ELIXIR Converge - Demonstrators Scenarios for Data Management

Harmonised FAIR plant genotype & phenotype data management toolkit for Europe

From H2020 AGENT project

Michael is coordinating the work package that supports the data management of the project. He and his partners are managing a data repository for plant genetic and genomic data.

They want to ensure that the partners of the other work package will generate data FAIR at their birth.

Michael interacts at the start of the project with Vojtech and Ron who are leading the work packages respectively in charge of the phenotypic and genotypic characterization of the panels of wheat and Barley genetic resources.

Vojtech and Ron need to survey the progress in the production by 10 different genbanks of the seed lots that are necessary for their activities.

They need first to identify the common list of accessions/plant material studied.

Michael wants to take this opportunity to collect and check the metadata of the accessions studied and also the metadata in relation with the planned phenotypic measurements.

The three researchers want to use a common environment facilitating the collection and storage of metadata along with the different steps of data production.

Scenarios from the ELIXIR Knowledge Exchange Scheme w/ Phenospex

Tom is feature manager in Phenospex, a biotech company that develops and provides hardware and software for automated plant screening and plant phenotyping for various conditions. He wants to ensure that the data outputs follow the standards recommended by the Plant Science community. He knows that this would facilitate the uptake of their developments by the academic world but also that it might facilitate data interoperability in their workflow on the long term for industrial clients. He contacts the ELIXIR Plant Science community to get some advice and guidelines. Cyril and Daniel are members of the Plant science community, specialized in phenotypic data standards. They are very interested to check and possibly improve the standards they contribute to develop against real life examples of high throughput data and to get feedback from Tom.

From COST Integrape community

- Jérôme is a researcher studying the influence of environmental parameters in relation with climate change on grapevine berry development and composition. He has set up an experiment in which he studies a small panel of genetic resources in two different conditions, with or without drought stress. He observes phenological parameters along the season and samples berries at different stages to perform transcriptomic and metabolomic analysis. He wants to ensure that he keeps the link between the plant material in the field and their phenotypes and the results obtained in transcriptomics and metabolomics along his multiple steps of data processing and data analysis and in the long term, for other users. He needs help to identify the best environment of both data management and data analysis. He wants to be sure that his data are published in recommended archives for the long term and will be reusable. He needs help for data submission to these archives. For the transcriptome analysis, he needs reference data set about gene annotation and gene sequence for grapevine. He would also need better metadata in transcriptome analysis that are in international archives to reuse them in an easier and better way.

- Camille is a teacher at the university studying resistance to diseases in grapevine. Natural resistance is found in wild grapevine species and Camille is therefore developing comparative genomics approaches for very complexe families of genes involved in responses to pathogens. She produces new or updates of whole genome sequences and annotates them. She curates the annotations for the gene families she is interested in and compiles knowledge about when and where they are expressed using published or *de novo* transcriptomic data. She needs a smooth and always up to date workflow for sequence and annotation submissions to ENA to facilitate the publication of her work. She struggles with the different versions of genomes, genome annotations for the reference grapevine genome(s). She would like to be sure that she uses the best tools for her analysis (genome assembly, genome annotation, sequence alignment, phylogeny, …).

- Tomàs and Jérôme are both researchers. They have set up for the grapevine community in the frame of the COST action a training on transcriptomic analysis. The training is dedicated in priority (but not only) to PHD and postdoc students. They have contacted specialists from other plant communities in terms of data analysis but would like to improve their training course in terms of data management.

Scenario by an academic genotyping facility

Marie-Christine is running a genotyping platform. She interacts with a lot of teams, working on different plants. She helps them to choose the best technical platform to answer their research question at the best cost, she discusses the protocol for the DNA preparation and organizes the quality tests, the genotyping and returns the results to the teams. She sometimes also hosts some partners to help them for the SNP calling. She would like to ensure that she contributes at the best to the FAIR data management of her users, from the collection of the metadata samples to the SNP publication along with their provenance. She is not sure that her guidelines for data standardization are correct for plant samples. The users of her platform seem to have different interpretations or practices and the submission guidelines are not very well specified in the international archives.

Common Data management plans for the marine metagenomics Community

John is an early career researcher who is writing his first grant to enable him to generate metagenomics (genome) from seawater in an area of suspected pollution. He wants to find out how composition of species changed in the area of pollution. This is his first grant as a PI, and he is required by the funding body to write a data management plan (no more than 2 pages of text) to accompany his grant. Ideally, he would like to see examples of what other researchers in the community have written, not only to be able to write the data management plan (DMP) but also so that he can cost data storage with University IT. He doesn't know where to start. From the guidance given by the funder, he knows he needs to consider data reuse and knows of a public repository but would like to see if there are others available. He knows that for the duration of the project he must annotate datasets with appropriate metadata but does not know which metadata standard is appropriate. He knows that at the end of the project, the data needs to be accessible by the public. The university in which he works does not provide a solution, so he would need to find an alternative.

- Are there example DMPs for researchers online? These would be specific to marine metagenomics or not
- How can he best access other people writing DMPs within his community? Networks?
- How can he find other published datasets in the field?
- How can he find ontologies/controlled vocabularies that can be used to annotated metadata?
- How can he estimated the size of final data, so that he can cost disk storage?
- Where are the public databases/repositories for marine metagenomics?

FAIR and Emerging organisation of biomolecular simulation information

Alex is building mathematical models and uses public databases to parameterise his models. He needs a platform where he can store both models and link those models with data.
But this already exists - FAIRDOMHub! Thus, Alex started to use FAIRDOMHub and started communication with FAIRDOMHub team on how additional features could be added to FAIRDOMHub to solve the following problems:

- He would like to link reaction rate constants in his models with transcription data which can be found on ELIXIR. The higher is the transcription of a gene coding the enzyme, the higher is Vmax of corresponding reaction in his models.
- He wants that consistency of data will be checked by incorporating them into a mathematical model, so to go for model-driven data management.
- He wants that his models and data would be not only Findable, Accessible, Interoperable and Reusable, but also Emergable, so they can be integrated together in the larger model, interact and allow to reconstruct the emergent behaviour of the bigger system.