Proposta di Programma di Formazione

Titolo: Estensione dell'OpenCitations Corpus con riferimenti testuali in linea

Tutor: Silvio Peroni <silvio.peroni@unibo.it>, che può essere contattato per ulteriori informazioni

Obiettivi

L'OpenCitations Corpus (OCC, http://opencitations.net) è una collezione aperta di dati citazionali, messi a disposizione nel Pubblico Dominio (https://creativecommons.org/publicdomain/zero/1.0/deed.it), che rende disponibili, in RDF, informazioni sulle citazioni presenti negli articoli accademici.

L'*Open Biomedical Citations in Context Corpus Project*, finanziato dal Wellcome Trust (Londra, Regno Unito), vuole rendere l'OCC ancora più utile per la comunità accademica estendendo significativamente le tipologie di dati citazionali presenti nel Corpus, in modo da fornire dati relativi ai riferimenti citazionali testuali in linea e al relativo contesto semantico, così da poter distinguere i riferimenti che sono citati solo una volta da quelli citati più volte, vedere quali riferimenti sono citati insieme (ad esempio, nella stessa frase), determinare in quale sezione dell'articolo i riferimenti sono citati (ad esempio, nell'introduzione, nei metodi), e, potenzialmente, estrarre la funzione della citazione – ovvero la ragione per cui un autore cita un altro lavoro. A questo proposito, l'*Open Biomedical Citations in Context Corpus Project* ha finanziato l'assegno di ricerca in oggetto. Il principale obiettivo del progetto da svolgere da parte dell'assegnista è quello di sviluppare tutto il software necessario per estrarre dati citazionali e il loro relativo contesto, e di salvarli all'interno dell'OpenCitations Corpus. Allo stesso tempo, l'assignista deve sviluppare appropriate interfacce utente per interrogare e navigare questi nuovi dati citazionali.

L'assegno associato al progetto *non è* di natura commerciale.

Piano di attività

Si prevede uno svolgimento di 12 mesi per il programma complessivo. Seppur il *Open Biomedical Citations in Context Corpus Project* è una collaborazione con l'École de Bibliothéconomie et des Sciences de l'Information (Université de Montréal, Canada), l'Oxford e-Research Centre (University of Oxford, Inghilterra), il Centre for Science and Technology Studies (Leiden University, Olanda), ed è formalmente supportato da Europe PubMed Central (EMBL-EBI, Inghilterra), l'assegnista lavorerà con il Dr. Silvio Peroni presso il Digital Humanities Advanced Research Centre (DHARC) del Dipartimento di Filologia Classica e Italianistica (Università di Bologna, Italia). Il centro è un ambiente vivo e stimolante, in cui l'assignista dovrà fornire il suo significativo contributo personale al progetto. In una prima fase di circa due mesi, l'assegnista dovrà impratichirsi delle tecnologie utilizzate per lo sviluppo dell'OpenCitations Corpus e dovrà altresì completare l'attività di studio dello stato dell'arte relativamente allo sviluppo di modelli per la descrizione dei riferimenti testuali in linea da aggiungere al Corpus. Nei restanti dieci mesi l'assegnista sarà responsabile:

- dell'estensione dell'OpenCitations Data Model in modo da descrivere come i dati dei riferimenti testuali in linea dovranno essere modellati in RDF così da essere inclusi nel Corpus;
- dello sviluppo di script per l'estrazione dei riferimenti testuali in linea dagli articoli contenuti dentro l'Open Access Subset di letteratura biomedica messo a disposizione da Europe PubMed Central;
- dell'estensione e creazione di script e applicazioni per popolare il Corpus con i nuovi dati relativi ai riferimenti testuali in linea e per meglio mantenerlo nel tempo;
- dello sviluppo di interfacce utente per interrogare e navigare i nuovi dati aggiunti.

Requisiti

L'assegnista deve avere ottime competenze di ricerca, di programmazione, e di comunicazione. Inoltre, deve essere in grado di scrivere e presentare oralmente i lavori svolti in Inglese. Passate esperienze in Python,

Web Interface Design, Information Visualization, tecnologie Web, Semantic Web e Linked Data sono valori aggiunti, così come la dedizione alle tematiche di Open Science e l'abilità di lavorare in gruppo. L'assegnista dovrà, altresì, spendere qualche settimana di lavoro presso i nostri collaboratori al Centre for Science and Technology Studies (Leiden University, Olanda) e a Europe PubMed Central (EMBL-EBI, Inghilterra). Il requisito minimo per applicare per la posizione è avere una laurea magistrale in informatica, ingegneria informatica, ingegneria delle telecomunicazioni, o equivalente, ma sono ben valutate e auspicabili eventuali esperienze di ricerca conformi con un possibile percorso di dottorato di ricerca.

Research programme

Title: Extending the OpenCitations Corpus with Individual In-text References

Academic supervisor: Silvio Peroni <silvio.peroni@unibo.it>, from whom further information may be obtained

Goals

The OpenCitations Corpus (OCC, http://opencitations.net) is an open repository of scholarly citation data made available under a Creative Commons public domain dedication (CC0, https://creativecommons.org/publicdomain/zero/1.0/), which provides in RDF accurate citation information (bibliographic references) harvested from the scholarly literature.

The *Open Biomedical Citations in Context Corpus Project*, funded by the Wellcome Trust (London, United Kingdom), wants to make the OCC more useful to the academic community by significantly expanding the kinds of citation data held within the Corpus, so as to provide data for each individual in-text reference and its semantic context, making it possible to distinguish references that are cited only once from those that are cited multiple times, to see which references are cited together (e.g. in the same sentence), to determine in which section of the article references are cited (e.g. Introduction, Methods), and, potentially, to retrieve the function of the citation – i.e. the reason why an author cites another work. To this end, it has funded the salary of a skilled computer scientist / research engineer, who will primarily be involved in the development of all the necessary software for extracting the citation data and their contexts, and storing them into the expanded OpenCitations Corpus, as well as developing appropriate user interfaces for querying and browsing these data.

The project related to this position is non-commercial in nature.

Activity plan

The Research Fellowship position has a duration of 12 months. While the *Open Biomedical Citations in Context Corpus Project* is a collaboration with the École de Bibliothéconomie et des Sciences de l'Information (Université de Montréal, Canada), the Oxford e-Research Centre (University of Oxford, United Kingdom), Centre for Science and Technology Studies (Leiden University, The Netherlands), and supported by Europe PubMed Central (EMBL-EBI, United Kingdom), the Research Fellow will work with Dr Silvio Peroni in the Digital Humanities Advanced Research Centre (DHARC) at the Department of Computer Classical Philology and Italian Studies (University of Bologna, Italy). This is a lively and stimulating environment, and the Research Fellow will be expected to provide a key personal contribution to the project. During the first two months, the Research Fellow will practice the technologies used for the development of the OpenCitations Corpus (OCC) and will study the state-of-the-art related to the development of the models for describing the in-text reference to be added in the Corpus. In the remaining ten months, the Research Fellow will be responsible for:

- the extension of the OpenCitations Data Model so as to describe how the in-text reference data should be modeled in RDF for inclusion in the OpenCitations Corpus;
- the development of scripts for extracting in-text references from articles within the Open Access Subset of biomedical literature hosted by Europe PubMed Central;
- the extension and creation of scripts and applications related with the population of the Corpus with the new in-text reference data and with its regular maintenance;
- the development of appropriate user interfaces for querying and browsing these new data.

Requirements

Applicants are expected to have excellent research skills, computer programming skills, and the ability to

communicate, undertake academic writing and make verbal conference presentations in good English. Expertise in Python, Web Interface Design and Information Visualization, and in Semantic Web technologies, Linked Data and Web technologies would be highly beneficial, plus a strong and demonstrable commitment to open science and team-working abilities. In addition, applicants are expected to spend a few weeks working with our partners at Centre for Science and Technology Studies (Leiden University, The Netherlands) and at Europe PubMed Central (EMBL-EBI, United Kingdom). The minimal formal requirement for this position is a Masters degree in computer science, computer science and engineering, telecommunications engineering, or equivalent title, but it is expected that the successful applicant will have had research experience leading to a doctoral degree.