genomeweb BioInform
The Integrated Informatics News Source

BOSC Keynote Highlights NIH Plan to Create Common Framework for Data, Software, More

July 18, 2014

By Uduak Grace Thomas

NEW YORK (GenomeWeb) – The National Institutes of Health has approved a proposal to develop a framework that will connect biomedical data, software, training materials, and other resources developed as part of NIH-funded research projects, making them more accessible and shareable.

Philip Bourne, the National Institutes of Health's associate director for data science, provided details of the planned resource, called the Commons, in his keynote address during the Bioinformatics Open Source Conference, one of the special interest group meetings held prior to the start of the Intelligent Systems for Molecular Biology conference held in Boston July 11-15.

The proposal has been approved by the NIH leadership and the heads of all 27 NIH institutes. Co-leading the Commons' development are Vivien Bonazzi, the genome informatics and computational biology program director at the National Human Genome Research Institute; and George Koumatsoulis, a senior bioinformatics specialist at the National Center for Biotechnology Information. Koumatsoulis was responsible for the NCI's Cancer Genomics Cloud initiative.

Bourne described the Commons as "a conceptual framework" where research objects — datasets, software, research papers, experiment descriptions, and so on — are readily accessible and shareable by various stakeholders, such as NIH awardees, the rest of academia, the private sector, and government agencies. This is not going to be a large database housed at a single location that brings together content from a series of smaller databases, he explained. What's being proposed is something that is analogous to the Internet. Essentially, a functional system that's not owned by a single entity but which works because all of the participants abide by pre-defined rules.

Right now, the 27 NIH institutes operate and manage multiple, independent resources and repositories that exist largely in their own private universes. Under the current system, Bourne said, "you can't find an individual data object in the NIH repositories unless you know what you are looking for" and what database to look in. Furthermore, some of the data and tools that biomedical research projects generate don't fit into existing resources.

Creating a common identification scheme would address both of the aforementioned issues, he said. It would make locating and accessing research objects less time consuming and it would offer a way to access resources that aren't captured by existing NIH categories. This system would also make it possible to access data and resources from projects funded by institutes that are external to the NIH, as long as those groups agree to use the Commons-defined rules.

At a minimum, the rules would include a unique research object identifier (ROI) for each object in the ecosystem; and provenance for each ROI including a minimal set of information to identify the object's creator and those who have made modifications to it. Other rules include an independent body that issues the ROIs and rules that specify ways to resolve ROIs to locate the source of the research object in question. Also included would be recognizable object types for different data formats, software objects, and narrative digital object identifiers (DOIs) for instance; and common data elements, which would be used to describe individual object types. With these rules, the actual location of the resources would not matter as long as they have identifiers and can be indexed, Bourne said.

Ideally, the biomedical community would be able to agree on the rules at the start of the Commons. But a more realistic initial scenario, Bourne said, would likely be several "emergent Commons," each with distinct rules that are later reconciled around a common rule set. Besides pre-defined rules, the Commons would offer access to public and private cloud resources to handle compute and storage needs.

Establishing a workable business model for the Commons is critical, Bourne said in his talk. At present, researchers include computation costs in their grant applications, but under a proposed broker model, they would not get actual dollars to purchase compute resources. Instead, they would get credits — or Commons dollars — to spend with whatever Commons-compliant vendor best fits their needs.

For example, if a researcher generates a lot of data but does not have the compute power needed to analyze it, Bourne explained to *BioInform*, they might choose to spend the credits with a cloud provider that offers cheap compute. Conversely, if a researcher has sufficient local compute but needs storage space for their data, they would select a provider with lower storage prices. Service providers participating in the Commons don't necessarily have to be commercial entities, he said, noting that any group whose resources meet the Commons rules can play in the broker space. For instance, academic institutions with Commons-compliant repositories can offer storage space as a service to the community.

The credits would be managed by an independent broker and purchased on behalf of researchers by federal funding agencies, foundations, and so on. The broker would also oversee all the vendors that provide services to the community. Vendors would bill the broker for whatever services the researchers use and the broker in turn bills the NIH. The NIH would then pay the broker, who in turn would distribute the money to the vendors.

The difference between this approach and the current system is that it ensures the funds for computation are actually used for that purpose, Bourne said. At present, sometimes funds earmarked for compute in grants end up being spent elsewhere. It also reduces waste. Under the current scheme, "each grant effectively buys some kind of compute resource," he said. But, in many labs most of the time those resources are idle. With cloud infrastructure, researchers access compute power as needed and only pay for what they actually use. A broker would also be able to arrange better deals for the community, as service providers would want to ensure a steady supply of business for their offerings.

Bourne told *BioInform* that he will meet with an *ad hoc* advisory group of consultants in September to begin hashing out the exact details of the proposed Commons including its key features, potential problems, and opportunities. After that, the NIH will issue a request for information asking for input on the proposal. The community will have a chance to weigh in, for example, on what the standard ROI for the Commons should be — there are several already in existence such as DOIs and uniform resource identifiers. They will also put out an RFI related to the suggested broker system, quite likely before the pilots meeting in September, Bourne said.

Next, the NIH will run a series of pilot projects involving public/private partnerships. In defining the criteria for the pilots, the Commons planning team will focus on open systems; support for basic statistical analysis and the ability to embed existing applications; as well as application programming support that will enable integration with existing resources. Proposed system and business models will be reviewed and then a 6-12 month pilot phase will be launched. The fruits of the phase will be evaluated on how well they worked and cost effectiveness, and then a determination will be made about whether a wider deployment makes sense. The plan is to present a workable plan to the NIH leadership by the summer of 2015.

Merits and challenges

Bourne began his talk at BOSC by providing a grant funder's perspective on the biomedical research market. Currently, he said, there are no good answers to questions about how much money actually gets spent on data and software-related activities in biomedical science, or how much should be spent to achieve maximum benefit to biomedical science relative to what's spent in other areas. Reproducible research is still a major concern at the NIH and more broadly in the community, he noted.

Furthermore, governments and research institutes all over the world have instituted policies requiring researchers to share data generated from their projects. The Commons is an attempt to address these and other research-related issues, Bourne said.

In general, reactions to the proposed plan from BOSC attendees were quite positive.

"I think it's a promising start and refreshing to see the NIH take such a public stance on issues such as lack of training for researchers, open research practice, and the commons," Kaitlin Thaney, director of the Mozilla Science Laboratory, said in an email to *BioInform*. "I'm most excited about the training and prototyping focus, and the NIH's ability to help close the feedback loop in testing out changes to the research system across organizations, research groups, and various stakeholders such as publishers, technology providers … funding agencies, and libraries."

John Greene, SRA International's senior director of bioinformatics, said, "We hope to support Dr. Bourne and the NIH for this important endeavor, as we have in bioinformatics for over 15 years." Furthermore, he added, "We do hope funding mechanisms will be used that will allow industry participation."

Carole Goble, a professor of computer science at the University of Manchester, said she shares Bourne's vision. "As a long-term advocate of research objects … [and] reproducibility, and as the developer of a Commons for Systems Biology … I really welcome a practical approach that doesn't buy into the single site/monolith approach and recognizes discoverability and citation as first-class properties," she told *BioInform*. "I suspect that a [public-private partnership] will be a practical approach. It also plays well with the so-called [findable, accessible, interoperable, reusable] principles of bottom-up initiatives like FAIRport."

"The thing that I like best about it is that it would let me build tools and work with other people's data without asking for permission, and with less friction," C. Titus Brown, an assistant professor of bioinformatics at Michigan State University, told *BioInform* "The ideas of frictionless data sharing and permissionless innovation are really appealing."

But there are reservations.

"The biggest worry I have about the Commons is that it will be overly centralized," Brown said. "The NIH has a top-down [approach], in that it tends to think that lots of funding and top-down management is helpful. I think in this case, the underlying stratum of the Commons needs to be laid out and then little top-down, field-specific castles can be built in it, as experiments in different kinds of commons use." Bourne "has thought about these problems deeply, and so the real question is whether or not it all comes together," he added.

Also, "this does not solve the thorny issues of the 'Tragedy of the Commons': cultural change and the challenges of sustained investment in infrastructures," Goble said. But, "it's a significant step forward … and I hope the 'keep it simple not complex, keep it open not closed, keep it commons not commodity' prevails."

"The big challenge, in my opinion, is going to be to execute on the plan [Bourne] has outlined, as much of it has been spoken of before but failed to be implemented due to competing interests, inertia, and a top-heavy way of crafting and pushing technical solutions to researchers and of funding research in this space," Thaney added.

There is also a risk of "perceived expert blindness," that is "not knowing when proposed solutions are catering to 'business as usual' or actually fostering more collaboration, innovation, and a sea change," she said. "Shifting practice is challenging, and the NIH could play a key role in stressing the fact that most of the technology needed to get us started already exists but isn't interoperable … [that] and the remaining gaps in software are not always being surfaced or executed in ways that are furthering open web-enabled science."

And as often happens, there is a risk of reinventing the wheel. As it stands, the proposal does not appear to be going down that path but there is some skepticism, Thaney said. For example, "will the funding models for the program be flexible enough to accommodate support of existing programs and new ideas that might help the NIH achieve this vision?" she said. Also, will the NIH build on open-source offerings and existing communities of instructors and participants such as Software Carpentry — which Thaney leads — or will they develop their own training programs, she added.

One question raised by a BOSC delegate was whether or not there is concern that participation in the Commons would create unwanted, extra work for researchers who, in addition to their other tasks, would now have to package their data and software in a Commons-compliant fashion. In his response, Bourne highlighted the value of the data and software that are collected as part of these projects and suggested improvements, such as better data citation formats, as incentives for researchers to add the extra step to their workflows.

In his conversation with *BioInform*, Bourne further noted that this would help with recurring reproducibility issues. Right now, data and tools are deposited in various local repositories, but they aren't maintained. "For data and software like this there is at least a 10 percent attrition rate per year. So after five years, effectively half of that is gone" and along with it the chances of replicating the associated research, he said. Putting it into the Commons "at least has the potential to provide longevity," he said. "There's still a cost of doing this but at least it's likely to be more stable."

How these resources will be maintained as part of the Commons when grant funds run out is still an open question. When the subject was raised during his BOSC talk, Bourne said that right now, to get renewed funding to cover the upkeep of a database, for instance, a researcher might provide data on how many users they have worldwide and how many petabytes are downloaded per year. But that says nothing about how the data is actually used, he said. And on closer analysis, the usage patterns are highly variable.

"We haven't really analyzed what that means and … looked at it in the context of using this information to determine what we keep and what we throw away," he said. But "how we think about this going forward has to change." The actual model for this "has yet to be worked out," he said, "but if there are data that are clearly used extensively by a community, when a grant runs out I think it's in the NIH's best interest to continue to preserve that, and we need to be able to do that in the most cost-effective way possible."