1. Open Research Fund application

Reference number	UNS75541
Applicant name	Dr Heather Piwowar
Title of application	Open, complete, disambiguated database of authorship metadata in biomedicine
Total amount requested	£50,000.00

2. Application summary

Application title
Open, complete, disambiguated database of authorship metadata in biomedicine

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

Proposed start date	01/11/2018

Is your application being submitted through a university?	No

Name of administering organisation
Impactstory

Lead applicant's address at administering organisation	
Department/Division	
Organisation	Impactstory
Street	c/o The Hive, 128 West Hastings St, Suite 210
City/Town	Vancouver
Postcode/Zipcode	V6B 1G8
Country	Canada

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 Heather Piwowar
Department	
Division	
Organisation	Impactstory
Address Line 1	
City/Town	
Postcode	
Country	
Telephone No.	7788484724
Email Address	team@impactstory.org

ORCID iD	
ORCID iD	0000-0003-1613-5981

Career history (current/most recent first)

From	To	Position	Organisation
05/2013	01/2050	co-founder	Impactstory
06/2010	05/2013	postdoc	Duke University
09/1996	09/2005	Software Developer	Various technology companies

Education/training

From	To	Qualification	Subject	Organisation
08/2005	05/2010	Doctor of Philosophy (PhD;DPhil)	Biomedical Informatics	University of Pittsburgh
05/1995	05/1996	Master of Science (MSc)	Electrical Engineering and Computer Science	Massachusetts Institute of Technology
08/1991	05/1995	Bachelor of Science (BSc)	Electrical Engineering and Computer Science	Massachusetts Institute of Technology

Source(s) of personal salary support
Impactstory

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.

Dr Heather Piwowar is a cofounder of Impactstory, the nonprofit company behind the Unpaywall database for Open Access discovery. Unpaywall's collection of 20 million Open Access links is used by Europe PMC, Web of Science, Scopus, the British Library, and thousands of libraries worldwide. The free browser extension is used by more than 150,000 people, and the API gets a million calls a day. A longtime advocate for Open Science, Dr Piwowar is also a leading researcher in research data availability and reuse, including a seminal paper measuring the citation benefit of publicly available research data.

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.)*

Priem and Piwowar. (2018). The Unpaywall Dataset. figshare. 10.6084/m9.figshare.6020078 and http://unpaywall.org/products/api (more than 1 million API calls per day; integration into Web of Science, Scopus, Europe PMC; 150,000 browser extension users)

Piwowar, Priem, Larivière, Alperin, Matthias, Norlander, Farley, West, Haustein. (2018) The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* 6:e4375 (43 citations)

Piwowar and Priem. (2016) Impactstory Profiles, based on ORCID http://blog.impactstory.org/new-better-freer/ (15,000 signed up users)

Priem, Piwowar, Hemminger (2012) Altmetrics in the wild: Using social media to explore scholarly impact. Preprint. *arXiv*:1203.4745 (301 citations)

Piwowar, Day, Fridsma (2007) Sharing detailed research data is associated with increased citation rate. PLOS ONE 2(3): e308. (603 citations)

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

My research and software development work has been on the topic of open science, and has also been disseminated through open science practices:

- all of my code is immediately open source with an MIT license: https://github.com/hpiwowar

and https://github.com/Impactstory

- all of my peer-reviewed papers are open access (usually through open access journals but through preprints otherwise)
- all of my data and statistical scripts are open (originally on my personal website, then for research done in the last 5 years through Dryad or Zenodo) and have used knitr to improve reproducibility.
- I blogged about open science issues for years: https://researchremix.wordpress.com/
- I've given many presentations about open science topics: https://www.slideshare.net/hpiwowar/presentations?order=popular
- I have an active twitter account with almost 7k followers, where I primary talk about open science topics: https://twitter.com/researchremix
- I sign my peer reviews and encourage them to be made open to the public
- I signed and adhere to the Cost of Knowledge boycott, and only review for open access journals

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.
Jason Priem (Impactstory) is a key collaborator for this project. As with all projects taken on by Impactstory, he will co-lead the project and share equally in all decision making and implementation.

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 aim of this proposal is to create an open, complete, disambiguated database of authorship metadata in biomedicine. We believe this data source will be rapidly integrated into open science toolchains, facilitating innovations not otherwise possible. Right now, all authoritative sources of author metadata are proprietary (Google Scholar, Scopus, Web of Science), not readily available for commercial use (Author-ity, Microsoft Academic Graph),

or too incomplete to be usable (ORCID).

ORCID will solve the author disambiguation problem eventually, but it will take many years for all funders and journals to require ORCIDs for all authors. Even then, the back catalogue of papers will not be disambiguated.

The current proposal addresses this problem by disambiguating author information in Pubmed and making this data wide open for integration and reuse through an open, fast API with high rate limits, an annual open dataset release, open source code, and a Data Feed contract available for purchase to ensure sustainability.

With an Open source of disambiguated author information, developers will have the missing link to create an Open Google Scholar, an Open SciVal, an Open ResearchGate. Librarians can identify publications belonging to their researchers. Bibliometricians can analyze changes in openness behaviour, providing an evidence base for future open research policies, using open, auditable data.

An open source of author metadata removes the need for expensive database subscriptions, which allows a broader base of participation (globally, and outside academic institutions), saves money, and produces a more nimble, active, competitive set of alternatives for scholarly communication.

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.

Aims

The aim of this proposal is to create an open, complete, disambiguated database of authorship metadata in biomedicine. We believe this data source will be rapidly integrated into open science toolchains, facilitating innovations not otherwise possible.

Target audiences

The target audience of this proposal is software developers in scholarly communication (both nonprofit and commercial), librarians, bibliometricians, and ultimately biomedical researchers themselves.

Current ORCID approaches will solve the author disambiguation problem eventually, but it will take many years for all funders/journals to require ORCIDs for all authors. Even then, the back catalogue of papers will not be disambiguated.

The current proposal addresses this problem by disambiguating all authors of papers in biomedicine and making this data wide open for integration and reuse.

Activities

We propose to disambiguate the authors of articles in PubMed, keep this data up-to-date as new works are published, and make this data openly available.

This task has been completed before, but the results have not been kept up-to-date or made open

for commercial use.

Our approach will be to use methods from the literature that have proven successful for author disambiguation (work by Torvik, Smalheiser, Liu et al, Lerchenmueller, Sorenson), leveraging metadata available in Crossref, PubMed, and ORCID.

We will assign an author cluster (named by an ORCID whenever possible) to each author position for each paper in PubMed. This means that for a given author we have a list of everything they've published that has been indexed by PubMed.

Institutional information is a useful attribute in author disambiguation. As part of the disambiguation process we will model employment history (institutional affiliations, with year ranges) for each cluster. This information will be part of the released dataset.

We will have a mechanism that makes it easy for authors or their delegates to correct mistakes made by the automated algorithms.

Influence of this work on open research practices

We expect the long term influence of this work on open research practices to be profound: we believe it will unlock a new wave of open research tools.

Right now, all authoritative sources of author metadata are proprietary (Google Scholar, Scopus, Web of Science), not readily available for commercial use (Author-ity, Microsoft Academic Graph), or too incomplete to be usable (ORCID). With an Open source of disambiguated author information, developers will have the missing link to create an Open Google Scholar, an Open SciVal, an Open ResearchGate.

In addition, this dataset will enable the evaluation of institutional Open Access policies using Open tools. Bibliometricians will be able to track the changes in openness behaviour, providing an evidence base for future open research policies, using open source and auditable data sources. Removing the need for an expensive subscription to Scopus or Web of Science allows a broader base of participation (globally, and outside academic institutions), saves money, and produces a more nimble, active, competitive set of alternatives.

This will complement the work being done in Open Citations (i40c) and Open Altmetrics (Crossref Event Data), leverage the standards that have emerged for Open identifiers for people (ORCID), institutions (GRID), and funders (Crossref Funder Registry), and build on publicly available metadata for publications (PubMed and Crossref).

This proposal is limited to biomedicine: the rich Pubmed metadata will help us get started. If the work in this grant is successful, in the future we will expand the author disambiguation service beyond Pubmed to all authors of all academic works.

Monitoring and Evaluation

The proposed work will be considered successful if it achieves high adoption. Within six months of release, we expect the community to have initiated multiple integrations, including the following:

- a "search and link" ORCID importer to allow authors to populate their ORCID profiles using this data
- software applications rolling out a "type in your name and we will pull in your publications" wizard

- a Europe PMC Open Author Profiles integration
- an R wrapper for analysts and bibliometricians
- several published analyses combining this information with other data sources, such as Unpaywall data to study Open Access adoption by region, institution, career stage, and other variables
- use by libraries to further populate their institutional repositories
- integration with wikidata
- multiple blog posts or preprints with proposed algorithm improvements (which we will then integrate into our code whenever possible!)

We will be monitoring to see if these integrations and uses take place. If they don't, there are two main possibilities:

1. The dissemination methods are too inconvenient
2. The resulting author metadata is not high enough quality

We should be able to determine which of these issues is the problem by conversations with potential integrators. If the former, we will refine our dissemination methods to better meet the needs of the community (different API endpoints, bulk releases of subsets of the data, ResourceSync, etc). If the data is not high enough quality, we will first understand what kind of errors are causing problems in integrations and then focus on a subset of the data, work on its quality until it is sufficient, and expand from there.

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

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).

Impactstory has a longstanding commitment to open source code and open access publication, and we will continue to honor this during the duration of this grant. We will make all software developed as part of this grant available as Open Source under an MIT License, hosted on GitHub. We will keep this code freely available on the internet as long as Impactstory exists. We don't expect to write any academic papers as a result of this grant, but if we do we will make them Open Access using a preprint archive like arXiv or Zenodo, under a CC-BY license.

The dataset developed through this grant will be made Openly available as a bulk download annually, and freely available via an open API with high access limits (eg 100,000 calls/day). The data will be licensed as CC0.

To ensure sustainability of the service developed in this grant and its ongoing open data dissemination paths, we will also make a Data Feed contract available for a fee. The Data Feed product will provide access to weekly update data in formats designed to easily allow customers to keep a full copy of the dataset up to date locally. The Data Feed product has proven an attractive option for enterprise clients using the Unpaywall dataset (paying customers include large organizations, startups, and a few libraries), and as such is a proven model for sustainability.

It is important to us the Data Feed fee does not prove a significant barrier to the data adoption; we stay in touch with the community to make sure this is true, and adjust as necessary.

9. Costs requested

Currency requested Select the currency in which you wish to apply.
GBP - Pound Sterling

Salaries Are you requesting salaries?	Yes

Salaries

Description	Total (£)
Heather Piwowar salary (10%)	10,500
Jason Priem salary (10%)	10,500
Software developer salary (25%)	25,000

Materials and consumables Are you requesting materials and consumables?	No

Equipment Are you requesting equipment?	No

Miscellaneous costs Are you requesting miscellaneous costs?	Yes

Miscellaneous costs

Description	Total (£)
Processing power for author clustering (cloud)	4,000

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

- Staff
  - Part of the salaries and fringe benefits (£105,000 @ 10% for the two Impactstory employees who will work on this project, Jason Priem and Heather Piwowar). We will use the same approach we have successfully employed on other grants, dividing our time between:
    - project management and administration
    - user interface design, testing, and development
    - software architecture, design, programming, and maintenance
    - outreach and marketing
  - In addition we will hire a software developer for 3 months to do author clustering.
- Miscellaneous costs include cloud-hosted processor time for running author clustering algorithms, estimated roughly at 2 months of a high-performance EC2 instance on AWS
- The costs for hosting the database, running the API, and continuous updating of the author clusters will be covered by existing Impactstory funds and future earned income from Data Feed contract agreements.

Summary of financial support requested

	Total (£)
Salaries / Stipends	46,000
Materials and consumables	0
Equipment	0
Miscellaneous other	4,000
Total	**50,000**