1. Open Research Fund application

Reference number	214471/Z/18/Z
Applicant name	Dr Silvio Peroni
Title of application	Open Biomedical Citations in Context Corpus
Total amount requested	€55,500.00

2. Application summary

Application title
Open Biomedical Citations in Context Corpus

Proposed duration of funding (months, this should be no longer than 1 year)
12

Proposed start date	01/03/2019

Is your application being submitted through a university?	Yes

Name of administering organisation
University of Bologna

Lead applicant's address at administering organisation	
Department/Division	Departiment of Classical Philology and Italian Studies
Organisation	University of Bologna
Street	via Zamboni 32
City/Town	Bologna
Postcode/Zipcode	40126
Country	Italy

Research funding area Please select from the drop-down list the funding area that you consider your research falls under
Genetics, Genomics and Population Research

3. Lead applicant

Lead applicant details

Full Name	Dr Silvio Peroni
Department	Department of Classical Philology and Italian Studies
Division	
Organisation	University of Bologna
Address Line 1	via Zamboni 32
City/Town	Bologna
Postcode	40126
Country	Italy
Telephone No.	+39 348 72 34 548
Email Address	essepuntato@gmail.com

ORCID iD	
ORCID iD	0000-0003-0530-4305

Career history (current/most recent first)

From	To	Position	Organisation
03/2018	02/2021	Assistant Professor	University of Bologna
10/2015	02/2018	Assistant Professor	University of Bologna
01/2012	10/2015	Research Fellow	University of Bologna
04/2008	10/2008	Intern	Open University (Milton Keynes)

Education/training

From	To	Qualification	Subject	Organisation
01/2009	05/2012	Doctor of Philosophy (PhD;DPhil)	Ph.D. (Computer Science)	University of Bologna
09/2005	03/2008	Master of Science (MSc)	Master degree (Computer Science)	University of Bologna
09/2002	06/2005	Bachelor of Science (BSc)	Bachelor degree (Computer Science)	University of Bologna

Source(s) of personal salary support

The salary of the lead applicant is fully paid by the University of Bologna.

Clinical status Do you have a medical/veterinary degree?	No

Career breaks Have you had any career breaks or periods of part-time work, for example parental or long-term sick leave?	No

Do you wish to undertake this award part time?	No

Career contributions
What are your most important research-related contributions to date? This may include contributions to health policy or practice, or to technology or product discovery and development.

Silvio Peroni holds a Ph.D. degree in Computer Science and he is an Assistant Professor at the Department of Classical Philology and Italian Studies, University of Bologna. Among his research interests are Semantic Web technologies, markup languages for complex documents, design patterns for digital documents and ontology modelling, automatic processes of analysis and segmentation of documents, and creation and maintenance of large RDF repositories. In particular, his recent works concern the empirical analysis of the nature of scholarly citations, the study of visualisation and browsing interfaces for semantic data, and the development of ontologies to manage, integrate and query bibliographic information according to temporal and contextual constraints.

Over the past eight years, he has been involved in the development of SPAR (Semantic Publishing and Referencing, http://www.sparontologies.net) Ontologies. The SPAR Ontologies are a set of complementary and orthogonal ontologies that can be used for the description of the main areas of the scholarly publishing domain, and are currently adopted by several parties that include the US Global Change Information System, the Working Group on Document Standards of the High Level Committee Management of the United Nations, various W3C Working Groups, Springer Nature, OpenCitations, and Wikidata.

He is also the Director (with David Shotton) of OpenCitations (http://opencitations.net), a scholarly infrastructure organization dedicated to the promotion of semantic publishing by the use of Semantic Web (Linked Data) technologies, and engaged in advocacy for semantic publishing and open citations. In particular, he is responsible for the technical development in terms of software and hardware infrastructure of all the services provided by OpenCitations, i.e. the OpenCitations Corpus, the OpenCitations Indexes, and the interfaces for querying and browsing the data (OSCAR, LUCINDA, RAMOSE).

He is also a founding member of the Initiative for Open Citations (I4OC, https://i4oc.org), and his work on Semantic Publishing topics has been recently published by Springer in a book entitled “Semantic Web Technologies and Legal Scholarly Publishing”.

Research outputs
List up to 5 of your most significant research outputs, ensuring that at least two of these are from the last five years. Provide a statement describing their significance and your contribution (up to 50 words per output).

Research outputs may include (but are not limited to):

- Peer-reviewed publications and preprints
- Datasets, software and research materials
- Inventions, patents and commercial activity

For original research publications please indicate those arising from Wellcome-funded grants in **bold**, and provide the PubMed Central ID (PMCID) reference for each of these. Please refer to guidance notes.

Publications should be in chronological order with the most recent first. Please give citation in full, including title of paper and all authors. Citations to preprints should state “Preprint”, the repository name and the articles persistent identifier (e.g DOI).*

*(*All authors, unless more than 10, in which case please use ‘et al’, ensuring that your position as author remains clear.)*

1. Silvio Peroni, David Shotton (2018). The SPAR Ontologies. To appear in Proceedings of the 17th International Semantic Web Conference (ISWC 2018). OA at

https://w3id.org/spar/article/spar-iswc2018/ – This paper is the canonical reference for the SPAR (Semantic Publishing and Referencing) Ontologies that we have developed over the last eight years, that permit the entire field of scholarly publishing and referencing to be described in RDF, as part of the Web of Linked Open Data.

2. Silvio Peroni, David Shotton, Fabio Vitali (2017). One year of the OpenCitations Corpus: Releasing RDF-based scholarly citation data into the Public Domain. In Proceedings of the 16th International Semantic Web Conference (ISWC 2017): 184-192. DOI: https://doi.org/10.1007/978-3-319-68204-4_19, OA at https://w3id.org/people/essepuntato/papers/oc-iswc2017.html – A progress report on the new instantiation of the OpenCitations Corpus, set up at the University of Bologna in 2016, which now contains >12.8 million citation records to >6.5 million cited resources recorded in RDF.

3. Silvio Peroni, Tanya Grey, Alexander Dutton, David Shotton (2015). Setting our bibliographic references free: towards open citation data. In Journal of Documentation, 71 (2): 253-277. DOI: https://doi.org/10.1108/JD-12-2013-0166, OA at http://speroni.web.cs.unibo.it/publications/peroni-2015-setting-bibliographic-references.pdf – A paper outlining the need for open citations as a fundamental requirement for open scholarship, and reporting the development in 2010 of the prototype OpenCitations Corpus at the University of Oxford.

4. Silvio Peroni, (2014). Semantic Web Technologies and Legal Scholarly Publishing. Cham, Switzerland: Springer. ISBN: 978-3-319-04776-8. DOI: https://doi.org/10.1007/978-3-319-04777-5 – A book describing the digital publishing revolution, explaining how markup methods and semantic web technologies can be used to develop semantic publishing for a particular scholarly domain (the legal domain), and presenting the first comprehensive description of the SPAR (Semantic Publishing and Referencing) ontologies.

5. Silvio Peroni, David Shotton (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Journal of Web Semantics 17: 33-43. DOI: https://doi.org/10.1016/j.websem.2012.08.001, OA at http://speroni.web.cs.unibo.it/publications/peroni-2012-fabio-cito-ontologies.pdf – A paper in the leading journal of the Semantic Web field describing FaBiO, the FRBR-aligned Bibliographic Ontology, and CiTO, the Citation Typing Ontology, the two most important of the SPAR ontologies, and giving examples of their applications for the semantic description of scholarly articles.

Principles of open research
Briefly outline how you have embraced and adopted the principles of open research during your career to date

I am directly involved in promoting open research and scholarship, particularly by providing access to the world's bibliographic metadata as Linked Open Data. I espouse the FAIR Data Principles that scholarly data should be **findable**, **accessible**, **interoperable**, and **reusable**. In addition, I hold to the position put forward by the Initiative for Open Citations (https://i4oc.org), of which I am a founder member, that bibliographic citations should be **structured**, **separate**, and **open**.

The main datasets made available by OpenCitations (http://opencitations.net), a scholarly infrastructure organization that I direct with David Shotton, have been released following strictly the aforementioned principles. All these data have been published under a CC0 waiver so as to enable their unrestricted reuse any purpose.

In addition, all the software (e.g. LODE, https://github.com/essepuntato/LODE), ontological models (e.g. the SPAR Ontologies, http://www.sparontologies.net), documents (e.g. the OpenCitations Data Model, https://doi.org/10.6084/m9.figshare.3443876), and other material produced during my

research activities have been made available in open forms under appropriate licenses – e.g. the ISC license for software and CC-BY for other kinds of entities – so as to maximise their reuse.

4. Team members and collaborators

Will you require any team members or key collaborators for this proposal?	Yes

Please list your team members or key collaborators (name and organisation) and provide a very brief outline of their role in the proposed research.
(Alphabetically) Vincent Larivière (École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal): Vincent's research aims at investigating new types of knowledge production and their use, as well as how new means of distribution enhance access to knowledge and speeds its dissemination. He holds the Canada Research Chair on the transformations of scholarly communication, is scientific director of the Érudit journal platform, and associate scientific director of the Observatoire des Sciences et des Technologies.Vincent is an expert on scientometric data sources; he will provide expert advice for this project and independent evaluation of the Open Biomedical Citations in Context Corpus. David Shotton (Oxford e-Research Centre, University of Oxford, Oxford, UK, Director of OpenCitations): OpenCitations is a scholarly infrastructure organization dedicated to providing bibliographic citation information as Linked Open Data, and engaged in advocacy for semantic publishing and open citations. It will provide both semantic web expertise and data hosting for this project. Its principal service, the OpenCitations Corpus, is already populated with citation data ingested from the Open Access Subset of PubMed Central, but currently it does not contain information about in-text references and their textual context. David will provide expert advice for the uses of existing OpenCitations resources and for the development of appropriate ontological definitions to describe citation contexts, and he will undertake user testing of the project outputs. Ludo Waltman (Centre for Science and Technology Studies (CWTS), Leiden University, Netherlands): CWTS is one of the world's leading bibliometrics research organizations, at which Ludo holds the Professorship of Quantitative Science Studies. Ludo will supply bibliometrics and context analysis expertise for the textual contexts of in-text references, and will adapt his VOSviewer software tool for bibliometric visualization to work with the data within the proposed Open Biomedical Citations in Context Corpus.

I confirm that the team members or key collaborators named above have agreed to be involved, as described, in the proposed research and are willing for their details to be included as part of this application.
Confirmed

5. Transparent decision making

Are you happy for us to share these details of your application on the Wellcome website?	Yes

6. Proposal summary

Provide an outline of what your successfully completed Open Research Fund activity will look like and what you will have achieved.
The Open Biomedical Citations in Context Corpus will offer an open database of citations between biomedical publications in which data are provided at the level of individual in-text references. Conventional citation indexes (Web of Science, Scopus) require subscription and cover only sources that meet criteria determined by the database producer. Moreover, they provide data only on the references in reference lists, not on individual in-text references. Building on the OpenCitations Corpus (http://opencitations.net/corpus), we will create a corpus of biomedical citations that provides 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. This will free biomedical researchers from the limitations of existing databases and from the traditional restrictive methods of literature searching and evaluation. We will also create APIs and visual interfaces that enable researchers to explore the Open Biomedical Citations in Context Corpus. Third parties will be able to develop additional search and browse interfaces and analytical services based on the Open Biomedical Citations in Context Corpus, assisting researchers in finding relevant information within an ever-expanding literature, and thus enabling them to carry out their work more efficiently and effectively.

7. Details of proposal

Provide details of your Open Research Fund proposal, including: (i) the vision for your proposal, including aims, target audiences, activities; (ii) how your proposal will influence open research practices in your field or more broadly; (iii) how you will monitor and evaluate your proposal, including success indicators.
Vision of the proposal Citations are primary scholarly data that provide both provenance and an explanation for how we know facts, and are an important vehicle for the discovery, dissemination, and evaluation of scholarly knowledge. However, citation data are not usually freely available to access, they are often subject to inconsistent, hard-to-parse licenses, and they are frequently not machine-readable. Furthermore, current citation indexes contain no information about the number of times a particular work is referenced in the citing work, nor about the structural and semantic context of such references, i.e. from what section(s) of the article they are made, and the textual contexts of these in-text references. The aim of the Open Biomedical Citations in Context Corpus is to create a new open corpus that contains individual in-text references in the biomedical literature. Having data at the level of individual in-text references offers many new opportunities. It will, for instance, make it possible to distinguish between cited works that are referenced just once in a citing publication and those that are referenced multiple times. In addition, it will be possible to see which in-text references occur together (e.g. in the same sentence or the same paragraph), to determine in which specific section of the publication these in-text references occur (e.g. Introduction, Methods, or Results), and potentially, by textual analysis of the citation contexts, to retrieve the functions of citation – i.e. the reason why an author cites another work. The target users of our project are primarily the biomedical researchers themselves, particularly those that need to study the literature in their field systematically. The following activities will be carried out: 1. We will harvest the full text and reference lists of articles within the Open Access Subset of biomedical literature hosted by Europe PubMed Central (letter of support attached as

additional information) that are available in XML by using their RESTful API.
2. We will extract the in-text references from the full text of each harvested publication, and store these in an expanded and extended version of the OpenCitations Corpus, which is an open database using Semantic Web technologies to record scholarly bibliographic and citation data in RDF.
3. We will develop a set of web interfaces to explore and query the proposed Open Biomedical Citations in Context Corpus. For this, we will adapt existing OpenCitations technologies to provide search and browse interfaces for humans, and a REST API and a SPARQL endpoint over the RDF data to permit programmatic access.
4. We will adapt the popular VOSviewer software tool for bibliometric visualization to make use of the proposed Open Biomedical Citations in Context Corpus. For example, it will be possible to visualize the strength of citations based on the number of their in-text references.

Proposal influence

Literature searching, literature reviewing, and bibliometric analyses are currently done mostly by using the data from proprietary closed databases, such as Web of Science, Scopus, and Google Scholar. The proposed Open Biomedical Citations in Context Corpus will enable the above-mentioned tasks to be carried out using an openly available citation corpus. This has several important advantages. First, the tasks can be carried out by anyone worldwide with no fee. Second, bibliometric analyses can be carried out and published in a fully reproducible manner, which is not possible using proprietary databases. Third, fully automated large-scale analyses can be carried out, which is not supported by proprietary databases because only small portions of data can be extracted from these databases at any one time, and even this requires a significant amount of manual work. Fourth, third parties can freely develop analytical services that make use of the proposed Open Biomedical Citations in Context Corpus.

Compared to existing citation indices (either proprietary or open) that work at the level of references in the reference lists of publications, the proposed Open Biomedical Citations in Context Corpus will reap important benefits by containing open information at the level of in-text references. For instance, when using citation links to search for relevant literature, it will be possible to filter by the section in the citing publication in which a reference occurs, enabling researchers to focus specifically on references related to, say, methods or empirical findings. These features will enable researchers to carry out tasks such as literature searching and systematic reviewing in a more effective and more efficient manner.

Monitor and evaluation

Monitoring and evaluation will be based on a number of indicators:
1. the fraction of the Open Access subset of Pubmed Central from which information on in-text references has been extracted and indexed;
2. the proportion of all publications indexed in PubMed for which in-text reference data have been made available in the proposed Open Biomedical Citations in Context Corpus;
3. the number of tools that make use of the proposed corpus; and
4. the number of times the proposed Open Biomedical Citations in Context Corpus is queried per month.

Since the Open Biomedical Citations in Context Corpus will build on existing working services and tools - the OpenCitations technologies and the VOSviewer software tool for bibliometric visualization - its sustainability will be high.

Additional information
You may submit up to two A4 pages of additional information (such as graphs, figures, tables and essential unpublished data).

Dr Johanna McEntyre
Head of Literature Services
EMBL-European Bioinformatics Institute
Wellcome Genome Campus
Hinxton, Cambridge CB10 1SD
United Kingdom
Tel.: +44 1223 492599
Email: mcentyre@ebi.ac.uk

Professor Silvio Peroni,
Department of Classical Philology and Italian Studies
University of Bologna,
Bologna, Italy

3 August 2018

Dear Silvio,

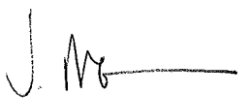
I am writing to offer my enthusiastic support of your application to the Wellcome Open Research Fund entitled "Open Biomedical Citations in Context Corpus" As you know, I have been the Team Leader for Literature Services at EMBL-EBI since 2009, and the PI for Europe PMC - the database of abstracts and full-text articles for the life sciences. As a part of these services we generate open citation networks.

The work you describe - essentially giving context to citations based on mining the full text literature - will provide a very interesting open dataset for further analysis. It could provide a basis for weighting or classifying citations - moving beyond simple counting to understanding why a given paper has been cited, opening interesting questions and opportunities for trends analysis. For example, it may provide insights into positive and negative citations, citation chains (where did the “fact” originate?) and possibly even shed light on concerns such as citation bias. As the corpus will be made open, this will maximise the opportunity for any researcher to analysis or further compute on the data, testing their own ideas. Therefore I think this could be a most interesting and useful resource for the future.

It is probably obvious to you from this letter of support that I look forward to investigating the corpus myself! I am therefore most willing to provide support to the team to use Europe PMC services to best effect in building it.

I wish you all the best with your application.

Best regards



Jo McEntyre PhD
Head of Literature Services, EMBL-EBI
Europe PubMed Central: http://europepmc.org

8. Outputs management and sharing

Will the proposed research generate outputs of data, software, materials or intellectual property that hold significant value as a resource for the wider research community?	Yes

Which approach do you intend to use to maximise the impact of your significant research outputs to improve health and benefit the wider research community?
Make research outputs available for access and re-use

Please provide an outputs management plan. Ensure this describes any significant data, software, materials or intellectual property outputs, their management, and resources required (refer to guidance).

1) For significant data, software and materials outputs

(i) What significant outputs will your research generate?

The project will generate:

1. A large RDF dataset containing information about in-text references and citations, compliant with specific Semantic Web models such as the OpenCitations Data Model (http://opencitations.net/model) and the SPAR Ontologies (https://www.sparontologies.net).
2. A series of software applications for creating the aforementioned data.
3. A series of software interfaces for human browsing, searching, querying, and visualizing the data.
4. Both SPARQL and REST API programmatic interfaces to the data.

(ii) When do you intend to share these outputs?

We will release all these data as soon as they are produced.

(iii) Where will you make these outputs available?

Dumps of the whole dataset will be made available on Figshare in CC0 at appropriate intervals, and all the resources of the Open Biomedical Citations in Context Corpus will also be accessible on the Web by using the OpenCitations infrastructure. All the software will be made available on GitHub and released with Open Source licenses - e.g. ISC.

(iv) How will they be discovered and accessed by others?

OpenCitations will make available several mechanisms to access all the resources in the dataset, in particular a REST API, a search application, and a SPARQL endpoint. In addition, all the resources will also be discoverable by means of their URLs, compliantly with the principles of Linked Open Data and the FAIR guidelines. In addition, the third-party platforms that will be used for the data (i.e. Figshare) and the software (i.e. GitHub) already implement mechanisms that facilitate the access and discovery of such items.

(v) Are limits on sharing required?

No.

(vi) How will these outputs be preserved?

All the outcomes produced are build on existing working services (OpenCitations) and tools (VOSviewer), and for this reason their sustainability will be high. In addition, the availability of these outcomes on Figshare and GitHub will guarantee long-term preservation, even in the extreme case that our own databases and services should be closed.

2) For intellectual property outputs

(i) What IP will your research generate?

n/a: all the metadata we generate will be placed in the public domain under a CC0 waiver, and the software will be made available under an open license.

(ii) How will you protect this IP?

n/a

(iii) How will the IP be used to achieve health benefits?

Open information improves knowledge dissemination. This will accelerate biomedical research, which, we trust, will result in health benefits.

(iv) Provide the name and contact details for the person in your organisation (e.g. Technology Transfer Officer or Business Development executive) who can act as a point of contact for Wellcome in connection with the protection and commercialisation of this IP.

n/a

3) Describe any resources that you will need to deliver your outputs management plan

We do not need any particular resource for the delivery of the outputs of the project, except the ones already mentioned, i.e. the OpenCitations infrastructure, VOSviewer, Figshare and GitHub.

9. Costs requested

Currency requested Select the currency in which you wish to apply.
EUR - Euro

Salaries Are you requesting salaries?	Yes

Salaries

Description	**Total (€)**
Used to appoint a Research Fellow (either with a MSc or ideally with a Ph.D.) for one year	33,000

Materials and consumables	Yes

Are you requesting materials and consumables?	

Materials and consumables

Description	Total (€)
Supplies and material, papers and ink for printing, and other stationery objects and writing materials	500

Equipment Are you requesting equipment?	Yes

Equipment

Description	Total (€)
Laptop and accompanying equipment (an external monitor, adapters, etc.) for the use of the Research Fellow	3,000

Miscellaneous costs Are you requesting miscellaneous costs?	Yes

Miscellaneous costs

Description	Total (€)
Travels, accomodations, meeting, workshop	19,000

Justification for costs requested
Provide a high-level budget breakdown and justification for costs requested.

The costs requested are organised as follows:

Salaries

The full amount specified will be used to appoint a Research Fellow (either with a MSc or ideally with a Ph.D.) for one year by the Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy. The person will have the key developmental role for this project under the supervision and guidance of the lead applicant, developing all the necessary software for extracting the citation data and their contexts, and storing them into the expanded OpenCitations Corpus, as well as developing appropriate user interfaces for querying and browsing these data. He/she will be based in Bologna, but will spend time with Joanna McEntyre at the EBI at Hinxton, learning about the bibliographic reference and full-text data available from Europe PubMed Central to be consumed by this project, and with team member Ludo Waltman adapting VOSviewer to work with the new RDF citation context data created.

Materials and consumables

The small amount specified will be used for buying supplies and materials such as posters for advertising the project in events and conferences, papers and ink for printing, and other stationery objects and writing materials.

Equipment

The requested money will be used to buy a new laptop and accompanying equipment (an external monitor, adapters, etc.) for the use of the Research Fellow, so as to allow him/her to work on the project with the appropriate tools.

Miscellaneous costs

The amount requested is necessary to cover the following expenses:

1. 500 EUR for travel expenses of the appointed Research Fellow to go from Bologna to Cambridge and return, and additional 2,100 EUR to allow him/her to stay in Hinxton for 2 weeks (~150 EUR per night) to work with Joanna McEntyre of the European Bioinformatics Institute on the Europe PubMed Central dataset.
2. 500 EUR for travel expenses of the appointed Research Fellow to go from Bologna to Leiden and return, and additional 4,200 EUR to allow him/her to stay in Leiden for 4 weeks (~150 EUR per night) to work with Ludo Waltman at the Centre for Science and Technologies (CWTS) of the Leiden University on VOSviewer.
3. 2,500 EUR for travel expenses and other 1,000 EUR for accommodation expenses of the team members (Vincent Lariviere, David Shotton, and Ludo Waltman) to attend a 2 day kick-off meeting in Bologna with the lead applicant and the appointed Research Fellow.
4. 4,000 EUR to organise a final workshop and an hack day in Bologna to disseminate the results and outcomes of the project to a large audience. This amount can be used for covering some travel and accommodation expenses of the team members and to cover other needs (food, beverage, advertising material, etc.).
5. 4,200 EUR to cover the expenses (travel, accommodation, food, beverage) to allow the lead applicant and the Research Fellow to participate in at least two European conferences or one major international conference, to present and share the results of the project.

Total cost

The final cost of the project will be 55,500 EUR (~49,450 GBP as of 3 August 2018).

Summary of financial support requested	
	Total (€)
Salaries / Stipends	33,000
Materials and consumables	500
Equipment	3,000
Miscellaneous other	19,000
Total	**55,500**