September 2020

Building Sustainable Data Services

RfII Discussion Paper on the Enhancement of Research Data Infrastructures

In 2017 and 2018, the German Council for Scientific Information Infrastructures (RfII) published discussion papers proposing the design of a National Research Data Infrastructure (NFDI).[1] The foundations of the NFDI are currently being laid, and the recent approval of the first successful consortia will chart a new course for building national data services from 2020 onwards, which are supported by the federal and state governments. In parallel to these developments, other information infrastructures are emerging in Germany and across Europe, including the transnational services offered by the European Open Science Cloud (EOSC).

The RfII was established by the Joint Science Conference (GWK) to provide advice on the enhancement of scientific information infrastructures and related topics of the digital turn in science and humanities. In producing this discussion paper, the RfII aims to encourage and nourish a wider dialogue on the development of research data infrastructures. In putting forward its reflections on these issues, the RfII attaches great importance to a well-balanced consideration of the sometimes overlapping needs of science, public administration, and industry, as well as relating aspects of the international cooperation.

SUSTAINABLE BENEFITS FOR SCIENCE

Research data services for public science should be designed to ensure that they can be used across scientific communities with long-term sustainability. In this sense, the use of data is not necessarily synonymous with exploitation in the commercial sense.[2] This broadly accepted perspective is reflected in the guiding principles of a range of initiatives calling for greater openness, sharing, and accessibility in science (Open Access, Open Science, or FAIR). Against this backdrop, the NFDI will be shaped by broad disciplinary communities as well as more specialised research fields to develop a portfolio of quality research data services for concrete scientific demands. Further public funding will support the creation of future data

1 Cf. RfII (2018) – Wide Impact for Research; RfII (2018) – Cooperation as an Opportunity; RfII (2017) – Step by Step.

2 Uses by other actors in a non-scientific context are not excluded in principle, but are usually not the main focus of the inner-scientific organisation of data.

collections and services for science, so the design will need to be flexible enough to be compatible with evolving needs.

An array of design alternatives is conceivable when considering sound solutions. This applies not only to service provision and quality assurance, but also to the "operating" and "business models" on which services and infrastructures are founded. The term "business model" is used here in the not-for-profit sense, with a focus on sustainability and resource conservation, the distribution of roles and responsibilities, and the rights and obligations of data or service providers. The model must also take into account the end-user of data, including regulatory norms and protections for data re-use or exploitation, which effectively guard against the appropriation of data for commercial purposes, the risk of commercial monopolies forming, or future vendor lock-in. To contribute to this debate, the RfII has formulated a series of basic design features and requirements covering data and data services that are deemed necessary from a scientific perspective.

DESIGN FEATURES FOR DATA (COLLECTIONS) AND DATA SERVICES

Targeting services and focusing on user needs. First and foremost, data or service offerings must reflect the needs of their (scientific) target groups. Negotiated processes for this purpose must be implemented systematically in an inclusive and fair manner by NFDI consortia, but also by transnational structures such as the EOSC or other large research data infrastructures.[3] In developing such processes, the focus should first be on the currently existing scientific demand of users and not based on projected (possible or conceivable) future demand.

Location-neutral access regimes. Large national research data infrastructures such as the NFDI are designed as supra-regional, transdisciplinary networks, with a mission to provide equal access and quality of service for researchers regardless of their location. Where access for users must be subject to local conditions (such as consent to data protection regulations or compliance with specific practices), internal processes should ensure that there are no regional or organisational biases within networks, or between actors that maintain the services. In cases where research data infrastructures are primarily designed for national needs, connectivity to supra- and international science structures must be built in by design at an early stage. In particular, the ability to cooperate effectively in the European Research Area must be ensured.

Ensuring broad scientific participation. Even data and data services from single scientific disciplines or specialised fields should be provided on the basis of ensuring the broadest possible scientific and transdisciplinary benefit for academia and beyond. Where this is possible and useful, metadata imported into multidisciplinary infrastructures such as the NFDI should be converted into community-agreed or internationally recognised standards and ontologies that are amenable to automated processing. In addition to labelling and tagging or

3 Cf. detailed information on user involvement: RfII (2018) – Wide Impact for Research; with regard to European design approaches, see RfII (2018) – Statement on the EOSC.

cross-referencing research data, perspectives for the development of transformed or enriched data sets or "data products" should be conceived to create opportunities for wider re-use in other disciplines, fields, and domains.[4]

Appropriate terms of use. Data services for research are characterised by certain quality features such as traceability, consistency, and independence of non-scientific interests. Users should be supplied not only with the best possible information on the provenance and scientific quality of data from the provider, but the long-term availability of content, access, and user rights must also be guaranteed, if required.[5]

Non-scientific use of data. As a matter of principle, services provided by the scientific community should also be accessible to industry and society. Nevertheless, the value proposition and conditions attached to using scientific data for non-academic purposes can be clearly different from those of data sharing within the scientific community. The conditions for data re-use in the socio-economic sphere should therefore be conceptually taken into account at an early stage when new services or data sets are created. Well-established and tested models, as well as empirical evidence from practice across a wide range of research communities and fields, could provide a basis for this.

Use and qualitative expansion of data in the research process. As a common resource, data services should aim to advance scientific research endeavours. It is therefore desirable that the outputs of processes that enrich, expand, or transform scientific data from open sources or archives should also be made openly available, and thus contribute to the growth and quality of shared data commons. In order to achieve this goal, scientific data service providers have a responsibility to conclude appropriate agreements with users requesting access to their data stocks.[6]

Scientific reputation. Data infrastructures and their services can make a crucial contribution to science through the extent to which they ensure the quality of research data, methodologies, and processes. Important reputational value can be gained in this manner. Hence, data infrastructures should integrate scientific reputation and prestige as intrinsic values in both the preparation of data for scientific use, and the development and operation of data services.[7] The NFDI, which is currently being established, can make a significant contribution

4 A cross-domain indexing of data can be associated with effort, cf. RfII (2020) – The Data Quality Challenge, p. 69 f. and literature quoted there. This is one of the reasons why the idea of an open provision may require professional efforts in the concrete case in order to guarantee the desired subsequent use (cf. RfII (2019) – Statement on Open Data and Open Access).

5 Research data infrastructures or their operators should in particular provide information on the continuity of services. According to investigations by the RfII, conditions of use and service contracts in the field of scientific data have often not yet reached the level of regulations that are customary, for example, for access in the field of digital scientific literature.

6 See also: RfII (2018) – Wide Impact for Research p. 4.

7 The RfII has commented on the need to enhance the reputation of scientific data work and the creation of scientific data products in its position paper on the data quality challenge, RfII (2020) – The Data Quality Challenge.

to this end, and also provide impetus to promote such values within the European Research Area.

Building skills and expertise. The RfII continues to note severe bottlenecks in training capacities to provide the skilled workforce needed at the science/information infrastructure interface. Research infrastructures have a special role to play here.[8] In addition to providing specialist, technical, and IT-related training, economic and legal skills are also acquired by coordinating and managing services and infrastructures, for instance, as part of developing operating and business models or in drafting contracts with commercial service providers or partners. In this respect, NFDI consortia should also consider themselves active drivers of training efforts to build a workforce with the required skill sets.

Technical and organisational interconnectivity. A wider dialogue should be initiated amongst scientific data services providers on a range of possible reference models for the operation of data infrastructures.[9] This applies in particular to the NFDI consortia, which must track developments in existing or emerging scientific IT infrastructures and their operating models at the local level of universities and research institutions and also in the wider field of European data infrastructures, supercomputing, and the EOSC. In this regard, developing interactions with initiatives in the private sector and public administrations to network databases would also be beneficial.

Willingness for reform and compromise. The creation of multidisciplinary and cross-domain research data infrastructures requires coordination and joint decisions between actors in terms of designing technical systems, exchanging know-how, converging on standards, sharing of services, cost absorption, and, where appropriate, resources. In some cases, this may mean abandoning or adapting well established models and procedures. To achieve the overarching goals, ensure the financial viability of the overall system, and overcome particular interests, the ability to compromise and an openminded culture based on mutual understanding should be promoted.

Principles of best suitability. The frequently collaborative nature of services entails a division of labour and tasks that must be negotiated between the scientific partners in a network, as well as with external scientific or commercial actors and their service offerings. In such circumstances, the transfer of a task to the most suitable actor from a professional and technical point of view should be a guiding principle, alongside safeguarding the sovereignty of science in this constellation. This should be based on predetermined, clear understandings of the tasks and responsibilities: which tasks should remain the remit of the scientific network or community, and which tasks may be served best through international frameworks or cooperation with external service providers and partners.

8 RfII (2019) – Digital Competencies.

9 For example, the differentiated reference architecture for data platforms of the International Data Spaces Association (Whitepaper 2019) or the OAIS reference model for long-term archiving. Maturity models for research infrastructures (e.g. RISE, RISE-DE) as well as standard clauses for scientifically adequate terms of use also appear to be useful.

"Paying with data". When building a service portfolio, decisions may have to be made on the integration of third-party providers. In such cases, specific attention should be paid to ensure that third-party business models which track and commercially market data and usage from scientific processes or interim results, do not impinge upon the rights of individual scientists, research in general, or the objectives of scientific institutions and infrastructure networks.[10] In individual cases borderline issues are likely to arise which have not yet been sufficiently explored or considered within the scientific community. Therefore communication processes aimed at achieving consensus on how to deal with "grey zone" situations in the overall system are to be resolved as transparently and openly as possible by the responsible actors.

Data sovereignty of researchers. When publicly funded scientific actors or the collaborative structures or infrastructure facilities they represent conclude contracts with (commercial) intermediaries, the preservation of data sovereignty rights by researchers, the scientific community and its institutions should be explicitly articulated – even in detail. Any contract with commercial entities should be publicly accessible, and should be drafted to specifically inhibit the formation of monopolies or vendor lock-in, which are detrimental to scientific progress.[11]

10 For "data in return", see also Datenethikkommission (2019) – Gutachten der Datenethikkommission.
11 This is in line with the key points of a data strategy of the Federal Government. See https://www.bundesregierung.de/resource/blob/975226/1693626/60b196d5861f71cdefb9e254f5382a62/2019-11-18-pdf-datenstrategie-data.pdf?download=1 (last accessed on 04.05.2020).

Imprint

Courtesy translation, September 2020, original text adopted in May 2020

German Council for Scientific Information Infrastructures (RfII) - Head Office
Papendiek 16, 37073 Göttingen, Germany
Phone +49 551 392 70 50
Email info@rfii.de
Web www.rfii.de

SUGGESTED CITATION
RfII – German Council for Scientific Information Infrastructures: Building Sustainable Data Services. RfII Discussion Paper on the Enhancement of Research Data Infrastructures, Göttingen 2020, 6 p.

The German National Library lists this publication in the German National Bibliography; detailed bibliographic data are available on the Internet at urn:nbn:de:101:1-2020052604.