Programma di formazione

(English version at page 5)

Titolo

Evoluzione dell'infrastruttura computazionale di OpenCitations come un cluster di servizi scalabili ed interoperabili

Responsabile scientifico

Professor Silvio Peroni <silvio.peroni@unibo.it>, Direttore di OpenCitations, che può essere contattato per ulteriori informazioni.

Obiettivi

OpenCitations (http://opencitations.net/) [1] è stata creata come una **infrastruttura aperta e completamente gratuita** 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**. OpenCitations è gestita dal Research Centre for Open Scholarly Metadata dell'Università di Bologna (https://openscholarlymetadata.org). Inoltre, è impegnata nella promozione delle citazioni aperte, in particolare nel suo ruolo di membro fondatore della Initiative for Open Citations (I4OC, https://i4oc.org) della Initiative for Open Abstracts (I4OA, https://i4oa.org).

Alla fine del 2019, OpenCitations è stata selezionata dalla Global Sustainability Coalition for Open Science Services (SCOSS, https://scoss.org) per il loro secondo ciclo di supporto per la raccolta fondi, visto che OpenCitations è ben allineata con gli obiettivi propri della Scienza Aperta ed offre un servizio innovativo. L'obiettivo di questo supporto di SCOSS è quello di permettere ad OpenCitations di continuare ad offrire e ad estendere i suoi servizi per i prossimi tre anni, garantendone la transizione in una infrastruttura accademica globale con una base finanziaria sicura. Inoltre, OpenCitations prende parte in altre importanti collaborazioni e progetti, che includono la European Open Science Cloud (EOSC), due progetti finanziati dall'Unione Europea – OpenAIRE Nexus (https://www.openaire.eu/openaire-nexus-project) e RISIS2 (https://www.risis2.eu/) – e altri progetti finanziati da enti nazionali come B!SON

(https://projects.tib.eu/bison/en/project/) e i progetti EXCITE/OUTCITE (https://excite.informatik.uni-stuttgart.de/).

Parte di questo lavoro riguarda il rafforzamento l'attuale infrastruttura tecnica e computazionale (server, processi paralleli, backup, etc.) di OpenCitations, che è attualmente ospitata presso l'Università di Bologna.

La mole e la complessità di dati messi a disposizione da OpenCitations aprono numerosi problemi relativi al miglioramento, la scalabilità e l'ottimizzazione della sua infrastruttura. È necessario quindi studiare e personalizzare le più recenti tecniche di gestione di Big Data per creare una base di conoscenza su cui costruire servizi decentralizzati, scalabili, duraturi e facilmente personalizzabili. L'obiettivo dell'Assegno di Ricerca, infatti, è da un lato studiare le limitazioni e i miglioramenti necessari all'attuale infrastruttura e, dall'altro, re-ingegnerizzare l'infrastruttura per renderla più scalabile, affidabile, interoperabile, riusabile e distribuita. Saranno quindi affrontati, tra gli altri, temi di erogazione di servizi Web, gestione di basi di dati distribuite ed eterogenee, processi di conversione e ingestione dati, computazione parallela e sicurezza. Particolare attenzione sarà dedicata alla preservazione dei dati e al processo di manutenzione e aggiornamento a lungo termine dell'infrastruttura.

Come tutte le attività relative a OpenCitations, il lavoro da svolgersi per questa posizione è di natura non commerciale.

Piano di attività

L'Assegno di Ricerca avrà durata di 12 mesi, a partire da Aprile 2023. La posizione è rinnovabile per ulteriori due anni. L'Assegnista di Ricerca lavorerà direttamente con il Professor Silvio Peroni e diventerà membro 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 multiculturale, vivo e stimolante, ed è atteso che l'Assegnista di Ricerca fornisca contributi personali centrali alle attività di OpenCitations, in particolare relativamente alla sua infrastruttura tecnica. La presenza di persona nel Centro di Ricerca è obbligatoria.

Durante i primi due mesi, l'Assegnista di Ricerca verrà introdotto/a a tutti i sistemi hardware e software usati da OpenCitations per fornire i propri servizi, così da capire lo stato corrente dell'infrastruttura tecnologica. Nei restanti mesi, l'Assegnista di Ricerca assumerà la responsabilità di gestione e sviluppo di questi sistemi, proponendo soluzioni innovative basate su pratiche di ricerca aggiornate, ove necessario. Questi sistemi includono:

- Un server fisico principale, che fornisce tutti i servizi di OpenCitations, e un server di test utilizzato per provare nuovi componenti software prima di mandarli in produzione;
- 30 Raspberry-Pis per permettere il processamento parallelo e l'ingestione di dati;
- Il *network-attached storage* (NAS) per backup di sistema programmati;
- Le basi di dati di tutti i dataset di OpenCitations, che includono l'OpenCitations Corpus [2], gli OpenCitations Indexes [3], il Open Biomedical Citations in Context Corpus [4], e il futuro OpenCitations Meta.

Durante il lavoro, l'Assegnista di Ricerca verrà anche coinvolto/a in altre attività atte a migliorare la performance e la sicurezza dell'infrastruttura tecnica di OpenCitations e i servizi che mette a disposizione in modo conforme ad un piano di sviluppo che è stato introdotto in [5]. Mentre il professor Peroni dirigerà e supervisionerà il lavoro, l'Assegnista di Ricerca avrà la libertà decisionale e di responsabilità di gestire in modo autonomo e sistematico queste attività.

Requisiti

Tutti/e i/le candidati/e devono avere abilità come sistemisti/e e come programmatori/trici, e devono essere in grado di leggere, scrivere e parlare in un buon inglese. Esperienze dimostrabili nell'usare e gestire server Linux (in particolare Debian e Raspbian), sistemi di orchestrazione di container (per esempio Kubernates), sistemi *infrastructure-as-code* (per esempio Terraform), server Web che implementano FastGCI (in particolare lighttpd), basi di dati basate su grafi (in particolare Blazegraph), programmazione in Python e utilizzo dei più comuni Python Web Application Framework (in particolare web.py e Flask), 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, e una buona conoscenza dell'italiano sono elementi favorevoli per la candidatura.

I requisiti minimi formali per la posizione sono il possesso di una Magistrale in Informatica, Ingegneria Informatica, Ingegneria delle Telecomunicazioni, Digital Humanities o equivalente. Il candidato deve avere un'esperienza adeguata e dimostrabile come sistemista e programmatore, comprovata dai documenti da allegare in fase di domanda. La candidatura (in Italiano o in Inglese) deve almeno includere:

1. una lettera di motivazione che introduca le ragioni del candidato nell'applicare per la posizione;

2. un Curriculum Vitae completo di informazioni riguardanti attività scientifico-professionali e relative alla produttività scientifica;

3. una descrizione di un portfolio pubblico di software, progetti e attività svolte negli ultimi dieci anni, indicando chiaramente quale ruolo è stato ricoperto in ognuno.

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 bibliografici

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. Peroni, S., Shotton, D., & Vitali, F. (2017). One Year of the OpenCitations Corpus. In C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, & J. Heflin (Eds.), The Semantic Web – ISWC 2017 (Lecture Notes in Computer Science Vol. 10588, pp. 184–192). Springer. https://doi.org/10.1007/978-3-319-68204-4_19, freely available at https://w3id.org/people/essepuntato/papers/oc-iswc2017.html
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
4. Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., Mayr, P., Romanello, M., & Zumstein, P. (2020). The OpenCitations Data Model. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020* (Lecture Notes in Computer Science Vol. 12507, pp. 447–463). https://doi.org/10.1007/978-3-030-62466-8_28, freely available at https://doi.org/10.48550/arXiv.2005.11981
5. Grieco, G., Heibi, I., Massari, A., Moretti, A., & Peroni, S. (2022). Enabling Portability and Reusability of Open Science Infrastructures. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, & A. Poggi (A c. Di), *Linking Theory and Practice of Digital Libraries—26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20–23, 2022, Proceedings* (Lecture Notes in Computer Science Vol. 13541, pp. 379–385). Springer. https://doi.org/10.1007/978-3-031-16802-4_36, freely available at https://doi.org/10.48550/arXiv.2206.03926

Research programme

Title

Evolving the OpenCitations infrastructure as a scalable and interoperable cluster of services

Supervisor

Professor Silvio Peroni <silvio.peroni@unibo.it>, Director of OpenCitations, from whom further information may be obtained.

Goals

OpenCitations (http://opencitations.net/) [1] has been established as a **fully free and open infrastructure** to provide access to global **scholarly bibliographic and citation data**, of quality and coverage to rival those from proprietary services, such as Web of Science and Scopus. OpenCitations is **not-for-profit** and all **its services are free**. OpenCitations is managed by the Research Centre for Open Scholarly Metadata of the University of Bologna (https://openscholarlymetadata.org). It is also engaged in advocacy for open citations, particularly in its role as a key founding member of the Initiative for Open Citations (I4OC, https://i4oc.org) and of the Initiative for Open Abstracts (I4OA, https://i4oa.org).

At the end of 2019, OpenCitations was selected by the Global Sustainability Coalition for Open Science Services (SCOSS, https://scoss.org) for its second round of crowdfunding support since OpenCitations aligns well with Open Science goals and is an innovative service. The goal of such support is to enable OpenCitations' operations over the next three years as it transitions into a global scholarly infrastructure organisation with a secure financial footing. In addition, OpenCitations takes part in other important collaborations and projects, including the European Open Science Cloud (EOSC), two EC-funded projects – OpenAIRE Nexus (https://www.openaire.eu/openaire-nexus-project) and RISIS2 (https://www.risis2.eu/) – and other national-funded projects such as B!SON (https://projects.tib.eu/bison/en/project/) and the EXCITE/OUTCITE projects (https://excite.informatik.uni-stuttgart.de/).

As part of this work, we need to strengthen the current technical and computational infrastructure (server, parallel processing, backup, etc.) used by OpenCitations, which is currently hosted at the University of Bologna.

The amount and complexity of the data made available by OpenCitations opens up to several issues related to the improvement, scalability and optimisation of its infrastructure. Therefore, it is crucial to study and customise recent approaches for managing Big Data to create a knowledge base to build decentralised, scalable, long-lasting and easily-customizable services. Indeed, the goal of the Research Fellowship is, on the one hand, to study the current limitations and possible improvements to introduce in the OpenCitations infrastructure and, on the other hand, to re-engineer the current infrastructure to make it more scalable, reliable, interoperable, reusable and distributed. Thus, the Research Fellow is expected to address issues related to the provision of Web services, the management of distributed and heterogeneous databases and services, data conversion and ingestion processes, parallel computing and security. Particular attention will be given to data preservation and the long-term maintenance and update of the infrastructure.

As with all OpenCitations activities, the work related to this position is non-commercial in nature.

Activity plan

The Research Fellowship will have a duration of 12 months, commencing in April 2023. The position can be renewed for additional two years. The Research Fellow will work directly with Professor Silvio Peroni and will become a member of the Research Centre for Open Scholarly Metadata at the Department of Classical Philology and Italian Studies, University of Bologna, Italy. This is a multicultural, lively and stimulating environment, and the Research Fellow will be expected to provide a key personal contribution to the OpenCitations activities, particularly as they relate to its technical infrastructure. Attendance in person at the Research Centre will be required.

During the first two months, the Research Fellow will be introduced to all the hardware and software systems currently used by OpenCitations to provide its services to become familiar with the current status of the technical infrastructure. Subsequently, the Research Fellow will assume responsibility for managing and developing these systems, introducing innovative solutions coming from current research practices when necessary. They comprise:

- The main physical server that provides the OpenCitations services and a testing server used to check new software components before putting them in production;

- 30 Raspberry-Pis for enabling parallel processing;
- The network-attached storage (NAS) for scheduled back-up of the system;
- The databases storing the data of all the OpenCitations datasets, including the OpenCitations Corpus [2], the OpenCitations Indexes [3], the Open Biomedical Citations in Context Corpus [4], and OpenCitations Meta [1].

During the course of the job, the Research Fellow will also be expected to engage in other activities and developments that improve the performance and security of the OpenCitations technical infrastructure and the services running on it according to a development plan that has already been sketched out in [5]. While Professor Peroni will direct and supervise his/her work, the Research Fellow will be given and will be expected to exercise a considerable degree of personal freedom and responsibility in undertaking these tasks.

Requirements

Applicants are expected to have sysadmin skills and computer programming skills, and the ability to read, write and speak in good English. Demonstrable expertise in using and managing Linux servers (especially Debian and Raspbian), container orchestrator systems (e.g. Kubernetes), infrastructure-as-code systems (e.g. Terraform), Web servers implementing FastGCI (especially lighttpd), graph databases (especially Blazegraph), Python programming and Python Web Application Frameworks (especially web.py and Flask), Git-based version-control systems (especially GitHub), plus a strong and demonstrable commitment to open science and team-working abilities are highly desirable. Demonstrable knowledge of Semantic Web technologies, Linked Data and Web technologies, and good knowledge of Italian would also be very beneficial.

The minimal formal requirement for this position is a Master's degree in Computer Science, Computer Science and Engineering, Telecommunications Engineering, Digital Humanities or equivalent. It is expected that the successful applicant will have had substantial and demonstrable experience in sysadmin activities and computer programming described in the documents to submit for the application. The Research Fellowship application (written either in Italian or in English) should minimally comprise the following:

1. a cover letter of application, motivating why the candidate is applying for this position;

2. a full Curriculum Vitae containing information about prior scientific and professional jobs and the scientific production of the candidate;

3. a public portfolio of software, projects and activities undertaken within the past ten years, with a clear indication of the applicant's role in each.

The University of Bologna is an equal opportunities employer, and selection for this post will be made solely on merit.

References

6. 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
7. Peroni, S., Shotton, D., & Vitali, F. (2017). One Year of the OpenCitations Corpus. In C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, & J. Heflin (Eds.), The Semantic Web – ISWC 2017 (Lecture Notes in Computer Science Vol. 10588, pp. 184–192). Springer. https://doi.org/10.1007/978-3-319-68204-4_19, freely available at https://w3id.org/people/essepuntato/papers/oc-iswc2017.html
8. 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
9. Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., Mayr, P., Romanello, M., & Zumstein, P. (2020). The OpenCitations Data Model. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020* (Lecture Notes in Computer Science Vol. 12507, pp. 447–463). https://doi.org/10.1007/978-3-030-62466-8_28, freely available at https://doi.org/10.48550/arXiv.2005.11981
10. Grieco, G., Heibi, I., Massari, A., Moretti, A., & Peroni, S. (2022). Enabling Portability and Reusability of Open Science Infrastructures. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, & A. Poggi (A c. Di), *Linking Theory and Practice of Digital Libraries—26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20–23, 2022, Proceedings* (Lecture Notes in Computer Science Vol. 13541, pp. 379–385). Springer. https://doi.org/10.1007/978-3-031-16802-4_36, freely available at https://doi.org/10.48550/arXiv.2206.03926