Euro-BioImaging – ELIXIR Image Data Strategy

1. Preamble

The following Image Data Strategy defines the Preparatory Phase recommendation for the collaboration strategy between Euro-BioImaging and ELIXIR to support access to biological image data and link it to biomolecular data.

2. Description of the partners Euro-BioImaging and ELIXIR

ELIXIR is a pan-European research infrastructure for biomolecular data. It builds on existing data resources and services and follows a Hub and Nodes model, with a coordinating Hub in Hinxton, Cambridge, and a growing number of Nodes located at centres of excellence throughout Europe. The goal of ELIXIR is to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments and provide open access to world-leading data, compute, tools, standards, training and industry services. It is at the heart of ELIXIR's strategy to provide common services in data management to the biological and medical research infrastructures.

Euro-BioImaging aims to provide open user access to a complete range of state-of-the-art imaging technologies in biological, molecular and medical imaging for life scientists in Europe and beyond. In addition, Euro-BioImaging plans to offer image data services and training for infrastructure users and providers. The research infrastructure will consist of a set of complementary, strongly interlinked and geographically distributed Nodes that provide physical access to European scientists in all Member States to produce image data. The pan-European infrastructure will be empowered by a strong supporting and coordinating Hub and focus on provision of common image data services, including image data repositories.

3. The added value of linking Image Data with Biomolecular Data

Image data are acquired for a large variety of research purposes on a large variety of biological samples with different degrees of molecular information. The strategy proposed here aims to provide maximal flexibility to each research community that systematically acquires image data sets. The strategy will provide a framework for the communities to make their data accessible and support each community in maturing their data sets by defining data standards and architectures for common repositories, while at the same time enable linking with the available biomolecular databases to the maximal extent possible.

To add value to both image and biomolecular data, it is vital to have a strong data interface between Euro-BioImaging and ELIXIR. This is evident where imaging provides structural and functional information to the used biomolecular probes or agents (e.g. gene silencing and editing reagents, protein markers, molecular reporters of physiological processes) or where biomolecular samples can be placed in an image-based spatial and/or temporal coordinate system, e.g. a cellular or organismal atlas. It is clear that image data provide major additional value to biomolecular data and vice versa.

It is expected that the majority of images produced at Euro-BioImaging Nodes are specific to a particular experiment and research question, and therefore should be stored temporarily at the Euro-BioImaging Node for quality control and analysis and then archived locally via appropriate mechanisms at the user's home institution in line with funders' requirements. Euro-BioImaging's general policy therefore is "data belongs to the user".

However, it is expected (and already the case) that a smaller proportion of the images constitute valuable resources for a broader community of users that will often be accessed as a reference or that will be recomputed to extract additional information and knowledge. We term this category of image data here "**reference images**". Their information content consists of the raw images, as well as of automatically and manually extracted data and annotations that parameterize the reference images with standard descriptors and ontologies (both quantitative and qualitative).

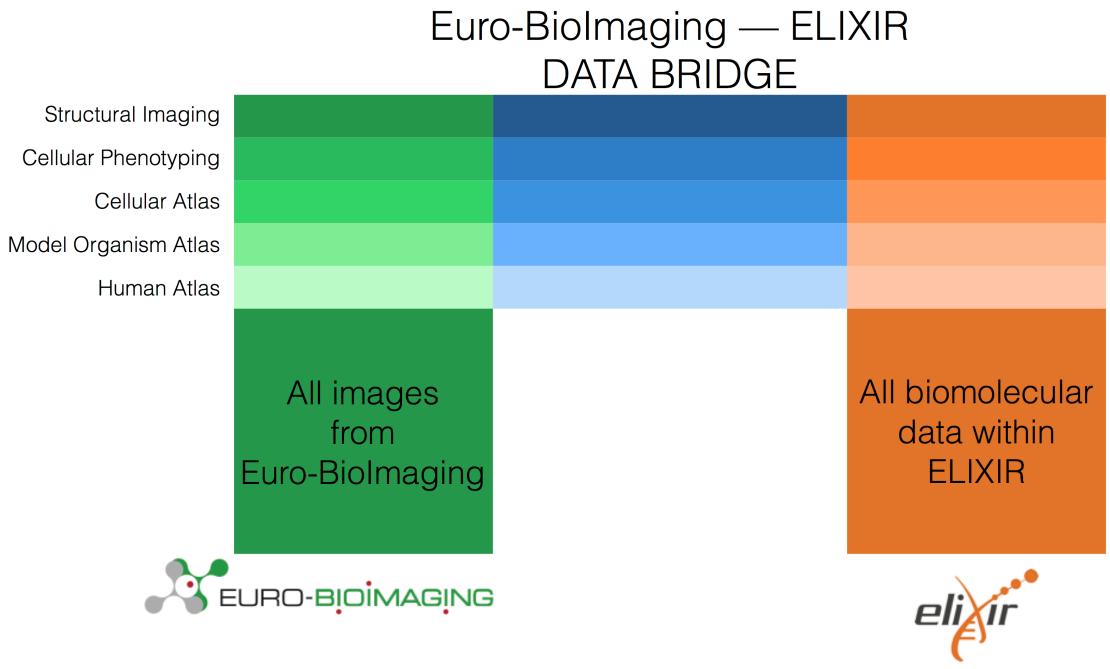

Figure 1. Scheme for the strategy to integrate reference image data sets from Euro-BioImaging with biomolecular data in ELIXIR.

Therefore as shown in Fig. 1, there needs to be a robust bridge component that links Euro-BioImaging reference images with ELIXIR biomolecular data, and

provides appropriate data standards for parametrization and annotation to ensure interoperability.

Such reference image data sets of Euro-BioImaging are conceptually bounded in different domains of biomedical research. Although complex to decide which image data constitute reference images for each domain and how the images should be parameterized, this task is similar to other challenges in biology such as sampling gene expression data systematically in cells and tissues. Five reference image domains have already been identified:

1. Cellular and molecular structure data (derived from super-resolution light microscopy and 3D electron microscopy)
2. Cellular Phenotypes
3. Cellular Atlases
4. Model Organism Atlases
5. Human Atlas

It is expected that more domains (e.g. image-based model organism phenotypes, medical imaging data, etc.) will arise in diverse communities in the future, and we will approach this with a model that allows capturing and linking to emerging domains early and allows for maximal flexibility in organizing the data storage appropriately (see below).

4. Organization of Image Data storage in established and emerging reference image domains

Euro-BioImaging would like to map potential reference image domains early to engage their user community, understand their data infrastructure needs and support high profile projects as early as possible. To establish the first link, each study that produces a large image data set would be registered in the BioStudy Object database at the ELIXIR Node in the European Bioinformatics Institute (EMBL-EBI) and obtain a unique accession number (Data DOI). The associated image data would be handled in different ways, depending on the maturity of each scientific domain in terms of data standards and accepted central or distributed domain specific repositories:

a. Emerging domains: BioStudy Object DB would link to the image data stored at the data producer who provides it as a clearly organized filesystem with minimal metadata standards.
b. Mature domains: Image data storage would be coordinated by the Euro-BioImaging Hub in appropriately organized databases with well-defined data standards, descriptors, annotations and links to the biomolecular data in ELIXIR.

As emerging reference image domains mature in terms of user community and data standards, data sets could be moved to centrally coordinated Euro-BioImaging databases and the BioStudy Object DB would be updated to point to the new storage location. Alternatively, and especially for very large volume data, the local filesystems would be converted into a repository type database with

appropriate standards based on the Euro-BioImaging :: ELIXIR bridge technology, but would physically be organized as a globally coordinated federated archive in the future.

5. Conclusion

The Euro-BioImaging :: ELIXIR Image Data Strategy will enable Europe's infrastructures on biomolecular data and imaging technologies to integrate and thereby add maximal value to their respective data. This will generate new information and knowledge and enable the infrastructures to jointly offer the urgently needed data services for imaging-based research in the life sciences.