Programma di formazione

Titolo

Sistemi e workflow per la ricerca, analisi e risoluzione di errori dei dati bibliografici e citazionali contenuti nei due database OpenCitations Meta e OpenCitations Index e implementazione di software per automatizzare la loro correzione

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 (http://opencitations.net/) [1] è una infrastruttura Open Science che mette a disposizione una grossa mole di metadati bibliografici e dati citazionali accademici, di qualità e copertura tali da competere con servizi proprietari, come Web of Science e Scopus. 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 è stato sviluppato e testato un nuovo software (lanciato nel dicembre 2022) per creare una nuova raccolta, chiamata OpenCitations Meta (https://opencitations.net/meta) [2]. I metadati esposti da OpenCitations Meta includono 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). Inoltre, OpenCitations Meta contiene 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.

Tutte le risorse bibliografiche incluse in OpenCitations Meta derivano da fonti/risorse esistenti (attualmente tali risorse includono Crossref, DataCite e PubMed). In particolare, esiste un record specifico per ogni entità (che cita o che viene citata) coinvolta in ogni citazione inclusa negli Indici di OpenCitations (https://opencitations.net/index) [3] - un indice di citazioni aperte contenente tutte le entità che citano e sono citate, identificate da identificatori persistenti usati dalle fonti (cioè DOI o PMID), che sono coinvolti nelle citazioni.

Uno dei principali scopi di OpenCitations Meta è quello di gestire il problema posto dalle risorse bibliografiche che non dispongono di identificatori persistenti, associando a ciascuna di esse un nuovo identificatore persistente locale - un OpenCitations Meta Identifier (OMID). In questo modo, OpenCitations Meta può contenere i metadati di qualsiasi pubblicazione accademica, senza la necessità obbligatoria che un identificatore persistente esterno sia fornito dalla fonte dei metadati (ad esempio, DOI).

In linea con questo scopo, l'obiettivo principale di questo lavoro è quello di migliorare la qualità dei dati di OpenCitations Meta e di OpenCitations Index andando alla ricerca di e correggere, con strumenti computazionali che prevedono lo sviluppo di software ad hoc,

eventuali errori presenti nelle sorgenti usate (Crossref, DataCite, National Istitute of Health Open Citation Collection, OpenAIRE, Japan Link Center) o ottenuti a seguito dell'elaborazioni di queste sorgenti per l'ingestione dei dati nelle collezioni di OpenCitations.

Piano di attività

L'Assegno di Ricerca avrà durata di 12 mesi a partire da Maggio 2024, eventualmente rinnovabile. L'Assegnista 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 dei software attualmente utilizzati per creare e modificare dati conformi all'OpenCitations Data Model, che è il modello appositamente sviluppato per descrivere i dati, ed in generale della infrastruttura OpenCitations. Dopodichè il lavoro dell'Assegnista di Ricerca può essere organizzato e riassunto in questi punti:

1) Sviluppo di script e sistemi per l'interrogazione dei database di OpenCitations basati su SPARQL come linguaggio di interrogazione di riferimento, in modo da avere gli strumenti per ricercare gli errori;
2) Analisi, attraverso gli strumenti sviluppati, dei database di OpenCitations per l'identificazione del maggior numero possibile di errori introdotti nelle collezioni;
3) Ideazione e implementazione di workflow computazionali, eventualmente basati su euristiche, per la risoluzione degli errori identificati;
4) Documentazione degli errori e delle relative risoluzioni adottate mediante l'utilizzo dei meccanismi di tracking di issue a disposizione nei repository software di OpenCitations;
5) Esecuzione e messa a sistema delle modifiche effettuate.

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 LM43 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. https://doi.org/10.1162/qss_a_00292
3. Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. Scientometrics, 121(2), 1213–1228. https://doi.org/10.1007/s11192-019-03217-6