

Programma di formazione

Titolo

Processo automatico per ingestione dei dati nelle collezioni di OpenCitations

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>). 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 [2] 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 in OpenCitations 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.

L'ingestione di nuovi dati proveniente da diverse sorgenti – tra cui Crossref, DataCite, NIH-OCC, OpenAIRE, JaLC – è un processo che prevede molti passaggi svolti in modo manuale o semi-automatico (assistito da un tecnico) che, necessariamente, porta via diversi giorni di lavoro prima di poter ingerire e quindi rilasciare i nuovi dati dell'infrastruttura. Inoltre, non permette al momento l'integrazione sistematica di dati provenienti da un meccanismo di crowdsourcing che è stato recentemente sperimentato.

Uno dei principali scopi di OpenCitations, che è anche l'obiettivo principale di questo lavoro, è quello di cercare di migliorare sempre di più la procedura di ingestione dei dati in modo da renderla completamente automatica sia per le fasi di preprocessamento delle sorgenti, della conversione e ingestione dei nuovi dati da inserire nei database di OpenCitations, sia della verifica della corretta esecuzione del processo e dei dati che questo ha prodotto, e infine alla pubblicazione delle nuove versioni delle collezioni in specifici repository e contesti applicativi. L'obiettivo, in particolare, è quello di implementare, mediante l'identificazione ed uso di uno specifico framework, una pipeline automatica, flessibile e estendibile integrata appropriatamente con le tecnologie correnti usate da OpenCitations per l'infrastruttura tecnologica in essere.

Piano di attività

L'Incarico di Ricerca avrà durata di 12 mesi a partire da Aprile 2026. La persona vincitrice dell'Incarico 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 la persona vincitrice 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 e del framework tecnologico. Dopodichè il lavoro della persona vincitrice del Contratto di Ricerca può essere organizzato e riassunto in questi punti:

- 1) Raccogliere ed organizzare i requisiti che deve implementare la pipeline automatica di ingestione dei dati;
- 2) Effettuare un'estensiva analisi dei framework esistenti per la creazione di pipeline software automatiche che si possano integrare in modo efficace e appropriato con la esistente infrastruttura tecnologica hardware e software di OpenCitations;
- 3) Lo sviluppo sistematico e agile di una prima versione della pipeline, al fine di identificare la sua efficacia e rilevare specifiche criticità.
- 4) Il test sistematico delle varie componenti software da sviluppare o già a disposizione per avere la sicurezza del corretto funzionamento della pipeline.
- 5) La messa in produzione della pipeline per l'ingestione sistematica delle sorgenti di OpenCitations.
- 6) Scrivere e pubblicare una documentazione appropriata per descrivere i risultati ottenuti da questo lavoro, inclusi lavori scientificamente rilevanti.

Mentre il professor Peroni dirigerà e supervisionerà il lavoro, la persona vincitrice dell'Incarico 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*, 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>

Training Programme

Title

Automatic process for data ingestion into the OpenCitations collections

Scientific

Supervisor

Professor Silvio Peroni silvio.peroni@unibo.it, Department of Classical Philology and Italian Studies, University of Bologna / Director of OpenCitations, who may be contacted for further information.

Objectives

OpenCitations (<http://opencitations.net/>) is an Open Science infrastructure that provides a large body of bibliographic metadata and scholarly citation data, with a level of quality and coverage comparable to proprietary services such as Web of Science and Scopus. OpenCitations is non-profit and all its services are completely free of charge. It is managed by the Research Centre for Open Scholarly Metadata of the University of Bologna (<https://openscholarlymetadata.org>).

Over the past three years, a new software system has been developed and tested within OpenCitations (launched in December 2022) to create a new collection called OpenCitations Meta (<https://opencitations.net/meta>). The metadata exposed by OpenCitations Meta include bibliographic metadata used to describe the basic information of a bibliographic resource. In particular, it stores identifiers of bibliographic resources (e.g. DOI, PMID, ISSN, and ISBN), the title, type, publication date, pages, and venue of the resource (including volume and issue numbers when the venue is a journal). In addition, OpenCitations Meta contains metadata concerning the main actors involved in the publication of a bibliographic resource, namely authors, publishers, and editors; each actor may be characterised by additional identifiers (e.g. ORCID) where available.

All bibliographic resources included in OpenCitations Meta are derived from existing sources/resources (currently including Crossref, DataCite, and PubMed). In particular, a specific record exists for each entity (citing or cited) involved in each citation included in the OpenCitations Index—an open citation index containing all citing and cited entities, identified by persistent identifiers used by the sources (i.e. DOI or PMID), that are involved in citations.

The ingestion of new data from different sources—including Crossref, DataCite, NIH-OCC, OpenAIRE, and JaLC—is currently a process that involves many manual or semi-automatic steps (assisted by a technician), which necessarily requires several days of work before new data can be ingested and released within the infrastructure. Moreover, at present it does not allow the systematic integration of data coming from a crowdsourcing mechanism that has recently been experimented with.

One of the main aims of OpenCitations, which is also the primary objective of this work, is to continuously improve the data ingestion procedure in order to make it fully automatic. This

includes the preprocessing of sources, the conversion and ingestion of new data into the OpenCitations databases, the verification of the correct execution of the process and of the data it produces, and finally the publication of new versions of the collections in specific repositories and application contexts. The specific goal is to implement—through the identification and use of a suitable framework—an automatic, flexible, and extensible pipeline, appropriately integrated with the current technologies used by OpenCitations in its existing technological infrastructure.

Work

Plan

The Research Fellowship will have a duration of 12 months starting from April 2026. The successful candidate will work directly with Professor Silvio Peroni within the Research Centre for Open Scholarly Metadata, at the Department of Classical Philology and Italian Studies of the University of Bologna (Italy). The Research Centre is a lively and stimulating environment, and the successful candidate is expected to provide central personal contributions to the activities of OpenCitations. Remote work may be possible if strictly necessary, but otherwise on-site presence at the Research Centre is preferred.

During the working period, an initial introductory phase is planned to become familiar with the application context and the technological framework. After that, the work of the successful Research Fellow can be organised and summarised as follows:

- Collecting and organising the requirements that the automatic data ingestion pipeline must implement;
- Carrying out an extensive analysis of existing frameworks for creating automatic software pipelines that can be effectively and appropriately integrated with the existing hardware and software infrastructure of OpenCitations;
- The systematic and agile development of a first version of the pipeline, in order to assess its effectiveness and identify specific critical issues;
- The systematic testing of the various software components to be developed or already available, to ensure the correct functioning of the pipeline;
- Deployment of the pipeline for the systematic ingestion of OpenCitations data sources;
- Writing and publishing appropriate documentation describing the results achieved by this work, including scientifically relevant publications.

While Professor Peroni will direct and supervise the work, the successful Research Fellow will be responsible for managing these activities autonomously and systematically.

Requirements

All candidates must have excellent programming skills and, as an added value, must be able to speak, write, and give conference presentations in good English. Demonstrable experience in programming with Python and the use of common Python libraries, as well as version control systems based on Git (in particular GitHub), is highly desirable. In addition, a strong and demonstrable commitment to Open Science and the ability to work in a team are also highly desirable. Demonstrable knowledge of Semantic Web technologies, Linked Data, and web technologies in general will be considered a favourable element for the application.

The formal minimum requirement for the position is a Master's degree (Laurea Magistrale) LM-43 or equivalent. The candidate must have adequate and demonstrable experience as a programmer, supported by the documents submitted with the application. The application (in Italian or English) must at least include a Curriculum Vitae containing information on scientific and professional activities and scientific productivity. Letters of recommendation are optional but strongly encouraged.

The University of Bologna is an equal opportunity employer, and selection for this position will be based exclusively on merit.

References

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>