The National Research Data Infrastructure (NFDI)- a Cornerstone for Biological and Biodiversity research

Position paper by the German Federation for Biological Data (GFBio)

In 2016, the National Council for Scientific Information Infrastructures (RfII)[1] introduced a far-reaching perspective for the development of a sustainable National Research Data Infrastructure (NFDI) in Germany[2]. In subsequent reports and publications specific aspects and a vision were outlined that meet the requirements of digital science for the next decades. Key recommendations are to organize the existing infrastructural landscape into consortia and to name scientific communities as the key drivers to gear the development of the NFDI.
One of the possible buildings blocks for NFDI is the German Federation for Biological Data (GFBio[3]), which is a consortium of 20 institutions in Germany comprising domain-specific data centers, museums, collections, and research facilities. The DFG funded project aims at establishing a federated infrastructure for biological data and follows a holistic approach encompassing technical, organizational, financial, and cultural aspects. GFBio is running for five years, is fully operational, and currently in a consolidation phase. The consortium has set up a charitable association[4] as legal entity. Services supplied include data submission, long-term archiving, publication, a data portal, and a web based tool for visualization and analysis (VAT[5]) as well as a terminology service (TS[6]). As a unique selling point GFBio enables uniform access to environmental (PANGAEA[7]), sequence (EMBL-EBI[8]/SILVA[9]), biodiversity and collection data (e.g. processed and manually curated by systems like DWB[10] and BEXIS 2[11]).

[1] http://www.rfii.de/en/home/

[2] Rat für Informationsinfrastrukturen (RfII) Recommendations 2016: Performance through Diversity, http://www.rfii.de/?wpdmdl=2075

[3] https://www.gfbio.org

[4] eingetragener Verein, https://www.gfbio.org/gfbio_ev

[5] https://www.gfbio.org/data/visualizeandanalyze

[6] https://terminologies.gfbio.org/

[7] https://www.pangaea.de/

[8] https://www.ebi.ac.uk/

[9] https://www.arb-silva.de/

[10] http://diversityworkbench.net/

[11] http://bexis2.uni-jena.de/

Federation and Fragmentation

Federated data infrastructures like GFBio are mostly building on existing structures and developments of various institutions. Naturally, they substantially conserve work that has been invested in the past and benefit from the expertise, innovations, and resources of consortium members. Moreover, each of the different partners is widely interlinked with international activities and developments. However, a substantial amount of time and effort has to be invested into the effective organization. Workflows, standards, interfaces, and resources have to be aligned and coordinated. Time is needed to cope with the fragmented landscape and to agree on and to develop the necessary commonalities - more time than is usually given by traditional project-based funding regimes. Therefore, the concept of NFDI as a long-term measure is most appreciated by GFBio.

Semi-Automated Workflows for high-quality Data

RfII clearly emphasizes the need for quality data and services. In particular, for the emerging landscape of cloud based data service platforms such as GBIF[12], DataONE[13], EOSC[14], or GEODAB[15] easy to use, integrated, and reliable high-quality data are needed. This requires certified services and harmonization of data structures and semantics. Along the same lines the FAIR Data Publishing group emphasizes machine readability of data as one of the major challenges[16]. This can only in part be achieved through sophisticated systems and automation. Predominantly, manual curation by domain experts is needed to meet the special requirements in the different research fields. This is a long-term investment that is currently not funded. Very positive in this respect is RfII's recommendation for a bold investment in human resources.

Integrating with Research Practice

In fact, data management should be seen as an integral part of research and research funding. However, in the past there was almost a polarization between science and the development of data infrastructures. To improve the context with science we need top down (e.g. policies) and bottom up (incentives) measures and developments on different levels including means to leverage a cultural change. Again, this needs significantly more time than the technical implementation of data services. Still, scientists have insufficient awareness of quality data infrastructures, instead, very often making use of simple not FAIR repository services. The insufficient awareness is also due to the fact that qualified personnel with expertise in data science is sparse. GFBio - like state-of-the-art research in general - needs these 'hybrids' linking the two worlds - science and IT. Data scientists not only have the capabilities to make the best use of supplied services but also push the development of research data infrastructures. The situation requires changes in curricula, which is out of scope for GFBio and other similar projects. Consequently, RfII recommends this to be covered by the NFDI.

---

[12] https://www.gbif.org/

[13] https://www.dataone.org

[14] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[15] http://www.geodab.net/

[16] The FAIR Guiding Principles for scientific data management and stewardship. – Sci. Data 3:160018, https://doi.org/10.1038/sdata.2016.18

Sustainability

The most urgent problem for project funded federated infrastructures like GFBio is sustainability. Permanent resources are not only needed for the operation of the common services, in particular curation, but also for the further evolution of the infrastructure. Regarding the fast developments in IT continuous adaptations are required. GFBio favours a mixed 'business' model composed of a fixed and demand oriented funding part, a model which has been shown to be successful in open source software development. Fixed funding is currently assumed to be compensated by in kind commitments of participating institutions (essentially basic services). The demand oriented part, that is data management as a funded part in research, is seen to be crucial for the data services to adapt to science needs[17]. However, although funding policies are in place[18] and supported by review boards as well as commissions, the implementation is still at its early stage. In this respect, there is an urgent need to develop reliable new funding models that do not impose practical hurdles.
Should GFBio fail, this will not be due to a lack of quality of services, but will rather be due to a lack of adequate long-term resources. The sustainability problem is addressed by GFBio but is unlikely to be finally solved within the remaining funding period. Here, we clearly see the responsibility and task of the NFDI. GFBio together with its partners and its community network is committed to strongly contributing to the success of NFDI.

Contact

Michael Diepenbroek, Coordinator GFBio - coordination@gfbio.org
Board of GFBio e.V. - vorstand@gfbio.org

[17] compare: Royal Society (2012) Science as an open enterprise: open data for open science, https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf

[18] e.g. DFG (2015) Leitlinien zum Umgang mit Forschungsdaten in der Biodiversitätsforschung, http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten_biodiversitaetsforschung.pdf