Proposta di Programma di Formazione

Titolo: Migliorare il processo di modifica e arricchimento dell'Open Biomedical Citations in Context Corpus

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

Obiettivi

OpenCitations (http://opencitations.net) è un'infrastruttura che promuove e mette a disposizione strumenti per la diffusione di dati bibliografici e citazionali aperti nel contesto accademico, messi a disposizione nel Pubblico Dominio (https://creativecommons.org/publicdomain/zero/1.0/deed.it). Utilizzando le tecnologie del web semantico - che permettono la creazione e la pubblicazione di dati FAIR (findable, accessible, interoperable, reusable) sul web in modo che siano facilmente processabili anche da agenti software - l'infrastruttura nasce per offrire tanto alle istituzioni che ai singoli la possibilità di analizzare e riutilizzare le citazioni scientifiche nelle raccolte bibliografiche e nelle attività di ricerca.

L'*Open Biomedical Citations in Context Corpus Project*, finanziato dal Wellcome Trust (Londra, Regno Unito), vuole rendere i dati citazionali in OpenCitations ancora più utili per la comunità accademica estendendo significativamente le tipologie di dati citazionali attualmente gestiti, 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 le borse di ricerca in oggetto.

Il principale obiettivo del progetto da svolgere da parte dei borsisti è quello (a) di sviluppare tutto il software necessario per la gestione delle modifiche dei dati salvati nell'Open Biomedical Citations in Context Corpus, e (b) di sviluppare meccanismi di *indexing* così da processare simultaneamente un grande numero di documenti in input localmente, senza avvalersi di servizi esterni.

Le borse di ricerca associate al progetto *non è* di natura commerciale.

Piano di attività

Si prevede uno svolgimento di 5 mesi per ogni borsista 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), il/la borsista lavorerà con il Dr. Silvio Peroni presso il Research Centre for Open Scholarly Metadata del Dipartimento di Filologia Classica e Italianistica (Università di Bologna, Italia). Il centro è un ambiente vivo e stimolante, in cui i borsisti dovranno fornire il loro significativo contributo personale al progetto. In una prima fase di circa un mese, i borsisti dovranno impratichirsi del software e delle tecnologie utilizzate per processare documenti scientifici in formato XML, per recuperare metadati bibliografici attinenti, e per gestire i dati RDF finali che verranno salvati nel Corpus. Nei restanti mesi, i borsisti saranno responsabili:

- dell'estensione di software esistente usato per la creazione di dati RDF in modo che possa venir gestita correttamente la modifica e cancellazione di dati presenti nel Corpus, organizzando la necessaria infrastruttura software per tenere traccia dei cambiamenti in modo coerente ai modelli già utilizzati;
- la creazione di uno o più indici di metadati e la conseguente estensione del software usato per il processo di creazione dei dati da salvare nel Corpus in modo da integrare gli indici sviluppati nel processo.

Requisiti

I borsisti devono avere ottime competenze di ricerca, di creazione e gestione di dati e modelli, di programmazione, e di comunicazione. Inoltre, deve essere in grado di scrivere e presentare oralmente i lavori svolti in Inglese. Passate esperienze in Semantic Publishing, Python, 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. Il requisito minimo per applicare per la posizione è avere una laurea triennale, con adeguato curriculum scientifico-professionale. Sono altresì ben valutate e auspicabili eventuali esperienze di ricerca conformi ai temi del progetto e con un possibile percorso di dottorato di ricerca.

Research programme

Title: Enhancing the workflow for modifying and enriching the Open Biomedical Citations in Context Corpus

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

Goals

OpenCitations (http://opencitations.net) is an independent infrastructure organization for open scholarship dedicated to the publication of open bibliographic and citation data made available under a Creative Commons public domain dedication (CC0, https://creativecommons.org/publicdomain/zero/1.0/). Using Semantic Web technologies – which enable the creation and publication of FAIR (findable, accessible, interoperable and re-usable) data on the Web to make them easily processable by software agents – the infrastructure allows single users and large institution to analyze and reuse scientific citations in bibliographic collections and research activities.

The *Open Biomedical Citations in Context Corpus Project*, funded by the Wellcome Trust (London, United Kingdom), wants to make citation data in OpenCitations more useful to the academic community by significantly expanding the kinds of citation data currently handled, 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, the *Open Biomedical Citations in Context Corpus Project* has funded the salary of two short-term research fellows.

Th main goals of the project the short-term research fellows have to address are (a) to develop the software for handling the modification of the data stored in the Open Biomedical Citations in Context Corpus, and (b) to develop indexing mechanisms to process simultaneously a large number of documents in a local environment, without using external services.

The project related to these positions is non-commercial in nature.

Activity plan

The short-term research fellowship positions have a duration of 5 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 short-term research fellows will work with Dr Silvio Peroni in the Research Centre for Open Scholarly Metadata at the Department of Computer Classical Philology and Italian Studies (University of Bologna, Italy). This is a lively and stimulating environment, and the short-term research fellow will be expected to provide a key personal contribution to the project. During the first month, the short-term research fellows will practice the software and technologies used for processing scientific documents stored in XML format, for gathering related bibliographic metadata, and for handling the resulting RDF data that will be stored in the Corpus. In the remaining months, the short-term research fellows will be responsible for:

- the extension of existing software used for creating RDF data to handle correctly the modification and removal of data in the Corpus, managing the necessary software infrastructure to keep track of the changes compliantly with the models in use in the Corpus;
- the creation of one or more metadata indexes and the related extension of the software used in the workflow of data creation, to integrate the indexes developed in the workflow.

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 Semantic Publishing, Python, 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. The minimal formal requirement for this position is a Bachelor degree with appropriate experience in the topics of the project. In addition, it is expected that the successful applicant will have had research experience leading to a doctoral degree.