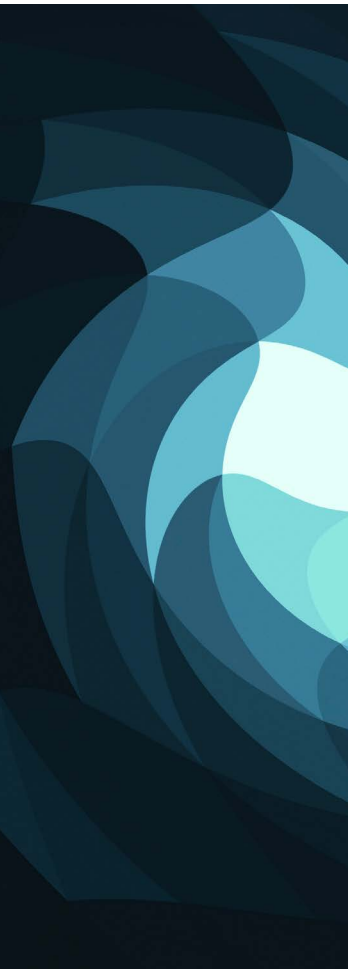
Enhancing data quality in knowledge bases: A FAIR approach

A FAIRy tale by **Manuel Lera-Ramírez** – a personal story about how I work with FAIR data



The FAIR principles are vital guidelines for data management and sharing across various scientific disciplines. While FAIR emphasises Findability, Accessibility, Interoperability and Reusability, correcting errors in (meta)data curation is an intrinsic part of these principles. FAIR data isn't just about adhering to standards; it's about ensuring that the data is accurate, consistent, and reliable. In this blog post, I describe the development of a software pipeline and API to fix and prevent errors in the context of a model organism knowledge base.

Genotype-to-phenotype annotations in knowledge bases

PomBase is the comprehensive model organism knowledgebase for the fission yeast Schizosaccharomyces pombe, which aims to standardise, integrate, display, and disseminate biological knowledge and datasets to the wider scientific community, making a wide range of datatypes from large and small-scale publications FAIR.

Among other things, PomBase hosts curated literature annotations, generated through community curation, a process that allows authors to curate their published research in an intuitive web interface. Annotations are then validated and discussed with professional curators. The involvement of both parties increases curation quality and allows researchers to become familiar with the supported data types.

One of those manually curated data types is Genotype-to-phenotype relationships, which capture the results of genetic experiments: they link the experimental conditions and the genotype of a query strain (the alleles present in that strain that are not present in the control strain) to a phenotype. These annotations contribute to our understanding of gene product functions, as well as the roles played by specific domains or residues within them, as they capture the phenotypes caused by specific residue modifications or truncations.

Challenges in manually input allele descriptions

In PomBase, allele descriptions that describe the sequence modifications present in an allele are not generated automatically, but manually input by curators as plain text. This manual input process, while essential for data curation, has historically introduced unique challenges to data quality. Two primary sources of errors have been prevalent:

1. **Syntax errors:** with allele descriptions input manually, the potential for human error naturally arises. Syntax errors and non-standard nomenclature often creep in, making it challenging to maintain data consistency and accuracy.

2. **Sequence inconsistencies:** beyond syntax issues, manual data input can result in discrepancies between allele descriptions and the actual gene sequences. These discrepancies may be a result of errors during input or can emerge over time as gene structures evolve. For instance, an allele described as "A123V" (alanine 123 replaced by valine) may be accurate at the time of input but then become incorrect if the gene structure is updated in the reference genome (e.g. if an intron is added). These inconsistencies can significantly impact the reliability of our data.

Transforming manual input into high-quality data

Recognizing the challenges introduced by manual input, we developed a software pipeline to address these issues head-on:

- **Leveraging historical data:** Our pipeline uses all previous S. pombe genome versions, which have a unique identifier and are publicly accessible (the first two FAIRs principles). This allows us to correct errors that were referring to old sequences, ensuring that allele descriptions map to updated sequences.

- **Standard file formats as inputs + adaptable grammar:** Our pipeline takes genomic data in standard formats as input, making it usable by others. However, it also relies on an adaptable grammar to identify, check, and format allele descriptions, since these are different for every organism and can evolve over time. As a proof of concept, we have run the pipeline on data from SGD (the Saccharomyces cerevisiae knowledge base). This demonstrates its applicability beyond our specific dataset, highlighting its potential to enhance data quality in other genomics projects.

- **Preventing future manual input errors:** in addition to correcting existing errors, we have reused the pipeline code to develop an API that checks allele descriptions against the latest gene structure information when they are added to the database, effectively preventing syntax errors and incorrect residue references from being included. This verification procedure represents a significant step toward maintaining data quality over time.

Unlocking new possibilities through improved data quality

The impact of these improvements extends beyond error correction:

- **Engineered allele sequence representation:** With standardised and correct allele descriptions, we know the positions in the protein sequence that produce a phenotype when mutated or removed. We now display this information in our protein feature viewer.

- **Standardisation to HGVS Nomenclature:** Our pipeline transforms our organism-specific allele descriptions, which are designed to be user-friendly for manual input, into the standardised HGVS (Human Genome Variation Society) variant nomenclature. HGVS nomenclature is hard to type by hand but is a recognised standard in genomics.

- **Structural changes in alleles:** AlphaFold has been previously used to predict the structure of naturally occurring protein variants and our next step is to use it to predict the structure of all our alleles, which are mostly produced via genetic engineering. This will be a large dataset that could be used to relate the changes in structure to the observed phenotypes.

In summary, we have improved the quality of PomBase data by correcting existing errors and setting up mechanisms to prevent the addition of new errors in the future. In doing so, we have made the data FAIRer by systematising allele descriptions and translating them into standardised HGVS variant nomenclature, overall making our allele dataset more Interoperable and Reusable. The software tools we have developed for this rely on standard file formats and leverage previous dataset versions, while allowing for organism-specific configuration, making them reusable by other genomics projects. The outcome of the quality control pipeline is a standardised error-free dataset that can be used to represent alleles in a comprehensive way in our website and for structural analysis.

Useful resources and training

1. PomBase site: https://www.pombase.org
2. SGD site: https://www.yeastgenome.org
3. HGVS variant nomenclature guidelines: https://varnomen.hgvs.org
4. Standardisation of allele descriptions for S. pombe: https://doi.org/10.1093/genetics/iyad143
5. Community curation in PomBase: https://doi.org/10.1093/database/baaa028
6. Canto, the community curation web tool used in PomBase: https://doi.org/10.1093/bioinformatics/btu103

Manuel Lera-Ramírez
An ELIXIR-UK Fellow from the 2nd Cohort (2022-2023).

As a biocurator in PomBase, I curate articles that use S. pombe as a model organism, often in collaboration with the authors. In addition, I develop the ontologies to capture that information and software tools to document historical changes and perform quality control in our data. In the last year, I have spent countless hours going back to old publications to fix errors in alleles in our database, so I decided to systematise the error-fixing process by building a software pipeline.

This work was funded by the ELIXIR-UK: FAIR Data Stewardship training UKRI award (MR/V038966/1) 6/1)