1 Binding letter of intent for the

Astro-NFDI Consortium

Correspondence:
msteinmetz@aip.de, mkramer@mpifr-bonn.de, s.wagner@lsw.uni-heidelberg.de, s.pfalzner@fz-juelich.de, henke@aip.de

2 Formal details

2.1 Planned name of the consortium

Astronomy, Astrophysics, and Astroparticle Physics within the National Research Data Infrastructure

2.2 Acronym of the planned consortium

Astro-NFDI

2.3 Applicant institution

Leibniz-Institut für Astrophysik Potsdam (AIP)
Prof. Matthias Steinmetz

2.4 Spokesperson

Prof. Matthias Steinmetz, msteinmetz@aip.de

3 Objectives, work programme and research environment

Physics (32)
Primary: Astronomy (311)
Primary: Particles, Nuclei and Fields (309)
Secondary: Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics (310)
Secondary: Mathematics (312)
Secondary: Computer Science (409)

Summary of the planned consortiums main objectives and task areas

Astronomy, Astrophysics and Astroparticle Physics (henceforth in short: Astrophysics) are discovery sciences with a long tradition of collecting extensive amounts of data. Data are combined from a multitude of observations produced by various facilities that cover the full electromagnetic wavelength range as well as non-electromagnetic information such as cosmic rays, neutrinos, and, more recently, gravitational waves. These data are provided by a wide and diverse range of facilities and organisations, ranging from international space agencies and large infrastructures to dedicated specialised but essential facilities on smaller scale. Simulations play a crucial role as the theory backbone as well as for campaign planning. Astronomical data are repeatedly mined, often with a multitude of algorithms, to derive observational constraints for a large variety of science cases. *The dominant requirement to reuse data in a diverse array of settings and to cross-correlate the multitude of data is a key feature of astrophysics*, which is much less common in the other empirical sciences driven by dedicated laboratory experiments. Modern facilities are often designed to allow reuse of data for future, often not yet known, science cases. The diversity of astronomical facilities and data archives has already been a driver to successfully develop global astronomy-wide common data formats and a tradition of open data. Data sharing, paired with limited restrictions of data ownership or commercial interests on the data, makes this domain an ideal field for development of scientific data workflows and methods in the data lifecycle.

Astro-NFDI is an initiative organized and supported by the astrophysical community and the observatory-oriented astroparticle physics community in Germany, represented by the Rat deutscher Sternwarten (RDS) and the Komitee für Astro-Teilchenphysik (KAT), together representing more than 3000 scientists covering all facets of the field: observers, theoreticians, computational astrophysicists, data astrophysicists, instrument developers, data users and data providers. Astro-NFDI is closely coordinated with the operators of large (mostly international) facilities in astrophysics. Building on a decades-long experience in providing and using data archives, standardizing them, and meshing them together, e.g., via the Virtual Observatory, our main objective is to establish a coherent data infrastructure that is distributed, scalable, adaptive and dynamic, with well-defined standards and interfaces, and with well-defined policies for data and software. The data infrastructure is envisioned to be very well intra- and interconnected, to obey, where feasible and reasonable, the FAIR principles, and to ensure long-term archiving. The need for such an infrastructure has been emphasized in national and international strategy documents, such as the EU Astronet science vision and infrastructure roadmap, or the RDS-Denkschrift 2017 „Perspectives of Astronomy and Astrophysics in Germany 2017-2030“.

A formidable challenge for the near future is that volumes and rates of primary data become so large that storage of all primary data will not only be unaffordable, but also simply technologically infeasible. Data will need to be compressed with irreversible losses. The challenge is to simultaneously retain the capacity to reproduce earlier results and to re-analyse the data to address new science questions. While there is considerable know-how in NFDI-related topics in the astrophysics community, *the aforementioned challenges require new concepts and new environments to ensure the successful science exploitation of the current and next generation of facilities*. The proper implementation and potential extension of the FAIR data standards for such an environment has implications far beyond the community served by Astro-NFDI.

Task area 1 - Governance: The *Executive Board* will be in charge of the oversight of Astro-NFDI. It is composed with representatives from the overall German community (Universities, WGL,

MPG, HGF, and IT centres). The executive board is also the formal link to other NFDI initiatives. The *Management Board* will include all work package leaders to coordinate and monitor progress. The Management Board would allocate resources as progress is made, new challenges are identified, or strong international developments have to be taken into account. Each WP leader will coordinate the work associated with the given tasks. Regular board meetings, as well as annual all-hands meetings will allow the whole community to interact. An external *Science Advisory Committee* of astrophysicists and computer scientists will provide advice on ongoing and future developments in the field that need to be addressed in the evolving Astro-NFDI. The *User's committee* with elected representatives from the RDS and KAT institutions ensures feedback and input from the user side regarding the services offered by NFDI. An *Infrastructure Control Board* will be responsible for synchronising the top-level requirements and deliverables of Astro-NFDI with the national and international data providers (observatories, data and compute centers). These governing agencies define the organisational structure of Astro-NFDI (in close cooperation with the NFDI directorate and all NFDI consortia), and they will shape the development (with the funding agencies) of long-term support structures (post 2028).

Task area 2 - Data infrastructures: International collaboration and data sharing in astronomy needs common standards and protocols beyond common file formats. Data archives will remain bound to observational and compute facilities. Astro-NFDI will contribute to the ongoing efforts of standardisation of interfaces and services for data archives, to enable the maximum benefits of the diversity of the astronomy community, and making the data FAIR. For smaller specialized observational facilities and data collections, we want to enable an easy management of their data with tools which are based on current standards and protocols, and develop these in parallel with new requirements. For demanding new applications we will form federated data centers.

Task area 3 - Data services: Beyond working procedures additional services are required to cope with current and future challenges. Distributed data from diverse sources and providers require common interfaces for universal access. For many future facilities a new paradigm is required. Data are often too large to be transferred to users, such that analysis needs to be carried out in central archives (moving data analysis to the provider). Using common interfaces and standardised services will enable all necessary treatment of data stored in distributed archives and from different data sources.

Task area 4 - Data workflows: Standard interfaces allow standardised workflows. This is essential for comparison of data (measurements) with simulations and data models. It includes standardised procedures for data curation and the design of processing frameworks for real-time analysis. Application of machine learning techniques need customized environments for accessing the data collections, retaining re-usable results and documenting decision criteria.

Task area 5 - Data irreversibility challenges: The large increase of data volumes has major impacts on the long-term use of data. In many future facilities, data have to be pre-analysed and reduced already during the data acquisition. This task area is dedicated to provide solutions for data, where such dynamic filtering and steering of data streams, or filtering and compression of data requires better or new solutions for the reproducibility and re-use of retained data, in light of the data irreversibility.

Task Area 6 - Synergies: Work in some of the above task areas have direct links to other consortia. Especially with the PAHN-PaN and DAPHNE communities, activities shall be tackled in matching work-packages early on. Synergies with other consortia will be explored in joint workshops and developed appropriately.

Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortiums objectives

The existing infrastructures fulfill three functions: data production, data analysis, and data sharing/publication. Together they form the digital landscape. Observational infrastructures are space or ground based. Space based facilities are operated by transnational organisations such as ESA or NASA. Large ground-based facilities are operated by international organisations such as ESO, CTAO, ILT or SKAO. Medium and smaller scale specialized facilities are operated by consortia of institutes. For computational astrophysics, high performance compute centers in Germany, Europe or in the US provide the compute power for large simulations, along with medium or smaller compute facilities being used for analysis. For legacy data, smaller archives in e.g. AIP or ZAH, larger archives such as the CDS in France, and the archives of large facilities host collections of carefully curated data. For each instrument and facility, the data is organized and often reduced according to their immediate goals. Data providers also have individual data access policies and procedures.

Astro-NFDI will cooperate with these facilities to enhance the efficient navigation of this multi faceted landscape. We will join forces with the Virtual Observatory and employ the collaborations in which the partnering institutes are involved in to improve the data flow through the digital landscape.

The diversity of data sources, as well as emerging new data analysis techniques require new structures to perform the data analysis. While there is already considerable progress made with standardised access to distributed archives (e.g., through SQL based interfaces), the distribution of data collections over many sites also hamper the application of advanced analysis tools. The need for efficient resource usage leads to forming federated archives, where harvested data collections are matched with sufficient computing power. The organisation of and access to these structures is one of the most challenging tasks for Astro-NFDI. The employment of now forming European structures such as the European Open Science Cloud (EOSC) can provide a means to overcome some scaling problems. We are participating in European ESFRI/H2020 projects, in particular ESCAPE, EUDAT, and work with other consortia to define interfaces and tools to this end.

For the frictionless flow of the data as part of the scientific analysis process, Astro-NFDI will improve tools for making our data FAIR. Starting from already available common formats (FITS, VOTable) and other VO standards and protocols, providing interfaces and extensions to widely used tools (e.g. astropy), and services to facilitate the collection of sufficient metadata is crucial to achieve this goal.

For published data, the use of persistent identifiers is one basic requirement, and we will work with IVOA, GeRDI, and TIB/DataCite towards deployment and efficient use of such services, thereby joining forces with all other NFDI consortia.

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

In preparation for data flows from observatories with petabyte scale data volumes, the astronomical community started cooperation with the particle physics community several years ago. Data management procedures for massive data rates have been addressed in this community since the advent of the LHC. Furthermore, the field of astroparticle physics, emcompassing high-energy observatories operating with detectors from particle physics and particle physics experiments testing physical properties of the universe, provides a science-topical connection between the two communities. We intend to use the opportunities provided by the NFDI to carry out joint activities which will help addressing the development of the increasingly demanding requirements in data management in the forthcoming decade.

Interfaces to many NFDI consortia in the fields of natural sciences, engineering and life sciences emerge from preexisting collaboration on data management aspects. One detailed example would be the interfaces with PAHN-PaN:

- the provision of format interfaces between different highly developed data and metadata formats that have emerged as standards in internationally densely connected research fields (here astrophysics and particle physics)
- the development of data management procedures that connect very voluminous simulations of the data acquisition process (measurement) for data calibration and curation as metadata sets
- the development of dynamical archives for data acquisition and storage for measurements of extremely large volumes. These require “on-the-fly” decisions about irreversible compression and instantaneous alterations of triggers that define the subsets of data retained for analysis and storage.

We have agreed to define matching tasks in both initiatives to address these challenges and to explore whether our respective links to other consortia provide a nucleus for further ties.

In a similar manner we agreed to cooperate with DAPHNE, whose data management efforts are, from a logistical point of view, comparable to that of an observatory (with users/observers carrying out measurements/observations at a photon- or neutron facility/an observatory with its instruments).

Scientific collaborations with colleagues from other disciplines often raise data structure related problems. These cooperations with neighbouring fields range from Physics, Informatics, Mathematics, Earth Science and Engineering to fields that develop advanced imaging processing with applications in Medicine, Biology and Genetics. The NFDI provides an opportunity to leverage mutual experiences from such cooperations on a larger scale. For example: NFDI4Earth has similar challenges bringing together data at different wavelength and the need to make simulation and observational results comparable.

We will boost joint activities and an early set of actions that promote the identification of common problems with increased spending for synergies in the first three to five years, anticipating that the knowledge transfer will ramp up gradually and assuming that experience with different measures shall be gained early on. We contacted several other forming consortia to explore common topics for collaboration within NFDI.

4 Cross-cutting topics

Many tasks foreseen in Astro-NFDI deal with topics that shall be developed jointly or in close contact with other NFDI consortia.

1. **Provenance and metadata**

   While classical metadata describing static data sets are sufficient to characterize the setup of a measurement for very diverse sets of data and data generating facilities in astrophysics, they prove to be insufficient for optimum use throughout the whole life cycle of a data set. Challenges emerge for combination of simulation and observational data, or processing of streaming data. While these challenges arise as domain specific problems, they are also known in other domains like e.g. in the geosciences or in climate research.

   Simulations of measurements provide essential metadata for data curation and data reuse but already today these metadata may exceed the volume and complexity of the data themselves, providing new challenges to data management. Simulations of physical processes that are matched against measured data generate data sets in their own right whose metadata need to incorporate all information required to associate the simulated data sets with measured data.

   Metadata shall provide links to all other elements of the data life cycle that is connected to specific data sets, in particular provenance. These include, e.g., documentary about motivation of measurements (e.g. proposals), choices for alterations of data taking of time domain studies as a result of changing parameters (dynamic filters), connections to different data sets in data re-use. The connection between data set(s) and scientific articles is a part of the data FAIRness.

2. **On-the-fly data mining**

   Time-domain studies often result in changes of the data taking process as a result of a particular measurement. This will also be an essential element of data management in the emerging big-data studies. On-the-fly decisions require prompt data mining tools that need solutions from many fields.

3. **Converters between internationally established data formats**

   While astrophysics has established common data formats for very diverse sub-communities, used globally by all data providers, cooperation with other communities that have their own well established formats requires conversion tools between widely used data formats. The generation of such converters will be relevant across the entire NFDI community.

4. **Open data**

   Open data access is a common and well-developed approach in astrophysics that will potentially be of interest in the NFDI community at large. Reuse of data after a proprietary period often generates many additional science products or even allow to address new science cases. The users of future large-scale facilities with big data potential are often organized in large international collaborations. Consequently, versatile modes of data access need to be enabled during the proprietary period. These modes may mirror data access management problems in many other fields, e.g.those that deal with personalized or commercial data resulting in restricted data access rights.

5. **Data management plans, tools and policies**

   The NFDI will further the usage of instruments which assist collaborations, organisations and projects in collecting and retaining information about the data flow throughout the lifecycle of their data sets, for improving the data management porcesses across all communities, and

encourage all consortia to employ them in their respective domains. In DFG funded projects, in ESCAPE and other interdisciplinary projects, and within the Research Data Alliance (RDA) we have already contributed to implement such instruments.

6. **Training**

A central aim of the NFDI will be to provide a stable level of expertise not only for the development of a data infrastructure, but also for user training and support at all levels, from Bachelor-level introductory IT courses to project-specific challenges.

To be prepared for the future challenges in these fields calls for integration of topics related to the research data management into the core curricula at universities. Such an integration is expected to profit from integrated activities across (at least) neighboring NFDI initiatives. While the level of expertise in IT technologies in general and data management in particular varies across the science landscape, a commonly recognized core curriculum is desirable.

7. **Legal barriers**

From former cross disciplinary and collaborative projects it is well known that exchanging resources across borders (institutional or regional) is difficult to organise, also in many more formal aspects (legal constraints, constraints by the funding agencies, national borders, to name a few examples). This is a topic which needs to be addressed within the framework of NFDI.

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

Astro-NFDI intends to contribute to all common activities through a range of measures ranging from dedicated cooperation with individual consortia in matching work packages to workshops that identify areas of synergy and providing expertise between different consortia. In all fields addressed a very close link to the physics community will be ensured though coordination with the DPG.

1. **Provenance and metadata**

Astro-NFDI partners have been working on a provenance standard and metadata formats for astronomical data within the international Virtual Observatory which connects hundreds of archives, repositories and data providers. We offer to share this experience and provide advice through workshops. We plan to contribute to further development and extension of metadata standards and provenance links within a common working group of experts and through cooperative work.

2. **On-the fly data mining**

The efficient managing of data generated by real time processing of large data streams emerges in all fields with time-dependent data (e.g. life science, climate research, geosciences, and others). Further developments will be carried out in collaboration with computer science and mathematics. This topic emerges in several task areas of Astro-NFDI and shall be addressed in a dedicated task that will work jointly with experts from other NFDI initiatives.

3. **Converters between data formats**

In the astronomical community an international data format is used world-wide (a subset has even been developed into an industry standard). Large data sets from other communities with common data formats need to be exchanged by providing converters.

Significant experience was gained in developing converters between standards in particle physics and astronomy in recent years. We will extend these converter tools to other widely used formats where a need for data exchange emerges and will define a protocol for establishing converters suitable for other applications. This will be carried out in joint work packages with DAPHNE and PAHN-PaN. Further cooperation will be organised through workshops with the communities particularly in engineering.

4. Open Data

The astronomical community has considerable experience regarding open data and has successfully dealt with many of the emerging issues. Expertise in this field and advice will be made available through workshops on different levels. Explicit work on further development will be carried out in a joint task with PAHN-PaN. It is foreseen to extend the cooperation to other NFDI consortia.

5. Data management, tools, policies

In a cross disciplinary context, our community has contributed to the development of tools, namely RDMO, which assist in e.g. creating data management plans for proposals according to the requirements of funding agencies. We know from several collaborations, e.g. NFDI4Life and NFDI4Ing, that it is planned to roll out RDMO as part of their work. We will continue to contribute to these cross disciplinary tools.

6. Training

Astro-NFDI partners intend to include courses on data science in their respective curricula. We contribute to the development of a joint curriculum through workshops and schools for establishing training concepts. An exchange on the formal treatment in physical sciences (and possibly all disciplines that include introductory courses of physics in their curricula) will be sought with the DPG.