NFDI Infrastructure for pre-clinical data

Acronym: DeBioData

Speaker:
Philip Gribbon, Fraunhofer Institute for Molecular Biology and Applied Ecology, ScreeningPort (philip.gribbon@ime.fraunhofer.de)

Details of the planned consortium

Which research area should be addressed?

DeBioData (DBD) is a consortium assembled from German Life Science Institutions and associated Research Infrastructures working in the field of pre-clinical research and human disease biology. The scope of the consortia's interests extends from basic scientific studies on molecular targets and their linkage to disease through to the analysis of late stage in-vivo proof-of-concept studies and their translation towards clinical investigations. The key objectives aimed for by the consortium in the context of NFDI program are to enable a robust, efficient and qualified network of data infrastructures to extend knowledge about disease relevant biological mechanisms by facilitating the sharing of relevant pre-existing qualified data (with DOI, metadata, validated workflows) which have yet to be elevated to FAIR standards. We will develop DBD into a qualified infrastructure which links together a network of pre-existing indication, chemical, biological, 'omics and target-centric databases as well as any novel upcoming relevant resources, which collectively will form a unified resource for University, SME and large pharma industry based researchers. Together with tool providers, data originators and data users we will define realisable standards for FAIR data from fundamental through to pre-clinical research domains by developing requirements and guidelines on the FAIRification processes aligned to emerging European standards which individual German Institutions can adopt. Data originators will be provided with secure and compliant solutions, by means of novel blockchain-like solutions. This will assure data integrity and allow scientific users to have a "one-stop shop" web-based service to search, find, collect and aggregate FAIR data important for their projects.

What kind of data are you dealing with?

Quantitative Bioassay related data analyses including but not restricted to chemical compound related primary, secondary and selectivity screening results. This would include pathogen-related profiling results (MIC and Time-Kill analyses, etc.); toxicity and liability results (Cyto-toxicity, etc.); in-vitro safety assays (P-450, HERG, Cardiac Ion channel panels); Chemical, physico-chemical and biological descriptors (structures, sequences, logP, etc.); in-vitro ADME studies; imaging data covering cellular phenotypes. There will also be selected computational, modelling and simulation including molecular structure results (docking, homology models, MD, conformational search, similarity, etc.) and data on binding mode and pharmacophore analyses. The types of data will cover: discovery (in-vitro) structural and 'omics related data including genetic data; proteomic and transcriptomic data as well as protein structural information.

Which essential measures do you plan to introduce for data management in your research area and which services do you want to offer?

Our concept will be to strategically align data providers, users, tool developers and relevant infrastructure platforms at major German Research sites, which share our common research

goals. Moreover, the network-focussed approach implies that all German Research Institutes in any region can be associated with our activities, which will serve to maximise the impact of DBD across the scientific community. Users will be able to address their scientific questions making use of a larger data repository than previously, and, within the network have access to a defined collection of machine-learning and artificial intelligence-based tools and workflows. This will allow users to generate, test and validate general prediction models and/or processes in their specific data domain. The higher aggregation levels achievable in this way will pave the way to more precise models and enhance our capabilities to probe and understand the fundamental determinants of cell and tissue function and the deviations associated with (pre-) disease states. Together with identified interested German stakeholders and in cooperation with other NFDI projects, the consortium will identify the technologies which best can be adapted and adopted by institutional users. Our services will retain data on owner servers whilst making it accessible through maintained and sustainable web services or documented application programming interface (API) for data retrieval. We will introduce common ontologies, standards and quality controls for their data (e.g. detailed metadata and resource description language (RDF)). A central access process to access pre-clinical data will enable users to search and interoperate/integrate their workflows. This will stimulate drug discovery by improving knowledge about biological mechanisms which will allow for development of novel therapeutic options.

What are the special requirements for research data management in your field of research and how do you want to address them?

Novel technologies, such as high-resolution microscopy and imaging, mass spectrometry and high-throughput sequencing, have yet to develop standardization tools for saving, export and sharing the immense amounts of data generated. Commercial providers have few incentives to define a common format for data files produced by their instruments. In general, there is also a limited availability of unique identifiers for non-regulate pre-clinical research results and a lack of qualified (ontologized) metadata. This leads to challenging data retrieval, inadequate “quality” assessments and comparisons/aggregations of results, even in open access repositories which are prevalent in this domain. For example, several open access repositories are available in different scientific areas from genes to cell lines to animal studies. However, few of those provide high degree of data integration and FAIRness. In collaboration with complementary NFDI partners, we will establish a resource which is by design FAIR and integrates pre-clinical data types over the lifetime of the infrastructure. To achieve this, we will make use of new and emerging concepts (i.e. microservice architectures, cloud, blockchain, web 4.0/5.0) to enable non-bioinformaticians to perform expert data manipulation and support hitherto challenging cross-resource analysis. The DeBioData Infrastructure will facilitate hypothesis generation to help elucidate the molecular determinants of diseases by providing interoperable FAIR data of assured quality which is suitable for using advanced algorithmic analyses of aggregated data sets. It will do this by providing documents, tools, workshops, e-resources, applications, etc., to make data accessible and FAIR for all participants of the network.

What experience/background does your group have in data management?

At the centre of the proposed DeBioData infrastructure is EU-OPENSCREEN (European Research Infrastructure for Chemical Biology), which coordinates the data infrastructure activities of 21 screening and medicinal chemistry infrastructures in 8 European countries, including four German sites. Partner TUM is strongly engaged in the national and European proteomic community, exemplified by its leading role within the DKTK Proteomics platform and its participation in the European network project EPIC-XS. Fraunhofer has extensive

experiences in data management within multiple national and European programs (e.g. Medicin Initiative, FhG Data scientist training). Partner UHH manages state of the art bioinformatics infrastructures including NERDD resource for drug discovery, the ProteinsPlus server for protein structures and SMARTS for visualization for SMARTS strings. The other partners (HZI, MPI, UFZ) are all involved in extensive institutional data management activities covering pathogens, cellular imaging and toxicological data resources, respectively.

Which relevant (international) partners and existing infrastructures do you want to bring together?
In addition to EU-OPENSCREEN network, TUM offers connections to national and international academic and industry partners of different disciplines such as the European Bioinformatics Institute (EBI), SAP and IBM. All of the partners have extremely strong international networks reflecting world-leading roles in data management related to their respective fields. Partner IME is strongly involved in the IMI-FAIRplus project as a work package lead as well as co-leading the cloudification work packages of the European Open Science Cloud (EOSC-) LIFE project, where EU-OPENSCREEN is also a partner.

Where do you see interfaces to the overall NFDI?
Initial discussions involving telephone conferences or face-to-face meetings have taken place to identify concrete options for cross-consortia collaboration. We plan to align the concept of DeBioData with that of one or more other prospective NFDI consortia. **NFDI4Chem** – The application of Medicinal Chemistry methodologies is a key part of small molecule drug discovery, pre-clinical research and chemical biology optimisation of compounds towards tools and leads. Discussions and exchanges of information between the respective consortium spokespersons have revealed multiple areas of common interest. These synergies will be further explored in discussions between the 2 consortia going forward **NFDI4Biological Imaging and Medical Photonics NFDI4BIMP**. DeBioData and NFDI4BIMP plan to collaborate on the definition and use of image data formats and ontologies shared across multiple NFDI consortia to facilitate image data exchange and integration and promote scientific collaboration.

Cross-cutting issues.
A substantial part of the overall DeBioData effort will be directed towards implementing FAIR approaches in chemical biology and pre-clinical research, by driving the successful application of ontologies and data / metadata standards. Therefore, the DeBioData consortium would stand to benefit from the work of crosscutting NFDI initiatives working on these topics.

Topics your consortium could contribute to and how.
The DeBioData consortium would primarily act as a scientific domain-related infrastructure and would not directly contribute to cross-cutting topics. We aim to operate in the most collaborative manner possible and would be very willing to contribute to the efforts of other consortia, should the need and opportunity arise.

What do you expect from your participation in the NFDI conference? We will interact with cross cutting and newly operational consortia to find any areas of overlaps and collaboration opportunities. Establish links to compute and storage based infrastructure based consortia and identify opportunities for cooperation.

Expected members of the consortium (co-spokespersons and the other participating institutions):

Co-Sprecher/in	**Zugehörige Institution**
Stephanie Heinzlmeir Chair of Proteomics and Bioanalytics stephanie.heinzlmeir@tum.de	Technical University of Munich Emil-Erlenmeyer-Forum 5 85354 Freising
Ursula Bilitewski Professor, Group Leader Ursula.Bilitewski@helmholtz-hzi.de	Helmholtz-Zentrum für Infektionsforschung GmbH Inhoffenstraße 7 38124 Braunschweig
Wolfgang Fecke Director General wolfgang.fecke@eu-openscreen.eu	EU-OPENSCREEN ERIC Campus Berlin Buch Robert-Rössle-Str. 10 13125 Berlin
Marc Bickle Head Technology Development Studio bickle@mpi-cbg.de	Max Planck Institute of Molecular Cell Biology and Genetics Pfotenhauerstrasse 108 D-01307 Dresden
Mathias Rarey Director Zentrum für Bioinformatik rarey@zbh.uni-hamburg.de	Universität Hamburg Bundesstraße 43, 20146 Hamburg
Wibke Busch Group leader "iTox - integrative toxicology" wibke.busch@ufz.de	Department Bioanalytical Ecotoxicology Helmholtz Centre for Environmental Research GmbH - UFZ Permoserstraße 15 / 04318 Leipzig