1. Open Research Fund application

Reference number	214478/Z/18/Z
Applicant name	Dr Thomas Krichel
Title of application	"bims: Biomed news" from machine learning to expertise sharing
Total amount requested	£22,800.00

2. Application summary

Application title
"bims: Biomed news" from machine learning to expertise sharing

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

Proposed start date	01/02/2019

Is your application being submitted through a university?	No

Name of administering organisation
Open Library Society, Inc.

Lead applicant's address at administering organisation	
Department/Division	
Organisation	
Street	34-20 78th Street #3D
City/Town	Jackson Heights, NY
Postcode/Zipcode	11372-2572
Country	United States

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

3. Lead applicant

Lead applicant details	
Full Name	Dr Thomas Krichel
Department	
Division	
Organisation	Open Library Society
Address Line 1	34-20 78th Street #3D
City/Town	Jackson Heights, NY
Postcode	11372-2572
Country	United States
Telephone No.	7(8)9137488056
Email Address	krichel@openlib.org

ORCID iD	
ORCID iD	0000-0002-8421-6356

Career history (current/most recent first)

From	To	Position	Organisation
01/2001	08/2012	Assistant then Associate Professor	Long Island University
02/1993	04/2001	Lecturer	University of Surrey

Education/training

From	To	Qualification	Subject	Organisation
03/1995	12/1999	MB/PhD	Economics	University of Surrey
10/1989	09/1990	Master of Arts (MA)	European Studies	University of Exeter
10/1986	06/1989	Magistère d'Économie	Economics	Université Paris 1 Panthéon-Sorbonne, l'École Normale Supérieure de la rue d'Ulm and EHESS
10/1984	06/1986	DEUG es science économiques	Economics	Université de Toulouse

Source(s) of personal salary support
I own some real estate and provide some consulting services.

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?	Yes

Please provide details
I stopped employed work in 2012. Since then I have been living on investment income and consultancy. All my consultancy has been related to my main accomplishment in life, which is the creation of the RePEc digital library.

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.

I am a veteran open-access activist. I have a track record of turning modest financial contributions into sustained services. In April 1993, when I was a lecturer in economics, I published the first online academic paper in economics. This was the start of my WoPEc collection of economics working papers. Between 1997 and 1999, WoPEc was supported by the Joint Information Systems Committee of the UK Higher Education Funding Councils (JISC). In 1997, I morphed WoPEc into the RePEc digital library. I continue to coordinate RePEc.

In the remainder, I survey twenty years of work leading up to this application.

In 1998 I created “NEP: New Economics Papers” at http://nep.repec.org. It aims to circulate information about new working papers to subject-specific audiences on a weekly basis. Its first issue had 24 papers. It was emailed to selectors. I instructed them to remove the non-relevant papers from the email and pass it on to subscribers. In 2002, I became concerned that the continued growth of RePEc would imply that an unaided selection of papers would become too heavy a workload. In 2004, I used left-over JISC funds to hire Roman D. Shapiro. He wrote the first version of “ernad”. At its inception, ernad was a specialised software suite for the production of NEP. It aids the selectors through machine learning. These days, selectors work through about 700 new papers every week. The median selector spends ten minutes on the job. This shows that the machine learning built into ernad is quite effective, but human input is still required. Since 2013, I enhanced ernad into a software that can handle services based on various collections. Each service requires a specific set of XSLT style sheets. These style sheets are called by common procedural code. I also refactored the machine learning code so that it is capable of supporting the near simultaneous release of a large number of report issues. Still, ernad has no facilities for opening reports. Work to be done, if this application is funded, would build a report-creation system.

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.)*

"Open Citation Content Data" by Mikhail Kogalovsky, Thomas Krichel, Victor Lyapunov, Oxana Medvedeva, Sergey Parinov and Varvara Sergeeva. To be presented at the Special Track on Open Access Repositories, Research Information Systems and Data Infrastructures of the 12th Metadata and Semantics Research Conference, 23-26 October 2018, Limassol, Cyprus, available at http://openlib.org/home/krichel/papers/limassol.pdf

Contribution: I did the data summaries and statistics.

"Towards Open Data for the Citation Content Analysis", by Jose Manuel Barrueco, Thomas Krichel, Sergey Parinov, Victor Lyapunov, Oxana Medvedeva, Varvara Sergeeva, accepted at the Data Analytics and Management in Data Intensive Domains conference in Moscow, Russia 10-13 October 2017, available at http://openlib.org/home/krichel/papers/vorobyovy.pdf

Contribution: I calculated the results and wrote half of the paper."Developing a predictive model of editor selectivity in a current awareness service of a large digital library" by Thomas Krichel and Nisa Bakkalbasi. Library & Information Science Research 2005, volume 27 pages 440–452, 2005, available at http://openlib.org/home/krichel/papers/boston.pdf

Contribution: I collected the data, defined variables and wrote the paper.

"Does Precommitment Raise Growth? The Dynamics of Growth and Fiscal Policy" by Thomas Krichel and Paul Levine,
The Scandinavian Journal of Economics, 2001, volume 103, number 2, pages 295–316, available at http://openlib.org/home/krichel/papers/grusas.pdf

Contribution: This summarises work that Paul and I did for years, so it's really hard to say who did what.

"Personal data in a large digital library" by José Manuel Barrueco Cruz, Markus J.R. Klink and Thomas Krichel, presented at the European Conference on Digital Libraries in Lisbon, Portugal 18-20 September 2000, available at http://openlib.org/home/krichel/phoenix.a4.pdf.

Contribution: I wrote the entire paper. This paper is not in the conference proceeding as I refused to hand over copyright to Springer on their terms.

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

I am a pioneer of open digital libraries. As a trained economist, I see achieving sustainability as the main problem of digital libraries. Therefore I have always tried to build interoperable systems that avoid centralisation. They build scale by freely exchanging data. This is best illustrated in RePEc. There are now over 2000 archives that collectively maintain 2600000 records in the core bibliographic dataset. We have user services that visualise this data. They contribute to a common logging service, LogEc. We have a CitEc citation data service. We have a database of economics research institutions. Authors register their publications with the RePEc Author Service. We can then aggregate citations and other usage data by authors. There is more ... but it all works because our data is open and being constantly reused. It is a system where the components operate independently but there is a strong positive feedback between them. There is no way that lack of funding will shut down RePEc. It will never be bought by Elsevier.

This application brings some of this spirit to the biomedical arena, again by building an aggregate service that will, over time, rely on the self-interested contribution of a cast of thousands.

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.
Gavin McStay, Ph.D., Senior Lecturer, Department of Biological Sciences, Staffordshire University, Stoke-on-Trent, United Kingdom Gavin is an active researcher in the field of mitochondria and cell death. He met Thomas Krichel at a social function in New York City. Initially, Gavin saw Biomed news as a way to improve the discoverability of new research articles without the repetitive nature of keyword searching and use of multiple sources. Thomas opened the "bims-cytox1" report for him. The first few issues (approximately 4-6 weeks) of the report required more searching through the ranked abstract list. However, after these initial issues the ranked list became a useful source of abstracts with most of the relevant abstracts in the top 50. This list also incorporated abstracts that were not on the topic, but related to the topic and of interest to him. Thus Gavin got a new source of relevant articles that would not have been discovered using keyword searches. Over time, Biomed news has freed Gavin from the need to carry out multiple searches and use different systems to keep up to date with abstracts to find those of interest. After using the system for over a year now, Gavin sees Biomed news as a way for individuals and organisations all around the world to be part of a community where scholarly biomedical research abstracts can be shared globally in an easily adoptable form. Gavin is now the director of Biomed news. Gavin has been working closely with the small number of selectors that have been recruited to iron out technical issues. The selectors have provided very positive feedback so far, mostly relating to the ranking of the abstracts and their relevance to the topic and the short amount of time required to process weekly report issues. Gavin directed Thomas to fine-tune the system based on selectors' feedback. This has made the Biomed news platform easier for selectors to use, an important factor for when we release more publicly. Gavin will advise the administrative assistant during the recruitment efforts of potential selectors and submission of advertising to biomedical research journals. In addition, he will be attending biomedical research meetings to increase visibility of Biomed news to potential selectors and subscribers. The complementary expertise that Gavin and Thomas bring in are a strength of this application. They are able to go back and forth as an end-user and developer, respectively. This interaction has been, and will continue to be, at the core of Biomed news to provide a service that is user friendly for the intended audience, based on user feedback.

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.

Progress in science depends on researchers being aware of new research results. Top researchers can rely on insider networks or colleagues. In the absence of insider networks, other researchers have to search PubMed. PubMed is a great product for discovery. But current awareness within a subject area is not well served. Repetitive searching is cumbersome and PubMed profiles are imprecise.

Open research communication can help. It can provide a system where expert selectors monitor the appearance of new papers on their topic of interest. The topic of the report is given by the selector and it stays constant. Machine learning can then be used to make the periodic selection tasks really efficient. The results are available to anybody who is interested in the topic, whether they are other researchers or the general public.

We see three benefits:

(1) it will be a step forward to democratise access to knowledge. Even today, most people can get access to papers even when they are behind a paywall, just by writing to the authors. But many people do not know what papers to get. Here is where an expertise sharing system can help.

(2) it will be a step forward for the emerging preprint efforts in the biomedical arena. Preprints have the disadvantage that they are not classified by topic specific journals. Our system will help them to find an audience.

(3) it can help in the fight against fake science. It is harder to dupe an expert.

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.

You may think that a system as outlined in the summary is a pipe dream. But it exists. Thomas Krichel created it 20 years ago, in an area few people are aware of: academic economics. It's called "NEP: New Economics Papers". It is for working papers only. In conjunction with other parts of RePEc it has done miracles to lift the working paper culture in economics to new heights.

To build a similar system for the biomedical sciences, we need a bibliographic database to watch. PubMed gives us a head start. Next we need software that produces all new papers from PubMed every week. Thomas built one starting in 2014. At this point, it runs like a clockwork. Then comes the most complicated part. It is to build an interface that allows selectors to build weekly report issues. Thomas started on this in 2015. He developed the ernad software originally written for NEP. Now, ernad can examine the 22573, on average, papers that are new to PubMed every week. Ernad uses machine learning to provide the selectors with the most relevant papers in a ranked list.

In early 2017, Thomas found the first selector, Dr Gavin McStay. He now directs the project. It is

called "bims: Biomed news" at http://biomed.news. In early 2018, we started to recruit selectors. We have about ten. With this few selectors, we can really only say that we have a service prototype. The good news is that selectors like the system. They appreciate that the system is less time-consuming than PubMed searches and more precise than PubMed profiles. The median selector spends about ten minutes on the weekly task of composing a report issue. The not so good news is that we have found it more difficult to convey the open science nature of the project. We do not hide the open science nature, but it is not obvious. The output data is already available by rsync, a protocol not wholly familiar to laboratory-based researchers. Selectors will be encouraged to disseminate information about their reports through personal web pages, social media links and via other scholarly communication outlets. The selectors' participation in Biomed news is intended to be part of the larger community movement to support open access to scientific literature.

The objective of the funding application is (1) to turn the prototype to a service, and (2) to morph it from a machine learning tool to an expertise sharing service.

To elevate the project from prototype to actual service, we need to expand the number and diversity of selectors. Currently our selectors cover 0.1% of PubMed every week. Gavin thinks that it is best to have highly selective reports. The more selective reports are, the more reports we have to create to cover all of PubMed. For example, if a report brings out seven papers a week, we need at least 3000 reports, but probably many more as there will be overlaps.

To recruit selectors, we want to hire an assistant who will look for email addresses of researchers to be potential selectors. A list of grantees provided by the Wellcome Trust would be a good starting point. We will focus on laboratory heads, post-doctoral researchers and experienced laboratory scientists at reputable biomedical research institutions all over the world, especially those who have demonstrated support for the open science movement. We will contact them via email and invite them to become a selector. We believe these individuals would show dedication to the service and spread visibility of the project. The support of the Wellcome Trust would increase credibility of cold-call email communications. Gavin will travel to specific international biomedical research conferences to announce and describe the service providing opportunity to open reports on site.

In order to elevate the project to become an expertise-sharing system, we will set up email distribution of reports. We will construct homepages of reports where web visitors will be invited to subscribe. We will encourage selectors to build a subscriber base. We will also monitor selectors to ensure that they are doing the weekly editing on time.

Our success indicators will initially be to recruit selectors of diverse topics to raise coverage. The number of recruited selectors from different research areas will increase the coverage of Biomed news selections. After this we will want to recruit as many readers as possible.

Critics may suggest that our project is flaky because we rely on volunteer selectors. This is only partly true. It depends on the selector. If the selector is a patient with a chronic disease, (s)he may be thought of as a volunteer. We will open reports for patients. But they are not our recruiting focus, academics are. Academics have to know the literature anyway. We give them a state-of-the-art tool. As readership of reports increases, name recognition benefits sets in. Selectors will be able to include selectorship as a service item in their CVs. With this in mind, we think of Biomed news as a rare constellation where pure self-interest, aided by sophisticated technology, maintains an information source that will be of great benefit to humanity.

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).
We run an almost complete open data operation. Almost the entire final output data of the project is already available via public rsync. We do not show things like editors' email addresses and ernad passwords. We aim to get third parties to make our output data available in their systems. Making the intermediate data that selectors' have provided is more difficult because we have to be concerned about selectors' privacy. Thus we do not plan, for example, to show that a selector placed a certain paper at the bottom of the report issue list. The data now shows that the paper is at the bottom of the list but we do not reveal that this was deliberate action. Our report creation and management tool will have a section to manage privacy settings to eventually allow us to publish such data as well. The Biomed news service belongs to the Open Library Society, Inc. It is a small US 501(3)(c) corporation founded by Thomas Krichel. The society is the legal body representing the service. Incidentally, it is also the organisation that represents RePEc. The PubMed data that is used in Biomed news has been compiled by the National Library of Medicine. Terms and conditions for the use of this data is the same as for PubMed. The ernad code, mainly Perl and XSLT, which makes the service run, belongs to Thomas Krichel. This funding application will not develop ernad. All code written under this application can be open-sourced and be published at the funder's request. The text on the biomed.news site is the property of whomever is mentioned as maintaining it. The inclusion of a paper in a particular report and the exclusion of the others is the intellectual property of the selector. The selector grants the Open Library Society a royalty-free, non-exclusive license to disseminate this information, together with the provenance information that the selector has agreed to. When we open a report, we take the provenance information that the selector gives us. This information may include the selector's name, email address, as well as affiliation data. Selectors understand that this data is disseminated as part of the report issues. Dissemination via rsync is instantaneous. If a selector wants to restrict the amount of personal data that has been disseminated, we honour that request. We can remove it from our data. “Our” data includes older report issues. But we cannot remove it from copies that third parties will have cached.

Let us turn to sustainability. It is exciting to develop a novel service. To make it sustainable we automate it as far as possible. It will be completely free for public reports. Still the service may raise money to allow for continued development. At this time we see three avenues.

First we may be able to gain sponsorship. If, for example, we have a large number of biomedical researchers who read our emailed reports, the Wellcome Trust may want to sponsor issues to advertise funding opportunities. Similarly, if we attract medical doctors, a pharmaceutical company may rent a space to praise the virtue of a drug. You may note that the current report issue format already features a space on the right of the selector information that can handle a limited amount of advertising. Of course nothing prevents other organisations from redistributing report data with their advertising. That is why it is important for us to develop our own base of subscribing readers. This is what we aim to start with this application.

Second, we can provide Biomed news on a software-as-a-service (saas) basis to persons and organisations who want to produce reports that are **not** public. We will always be free for selectors' use if the results are public. We intend to charge if they are not. Thus closed information will subsidise open information.

Third we could use ernad to support organisations to run current awareness reports on their proprietary databases, and charge for that. It is another saas revenue model. Note that we do not propose, in this application, to work on ernad itself. The software work we describe here can be open sourced without risking saas revenues from ernad.

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 (£)**
Work by Thomas Krichel for one year £1000/month	12,000
Work by assistant for 7 months at £800/month	5,600

Materials and consumables Are you requesting materials and consumables?	Yes

Materials and consumables

Description	**Total (£)**
2 years of a Hetzner EX41 server	1,200

Equipment Are you requesting equipment?	No

Miscellaneous costs	Yes

Are you requesting miscellaneous costs?	

Miscellaneous costs

Description	Total (£)
Attendance at international conference - USA	1,800
Advertising cost	1,000
Attendance at international conference - Africa	1,200

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

Thomas Krichel will do the bulk of the work under Gavin McStay's direction. He will work full-time on the project. He will do all the computer programming and all the system administration. He will set up the server and migrate Biomed news and the PubMed indexer to it. He will build the email distribution system as well as the report creation system. He will work on monitoring selectors. Thomas will ensure that the systems are kept running after funding ends.

Thomas will have two main development tasks. The first is to set up the dissemination system via email. Ideally we would use an email marketing solution like Mailchimp or Constant Contact. But in the absence of continuous funding, we feel it is best to roll out a home-grown system based an a customised installation of Mailman3. We aim to have a system where a user, with a single login can select reports from a single list of all reports. Eventually, we would like to have a system that will recommend to users new reports based on their existing subscriptions. We doubt that an outsourced system will have that, so that is one reason to roll out our own.

The second development task is to build a site where, within some limits, anybody can create a report. One limit is that we like report codes to be pronounceable to make them more memorable. At this time, we are not sure how far we can automate that. The other one is that we need to check for overlaps. Many basic checks can be automated, though probably not while the user waits. If the automated check rejects the report as a duplicate, the applicant may want a personal appeal against our software. That will have to be carried out by Gavin McStay.

Thomas will hire and train the assistant. Thomas will monitor the project assistant and report on the progress. The assistant will mainly work on the recruiting effort. (S)he will scan web sites of important departments to find suitable candidates to approach individually for becoming selectors. This will be a one-off campaign for us. Ideally, we will hire somebody with a biomedical degree. If Thomas does not find somebody with a biomedical education and sufficient English skills, we have to settle for a language graduate.

Gavin McStay provide his leadership efforts pro bono. He will travel to two international scientific meetings to increase visibility of Biomed news and recruit selectors. The two meetings (Experimental Biology in Orlando, USA - April 2019 and Africa International Biotechnology & Biomedical Conference in Mombasa, Kenya - August 2019) attract biomedical researchers from all over the world and from diverse biomedical research fields. He will submit abstracts for poster and oral presentations to describe the service and the benefits of it. At these meetings we will be able to collect details for prospective selectors, new topic reports and find readers for existing topics. An addition, we request a £1000 budget is for advertisements in professional journals. Any remaining travel funds will augment that advertising budget.

We will rent a Hetzner dedicated root server EX41. It can be paid in advance for one year only, so we will be able rent it for two years on project funds. As of 23 July 2018 it costs €46.80 a month, including VAT. We count €46.80 × 24 months = €1123.20 = £1002.4, at €1.12 to the pound. We request £1200 to accommodate for exchange rate fluctuations as well as a potential setup fee that Hetzner waived in the Summer.

We understand that the amount requested is small. There are several reasons behind this. (1) Thomas and the assistant will work from Novosibirsk, where the cost of living has fallen with the Ruble rate. (2) We understand that the value of the project is hard to communicate, even though we already have a running service. Therefore we intend to compete on costs. (3) As a small starting venture, we are more concerned about getting any funding that gets us started. As a small team, we would find it hard to quickly **and** productively spend a lot of money anyway. For us the credibility value of getting Wellcome Trust funding matters most.

Summary of financial support requested	
	Total (£)
Salaries / Stipends	17,600
Materials and consumables	1,200
Equipment	0
Miscellaneous other	4,000
Total	**22,800**