



# Bioinformatics Pros: Labs Should Include Data Management in All Project Plans

August 14, 2009

**Byline:** Vivien Marx

**Newsletter:** [BioInform](#)

[BioInform - August 14, 2009](#)

**With large-scale second-generation** sequencing-based projects taking hold, labs need to budget and plan for data management throughout a project's timeline, an NIH official told *BioInform* this week.

Many researchers "totally underestimate issues around data management," Joni Rutter, associate director of Human Population and Applied Genetics at the National Institute on Drug Abuse's Division of Basic Neuroscience and Behavioral Research, told *BioInform*.

Rutter is the project officer for the Epigenomics Data Analysis and Coordination Center and spoke about general issues connected to data management, not issues particular to EDACC.

Scientists writing grants think first and foremost about their science and when fitting a project to a particular NIH budget, data-management resources "tend to be the first things to go" when cutting costs, she said.

With centralized data repositories such as EDACC, NIH is helping to be "more efficient" in managing scientific data, since repositories assist in fostering standards and formats for data deposit, she said.

At this year's Intelligent Systems for Molecular Biology conference, Owen White of the University of Maryland School of Medicine, who directs the Human Microbiome Project's Data Analysis and Coordination Center, shed some light on how such a center operates.

White noted that the DACC is not responsible for primary data submission, but helps each of the HMP centers handle -omics and clinical data in a "rapid response" fashion.

The data center's "position," White said, is to have the sequencing centers deposit data and metadata, to "not interfere, and [to] let the pipeline stream along." Over time, "we will be structuring it" and doing "lots and lots of clean-up of the data," he said.

Metadata is being collected, he said, for example, on sample prep methods, since that will "impact how the data will cluster" in downstream analysis, he said.

He and his group are also "encouraging, or some might call it policing," which means monitoring data quality, assuring data is documented, and helping centers to resolve data-management issues.

"We're constantly evaluating tools used by the centers and developing standards," White said, adding that his group is developing a software repository and pipelines and making them available to scientists.

### **Change the World?**

Data management is not only an issue for large, multi-center projects, however. Small labs, in particular, are quickly finding that data management is a core requirement for high-throughput experimental platforms.

While ads for second-generation sequencers tell scientists that the instruments are "going to change the world," that change won't happen unless researchers in traditional biochemistry or molecular biology labs can use the data, "and they can't," Anton Nekrutenko told *BioInform* during ISMB.

Nekrutenko is associate professor of biochemistry and molecular biology at Penn State University and co-PI for Galaxy, an open source bioinformatics data integration and analysis platform.

Data-intensive bioinformatics tasks that were once relatively rare are now "permeating every aspect of biology," said James Taylor, a computational biologist at Emory University and Galaxy co-PI.

That development calls for "effective" methods of managing data, as well as introducing "more control, reproducibility, [and] transparency to data analysis," he said.

Nekrutenko and Taylor organized the data and analysis management special interest group, or DAM-SIG, meeting at ISMB/ECCB. One focus of the sessions was on the need to standardize metadata. "I think that's a very important aspect" of data management "to allow for "better, open interchange formats for understanding and querying experimental metadata across experiments," Taylor said.

Metadata management plays a "very important" role in microbial and metagenomic projects and is still often an unsolved challenge, Nekrutenko said.

Some tools are coming online to help researchers handle metadata, though. At ISMB, Philippe Rocca-Serra, a researcher at the European Bioinformatics Institute, outlined his group's "standards-supportive infrastructure," for annotating metadata in the ISA-Tab format with ontologies and according to standardized reporting guidelines so scientists can manage multi-domain experimental metadata.

A beta version of the software suite, called ISA Tools, was released in late July [here](#).

The suite includes ISAcreeator for annotating and editing metadata and an app called ISAconverter for converting ISA-Tab files to formats suitable for submission in public repositories.

## Annotation is Key

Other observers view accurate and thorough annotation as the key to effective long-term data management.

The research value of -omics data "correlates directly with how precise, exhaustive, and consistent" its annotation is, Tom Beatty told *BioInform* via e-mail this week.

Beatty, a principal researcher at business technology consulting firm CSC who specializes in life science and healthcare consulting within the firm's Emerging Practices Group, said that a community-wide, wiki-style approach to annotation could be useful.

Likewise, Sanjeev Wadhwa, director of CSC's Life Sciences R&D Practice, told *BioInform* via e-mail that semantically enriched wiki content, which can be processed and interpreted by wiki-based ontology infrastructures, "will provide intuitive means to collaboratively create, organize, and retrieve knowledge."

In the bioinformatics community, several projects are using wikis to crowd-source data annotation, including as [WikiGenes](#), [GeneWiki](#), and [WikiPathways](#).

Nevertheless, "while the 'many hands make light work' approach will speed a given wiki's annotation of the universe of genes and proteins for a given species, problems of ontological standards and method consistency loom large," Beatty said.

Wadhwa agreed, noting that wikis can serve as "a jumping off point" for deeper, more rigorous analysis and annotation, "which would then be assimilated into a more formal repository."

The usefulness of gene and protein annotations is directly correlated to data quality, Beatty said. "Data quality, in turn, is a function of consistent analysis methodology, standard ontology, vocabularies, and dictionaries, and vetting/approval of annotations, not to mention the all-important pruning of bad content."

## Still No Standards

Data management and frameworks for data integration are difficult in biology, said Nekrutenko, "because the science is not like light-bulb manufacturing where there are standards." In a scientific enterprise, standards are not foremost on the agenda, he said.

One challenge is that established protocols and robust analysis pipelines for emerging second-generation sequencing applications such as ChIP-seq or RNA-seq are lacking, Nekrutenko said.

The need to standardize ontologies, methodologies, and vocabularies will emerge from the science, "especially since countless person-years of research will eventually be conducted using these fundamental annotations," Beatty said.

But first larger issues of data governance have to be addressed, he said. While ontologies and controlled vocabularies are "vital," they are "a subset of the larger issue of governance. It would be better to approach it 'top down' and address ontology and vocabulary/dictionaries after larger governance issues have been locked down," he said.

In large-scale -omics projects, even if standards adherence is not foremost on every researcher's mind, there is an "invisible hand" effect, Beatty pointed out, since the commercial or scientific value of an individual scientist's contribution depends "completely on his or her work being accepted as valid by some larger group."

Computing is helping scientists "discover scientifically valuable connections and correlations across massive — and previously unlinked — datasets," Beatty said. However, while the biomedical community realizes the power of large data sets, it "is struggling with the semantic standards that will be used to unlock its value," he noted.

There are some efforts underway to establish standards for the field, however. For example, Geospiza is working to adapt the Hierarchical Data Format, originally developed for physics, to create a "domain-specific" version called BioHDF. They are identifying "key architectural elements" as they prototype the file format to optimize it for data storage and computation. [*BioInform*, March 20, 2009].

Geospiza's senior scientific developer Mark Welsh said at ISMB that DNA sequences that gobble up to tens of gigabytes as text-format files are reduced to "a few percent" of that size when stored in BioHDF.

Welsh told *BioInform* at ISMB that the group is currently testing BioHDF's robustness to make sure it can grow as sequencing technologies and bioinformatics demands evolve.

### Uncertain Future

But despite some advances in tools to help researchers handle their data, the long-term sustainability of data resources is an ongoing concern. A recent survey found that of more than 530 -omics databases in 33 European countries, more than 67 databases, or 12 percent, were not live or had not been updated since 2005, and the "update status" was unclear for another 78 databases, or nearly 15 percent of the total.

The study, led by Christopher Southan in the context of the European Life Sciences Infrastructure for Biological Information project, was published this spring and can be found here.

In addition to the relatively high number of inactive databases, 66 percent of survey respondents said they had one year or less of "assured" funding for their database projects, while 32 percent of respondents reported they were "very concerned" about the long-term sustainability of their data resources.

While nearly 24 percent of providers held under 0.5 gigabytes of data, 16 percent held between 0.5 and 1 GB, 11 percent held between 1 and 2 GB of data, around 20 percent held between 2 and 10 GB, and the remaining approximately 30 percent held more than 10 GB of data.

#### Genomeweb system

These settings are generally managed by the web site so you rarely need to consider them.

**Issue Order: 3**

