Genestack

Metadata, FAIR principles, and their importance in genomics

2019

There is an urgent need to improve the infrastructure that supports the reuse of scholarly data. In response to this need, the scientific community has put together the FAIR guiding principles of data management: that data should be findable, accessible, interoperable, and reusable.

In this e-book, you will learn what metadata are, why they are important, the challenges associated with metadata, more on the FAIR principles, and why, as a scientific community, we need to improve the way we take care of public and proprietary data.

What are metadata?

Metadata are data that provide information about data. Metadata exist to give data context. For example, a photograph is the raw data. Additional parameters, like the date and location the photo was taken, the color profile, lens, and aperture used, constitute the metadata.

In genomics or transcriptomics, metadata describe the sample that the DNA/RNA sequence was obtained from, e.g. the organism, the cell line, and the library-preparation method.

Why are metadata important?

Metadata allow you to make better use of your data. With omics data, for example, metadata are essential to correctly interpret the data and carry out meaningful comparisons with data from other samples or studies.

There is almost never a good reason not to save metadata – there are minimal storage requirements compared to the data they describe, and they offer a number of advantages:

- **Data reusability and reproducibility**
  You need metadata to reproduce your data correctly. For example, if you want to decode a picture with colors as they were intended, you need to use the right color profile (e.g. sRGB). If you want to reproduce DNA/RNA sequencing data or compare it with data from experiments in other studies, you need to know what cell line and experimental conditions were used, what treatment the cells or tissues were subjected to, etc.
- **Search for and find stored data**
  You may want to find a picture of interest, say from a holiday trip last year. Most likely, you've copied the picture over from your smartphone along with a hundred others. If you can remember the date the picture was taken, you can use that as a criterion in your search.

Similarly, metadata allow you to find and retrieve your data by searching for experiments by criteria like the cell line or organism, a certain sample treatment, or by a specific person or department involved. As long as the metadata have been input and saved, you can retrieve the associated data.

- **Data integration**
  Metadata also provide an easy path to integrate your omics data into your proprietary or public repositories. Repositories often rely on fields to be filled in by the researchers adding their omics data. Metadata fields, when accurately and consistently entered into your repository, allow users, and data management software to index, access, and recall the data entered.

- **Interpretation of data**
  You may be able to draw conclusions on the contents of the data without needing to access and analyze it directly, or you can make use of the full context to be able to interpret the data correctly. If you took a lot of pictures around the time of sunset you can make a reasonable guess that the weather was clear that particular evening. If you know the sample came from a cancer cell line, you can relate the data to a pathological state.

Metadata in omics

The omics industry (i.e. genomics, transcriptomics, proteomics) is rather unique when it comes to the data it uses. Data can be usually classified by the effort required to produce a unit and the reusability of those units, i.e. data are easy to produce and easy to reuse, vs data are difficult to produce and difficult to reuse. Most high-volume industry data fall into one of these categories:

The data are easy to produce and easy to reuse

A database of customer purchases for a store loyalty card can intuitively be broken down into rows. Each row is a unit within the database, which is easy to extract and manipulate. The same data could be filtered for a study into seasonal buying patterns or used to build detailed customer profiles. Likewise, each unit is produced automatically and with little intervention whenever a purchase is made by a customer. Data like these are easy to produce and re-use.

The data are an effort to produce and difficult to reuse

An example of this would be a complex simulation for a new piece of equipment. It takes thousands of man hours to produce and many hours of processing time to run. Since accuracy is so important in such simulations, and the end result is often both complex and bespoke, the reuse potential for data like these is limited.

Omics seems to be an exception to the rule: **data are an effort to produce, but in theory, easy to reuse.** The effort to complete a biological experiment is relatively high. For example, to find out the side effects of a new compound, researchers often employ differential gene expression experiments on cell lines. These are time- and resource-intensive to perform, but once you have

the data, you can reuse them repeatedly. Using these data when investigating compounds with related molecular structures, you could narrow down the scope of additional experiments you need to perform in a new study, thereby saving money and time.

Absolutely vital to making omics data reusable, is retaining the full experimental details used to generate the data – and this is why metadata in genomics are so important.

Consistency and accuracy: challenges with metadata

Naturally, metadata are not a cure-all for data management if not used correctly. There are two main problems associated with metadata that stand in the way of data discovery and reusability: issues with consistency and accuracy.

Accuracy is self-explanatory – a mistake when entering metadata, means you no longer have an accurate description of your data. Inaccurate metadata cannot be reliably searched for, which can make retrieving the associated data significantly harder, or potentially impossible.

Consistency here refers to consistent metadata fields for similar data, and consistent vocabulary within a field. Ideally, you want to describe your data with the same set of characteristics (fields) so you can query all of your data at once when performing a search. At the same time, you want to always use identical strings to describe a feature of your data, e.g. if you store half your photos as .jpeg and the other half as .jpg extensions, a search for .jpg will only return half of your data. An example with biological data is the species name: if you describe human data as "Homo sapiens" once, and "human" another time, you are going to run into problems with data search later on.

Ontologies

You can use an ontology to ensure your metadata are consistent. An ontology is a structured collection repository of possible entries along with the properties and relationships between them. The aim of an ontology is to unambiguously define the terms you use to describe objects. It is also a way to aggregate the knowledge you have on that object from many disparate sources and presenting it in a format that is easy to read for both humans and machines.

An example of an ontology is Cellosaurus, which describes cell lines used in biomedical research. Each entry contains several key pieces of information.

1. Name: recommended and synonymous.
2. Origin: species, tissue, disease, details from the original paper, etc.
3. Structured comments:
    - Common problems, such as contamination
    - Genetic modifications
    - Groups the cell may belong to, e.g. cell catalogs or cell panels
4. Cross-references: links to other databases, e.g. other cell line catalogs or resources that use cell lines in their definitions, such as ChEMBL for the effects of compounds.

Tagging a research sample with an ontology term allows you to efficiently integrate with a multitude of other resources. For example, synonymous names are a common headache when searching experiment databases – a pain that's quite easily solved by implementing an ontology, which allows searches to automatically return results for all synonyms.

The ontology can also be browsed to find cell lines that could be of interest. You could filter ontology terms by criteria like tissue and disease of interest, and then use the corresponding cell lines when searching for experiments in relevant repositories.

Finally, ontologies are fully open access, with a core group responsible for their curation. This means that anyone can suggest new entries and point out errors or additions, with quality controls in place for consistency.

Maintaining consistency

In theory, maintaining consistency and accuracy sounds simple, and to an extent it is – at least, it is at the start. When you first decide on a metadata vocabulary, you are looking at data you already have. You are likely able to intuitively predict the sorts of fields and vocabulary you will want to use. But, thinking about the future, as additional data become available, it may become obvious that a field you previously thought was unnecessary (or perhaps didn't exist) becomes needed. Likewise, older vocabulary may no longer provide the granularity you require – terms may need to be split or merged. Managing this branching structure becomes more and more difficult over time.

Mistakes are increasingly likely to slip in when you have more fields to fill and vocabulary terms to choose from, especially if your pipelines aren't automated. In everyday life these problems are often mitigated by the nature of the data we use. Metadata are usually provided automatically by the services and capture devices we use. In most situations outside of research, it's not much of a problem if our metadata aren't perfect – needing to spend an additional five minutes searching for the particular piece of music you wanted to listen to isn't going to ruin your day. However, the stakes are a lot higher when data volumes, accuracy, and speed requirements increase to the levels required by researchers in industry.

Public data and the metadata inconsistencies.

There has been a great push to encourage scientists to publish their raw data, especially their genomics data, online in recent years. Most major journals require authors to deposit their sequencing data into one of the major genetic sequence archives. GenBank, EMBL-Bank, and DDBL are examples of repositories for processed sequence data, while the Sequence Read Archive (SRA) stores raw sequencing data. To get a sense of the scale of these databases, as of August 2019 GenBank stores more than 213 million sequences, ranging from individual viral genes to whole eukaryotic genomes, and has historically doubled in size every 18 months.

Conflicting standards

It's important to note that while the major repositories are well-curated and deeply integrated, they do not accept all datasets or even all data types. Many other special-purpose repositories have emerged that capture, for instance, traditional low-throughput bench science data. The standards of metadata accepted here are less strict and the repositories place fewer restrictions on what data can be deposited. For instance, guidelines from the BBSRC, which is the lead UK life sciences funding agency for non-human studies, state:

"Data should be accompanied by the contextual information or documentation (metadata) needed to provide a secondary user with any necessary details on the origin or manipulation of the data in order to prevent any misuse, misinterpretation or confusion. Where standards for metadata exist, it is expected that these should be adhered to."

The standard of metadata required is therefore left up to reviewers and funding body advisors, which will inevitably vary both within and between organizations. Database providers, scientists, and the funding institutions they all ultimately answer to have different roles when it comes to managing public research data. Scientists are the ones who produce and use the data, so it is them who most often define metadata standards.

Yet, with the study fields so diverse in topic and scale, there are many competing and overlapping standards. Database providers have the most potential reach when it comes to enforcing metadata requirements, but do not have the resources to police them. Research funding agencies are in a much better position to ensure that the right metadata standards are used. Such funding agencies have access to the large and varied networks of scientists that are required to vet standards and have the power to enforce them through grant conditions. However, even here, the diversity of standards and individual use cases means that guidelines are not always strictly worded.

Efforts to improve standards

The Minimum Information About a Microarray Experiment (MIAME) standard for recording and reporting microarray data, first described in a 2001 paper (Brazma A., et al. Nature Genetics, 29, 365–371) and currently the dominant standard for microarray experiments, shows that a different, better scenario is possible. It has six critical elements:

1. Raw Data (Raw): data extracted from imaging files.
2. Processed Data (Processed): final normalized data.
3. Sample Annotation (Experimental Factor Value / Variables): what experimental conditions were each sample subjected to?
4. Experimental design (Experimental Factor Value / Variables): why were these particular samples processed?
5. Array design details (Platform): e.g. probe sequences and where they hybridize.
6. Protocols (Protocol): what experimental protocols were used in the laboratory and for data processing?

The standard outlines both data and metadata that should be recorded and subsequently reported. The creators of MIAME hoped that this consistent reporting would help to develop databases, public repositories, and data analysis tools.

An approach to managing metadata consistency and facilitate adherence to the MIAME standard is the MicroArray Gene Expression Tabular (MAGE-TAB) tool. Designed for both data collection and exchange between other databases and tools, MAGE-TAB aims to help scientists communicate functional genomics experimental data, in a standardized manner. There are many potential fields in MAGE-TAB that can be filled, and it is often up to reviewers to decide if submitted data sufficiently answer the questions set out in the rating criteria. Users of private or public repositories where such data is stored can choose to filter experiments based on raw metadata and/or on the reviewers' scores of submitted data.

Despite MIAME being a good effort to standardize data recording, there are still inconsistencies in data standards and varying ability to enforce them. This makes the overall genomics data ecosystem somewhat disorganized – which in turn makes data discovery and reusability, by both humans and computers, increasingly difficult. This is in contrast with, for example, environmental sciences, where metadata standards are defined more strictly (e.g. for NERC). However, there is a real drive to improve this and a great deal of work is being done to develop more robust and standardized systems for recording and reporting metadata.

The need for a genomics data and metadata management system

In order to improve the current metadata situation, a diverse set of stakeholders – including academia, industry, funding agencies and scholarly publishers – have come together to formulate a set of principles that will serve as guidelines for researchers wanting to enhance the reusability of their data.

The FAIR principles

The FAIR principles aim to overcome data discovery and reuse obstacles by ensuring that data and metadata are (F)indable, (A)ccessible, (I)nteroperable, and (R)eusable. The principles should apply not only to data in the conventional sense, but also to the algorithms, tools, and workflows that lead to the data. A particular emphasis has been put on making the data discovery and reuse easier not only for humans, but also for machines.

The public data landscape is changing, with many repositories already showing some degrees of "FAIRness", each with its own technological implementation for the different aspects of FAIR, e.g. Dataverse, FAIRDOM, and UniProt. There are even emerging projects in which FAIR is a key objective, e.g. bioCADDIE (a project to index biomedical data across data repositories and aggregators) and CEDAR (tools enabling the creation of metadata templates that follow community standards with vocabularies and ontologies integration).

When FAIRness falls down

Despite best efforts, some of these new repositories fall short in their "FAIRness" goals in one aspect or another, particularly when it comes to core challenges in multi-omics research: integration of external and internal data and computational pipelines. Not only do we, as researchers, want to easily access data, we also want to be able to easily process them. And we want to do this in combination with private or public datasets using a mixture of private or public tools, while keeping track of their complete provenance. The provenance of processed data not only facilitates reproducibility and reusability, but it also lends itself to the aggregation of knowledge, i.e. meta-analysis.

This is less of an issue for researchers in academia since they operate on a scale that can tolerate a degree of inefficiency in their integration. However, that does not hold true as data volumes and collaboration requirements increase. Industrial R&D departments have to focus on maintaining existing tools and creating point solutions, but often lack the time and resources to fully integrate all of their resources. At Genestack, we help companies to do just that through off-the-shelf software as well as a suite of flexible custom-development services. Our areas of expertise extend from data management to system integration, visual analytics, and scientific consultancy.

A software solution – Omics Data Manager

We recently launched the enterprise software Omics Data Manager (ODM) that lets you record linked multi-omics data along with rich metadata and relationships – ODM becomes the Single-Point-of-Truth for all of your multi-omics data management needs. Data is imported into ODM with rich metadata following strict templates with support for public and private ontologies to ensure consistency. Users can seamlessly search and query diverse private and public data using features such as ontology-based autocomplete and suggested common field values while maintaining strict access control. ODM records and retain relationship information and tracks file provenance.

We've also covered the essential integration aspects. ODM easily integrates into individual environments, like LIMSs, as well as other common platforms. The entire system is supported by our expert team of bioinformaticians and software developers.

To learn more about Omics Data Manager
visit www.genestack.com/products/omics-data-manager

The future is FAIR

Data only have value in the context of what we already know. Metadata allow us to build links between disparate blocks of detailed information and put it into perspective. However, creating a unified framework for metadata is challenging as our knowledge as a whole is constantly changing. Life science research and healthcare stand to benefit greatly from the opportunities afforded by this integration, but also suffer from the challenges associated with the fast-paced evolution of knowledge that NGS and multi-omics has recently injected into such fields. Di-

verse databases and data standards are competing and evolving in the omics space while users struggle to build compatibility between them. To tackle this and advance the field, we need to embrace the FAIR principles, so that, as a community, we can benefit from power that metadata offer.

Software built with intuitive integration in mind will allow large-scale omics datasets to be harnessed efficiently and integrate proprietary data with the wider omics landscape. Genestack exists to do just that.

Acknowledgements

2017 version written by Nemunas Antanavicius, edited by Kalina Cetnar and Kevin Dialdestoro.
2019 version updated by Davide Mantiero and Bill Hinchen.

At Genestack we develop software to help life-science researchers and R&D scientists manage multi-omics data. To learn more about our products and services visit www.genestack.com

Check out our content library

where you can find more guides, ebooks, and white papers on data management.

https://genestack.com/resources/library/

Contact us

info@genestack.com

www.genestack.com