1 - Binding letter of intent or non-binding letter of intent

This is a non-binding letter of intent (anticipated submission in 2020)

2 - Formal details

Planned name of the consortium

National Research Data Infrastructure for Adaptive Immune Receptor Repertoires

Acronym of the planned consortium

NFDI4AIRR

Applicant institution

Deutsches Krebsforschungszentrum (DKFZ)

Stiftung des öffentlichen Rechts

Im Neuenheimer Feld 280

69120 Heidelberg

Vorstand: Prof. Dr. Michael Baumann, Prof. Dr. Josef Puchta

Spokesperson

Dr. Christian Busse

christian.busse@dkfz-heidelberg.de

Deutsches Krebsforschungszentrum

3 - Objectives, work programme and research environment

Research area of the proposed consortium (according to the DFG classification system)

22 - Medicine

Concise summary of the planned consortium's main objectives and task areas

The adaptive immune system plays a fundamental role in health and disease and normally efficiently protects vertebrate hosts from infections and cancer. On the downside however, failures in its regulation are causative for autoimmunity, allergy, immunodeficiencies and lymphoid malignancies. To perform the critical task of self/non-self recognition, the adaptive immune system utilizes millions of randomly generated immunoglobulins/antibodies and T-cell receptors. The entireness of this highly dynamic set of receptors present in a host at a given time is referred to as Adaptive Immune Receptor Repertoire (AIRR). Due to the fixed (i.e., genetic) linkage of receptor reactivity to individual cell or clones, the structure of the AIRR represents key processes of the adaptive immune system: Diversification, selection, antigen recognition and clonal expansion. A comprehensive understanding of these processes will facilitate mechanistic insights and allow the development of diagnostic markers and novel therapeutic strategies. To this end, it is necessary to obtain the capability to combine data and metadata from diverse experimental technologies that provide complementary viewpoints on these processes. Thus, **the main objective of NFDI4AIRR is to build together with the German immunological community a network of federated repositories for data describing the state of the adaptive immune system and to provide tools and services that will facilitate integrated data analyses across these repositories**. In the first funding period of the NFDI program, NFDI4AIRR will focus on integrating data from AIRR-seq, flow cytometry (FC) and microscopy as central experimental platforms in immunology. The resulting network can be expanded to handle additional data types, as determined by the then-current needs of the community, in the subsequent funding round. Therefore, the main objective will be broken down into the three following sub-objectives:

Build a network of federated AIRR-seq repositories: "AIRR-seq" is a heterogeneous set of NGS-based technologies that provide information on the AIRR, i.e., the highly variable regions of adaptive immune receptors. These regions not only determine the antigen-reactivity of a given receptor but, due to their high entropy, can also act as endogenous genetic barcodes, allowing to track individual cells and their descendants over time. The unique capability of AIRR-seq to link reactivity and cell fate is the main driver behind the exponential growth of AIRR-seq data since its introduction in 2009. Members of the consortium have a long-standing expertise in the generation and analysis of this data type and have been involved from the start in the AIRR Community's efforts to provide standards and promote FAIR practices for AIRR-seq data. Building on this expertise and the existing software stacks for sharing AIRR-seq data, we plan to roll-out of the initial set of data repositories during the early phase of the consortium. This will create the basic infrastructure for addressing the other sub-objectives. Data - both from the consortium members as well as public data sets from third parties - will be curated and made publicly available throughout the funding period. While we expect that the largest proportion of this data will come from humans and mice, there is no technical restrictions to these species. It should further be noted that due to the narrow genome coverage and immunological diversification processes, AIRR-seq data can be anonymized and is therefore subject to the exemptions for scientific research provided by European and German data protection regulations.

Expand the repositories to handle flow cytometry (FC) data: FC is the current gold standard in immunology to describe, define and isolate immune cells and provides a rich phenotypic description at single-cell resolution. Nevertheless, there is currently no generally accepted way to share these often large and high-dimensional data sets. To fully utilize the wealth of information in FC data, NFDI4AIRR will develop and deploy software stacks that facilitate its storage, annotation and analysis while providing for close integration with associated AIRR-seq data sets. This is a critical and future-oriented activity, as with the current advent of commercially available platforms that use DNA barcoding of cells and surface markers to combine FC with AIRR-seq (e.g., 10X Chromium), the borders between these two technologies become increasingly fluid.

Facilitate access to and use of data: The consortium expects that there will be two main user groups of the content provided by NFDI4AIRR repositories. Data scientists who want to build advanced computational workflows, will require application programming interfaces (APIs) to access data, metadata and pre-processed results, while experimental scientists will need user-friendly applications to perform basic queries and analyses across repositories. In addition, the latter ones will also require training and support to gradually become more proficient users of data and services. To serve these needs, the consortium will address them appropriately within its task areas.

These objectives will be translated into the following eight task areas:

- *General administration:* Manage organizational, financial and legal concerns of the consortium and act as point of contact for the NFDI Directorate. This includes hosting of the NFDI4AIRR User Council, a committee representing the diverse NFDI4AIRR user communities and their needs.
- *Repository DevOps:* Develop and operate the consortium's federated repository infrastructure.
- *Data curation:* Provide consistent metadata annotation to existing data sets and make them available through the consortium's repositories. Additionally, organize the consortium's Quality Assurance (QA) Panel, which will develop and maintain guidelines for data and metadata quality and perform periodic audits among the federated repositories.
- *Application development and support:* Build novel user-friendly applications both for data analyses utilizing NFDI4AIRR's comprehensive federated infrastructure as well as simplified (meta-)data submission to NFDI4AIRR repositories. Provide support and training to users with a primarily experimental background to enable them to appropriately utilize the resources provided by the consortium.
- *FAIR, Open and Sustainable:* Implement a unified strategy to ensure the long-term availability and usability of the consortium's data and code. Furthermore, disseminate and support the adoption of these and related strategies by the immunological community to promote the cultural change towards Open Science.
- *Metadata and ontologies:* Harmonize metadata descriptors and ontologies with other NFDI consortia to facilitate fast and easy cross-consortia queries for data sets.
- *Data access and representation:* Develop and standardize programmatic interfaces (APIs) for exchange and interconnection with other NFDI consortia.
- *Beyond NFDI:* Engage stakeholders outside of the national scope of NFDI with the long-term goal to connect NFDI4AIRR at the international level, e.g., with EOSC, ImmPort, Human Vaccines Project and the AIRR Community.

Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfill the planned consortium’s objectives

Hardware and network infrastructure: In general terms, NFDI4AIRR is designed as a federated repository infrastructure, which will substantially lower the requirements for storage and compute capacity at the individual partner sites. Therefore, it is by default expected that partners can provide these resources as part of their contribution to the consortium. However, especially for large third-party data sets, we are evaluating the possibility to utilize centralized storage services like the Helmholtz Data Federation (HDF). These data sets might also require additional network bandwidth as well as some compute resources. Therefore, the consortium is currently in the process of estimating the potential resource requirements and is in contact with the Helmholtz Infrastructure for Federated ICT Services (HIFIS) for further input regarding this topic.

Software: NFDI4AIRR will build on the freely available iReceptor software stack, the core of which is licensed under LGPL3 (GNU Lesser General Public License). Notably, the consortium partners at DKFZ are currently part of the Horizon 2020-funded "iReceptor Plus" project, that aims to substantially enhance iReceptor's capabilities over the next three years. Therefore, the necessary know-how for further development and operation of the platform will already be present in NFDI4AIRR.

Interfaces to other proposed NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration

NFDI4AIRR’s "inwards" activities aim provide deep integration of immunological data sets with rich, domain-specific metadata. To embed this information within the efforts of other NFDI consortia, we are planing the following "outwards" activities:

NFDI4Microbiota: The continuous interaction between the adaptive immune system of vertebrate hosts and their microbiota makes this consortium a natural partner. To facilitate the parallel analysis of both, we plan to harmonize metadata and ontologies describing the host as well as identifying and enhancing formalized descriptions of sampling procedures.

NFDI4Medicine: The centralized metadata storage and the distributed search functionality proposed by NFDI4Medicine could be a complementary top-layer to NFDI4AIRR's federated data repositories. To this end, we plan to develop the required interfaces for data exchange and will make our expertise in domain-specific metadata and ontologies available to the NFDI4Medicine consortium.

NFDI4BIMP: Microscopy data is one of the key primary data types in immunological research. As NFDI4BIMP aims to provide generic and domain-spanning tools and services for the storage and management of microscopy and photonics-based imaging data, NFDI4AIRR is interested in utilizing these resources. To ensure the findability of imaging data describing the immunological processes that are the central focus of NFDI4AIRR, we will cooperate in developing and harmonizing metadata structures and ontologies.

NFDI4NanoSafety: NFDI4NanoSafety is the other consortium besides NFDI4AIRR that plans to manage significant amount of flow cytometry data. Therefore we aim to establish a platform for regular knowledge exchange in respect to the technical aspects of data management and storage. In the best case, this could lead to a shared backend storage infrastructure with domain-specific metadata layers on top of it. In addition, we will evaluate the potential shared requirements for domain-specific (i.e. immunological) metadata.

NFDI4RSE: The development of novel applications for data analysis and management and the continuous maintenance of existing software tools are central components of NFDI4AIRR. Therefore we plan to interact closely with NFDI4RSE to implement best practices ensuring the quality and sustainability of our code base. Furthermore, we welcome NFDI4RSE's proposition to establish resources to teach "Software Carpentry", i.e., basic programming skill to experimental scientists, as we consider this a critical step to raise the overall bioinformatic capabilities the immunological community.

NFDI4Life Umbrella: We share NFDI4Life's assessment that the creation of a intermediary structure for life science-specific coordination will be of benefit to the consortia involved and NFDI as a whole. Therefore we support the proposal to establish such a coordinative council.

Biobanking: Many pathological processes involve the adaptive immune system, thus biobanked samples of diseased tissues are an attractive resource for immunological analysis. We therefore consider it valuable to link immunological data at NFDI4AIRR that derives from samples of the German Biobank Alliance (GBA) with its respective metadata stored by the GBA. To this end, we will work together with the German Biobank Node (GBN), which coordinates the GBA, to develop and implement such linkage on the sample level. Due to the substantial benefit for translational research, we aim to pursue the interaction with GBA/GBN (both BMBF-funded projects), irrespective of their potential future integration into NFDI as a separate consortium.

4 - Cross-cutting topics

Please identify cross-cutting topics that are relevant for your consortium and that need to be designed and developed by several or all NFDI consortia.

- Centralized authentication and authorization infrastructure
- Centralized ontology services
- Open data and metadata standards
- Best practices for quality assurance (QA) of data and metadata
- Tools and resources for software sustainability
- Software and Data Carpentry

Please indicate which of these cross-cutting topics your consortium could contribute to and how.

NFDI4AIRR can contribute its expertise regarding the development of open community data standards as well as the extension and harmonization of domain-spanning ontologies. Several members of the planned consortium have already engaged in similar activities in the context of data standards development by the AIRR Community.