'assegno di ricerca è stata di 12 mesi. In tale periodo il mio lavoro è stata fatto in stretta collaborazione con il Dr. Silvio Peroni presso il Digital and Semantic Publishing Laboratory (DASPLab) dell'Università di Bologna, e successivamente all'interno del Dipartimento di Filologia Classica e Italianistica (FICLIT). Le fasi di svolgimento possono essere racchiuse nei seguenti passi:

1) Documentazione e pratica delle tecnologie utilizzate per lo sviluppo dell'OpenCitations Corpus, in particolare le tecnologie di Semantic Web, in aggiunta ad una revisione dello stato dell'arte riguardo alle tecnologie da adottare nello sviluppo di interfacce web user friendly per la ricerca e l'interrogazione dei dataset.

2) La effettiva realizzazione delle interfacce e la loro integrazione all'interno del sito web di OpenCitations (http://opencitations.net)

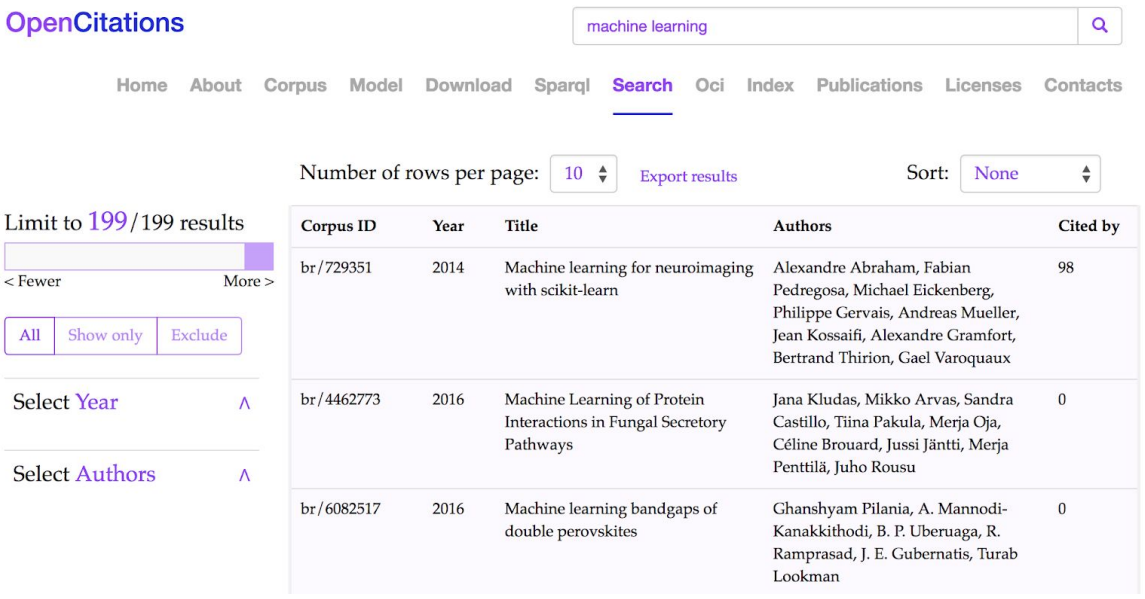
Fase di documentazione

Questa fase di circa 2 mesi, è stata principalmente dedicata allo studio delle tecnologie “Semantic Web”. Il fatto di non aver mai trattato in precedenza queste tecnologie, ha richiesto uno sforzo iniziale di documentazione e di studio dello stato dell’arte. Questo procedimento è stato notevolmente agevolato da l’aiuto ricevuto dal Prof. Silvio Peroni. Una ricca applicazione con esempi e esercitazioni riguardo l’uso corretto di queste tecniche a portato in oltre ad una preparazione più veloce in vista della seconda fase.

Fase di sviluppo delle interfacce

Dopo il termine della prima fase, il resto del periodo (circa 10 mesi), è stato dedicato alla creazione delle interfacce da poter integrare successivamente all’interno del sito web.
L’esigenza di avere tali interfacce è frutto del fatto che i dati presenti all’interno dell’OpenCitations Corpus fino a quel momento potevano essere interrogati solo da utenti esperti di Semantic Web e del linguaggio di interrogazione utilizzato: SPARQL. Per rendere questo servizio accessibile anche al resto degli utenti, privi di conoscenze nel campo del Semantic Web, si è voluto sviluppare delle interfacce User-Friendly di facile comprensione e utilizzo da parte di qualsiasi utente con un minimo di conoscenza riguardo alle classiche ricerche testuali fatte da i più comuni e diffusi sistemi, per esempio Google.
Alla fine di questo periodo sono state sviluppate due applicazioni, che successivamente furono integrate in OpenCitaions:

1) OSCAR (https://github.com/opencitations/oscar):
   Questa applicazione permette la creazione e customizzazione di una piattaforma di ricerca su dati in formato RDF, interrogabili tramite il linguaggio SPARQL. Nel caso specifico di OpenCitations, questa applicazione è stata configurata al fine di interrogare e visualizzare i risultati dell’OpenCitations Corpus. Questa configurazione di OSCAR è stata successivamente integrata all’interno del sito web, ed è accessibile tramite il seguente link: http://opencitations.net/search. Di seguito viene mostrato uno screenshot del risultato di una ricerca fatta dal sito di OpenCitations inserendo il testo “machine learning” come input:



2) LUCINDA: una volta fatta la ricerca i risultati mostrati possono essere accessibili e visualizzati separatamente utilizzando un software di browsing delle risorse, creato e customizzato tramite LUCINDA. Così come per OSCAR anche LUCINDA è stato creato affinché possa essere utilizzato su qualsiasi dataset in formato RDF, interrogabili tramite il linguaggio SPARQL. La versione integrata nel sito è ovviamente anche in questo caso dedicata all'OpenCitations Corpus.

Per entrambe le applicazioni il codice e liberamente accessibile tramite la Git repository ufficiale di OpenCitations (https://github.com/opencitations).

Convegni

- Nome: **The Web Conference,**
  Periodo: **23-27/4/2018**
  Luogo: **Lione, Francia.**
  In questa conferenza è stato presentato il web-tool sviluppato: OSCAR, facendo particolare riferimento alle sue caratteristiche e potenzialità. Viene inoltre fatto vedere un effettivo caso d'uso, ovvero il suo utilizzo all'interno di OpenCitations.

Pubblicazioni

- Heibi, I., Peroni, S., & Shotton, D. Enabling text search on SPARQL endpoints through OSCAR. Data Science, 1-23.
  DOI: https://doi.org/10.3233/DS-190016

- Heibi, I., Peroni, S., & Shotton, D. (2017). OSCAR: A customisable tool for free-text search over SPARQL endpoints. In Semantics, Analytics, Visualization (pp. 121-137). Springer, Cham.
  DOI: https://doi.org/10.1007/978-3-030-01379-0_9