Programma di formazione

Titolo
Ontologie per la descrizione FAIR-based di menzioni in articoli scientifici e software per l'estrazione di menzioni di dataset e software da PDF

Responsabile scientifico
Professor Silvio Peroni <silvio.peroni@unibo.it>, Dipartimento di Filologia Classica e Italianistica, Università di Bologna / Direttore di OpenCitations, che può essere contattato per ulteriori informazioni.

Obiettivi
OpenCitations (https://opencitations.net) [1] è una infrastruttura Open Science supportata dalla comunità scientifica e accademica che fornisce informazioni sulla ricerca aperte, accessibili e riutilizzabili per favorire processi trasparenti e riproducibili a supporto della ricerca, delle decisioni politiche e della reperibilità del dominio accademico. Questo si realizza attraverso la raccolta e la pubblicazione aperta di metadati accurati e completi che descrivono le pubblicazioni accademiche e le relative citazioni che le collegano. OpenCitations rende disponibili i dati delle citazioni e i metadati bibliografici in modo che chiunque, ovunque nel mondo, possa utilizzarli per qualunque scopo. OpenCitations è no-profit e tutti i suoi servizi sono completamente gratuiti. Essa è gestita dal Research Centre for Open Scholarly Metadata dell'Università di Bologna (https://openscholarlymetadata.org).

Negli ultimi tre anni, all'interno di OpenCitations sono stati sviluppati nuovi software utilizzati per creare due collezioni di metadati bibliografici e citazionali, ovvero OpenCitations Meta [2] e OpenCitations Index [3]. I metadati esposti da OpenCitations includono, oltre che ai collegamenti citazionali, i metadati bibliografici utilizzati per descrivere informazioni di base di una risorsa bibliografica. In particolare, conserva gli identificatori delle risorse bibliografiche (ad esempio, DOI, PMID, ISSN e ISBN), il titolo, il tipo, la data di pubblicazione, le pagine e il luogo della risorsa (con i numeri di volume e di fascicolo se il luogo è una rivista). Sono inclusi anche i metadati riguardanti i principali attori coinvolti nella pubblicazione di una risorsa bibliografica, ovvero gli autori, gli editori e gli editori, ogni attore può essere caratterizzato con altri identificatori (ad esempio ORCID) se disponibili. Tutti i dati in OpenCitations derivano da fonti/risorse esistenti (attualmente tali risorse includono Crossref, DataCite, PubMed, OpenAIRE e JaLC).

Uno dei principali scopi di OpenCitations è quello di processare, con sistemi automatici basati su tecnologie di machine learning e su altri approcci di intelligenza artificiale, le porzioni di testi di articoli scientifici che contengono puntatori a riferimenti bibliografici e a menzioni non formalizzate a dataset e software, con l'obiettivo di esporre dati di questa natura utilizzando (e, in caso di necessità, sviluppando) ontologie specifiche. L'obiettivo di questo lavoro di ricerca è quello di implementare, all'interno dell'infrastruttura di OpenCitations, dei meccanismi computazionali che permettano l'estrazione automatica di queste menzioni a software e dataset, e sviluppare/mantenere le opportune ontologie necessarie per la descrizione di questi dati.

Piano di attività
La Borsa di Ricerca avrà durata di 7 mesi a partire da Aprile 2025. Il Borsista di Ricerca lavorerà direttamente con il Professor Silvio Peroni nel contesto del Research Centre for Open Scholarly Metadata, presso il Dipartimento di Filologia Classica e Italianistica dell'Università di Bologna (Italia). Il Centro di Ricerca è un ambiente vivo e stimolante, ed è atteso che il Borsista di Ricerca fornisca contributi personali centrali alle attività di OpenCitations. Il lavoro a distanza può essere

possibile se strettamente necessario, ma altrimenti la presenza di persona nel Centro di Ricerca è preferibile.

Durante il periodo di lavoro è prevista una fase iniziale introduttiva e conoscitiva del contesto applicativo. Dopodichè il lavoro del Borsista di Ricerca può essere organizzato e riassunto in questi punti:

1. Sviluppare e/o arricchire ontologie per la descrizione delle menzioni in documenti scientifici e accademici.

2. Raccogliere ed organizzare un corpus adeguato di documenti (in formato PDF) contenenti menzioni a software e dataset da utilizzare come training data e gold standard.

3. Sviluppare uno o più sistemi diversi, ad esempio basati su language models e knowledge graph embeddings, per l'estrazione delle menzioni da articoli accademici.

4. Testare i vari sistemi per identificare i più promettenti – eventualemente anche combinandoli tra loro.

5. Sviluppare un software web-based e delle API che permetta utilizzo di questo sistema di estrazione in modo programmatico.

6. Scrivere e pubblicare una documentazione appropriata per descrivere i risultati ottenuti da questo lavoro.

Mentre il professor Peroni dirigerà e supervisionerà il lavoro, il Borsista di Ricerca avrà la responsabilità di gestire in modo autonomo e sistematico queste attività.

Requisiti

Tutti/e i/le candidati/e devono avere eccellenti abilità come programmatori/trici e, come valore aggiunto, devono essere in grado di parlare, scrivere, e presentare verbalmente a conferenze in un buon inglese. Esperienze dimostrabili di programmazione in Python e utilizzo dei più comuni librerie Python, e sistemi di versionamento basati su Git (in particolare GitHub) sono fortemente desiderabili. In più, è altresì fortemente desiderabile che il/la candidato/a abbia una forte e dimostrabile attitudine verso la Scienza Aperta e la capacità di lavorare in gruppo. Conoscenze dimostrabili nelle tecnologie del Web Semantico, Linked Data e tecnologie Web in generale sono elementi favorevoli per la candidatura.

I requisiti minimi formali per la posizione sono il possesso di una Laurea Magistrale o equivalente. Il candidato deve avere un'esperienza adeguata e dimostrabile come programmatore, comprovata dai documenti da allegare in fase di domanda. La candidatura (in Italiano o in Inglese) deve almeno includere un Curriculum Vitae completo di informazioni riguardanti attività scientifico-professionali e relative alla produttività scientifica. Eventuali lettere di raccomandazione sono opzionali, ma fortemente consigliate.

L'Università di Bologna è un'istituzione che da pari opportunità di impiego, e la selezione per questa posizione verrà fatta esclusivamente sul merito.

Riferimenti

1. Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies, 1(1), 428–444. https://doi.org/10.1162/qss_a_00023

2. Massari, A., Mariani, F., Heibi, I., Peroni, S., & Shotton, D. (2024). OpenCitations Meta. Quantitative Science Studies, 5(1), 50–75. https://doi.org/10.1162/qss_a_00292

3. Heibi, I., Moretti, A., Peroni, S., & Soricetti, M. (2024). The OpenCitations Index: Description of a database providing open citation data. Scientometrics, 129(12), 7923–7942. https://doi.org/10.1007/s11192-024-05160-7