Genestack

The importance of metadata in genomics and the FAIR principles

2017

The urgency to improve the infrastructure supporting the reuse of scholarly data has been increasingly acknowledged by the entire scientific community, which is evidenced by the recent formulation of the FAIR principles. In this e-book, we'll first explain what metadata is and why do we bother with it. We'll then discuss the problems associated with metadata, the FAIR principles, and the need for improvement of how do we, as a scientific community, take care of public data.

What is metadata?

To put it simply, metadata is information... about information. For instance, a conversation could be described with metadata. Where did the conversation happen? Who was talking? How loud was the conversation?

We are most used to talking about metadata in the context of stored information. When was this picture taken? By whom? What was the color profile used to encode the image?

In short, metadata exists to give data context.

Why do we bother with metadata?

The primary reason for saving metadata is to enable you to use your data better.

- Sometimes you need metadata to reproduce your data correctly. For instance, if you want to decode a picture with colors as they were intended, you need to use the right color profile (e.g. sRGB).
- Alternatively, you may want to find a picture of interest, say from a holiday trip last year. Most likely you have copied the picture over from your s martphone along with a hundred others. If you can remember the date you can use it as a criterion in search. If you can't, you probably have it in your calendar somewhere. As you can see, metadata can enable you to find a large unit of information based on small hints.
- It also provides an easy path to integrate your data into your repositories.
- In addition you may be able to draw conclusions on the contents of data without needing to access and analyze it directly. 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.
- Finally, there is almost never a good reason not to save metadata - it has minimal storage requirements compared to the data it is describing.

Metadata in genomics

Genomics industry is rather unique when it comes to the data it uses. Genomics data can be classified on two scales - effort to produce a unit, and re-usability of units.

Let's look at examples of both scales. 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 for seniors, 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. This sort of data is easy to produce and re-use.

The polar opposite of this is data that takes effort to both produce and 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, reuse potential is limited.

Most high-volume industry data falls into one of these two categories. Genomics seems to be an exception to the rule: it's an effort to produce, but in theory, it's 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 effort-consuming to perform, but once you have the data, you can reuse it over and over again, to help you in the future. Using the old data for related compound structures, you could narrow the scope of the experiments you now need to perform.

The absolute necessity in making genomics information reusable is retaining full details of the experiments used to generate the data – and this is why metadata in genomics is so important

Consistency and accuracy: problems with metadata

Naturally, metadata is not a cure-all for data management if it isn't used correctly. There are two main problems associated with metadata, standing in the way of data discovery and re-usability: consistency and accuracy problems.

Accuracy is self-explanatory - if you make a mistake when entering metadata, you won't have an accurate description of your data.

Consistency comes in two flavours - 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 your data at once. At the same time, you want to make sure you are always using the same strings to describe a feature of your data, e.g. if you store images with .jpeg and .jpg extensions, a search for .jpg will not return all the pictures stored. When it comes to biological data, you have be careful when entering, for example, the species name – if you describe human data as "Homo sapiens" once and "human" some other time, this will create problems with data search.

One way to ensure metadata consistency is via the use of a metadata ontology. An ontology is a structured collection repository of possible entries along with the properties and relationships between them. The aim is to unambiguously define the terms we use to describe objects. It is also a way of aggregating the knowledge we 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 is used to describe cell lines used in biomedical research. Each entry contains information such as :

1) Name - recommended and synonymous.

2) Origin - species, tissue, disease, details on the original paper, etc.

3) Structured comments -

- Common problems such as contaminations
- Genetic modifications
- Groups the cell may belong to, e.g. cell catalogues, cell panels

4) Cross-references - links to other databases e.g. other cell line catalogs, or resources which 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 frequent headache when searching experiment databases – a pain that's quite easily solved by implementing an ontology, which means the searches return results for all synonyms automatically. The ontology can also itself be browsed to find cell lines that could be of interest. You could filter ontology terms by e.g. tissue and disease of interest, then use the resulting cell lines when searching for experiments in relevant repositories. A final benefit of ontologies is that they are available publicly in full, with a core group responsible for curation. This means that anyone is able to make suggestions for new entries and point out errors or additions for existing ones with controls in place for consistency.

In theory maintaining consistency and accuracy sounds simple, and to an extent it is, at least 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 becomes 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 needed - terms may need to be split or merged. Managing this branching structure becomes more and more difficult over time.

Mistakes are more likely to slip in when you have more fields to fill in 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 is usually provided automatically by the services and capture devices we use. In everyday life, it's not that much of a problem if our metadata isn't perfect - needing to spend an additional five minutes searching for the particular piece of music we wanted to listen to isn't going to ruin your day. However, the stakes are a lot

higher when data volumes and accuracy and speed requirements increase to the levels required by industry.

Public data and the metadata inconsistencies.

There has been a great push to encourage scientists to publish their raw data, especially the genomics data, online, in recent years. Most major journals require authors to deposit their sequencing data in 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, currently GenBank stores just under 200 million sequences, ranging from individual viral genes, to whole eukaryotic genomes and has historically doubled in size every 18 months.

It's important to note that whilst the major repositories are well-curated and deeply-integrated, they do not accept all datasets or even data types. Many other special-purpose repositories have emerged capturing, for instance, traditional, low-throughput bench science data. The standards of metadata accepted 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 organisations.

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 define metadata standards. However, 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 man-power to police them. Research funding agencies are in a much better position to ensure that the right metadata standards are used. They have access to the large and varied networks of scientists required to vet standards, and have the power to enforce them through grant conditions. However, even here, the diversity of standards and use cases means that guidelines are not strictly worded. As a result, the overall genomics data ecosystem is a hot mess, which discovery and reusability of data by both humans and computers being increasing hard. This is in contrast with, for example, environmental sciences, where metadata standards are defined more strictly (e.g. for NERC).

The example of the Minimum Information About a Microarray Experiment (MIAME) standard, first described in a 2001 paper and currently dominant 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 normalised data.

3) Sample Annotation (Experimental Factor Value / Variables) - what experimental cond tions was each sample subjected to?

4) Experimental design (Experimental Factor Value / Variables) - why were these particular samples processed?

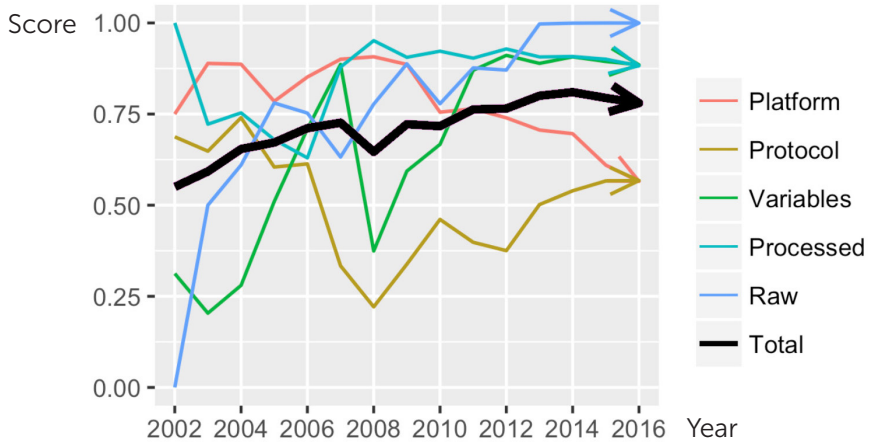
5) Array design details (Platform) - e.g. probe sequences and where they hybridize.

6) Protocols (Protocol) - what experimental protocols were used in the laboratory and for data processing?

As you can see, the standard outlines both data and metadata submission. The six criteria outlined in the original MIAME paper roughly map to the five scored fields shown in brackets. A suggested metadata format for ideal adherence is MAGE-TAB. You can read the specification for MAGE-TAB here. There are many potential fields that can to be filled and it is up to reviewers do decide how well submitted data answers the questions set out in the rating criteria. Users of the database can choose to filter experiments based on raw metadata, or on the reviewers' scores.

Average MIAME total quality scores have increased over time for experiments submitted to Array-Express. The graph illustrates some of the problems to do with designing a long-term metadata standard from the previous section. For example, there is a dip in scores around the year 2008 which roughly coincides with the introduction of the MAGE-TAB format in 2007. This likely lead new submissions to be rated more strictly to encourage adoption, while old submissions were unlikely to be re-rated to match the new standard.

Arrey Express MIAME scores



The need for genomics data and metadata management system

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

The FAIR principle 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 leading 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 are already showing some degrees of "FAIRness", each with its own technology 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 creation of metainfo templates that follows community standards with vocabularies and ontologies integration. However, each displays immaturity in its "FAIRness" trajectory in one aspect or another, particularly when it comes to a core challenge in genomics research: integration of external and internal data and computational pipelines. Not only do we want to easily access data, we also want to be able to easily process them, in combination with private/public datasets using a combination of private/public tools, whilst keeping track of their complete provenance. The provenance of processed data not only facilitate reproducibility/ reusability, but it also lends itself naturally to the aggregation of knowledge, a.k.a. meta-analysis.

This is not as much of an issue for researchers in academia since they operate on a scale that can tolerate integration that is inefficient. However, that no longer holds true as data volume and collaboration requirements increases. Industrial R&D departments focus on maintaining existing tools and creating point solutions, but do not have the time and resources to fully integrate all of their resources. Genestack develops a platform which allows companies to do just that.

Genestack provides users with a single point of truth for all of their genomics data management and processing needs. Our platform is built around the central idea of treating genomics data as meaningful objects. Data is imported onto the platform following strict metadata 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. A variety of common and custom bioinformatics tools are available through intuitive visual interfaces. Where possible, input file type and metadata is used to suggest which applications to use and to automatically set suitable parameters. Finally, the platform retains the detailed file provenance necessary for you to be able to reliably reproduce and re-use your results

To learn more about the Genestack platform, click here.

Conclusion

Data only has value in the context of what we already know. Metadata is a tool that allows us to build links between disparate blocks of detailed information and put it into perspective. However, creating a unified framework for metadata is challenging since our knowledge as a whole is constantly changing. The field of genomics stands to benefit greatly from the opportunities afforded by this integration, but also suffers from the challenges associated with the evolution of knowledge in a relatively new field. Diverse databases and data standards are competing and evolving in the genomics space while users struggle to build compatibility between them.

A platform built with intuitive integration in mind is needed to finally allow large-scale genomics datasets to be harnessed efficiently. Genestack works to be just that.

References: Written by Nemunas Antanavicius. Edited by Kalina Cetnar and Kevin Dialdestoro.