How can we make data sharing less daunting in order to address the scientific reproducibility problem?

Biomedical data is undergoing an identity crisis. “How can that be?” you may ask. It’s data: bits of information stored on servers somewhere; sequences of nucleotides in a genome; levels of gene expression in lots of different cells and lots of different organisms; images of brains and lungs and hearts; and all of these things tied to particular health problems.

How lost can data be?

Quite lost, in fact. Datasets are often unnamed, undescribed, homeless, and unpublished. And the sheer quantity of biomedical data generated by diverse labs all over the world makes the problem worse: How can you find a needle in a pile of needles if the individual needle can’t even be described in a unique way?

As a result of data’s identity crisis, researchers can’t find or access it in order to reproduce published work or use it in new ways.

“We have to solve reproducibility,” says Anita Bandrowski, PhD, a specialist at the Center for Research in Biological Systems at the University of California, San Diego. “We’re scientists for gosh sakes.”

If data had a will of its own, perhaps it could be coaxed to declare its identity, tell us where it came from, where it lives and what knowledge it holds. But data does not have a will of its own. Instead, researchers carry the burden of assigning data identifiers to their data, connecting metadata descriptors to it, putting it in a reliable repository or other predictable location, and publishing it. Unfortunately, the reward system for researchers conspires to keep data in the dark.

“We have to create a culture saying that data sharing is really important,” says Vivien Bonazzi, PhD, senior advisor for data science technologies and innovation in the Office of the Associate Director for Data Science (ADDS) at the National Institutes of Health (NIH).

Changing the data-sharing culture means changing the incentives—including the reliance on publication in a scientific journal as the sole currency of biomedicine. There must be room for recognizing the value of shared data and software. “Currently, you don’t get tenure from data and software,” Bonazzi says. “We have to find a way to give credit to the people doing the data wrangling.”

There is hope: The NIH is pushing an overarching philosophical shift toward making datasets FAIR—findable, accessible, interoperable and reusable. “Everyone agrees on that,” Bonazzi says. As a result, the NIH is funding the development of tools, software and systems to make data sharing and data discovery easier: systems for giving datasets an identifier, simplifying data annotation, and registering datasets in a searchable index. And to make it more real, the NIH wants to connect these efforts with software and supercomputing in an ecosystem called the NIH Commons. And all of those efforts will make it easier to track datasets and give researchers credit for them.

On the publishing end, there are changes afoot as well—changes designed to incentivize linking data to publications, publishing data and metadata in data journals, and even redefining what it means to
publish scientific results. Some of the most interesting thinking in this area is coming out of FORCE11, the Future of Research Communications and e-Scholarship, a grassroots organization dedicated to transforming scholarly communication through technology, where people are floating lots of ideas for changing the incentives around data sharing.

With projects launched on all fronts, it’s still unclear how things will play out. If the work in progress can solve data’s identity crisis, it just may have a significant effect on the reproducibility of scientific research as well.

Who Am I?
Giving Data a Name

For datasets to be FAIR, they have to have a name—a way to distinguish them from all other datasets. For some digital objects, such as scientific publications, the DOI (digital object identifier) has become standard. Some research groups have also started issuing DOIs for data and software. In Europe, researchers often use URIs (Uniform Resource Identifiers) issued by identifiers.org under the auspices of the European Molecular Biology Labs (EMBL–EBI) as part of the MIRIAM (Minimum information required in the annotation of models) Registry, a catalog of data collections. But DOIs and URIs are not the only globally unique identifiers out there.

“There are various camps,” says Bandrowski. Those in what she calls the “ontology camp” would like to see identifiers for each version of a software tool or dataset. Such an approach would be beneficial when researchers want to reproduce another group’s research results—i.e., they might need to know the version of a dataset or software tool that was used. But it could also get cumbersome pretty quickly.

There’s also “a less granular camp,” Bandrowski says, that would argue for a simpler system—giving unique identifiers to the data associated with particular funding efforts. This is the approach Bandrowski’s group has taken with RRIDs (Research Resource Identifiers). They provide a funder, such as the NIH, with a way to track the impact of a project.

It’s also possible that multiple options can co-exist. “NIH wants to foster the community discussion and watch for coalescing around it,” Bonazzi says.

Regardless of where the community ends up, all agree on the need to incentivize researchers to identify their datasets. The question is: What incentives will actually work? “Maybe if all the publishers said, what’s your RRID, then data would be more trackable,” Bandrowski says. She collaborated with others to run a pilot project through the Resource Identification Initiative—a Working Group that is linked to FORCE11—to test that idea. They convinced 25 editors in chief of scientific publications to require authors to use RRIDs (unique identifiers for the reagents, tools, data, software and materials used to perform experiments) in the methods section of their research papers. “Journals were willing to buy in because they care about reproducibility,” Bandrowski says. It worked: Authors obtained the required IDs, especially when the journal editors were persistent about checking the IDs; and additional publishers signed on to the requirement.

Why was it successful? Bandrowski has a theory:
RRIDs are kind of like an H index—the system used to measure an individual’s impact as a researcher. They offer a way to give credit to researchers for creating datasets and software.

The RRID system is also robust, with a strong registry stamp behind it as well as long-term financial backing, Bandrowski says. “There has to be a living entity that takes care of these things,” Bandrowski says. The RRID is an accession number that points to a registry page that lists the digital object’s funding, description, and people in charge. Automated checkers determine if a link is dead or live. “If it’s down for two to three weeks then a human looks,” Bandrowski says. If it’s permanently gone, then the registry page is changed to say that—“so you don’t get the 404 error message,” she says.

Members of FORCE11 are still laying the groundwork for data referencing to be done consistently across all the different journals. “We’re trying to see if all you need is a number but there may be other things that would make it a lot easier in the future,” Bandrowski says. “We’re bringing people together to see what they come up with.”

**Where Did I Come From? Metadata Made Simple**

A data identifier allows a dataset to say, in essence, “Here I am, I am unique.” But it doesn’t describe what the researchers did to gather the data. What laboratory procedures did they use? What machines took the measurements, determined sequences, or collected images? What do certain data fields or acronyms mean? All of that is opaque to the viewer of the data itself unless someone has annotated the dataset with metadata—a detailed but concise description of how the data were collected and what they represent.

But if researchers haven’t jumped at getting data identifiers, just imagine their reluctance to create metadata in a standard format. Again, incentives matter. “There’s no great reward for doing a good job of annotating the data to be useful for others, says Mark Musen, MD, PhD, professor of biomedical informatics at Stanford University. In fact, he says, there’s a disincentive—the fear of being scooped, or of others finding results you could have gotten yourself. So what could be done to change that?

Again, publishers are playing a role. “They are trying to be agents of change,” says Susanna-Assunta Sansone, PhD, associate director of the Oxford e-Research Center at Oxford University. One option is the so-called “data journal,” which may take many forms. The open source journal *GigaScience*, for example, requires that all supporting data and source code be publicly available and hosted in the journal’s database and cloud repository. And the primary article type in *Scientific Data*, a data journal from Nature Publishing Group, and Elsevier’s *Genomic Data*, is a data descriptor, designed to make data more discoverable, interpretable and reusable. Some data journals also include a machine-readable description of the dataset in addition to text. These efforts incentivize researchers to publish clear data descriptions—and since publication remains the currency of science, they also spread a little bit of the wealth to those who gather, curate, and wrangle data.

Funding agencies could also play a role in shifting incentives. “What if, to submit a new grant application you have to document that you did the right things with your old data?” Musen wonders. “That might have some teeth!”
So far, granting agencies haven’t taken that approach—yet. The NIH is, however, investing in metadata infrastructure as part of its push for data to live up to the FAIR principles. For example, the Center for Expanded Data Annotation and Retrieval (CEDAR), a Big Data to Knowledge (BD2K) Center of Excellence for which Musen serves as principal investigator (PI), is building tools to streamline metadata creation.

After one year in business, CEDAR has a prototype of a user interface. “We’re creating a library of hundreds of templates, each for a specific kind of experiment or experimental subject or specific instrument,” Musen says. The templates are designed to incorporate standards established for a particular field—standards that have been curated at BioSharing.org, a registry of more than 600 domain-specific minimal information checklists that is run by Sansone’s group at Oxford. But, importantly, researchers will be insulated from the technical details of the standards.

“The idea is that you will be guided,” says Sansone, a CEDAR co-investigator. “The system will intelligently create the template, customized to the needs of the researcher and the dataset, with the standards hidden from view.”

How Can You Find Me?
Data Discovery via DataMed

Good metadata is a first step toward data discovery. The next is an index and a search engine that can find that metadata in response to a researcher’s query.

Plenty of domain-specific indices have been created over the years, and many more are still being built and supported. “Unfortunately,” says Bandrowski, who helped develop an index called the Neuroscience Information Framework (NIF), “nobody comes to these things.” She and her colleagues received positive feedback from researchers whenever they publicized NIF at neurosciences conferences. Yet the next year, they would realize NIF had been forgotten. Her thinking now: “You have to meet the biologists where they are, which is PubMed.”

The latest data indexing effort funded by the NIH may do just that. It’s a data discovery index being developed by bioCADDIE (biomedical and healthcare Data Discovery Index Ecosystem), a BD2K Center of Excellence. “bioCADDIE is doing for data what PubMed is doing for literature,” says Sansone, who is on the bioCADDIE executive and steering committees. “We’re calling it DataMed.”

Lucila Ohno-Machado, MD, PhD, professor of biomedical informatics at the University of California, San Diego, who is principal investigator of bioCADDIE, calls the center “an integrator of multiple indexing efforts.”

Currently, if researchers go to PubMed to look for research in a specific subject area, they can explore all the publications out there. They cannot, however, go to a single place to explore all the datasets in the world, says Sansone. DataMed will be designed to allow that kind of exploration—with richer filters and data-specific browsing fields than the ones that are currently available in PubMed, she says.

It’s an ambitious goal. For the first phase, bioCADDIE has developed a unified way of describing datasets that connects up nicely with the CEDAR metadata templates. In order to map it to the databases that already exist, the team is working with the largest data repository managers. “It needs to become a cultural thing like PubMed indexing,” Sansone says. Just as journals create a JATS (Journal Article Tag Suite) file for indexing in PubMed, database creators would create a DATS (Data Tag Suite) file for indexing in DataMed.
Don't expect DataMed to allow searches at the level of molecular queries. "It will be able to retrieve datasets or point to another index, but not able to query on gene expression," Sansone says. "You will be able to narrow things down, but you still have to go out to the actual datasets." For example, a researcher might ask for all datasets of Alzheimer's patients that have RNA-seq, behavioral and imaging data available. Or they might ask for all proteomics and metabolomics datasets related to a specific biological process. Or for all data related to the effect of stress on health.

The bioCADDIE data discovery index is very much a work in progress. "It's all in discussion. It's all happening right now," Sansone says. The team expects to release a prototype in the summer of 2016.

Some researchers argue that DataMed should just be inside PubMed. "Scientists live in the literature almost every day," says Maryann Martone, PhD, past president (2014-15) of FORCE11. Much less frequently, they might be looking for datasets or software programs. "Let's start projecting things into where people actually are as opposed to expecting them to know that we exist," she says.

**Why Do I Matter?**

**Linking Data to Publications**

At some point in the future, perhaps datasets will have identifiers and associated metadata, be located at reliable addresses, and be findable in DataMed. But there remains the question of what those data are telling us. What knowledge has already been extracted from them?

Essentially, datasets need to be linked to publications. All the data identifiers and metadata in the world can't make that happen unless journals require authors to build those links and share their data. As mentioned above, journals have been stepping up their data-linking requirements. But many are not stringent enough about checking that datasets have been submitted to a repository that will live on after the project, Sansone says. "It's a slow process," she says; "but it is happening."

Some publishers and researchers, including a working group at the Research Data Alliance, are also pushing beyond bi-directional linking from data to publications. They'd like to see an overarching service that can combine links from different sources into a common "one-for-all" service model.

Links between publications and data could also cause a beneficial side effect: The ability to give people credit for the value of their data. "If someone generates data that got used 5,000 times or was cited 300 times, there will be a way to recognize that," Bonazzi says.

Some researchers think more significant changes are afoot. "In a recent perspective article in *Nature*, Philip Bourne, PhD, Associate Director for Data Science at the NIH, and his co-authors wrote: "There is an unnecessary cost in a researcher interpreting data and putting that interpretation into a research paper, only to have a biocurator extract that information from the paper and associate it back with the data. We need tools and rewards that incentivize researchers to submit their data to data resources in ways that maximize both quality and ease of access."

Musen would go further. "Ultimately, publication in science will have to move from prose to something machine processable," he says. "People don't pick up journals anymore and get cozy with them."

While the tools to make this transition do not yet exist in a realistic way, ideas along these lines have been percolating for a long time, especially within FORCE11.

One reality of the current publication system is its inability to deal with change over time, Bandrowski notes. "Essentially, you have these immutable objects [papers] that are referencing things that are changing all the time," she says. Databases grow, knowledge shifts, but papers remain static. There's a disconnect between "the flowing river of the Web and these stable objects that are publications—like rocks in that river," she adds.

Martone agrees: "The minute you publish something or put a dataset out there, there's already something you can say that you didn't say." These days, she's working on an effort to allow instantaneous annotation of anything on the Web using an open

---

The publishing data services working group at the Research Data Alliance is promoting a move from bilateral arrangements between data centers and publishers toward common standards and one-for-all services.
source tool created by Hypothes.is, a nonprofit for which Martone serves as director of biosciences and scholarly communications. Upon selecting text, on an existing Web page, users of the plug-in open a dialog box where they can enter whatever they want—a hyperlink, updates, additional information, tags. “It gives you the capacity to create searchable knowledge,” Martone says. She thinks the plug-in can help fix some of the structural problems in biomedicine. For example, it allows people to open up independent communication to update the literature. “Hypothes.is allows scaling of content at the time rate that science happens,” she says.

Martone imagines that eventually the links to and from various updates and tags will be data themselves. “We have to be able to read these signals much as Google reads the signals of links. We just have to figure out what those signals mean.”

For now, users can install the Hypothes.is plug-in in their browser. “It should be built into everything we have,” Martone says. “We’d love it built into PubMed and other browsers.” In fact, Hypothes.is is organizing a new coalition: Annotating All Knowledge (https://hypothes.is/annotating-all-knowledge/), that is bringing together publishers and other stakeholders to bring this capacity to all scholarship. “The challenge now is letting people know these capabilities exist,” she says.

The Commons
An Ecosystem for Data Sharing

In October 2014, Bourne announced plans to create the “NIH Commons” to catalyze the sharing, use, reuse, interoperability and discoverability of shared digital research objects, including data and software.

Bonazzi diagrams the Commons as a layered system consisting of three primary tiers: high performance and cloud computing (at the bottom); data, including both reference datasets and user-defined data (in the middle); and (at the top) services and tools, including APIs, containers, and indexes (DataMed, for example), as well as scientific analysis tools and workflows and—eventually—an app store and interface designed for users who are not bioinformaticians.

To be eligible for use in the Commons, data and software will have to meet the FAIR principles. To make that easier for researchers, the products of all of the BD2K centers will be part of the Commons ecosystem, including DataMed from bioCADDIE and streamlined metadata templates from CEDAR. And to incentivize participation in the Commons, the NIH plans to offer cloud computing credit vouchers that researchers can use with a provider of their choice, so long as the provider complies with the FAIR principles.

The Cloud Credits Model, as it’s being called, “democratizes access to data and computational tools,” said George Komatsoulis, PhD, (acting) chief of the informatics resources branch at the National Center for Biotechnology Information (NCBI), when he spoke at the BD2K All Hands Meeting. Right now, researchers access cloud computing with a credit card or through a university, he said. Komatsoulis anticipates that the voucher system will be more cost effective by creating a competitive marketplace for biomedical computing services and reducing redundancy. The voucher system is now being piloted in specific research areas—the Genomic Data Commons, for example. “Credits will be distributed the way the National Science Foundation distributes access to specific facilities such as light sources,” Komatsoulis said. “But having an existing NIH grant will be a precondition.”

With the Commons, the NIH is feeling its way toward a viable ecosystem for the sharing of big data. “We’re testing pieces of the ecosystem out,” Bonazzi says. “Does this make sense? What are the pieces that are missing? What still needs doing? And how do we facilitate the community to do those?”

The NIH doesn’t want to be in a position of saying here’s the infrastructure. “That’s not going to work,” Bonazzi says. “I’m not claiming this is it. I’m saying this is what I’m seeing the community is doing. If this is one step toward coalescing a concept around how we do biomedical science in the future, and that’s useful, then let’s use it as a point of discussion.”