#\*NFDI Small Disciplines (NFDI4SD)\*\*

## The task

The NFDI4SD consortium has the potential to solve a serious problem in the current state of research in the so-called small disciplines. The latter have achieved much in terms of the social and scientific significance of the disciplines’ findings: small disciplines number highly in the DFG's GEPRIS database, while the number of publications in these subject areas is considerable. The data, digitised findings and digital repositories that have been acquired through the research process are also impressive. However, there is a notable lack of secondary research data, that is, data from digital sources, the research findings derived from them and the data generated specifically for research purposes; the volume of FAIR-published secondary research data is negligible in relation to the track record of the small disciplines. **This contrasts starkly with its enormous importance**. It is these data that make it possible for researchers to access other data and repositories. Secondary data are the real engine of future digitisation. They connect the data of different collections and combine the data in meaningful ways.

Researchers cannot be held responsible for this state of affairs. Rather, it is the absence of an adequate research data infrastructure that is to blame. This lack of infrastructure **has the following consequences** for the NFDI4SD. We propose the following infrastructure:

- \*\*The NFDI4SD will integrate its numerous research data services into ongoing research projects, which will come into effect in the early stages of a project. The NFDI4SD’s services will become part of the scientific workflow and will not just be used for administrative and documentation purposes.\*\*

- \*\*Research data basically consists of its content (the actual data) and its paratext. APIs and other computational operations always refer to the entities of research data and metadata.\*\*

- \*\*The researchers will systematically guide the development of the NFDI4SD’s services according to their needs and practices through their close involvement in the research projects.\*\*

- \*\*The NFDI4SD aims to publish the research data and intends that the data will be reused following FAIRe principles by the scientific community. Data will be published according to tried-and-tested scientific publishing conventions.\*\*

- \*\*Computations will be part of the scientific work process. They will be organised on the basis of ‘computational notebooks’, which are based on the concept of ‘scientific literacy’: the notebooks can be run like computer programs and they will at the same time document in detail, by means of research data, the various work procedures - and so make them understandable to other researchers. \*\*

Although the infrastructure of the NFDI4SD is conceptually oriented towards the research activities of the small disciplines, thanks to its services it is open to any scientific project, regardless of its disciplinary affiliation. Research groups or projects in the small disciplines will need to register their infrastructure service needs as early as possible (ideally before the start of a project) and coordinate the required services with the NFDI4SD. This will enable the consortium to play a constructive role in the research process by supporting the generation, systematisation, processing and evaluation of the research data right up to data publication and the subsequent reuse of the data.

The NFDI4SD uses modern computational tools to make research data usable: the data become ‘computational objects’ and thus lay the groundwork for making previously unavailable research accessible. The NFDI4SD has thus introduced a new research concept, albeit one that is not primarily concerned with technical aspects and requirements. We need to look at research data in a new way and to treat it **operationally**. The research data combine(1) the content of the data (text, images, 3D objects, databases, programs) and(2) the data’s paratext, that is, information on the contextual environment of the research data (such as metadata or information on the data’s provenance).

This bundle of properties and information defines the research data that is generated, published and reused in the scientific research process.

The procedure derived from **it** will enable all the internal and external participants of the research data infrastructure to collaborate fruitfully within their respective task areas (TAs). The research data will be imported during the project funding stage and will then be processed further, finally being published (as ‘computational notebooks’, for instance) as quickly and sustainably as possible. The registration of the NFDI4SD’s research data will enable us to achieve another important objective: the proposed workflow will make the interdependence of the research data transparent. Which research data will influence the calculations of other data? Scientists or other users will thus be able to recognise the critical dependencies of their own data and will therefore be able to assess more effectively the scope of the research data in the light of a dynamically changing research landscape.

The terminology (that is, the technical language, the disciplinary categorisations or conventions) that relates to the research objects will be largely developed by the project members from the language used in their disciplines and will be standardised by communicating with each specific community. In most cases, the research projects will already have an appropriate nomenclature in place, which will be used by the NFDI4SD to prepare all the data. The members of each additional research project will contribute to building a network of semantically related research terminology. With the help of this terminology, the research data will describe the properties of the group of objects under investigation and the data will therefore become the empirical basis for the evaluation of general scientific findings.

The NFDI4SD’s services aim to be easy to understand and user friendly. Once registered, all researchers will receive authorisation to access the NFDI4SD’s cloud services.

- Researchers of all qualification levels and group sizes will control the requirements of the NFDI4SD’s services through the needs of their research projects. A group of spokespeople from task area one – \*\*TA1 (research fields)\*\* – will mediate and monitor the coordination process and will introduce the interests of the researchers to the NFDI4SD’s steering committee. As demand grows and the research fields increasingly differentiate, membership of the group of spokespeople is likely to be extended.

- It will be possible to access NFDI4SD’s services from, for example, the working environment of a ‘computational notebook’. Simply by inserting an additional line into the notebook, the functions can be switched from a local installation to the use of NFDI4SD’s cloud services. \*\*TA2\*\* will develop the resources for the NFDI4SD’s cloud service, in particular the innovative components of machine learning and big data.

- \*\*TA3\*\* will develop and operate all the services and offers that are needed to ensure high visibility for the reuse of the research data and collaborative teamwork. The publication of research data is as much a part of this field of activity as are further education and training.

- \*\*TA4\*\* The collaborative reuse of data will be based on normative data and standards. Usage metrics and ongoing quality assessment of the NFDI4SD will provide immediate feedback, which can be used to improve future research needs.

- \*\*TA5\*\* This task area is a vital interface for the institutional partners, the external data sources and the research repositories. It will draw up the licence agreements with the partners and will also coordinate the dissemination of the publications and the NFDI4SD’s services to the scientific community and to the interested but non-technical general public.

- \*\*TA6\*\* Every scientific publication and every reuse of research data is protected by copyright. Researchers should, from the start of a project, be aware of and comply with copyright issues; NFDI4SD's deep learning-based virtual assistants will provide support in this area. This task area will ensure that researchers know and follow the legal requirements in good time as well as help them submit proposals for the benefit of researchers and science in general.

## Additional value

### Researcher

Conceptually, we will be using research data in an entirely new way. Not unlike the process of transforming a manuscript into a book, the NFDI4SD will turn computational content into scientifically usable research data. Scientists will be able to publish their key findings and introduce them to the community while they are still conducting their research: the findings will be immediately documented and they can then be incorporated into the researchers’ own work as well as into the project. The NFDI4SD consortium will ensure that valuable research findings are preserved and remain accessible in the long term without any additional effort or infrastructure costs.

### Young scientists

The NFDI4SD supports budding researchers from the very beginning of their research, and frees them from any additional expenses. While a dissertation is being written up, any essential research data can be simultaneously compiled in a format that enables the data to be cited in the paper and, after it has been successfully completed, **published as research data together with the dissertation**. This will considerably increase the scientific impact of the research work.

### Content providers, libraries, archives and repositories

The digital resources of libraries, archives, repositories and research institutes that are now accessible have opened up a substantial part of our cultural heritage. These sources should not only be **passively perceived.** According to the NFDI4SD’s new concept, the content of these sources becomes research data that can be directly incorporated into the research process with very little effort. The author and funding institution details are entered into the paratext database, while, thanks to being fully integrated, the content of the research data will also be included in all future research data. The value of the research data of these institutions will also be enhanced by the subsequent publication of the data. Small and specialized collections, in particular, are becoming more important as they are now being integrated into all the data collections. Future researchers will be able to use the digital collections of all research institutions to expand their empirical base. The latest machine-learning techniques and global connectivity will enhance the effectiveness of all digitised content, which will enable researchers to access their own data stores without having to finance their own research projects.

### Community

Unpublished research data that was previously lost after the completion of a project will now be made available. Research data used to be understood as the material commonly accepted as necessary in the scientific community to validate the resulting research findings of a scientific investigation; once the studies had been published, though, there was no infrastructure to make the data usable or to present them as independent research findings. Now, however, it will be possible to publish independently the research data of all subject areas, but especially those of the small disciplines. We should therefore see a rise in the quality and importance of special research areas in the **sciences** generally.

### Funding institution

The findings of funded projects will be able to reach the scientific community more swiftly; this will help the actors to identify new trends as quickly as possible and to establish the needs of newly oriented research fields. The value of research funding is greatly enhanced thanks to the support provided by the research data infrastructure. Previously unused resources of research data will be made accessible with little additional effort and they will significantly increase the value of the existing data that is to be reused.