bio1NFORM
The integrated informatics news source

BioTeam Turns to Wiki Technology to Build Flexible LIMS for Next-Gen Sequence Data

February 1, 2008
By Bernadette Toner

Bioinformatics consulting firm the BioTeam is turning to the same collaborative software that underlies Wikipedia and numerous other web-based resources to help address the formidable data-management challenges of next-generation sequencing systems.

The company has used the MediaWiki software package as the basis for a flexible data-management system it has dubbed wikiLIMS. The system was initially developed for the Navy Medical Research Center in Rockville, Md., which installed two 454 Life Sciences' FLX sequencers earlier this year and hired the BioTeam to build its IT infrastructure.

The NMRC had "been pumping out data for about six months, but the data was still living on the machines and there was no real visibility around what was produced and what was run when and where," Michael Cariaso, a scientific consultant with the BioTeam, told *BioInform*. "We needed a solution for that environment and we [used the wiki technology to] built something pretty general."

Since then, he said, the company has opened projects with Cold Spring Harbor Laboratory and the National Cancer Institute, about using the system to handle data from other next-gen sequencers.

Cariaso noted that while wikiLIMS was initially developed to handle information from high-throughput sequencers, its applicability extends well beyond such systems. The company is also in discussions with an undisclosed core facility about using the system to manage data from a protein-analysis system, he said.

Because it allows users to easily create and edit web pages using any browser, wiki software is a natural front end for rapidly changing research environments, Cariaso said.

Most LIMS platforms are built on database technologies like Oracle or MySQL and are carefully structured and hard-coded to ensure that certain types of queries run very quickly. However, Cariaso said, "that doesn't really let you cope with the research environment that we find ourselves in so often, where … we don't really have any sense of the questions we're going to be asking."

In such settings, he said, "we don't really know how to design and model all the information and the relationships between it. We just want to start capturing immediately and then begin to figure it out a little more as we work with it, because very likely the question that we want to ask two weeks from now is not the same question we want to ask two months from now."

These issues are particularly acute for many customers of next-gen sequencing instruments, Cariaso noted, because they are still familiarizing themselves with the type — and volume — of data being generated.

Dick McCombie, a Cold Spring Harbor researcher participating in the lab's partnership with the BioTeam, agreed. "The data management with these new systems is the single biggest challenge, so we wanted to act pretty aggressively and go with people who we thought could speed up our getting a handle on this," he said.

McCombie said that his group developed an in-house LIMS system for handling gel and capillary electrophoresis sequencing data, but when it came time to select a similar system for the lab's Illumina Genome Analyzers, he decided to work with the BioTeam for several reasons.

First of all, he said, his team is still manually analyzing weekly production statistics and tracking the performance of its machines — a task that he'd like to automate as soon as possible, and thinks the BioTeam will likely "get that in place quicker than we would have by hiring people in-house."

He also said he's "encouraged" by of the kind of data-management capabilities wikiLIMS has produced at CSHL so far, but stressed that the lab is still in the "very early stages" of the project.

Cariaso said that wikiLIMS is integrated with lab instruments to automatically capture research data and also allows researchers to create and edit their own entries on the fly. This dynamic approach is well-suited to research groups that are still evolving their data models, he added.

The wikiLIMS system is "something of a shared brain" for a lab, "where we can all agree on a common dialog, a common vocabulary, common semantics, and we have the infinite memory of a hard drive."

For example, in the case of a next-generation sequencing run, "we hook a little bit of extra code in there, we have it read over the data it just processed, and it posts a summary into the wiki," he said. "So you don't need to have a researcher at the end of the run going off and recording into their lab notebook, 'Here's what the data said.' The machines record that data automatically and all the researcher has to do is point over to it and then focus on making their own sort of higher-level summary of the information if they want to."

Cariaso said that the BioTeam and its customers are currently developing a series of "templates" to serve as a standard model for entering information about samples in a structured format, though he noted that he and his colleagues "don't do much in the way of enforcement" of standards or common terminologies for the system.

"We're starting to emerge a standard representation of what a sample looks like, because that's a really general concept and every lab thinks about it a little bit differently," he said. Beyond that, "we rely on individual labs to establish their own protocols so far because the scale we've worked on so far has allowed that."

One important feature of the system is its ability to track every single change to the data — just as Wikipedia does — which "helps to prevent accidents" in data entry or data analysis. This self-healing aspect of wikiLIMS could ensure that the system evolves to meet the needs of individual labs. Cariaso likened the system to "a way of creating a book, where the topic is your laboratory, and every time you've got an observation, you can just go to your web browser and record a simple comment."

In another analogy, he described wikiLIMS as "something of a shared brain" for a lab in which "we can all agree on a common dialog, a common vocabulary, common semantics, and we have the infinite memory of a hard drive."

In addition to the data-capture and data-entry aspects of wikiLIMS, the system includes a number of "interactive" features, including options for setting up analytical workflows, that set it apart from the "static pages" in Wikipedia and other wikis, Cariaso said.

"There are buttons, check boxes, all kinds of stuff that you're used to seeing in more interactive web pages that are embedded directly into the wiki," he said. "And you can check-box off a couple of bits of data and say, 'Please begin this next step of analysis.'"

Cariaso said that the company is "experimenting" with the idea of integrating the system more tightly with workflow software such as Taverna or Pipeline Pilot, but hasn't yet taken any definitive steps in that direction.

Stan Gloss, managing director of the BioTeam, said that the company is currently installing the system solely through custom engagements, but may consider developing the system into more of an off-the-shelf package if there is enough demand.

Gloss said that the company is training its customers to use the system so that they can extend it entirely on their own with minimal assistance from the BioTeam. "It's not a captive model," he said.

Ultimately, he said, "it would be great if wikiLIMS became popular enough to have a community of open-source users" that could sustain itself and share tips and advice about the system.