# Metadata Repository



## A scientific software ecosystem metadata repository

The development of a shared cyberinfrastructure depends on insight into the development, use, diffusion and status of software created with NSF funding. The CIF21 framework speaks of the importance of “leveraging ongoing investments and using common approaches and components” and one crucial element in achieving this is ensuring that tools developed with science funding have the wide visibility and transparency leading to increased use, increased community coalescence and increased sustainability (see recent NSF workshops on this topic)[1]. Currently there is no easy way for scientists or science funders to ask about the portfolio of tools developed, to see what researchers are using, or to know where and whom to turn to for support and insight on cyberinfrastructure tools.

To this end, we propose to design, prototype and test a community process, improved reporting architecture and a working repository system that will help develop these needed insights. The guiding principles are that these ought to be community based, lightweight and designed to be kept up to date with a minimum of ongoing effort from already busy participants. Accordingly we will facilitate a design, drawing on research and experience in the scientific software ecosystem and open source software ecosystems. We will work to gather input and build support from the PI community through the existing periodic NSF PIs meetings, starting with the SI^2 meeting scheduled for January 2013 where we will gather input and reaction through questionnaires as well as presenting possible designs including storyboards and interface mockups.. Based on the feedback we receive from the SI^2 community, we will post the designs and information online and advertise the materials through community mailing lists. The community will be able to comment and provide feedback. After a period of one month, we will refine the design based on the feedback received.

Other communities have faced a need to gather data about software projects and we will leverage their experience by drawing inspiration and standards from similar systems implemented by community software projects, especially that of the Apache software foundation (see http://projects.apache.org/doap.html) as well as projects such as the VIVO research metadata repository[2] (Holmes et al, 2012). In addition to gathering and presenting data, systems such as these provide a template that prompts projects to consider their approach to the overall management of their tool development and diffusion: reporting requests provide gentle guides towards better practices and improved sustainability. We anticipate that the consultation with PIs will prompt an independently valuable exchange and documentation of best practices, similar to that which has occurred in open source communities.

Systemic insight is required in four areas: tools, development community, user community and use and plans for sustainable production.

Tools—It is important to know what has been developed. To this end we will develop a community process and reporting architecture will ask funded projects to identify infrastructure tools they have developed and provide a location where updated, machine-harvestable descriptions are published. Projects will be asked to identify and update specific download locations, the location of their source code repository and their plans for backup and archiving. Stretch elements here could include the location of test suites and continuous integration infrastructure (such as use of the National Middleware Infrastructure Build and Test labs) and the identification of major dependencies, especially those also funded through infrastructure funding.

Development—It is important to understand the development community for the software. To this end metadata should be gathered about who has undertaken the development and who is responsible for ongoing work on the software. Projects could be asked to provide information about which scientific publications describe their tools and approaches. In many circumstances it is desirable that projects be encouraging of contributions from outside the funded development team, projects could be asked for the location of documentation of their approach to this and perhaps their procedure for outside contributions.

Users—It is important to understand the user community for the software. To this end the community process should include gathering metadata about the location and nature of community infrastructure, such as mailing lists, forums, and structured interactions such as trackers or support ticketing systems. Projects could be encouraged to gather more data on their user community by asking how and where they track users, downloads and other project metrics, perhaps being guided to options for more automated usage collection (e.g., Thain et al, 2006.) Since acknowledgement of software in the published scientific literature is important, projects could be asked how their users are asked to provide acknowledgement, often undertaken through requests for specific citations that can be tracked in published literature.

Sustainability—It is important to understand plans and models for sustainable production. Since tools are often supported by multiple grants, each with different levels of contribution and start/end dates, projects ought to link between tools and grants, allowing an assessment of the extent and future of funding. In a manner similar to the specific requirement in the GPG for a data management plan, projects could provide links to their sustainability plans. To aid in the formation of sustainability plans projects could be asked to provide updated answers to questions like "How does the use of your software promote its sustainable development?"

Our proposal envisages a community based process and we expect to find many enthusiastic contributors who are keen to maximize the visibility of their tool and development process. Yet developers of scientific software are very busy and have conflicting demands on their time, including unclear rewards for time spent building tools compared with those from publishing substantive scientific results (Howison and Herbsleb, 2011). Moreover we believe that the NSF has a clear interest in the visibility of funded programs and thus that there is a role for policy to play. A clear sign of success of our project would be a process and architecture that the community supports building into the evolution of NSF required grant reports.


People involved: Ewa Deelman (USC, http://www.isi.edu/~deelman), James Howison (UT Austin, http://james.howison.name)


This project is funded by the National Science Foundation, under grant number OCI-1148515

[1] “Challenges of Scientific Workflows” in 2006 and “Cyberinfrastructure Software Sustainability” in 2009
[2] http://vivoweb.org/