ALFRED P. SLOAN FOUNDATION

www.sloan.org | proposal guidelines

PROPOSAL COVER SHEET

Project Information

Principal Investigator

Benjamin Nickolls
<REDACTED>

Grantee Organization:	Brave New Software (Sponsor)
Amount Requested:	$ 124,770
Requested Start Date:	1st November 2016
Requested End Date:	1st January 2018
Project URL (if any)	https://libraries.io

Project Goal

To raise the quality and robustness of all software by accelerating the natural selection processes that occur in the adoption of new frameworks, plugins and tools we collectively call libraries through the adoption of a process for indexing search inspired by Google's PageRank algorithm.

Objectives

Libraries.io evaluates individual pieces of software and then federates a portion of this value to frameworks, plugins and tools upon which it depends as detailed in a project manifest following the principle of Google's PageRank algorithm. Our proposal expands the scale of data gathered by Libraries.io using this approach from its current 1.7 million projects to over 3 million. We will create a more consistent base of data gathered across the 33 software packaging and distribution tools currently supported by employing approaches such as code compilation, containerisation and static analysis. We will improve access to and reuse of this data by improving documentation, enhancing programmatic interfaces and finally through the free and open publication of dependency graph and metric data.

Proposed Activities

The principal activities to be undertaken within this proposal are those of simple, user centred software development. To continue to develop the software necessary to gather data from platforms not currently supported and to expand the breadth of data gathered across the 33 software packaging and distribution tools already supported. This will involve face to face and virtual user research, user testing, prototyping, A/B testing, analytics and community management as needed.

Expected Products

Added support for projects hosted on collaborative coding platforms GitLab and BitBucket to indexes. Parity of data gathered across the 33 software packaging and distribution tools currently supported. Improved documentation. Extended programming interfaces. Release of dependency graph and metric data under and open data licence.

Expected Outcomes

Improved search facilities for developers seeking to solve common problems using Free and Open Source Software. To enable users such as Depsy.org, Linux Foundation supported coreinfrastructure.org and Open Tech Fund supported openintegrity.org to better traverse, rationalise and leverage this data.

Proposal

1. What is the main issue, problem, or subject and why is it important?

Over the last five years Free and Open Source Software has seen a near exponential increase in projects exploring technical and commercial niches. This is due in no small part to the success of the distributed revision control management software Git and GitHub — a commercial platform for hosting and working on Git-based software projects. The financial and cognitive cost of starting a project built upon Free and Open Source Software is now so low that we have seen the advent of seed, micro and accelerator based investment models *only* possible due to the considerable value extracted by industry from Open Source Software.

The proliferation of projects solving increasingly niche problems has led to less potential reputational return for individuals—a key factor in attracting contributors to an Open Source project. We also see evidence of a shift, presented in the Ford Foundation's landmark study *Roads and Bridges*[1]*,* from the egalitarian relationship between users and contributors once enjoyed to one in which too few contributors are overwhelmed by demands from developers more akin to a commercial consumer than a peer[2]. The proliferation of Free and Open Source Software, and value extraction by industry has lead to three issues addressed in this proposal:

Developers do not currently have a reliable way of judging which project offers the most commonly utilised approach to solve their problem. Developers often solve identical problems by 'scratching their own itch', when presented with no obvious or established convention for doing so, often publishing the results under an Open Source licence. This breadth first approach to problem solving exacerbates issues of sustainability, central to this proposal.

[1] http://www.fordfoundation.org/library/reports-and-studies/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure

[2] https://peerj.com/preprints/1233.pdf

Compartmentalisation of software and the application of technology to increasingly niche problems has resulted in an expansive ecosystem of interdependent software packages that are difficult to maintain. For a developer working on even a small scale application the overhead of managing and maintaining a piece of software and its dependencies — the frameworks, plugins and tools upon which it relies — can be remarkably high. Case in point: Facebook's latest JavaScript library for building data-driven interfaces *React*[3] has over 700 dependencies.

Key pieces of Open Source Software, those most frequently used in technology projects within academia and industry, are often supported by small groups of individual contributors with no financial support or contractual obligation to do so. The landscape in which we find ourselves in is shifting, we are stood at a point in history where the good-will of a few can no longer sustain the increasing demands the ecosystem places on them. Our collective challenge is to support those who are working on the essential 'digital infrastructure' as a public good. But what are we already doing to support them?

Commercial organisations are beginning to see the value in giving back to the community, but they are not immune to influence. Black Duck's 2015 survey found that 27% of companies now have a formal policy on contributing to Open Source Software and many companies employ staff to work all, or a considerable portion of their time on Open Source projects. Commercial incentives however, have a habit of perverting a project's focus for the benefit of its host or possibly destroying it entirely. The 'sale' of popular, Open Source JavaScript application engine Node.js to Joyent, its subsequent mismanagement and the near fatal effect this had upon the community provide us a stark example of how commercial influences —intentional or otherwise— can wreak havoc upon a project[4].

[3] https://facebook.github.io/react/
[4] http://readwrite.com/2014/12/08/node-js-fork-io-js-isaac-schlueter-statement/

Academic institutions have historically proven they are one of the few organisational structures provide the right environment for nurturing new projects. A successful relationship between an organisation and an Open Source project is one based on stewardship rather than control. There are numerous examples of this: The University of Illinois' launch of next-generation program compiler LLVM and The University of Auckland's development of R, now the preeminent statistical analysis tool in academic and industry[5].

Academic institutions can also provide the environment to sustain them. Their endowment-based funding models could be argued to be the predominant source of protection from commercial influence but they lack a robust way of demonstrating and *crediting* the value they create within the wider scientific and research community and industry. The philanthropic community could equally provide a similar environment for Open Source infrastructure.

2. What is the major related work in this field?

Nadia Eghbal's 'Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure'[6], offers a concerned view of the Open Source ecosystem. Commissioned by the Ford Foundation, the study was published July 2016 and provides a comprehensive overview of how Free and Open Source Software is developed, maintained and —it is argued— exploited. It stops short of providing a roadmap for solving the cultural, financial and institutional issues discussed, but suggests that measuring the usage and therefore impact of projects is one of the necessary pre-conditions to identifying digital infrastructure and catalysing real change.

Following the publication of the Heartbleed vulnerability in April 2014 the Linux Foundation established the Core Infrastructure Initiative, a project focussed on identifying and supporting undervalued,

[5] https://www.r-project.org/ contributors.html

[6] http://www.fordfoundation.org/library/reports-and-studies/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure/

under-funded and under-scrutinised open source projects, specifically those critical to internet infrastructure. Having galvanised and raised financial support from a select group of concerned industry leaders The Linux Foundation established the Census Project[7] to produce a hit-list of projects for the initiative to recommend to the group for support. The Census Project uses data from Black Ducks' Open Hub (formerly Ohloh.net), a free service that gathers data about open source projects by analysing factors concerning the source code alongside user-submitted wiki-like project data. To this it adds a measure concerning the project's potential impact upon the whole ecosystem based on its popularity. It does so using data from Debian's Popularity Contest[8] project which gathers data on a weekly basis about software installed on servers upon which the administrator has chosen to install a package specifically for this purpose. Popularity Contest has collected 184,146 such reports at the time of writing and this approach has been copied elsewhere at similar scale.

The Depsy project[9] seeks to address the endemic lack of citation of software used in and created by the research and scientific community. The citation has and continues to be the currency of these communities yet frequently it is found that software is under cited. Howison and Bullard[10] found that in 90 randomly selected biology articles only 37% of mentions involved formal citations to domain papers or to "software papers" written to describe software. Depsy mines papers to find full-text mentions of software referenced, crediting those involved and federating this credit through the ecosystem using contribution metrics where links between references and source code exist.

3. Why are the proposers qualified to address the issue or subject for which funds are being sought?

Our proposal supports the work of two key contributors, Andrew Nesbitt and Benjamin Nickolls.

[7] https://www.coreinfrastructure.org/programs/census-project
[8] http://popcon.debian.org/
[9] http://depsy.org
[10] http://onlinelibrary.wiley.com/doi/10.1002/asi.23538/abstract

Andrew Nesbitt is Libraries.io's founder and core developer, developing much of the platform as it is today (1.7m projects monitored in 160 languages) over the last two years. He is highly regarded and respected both as a talented developer and a lynchpin to the open source community through his work on projects such as 24 Pull Requests[11], an annual 'give back to Open Source' movement that encourages users to contribute to key Open Source projects over the holiday period. 24 Pull Requests is entering its fourth year in 2016 having inspired 2,691 individuals to make 15,055 contributions to 6,245 projects in 2015. Andrews's career has seen him move steadily from user experience design through building transactional services operating at considerable scale with Forward to developing discovery services and improving web performance at GitHub. Andrew holds a BEng in Robotics & Automated Systems from Uni. Plymouth and will act as head of engineering, overseeing development of Libraries.io and acting as community manager for volunteer contributors.

Benjamin Nickolls has acted as product advisor to Libraries.io for eighteen months, having seen the value that the platform could provide to projects he was developing within the Core Infrastrucuture Initiative. Ben's career has seen him move from engineer and cryptanalyst to sit squarely between engineering, product development and management. Today Ben heads mySociety's not-for-profit commercial subsidiary, building products and services for Government, charities and commercial clients. While not directly involved in mySociety's broader philanthropic support Ben does have oversight of how mySociety works within this community and has himself successfully raised funding from Government, NGO's, Charities and industry to support projects. He was also instrumental in securing an initial grant of $200,000 from the Open Technology Fund to support the Open Integrity Index, a project that seeks to provide an open framework to measure and categorise Open Source projects with respect to end-user privacy and security, using Libraries.io data. Ben holds a BSc in Computer Science from the Uni. Liverpool and a MSc in Computer Security from Uni. Birmingham, he will act as head of product,

[11] http://24pullrequests.com/

leading on product design, user research and communication. Ben will also forward further funding applications and manage administration.

4. What is the approach being taken?

Libraries.io's exists to raise the quality of Open Source Software by accelerating the the factors that affect its evolution. It does so by understanding the value an individual piece of software contributes to the entire ecosystem. Libraries.io does this by understanding the complex relationships between software and the frameworks, plugins and tools they depend upon. Libraries.io follows the principles of Google's PageRank which works by valuing a single page or site based upon an assessment of factors then federating some of this value to any site it links to, in effect treating these links as an endorsement of the content contained within the target site. Libraries.io takes this approach and extends it to the world of software by valuing a speficic project based on factors similar to the Linux Foundation's Census project (we unimaginatively call it SourceRank) and federating this to any software that it utilises as a dependency. It does so by traversing the network created by configuration management files that tell a developer which software needs to be downloaded and installed in order for a project to compile and/or run successfully. Most commonly this configuration management is handled by a package manager, a hosted service that provides indexes of versioned software releases for installation programmatically and similarly offers a framework for publishing one's own work for use by others. Package managers typically exist for an operating system (Debian Linux or OS X) or language (Ruby or Python) but they also exist for frameworks and applications (Wordpress or Chrome). Currently Libraries.io covers 33 package managers recording data from close to 1.7m projects hosted on GitHub, written in over 160 languges. It is, to our knowledge, the largest dataset of its kind in the public domain. Libraries.io's strategy is to use this network graph to address the three problems established in section one. Currently we address two of these issues directly with two tools:

Libraries.io's central, public facing website[12] is an open and free discovery service, serving search results for Open Source projects based upon the SourceRank score, weighted predominantly by the frequency of their use in other's software as a dependency. This is a strong indicator of reliability as a project that has a large community of developers invested in its success is more likely to be maintained by the creator and or a team of contributors, though this does not imply a sustainable project in the longer term. By indexing search results this way we accelerate the natural selection process, establishing convention or consensus which can improve productivity for the individual and provide some of the preconditions for a sustainable project.

DependencyCI[13] is a subsidiary project, which addresses the issue of maintaining increasingly complex projects. It is a service for developers to build quality checks into their workflow for their complete software dependency tree. It uses Libraries data to notify developers of updates to software they depend upon, license incompatibilities, deprecated (abandoned or unused) libraries and other factors that negatively impact a team's productivity due to the overhead required to maintain a project. DependencyCI was launched to the public in July 2016. With 1,663 users tracking 3,718 projects it encapsulates the early sustainability strategy for Libraries.io, with a target to generate $15,000 of funding for the project within the term of this proposal. No funding applied for within this proposal will be used to further DependencyCI.

The sustainability of Open Source Software is an open issue, one which Libraries.io is keen to continue to explore as part of the emerging communities in philantrophy and industry studying this problem. The value Libraries.io can provide today is to share with these groups the data collected about this complex, interdependent landscape. This is the basis for our proposal which focuses on expanding coverage,

[12] https://libraries.io
[13] https://dependencyci.com

creating parity of service across multiple languages and frameworks and improving responsiveness for users and re-users.

5. What will be the output from the project?

Libraries.io data has been gathered over the course of nearly two years, as such it contains a historical record of data concerning deprecated software irrecoverably removed from package managers along with versioning histories that may have been re-written by contributors. It is therefore a dataset that is already invaluable to our users.

Despite our current scale we are limited. Libraries currently supports 33 package managers, covering 1.7m projects hosted on collaborative coding platform GitHub. Sloan's support will enable Libraries.io to increase the breadth of data it collects, adding support for projects hosted on second-tier platforms GitLab and BitBucket. Predictions are that this will yield a further 1-2m projects tracked in Libraries.io.

Sloan's support will also expand the scale of data gathered, by creating parity in data gathered across the 33 package managers (tools used to declare dependencies and distribute sourcecode) currently supported, many of which require compilation in order to resolve the dependency tree, something we have yet to explore.

Finally Sloan support will improve access to Libraries.io data, extending the programming interfaces for search, data retrieval and traversal. This will include the publication of network graph and SourceRank data under a Creative Commons licence.

This work will increase the value of Libraries for users and reusers working in this field. Finally it will provide the necessary runway to engage the funders identified in section 7, better market and communicate the project's mission and to continue to work alongside funders and the wider community to explore how best to support our shared digital infrastructure.

6. What is the justification for the amount of money requested?

Our application for $124,770 represents a minimum level of support for both core contributors at a salaried rate comparable or below their current and historical income plus infrastructure necessary to allow Libraries.io to continue exapanding its scope. It provides funding for 12 months inclusive of a budget of $5,000 to cover travel and subsistence necessary to attend face to face meetings with the Sloan Foundation and $5,950 in ancillary costs including flexible, coworking office space for two staff. Finally it includes an administration cost of $15,000 (12% of request) which will cover operating costs of our fiscal sponsor Brave New Software for the duration of this proposal.

7. What other sources of support does the proposer have in hand or has he/she applied for to support the project?

Libraries.io curently generates circa $1,500 per month of AdWords revenue. Immediately we will seek to secure a grant of $75,000 offered by the Ford Foundation, providing the necessary funding to cover core staffing and infrastructure to 1st January 2018. Throughout the year we will seek support from (but not limited to) the Mozilla Foundation’s MOSS program, the Thiel Foundation’s Breakout Labs project, the Open Tech Fund, Google.org and finally the Linux Foundation whos work in the area of core internet infrastructure could benefit greatly from the support of Libararies.io data.

While we will be seeking further support from philanthropic foundations our long-term strategy is to to create a sustainable source of funding for the project through revenues collected through DependencyCI. Our intention is to formalise the current holding company (UK registered DependencyCI Ltd) as a subsidiary of Libraries.io. DependencyCI.com offers developers a range of workflow tools built upon Libraries.io data, launched in July 2016 and has received an encouraging response from industry. We have promising leads concerning company-wide usage at Sky, Goldman Sachs and Accenture. We have forecasted a conservative $7,000/year in revenues generated through the sale of enterprise services which will be used to fund group-wide infrastructure costs and/or staff salaries as required. Our intention is to re-invest any funds generated above this level to grow the customer base and develop the product until such time as DependencyCI is able to comfortably donate a significant sum to Libraries.io.

Appendices

Curriculum Vita

Andrew Nesbitt

Community

24 Pull Requests, http://24pullrequests.com/
Annual ‘give back to open source’ campaign inspiring 2,691 users to make 15,055 contributions to 6,245 projects in 2015.

London Node User Group, http://lnug.org/
Founding member of this a free, monthly meetup featuring talks and networking for those using Node.js

Industry

Founder and director, Dependency CI, 2016-Current
Continuous integration service using Libraries.io data, providing sustainable income to Libraries.io

Software Engineer, Independent Contractor, 2015 - Current

Software Engineer, GitHub, 2013 - 2014
Working on Discovery tools, New User Experience and web performance

Software Engineer, Independent Contractor 2012 - 2013

Software Engineer, Forward Internet Group, 2010 - 2012
Working on Ecommerce, Internal tools, Monitoring and Web Development

Web Developer, Rawnet , 2009 - 2010
Web Development for a variety of clients

Web Developer, Econsultancy, 2008
Redeveloping and upgrading a large classic ASP website in Ruby on Rails

Web Developer, Greenvoice, 2007
Developing a new social networking platform in Ruby

Web Designer, Ziymoo, 2006
Designing web interfaces for a video auction startup

Academia

BEng Robotics & Automated Systems, University of Plymouth, 2003 - 2006

Benjamin Nickolls

Community

Contributor, Core Infrastructure Initiative, 2015-Current
Focussing on the intersection of security and user-driven design, dissemination and fundraising.

Industry

Founder and director, Con Gas Games, 2015-Current
Physical game design, development and distribution.

Head of Services, mySociety, 2013-Current
Building sustainable income for parent charity from digital products based upon our open platforms.

Founder and Director, The Dot Consulting, 2011-2013
Open digital product agency focussing on mobile applications using web technologies.

Solutions Consultant, Ribbit Corp, 2009-2011
Working between engineering and business development to design propositions for programmable voice applications.

Information Technology and Research Graduate, BT, 2007-2009
Various roles in network security, applied research and corporate risk and continuity.

Academia

MSc Computer Security, University of Birmingham, 2006-2007

BSc Computer Science, University of Liverpool, 2002-2005

Attention to Diversity

While Libraries.io is currently a team of two, and this propsal sets out to support these two core contributors we will, in the future, foreground issues of diversity and seek demographic representation in the event of any later hire and subsequent employment.

Conflicts of Interest Disclosure

There are no known conflicts of interest between the project principals, the fiscal sponsor Brave New Software and the Sloan Foundation. Explicitly there is also no intellectual property contained within any of the products discussed in this proposal and any previous employer of either principal.

Information Products

The product this proposal seeks support are:

1) *Libraries.io*, that is: the service at https://libraries.io licenced under an Affero GPL licence and published at https://github.com/librariesio.
2) *The data set*', that is: a number of tables holding meta data concerning each of the projects indexed in Libraries.io and a description of their depdendencies, the format of which has yet to be formalised. The data set shall be released to the public under a Creative Commons 4.0 licence. This licence was selected after consultation with the Open Data Institute http://theodi.org.

This proposal does not seek support for DependencyCI a for-profit, commercial product available at https://dependencyci.com which forms part of a sustainability strategy for Libraries.io. The source code for Dependency CI is closed and not licenced to any third party under this proposal. A data sharing agreement shall be established between Libraries.io and DepedencyCI at such time as Libraries.io becomes a registered legal entitity. At this point it is our intention to create such an organisation as a parent of Dependency CI, enforcing a legal precedent of financial support through article of association or similar.

Budget

Project budget follows this page.