Concept for Setting up a Working Group in the NFDI Section "Common Infrastructures"

Name of the working group

Long Term Archival and Access

Acronym

LTA

Contact (persons)

Universitätsbibliothek der Humboldt-Universität zu Berlin
Andreas Degkwitz
andreas.degkwitz@ub.hu-berlin.de

Authors

Felix Bach
Andreas Degkwitz
Wolfram Horstmann
Peter Leinen
Michael Puchta
Thomas Stäcker

Version 1.01

Date 12.04.2022

Abstract

NFDI consortia have a variety of disparate and distributed information infrastructures, many of which are as yet only loosely or poorly connected. A major goal is to create a Research Data Commons (RDC)[1]. The RDC concept includes, for example, shared cloud services, an application layer with access to high-performance computing (HPC), collaborative workspaces, terminology services, and a common authentication and authorization infrastructure (AAI). The necessary interoperability of services requires, in particular, agreement on protocols and standards, the specification of workflows and interfaces, and the definition of long-term sustainable responsibilities for overarching services and deliverables. Infrastructure components are often well-tested in NFDI on a domain-specific basis, but are quite heterogeneous and diverse between domains.

LTA for digital resources has been a recurring problem for well over 30 years and has not been conclusively solved to date, getting urgency with the exponential growth of research data, whether it involves demands from funders - the DFG requires 10 years of retention - or digital artifacts that must be preserved indefinitely as digital cultural heritage. Against this background, the integration of the LTA into the RDC of the NFDI is an urgent desideratum in order to be able to guarantee the permanent usability of research data.[2] A distinction must be made between the archiving of the digital objects as bitstreams (this can be numeric or textual data or complex objects such as models), which represents a first step towards long-term usability, and the archiving of the semantic and software-technical context of the digital original objects, which entails far more effort. Beyond the technical embedding of the LTA in the system environment of a multi-cloud-based infrastructure, a number of technically differentiated requirements of the NFDI's subject consortia are part of the development of a basic service for the LTA and for the re-use of research data.[3]

The need for funding for the development of a basic LTA service for the NFDI consortia results primarily from the additional costs associated with the technical and organizational development of a cross-NFDI, decentralized network structure for LTA and the sustainable subsequent use of research data. It is imperative that the technical actors are able to act within the network as a technology-oriented community, and that they can provide their own services as part of the support for also within a federated infrastructure. The working group "Long Term Archiving" (LTA) is to develop the requirements of the technical consortia for LTA and, on this basis, strategic approaches for the implementation of a basic service LTA.

The working group consists of members of various NFDI consortia covering the humanities, natural science and engineering disciplines and experts from a variety of pertinent infrastructures with strong overall connections to the nestor long-term archiving competence network.[4] The close linkage of NFDI consortia with experienced partners in the field of LTA ensures that a) the relevant technical state-of-the-art is present in the group and b) the knowledge of data producers about contexts of origin and data users interact directly. This composition enables the team to take an overarching view that spans the requirements of the disciplines and consortia, also takes into account interdisciplinary needs, and at the same time brings in the existing know-how in the infrastructure sector.

[1] cf. NFDI Cross-cutting Topics Workshop Report, https://doi.org/10.5281/zenodo.4593769

[2] cf.”Leistung aus Vielfalt“ (2016), Empfehlung 4.3, S. 45ff. http://d-nb.info/1104292440/34

[3] cf. Specification ofthe NFDI-Section „Common Infrastructures“ (09/2021) orientiert.

[4] https://www.langzeitarchivierung.de/Webs/nestor/DE/Home/home_node.html

Motivation and Objectives

Since the inception of the digital age the creation of digital data and documents has been accompanied by problems of accessibility, obsolescence or even loss. Valuable data has disappeared from servers because of negligence or technical failure. Well known examples are corrupted tapes and loss of data of NASA's first space missions. But even if data is extant after a couple of years, quite often accessing the data can be difficult or almost impossible due to obsolete formats, lack of documentation or missing tools.

The problem is not confined to some rare instances, it permeates all areas of research where data is produced and stored on e.g. institutional servers that nobody takes care of once the project has ended and the researchers involved have moved away to new jobs and projects. Data that is perhaps useful for the next generation of researchers is no longer available, or in some cases research results cannot be reproduced as the pertinent information is missing. Preserving research data, therefore, is paramount to avoid repeating expensive experiments and wasting valuable resources, to allow re-purposing data and to obtain proof of the validity of former research results. LTA, however, is more than just creating backup files or taking precautions against data loss or intruders being the typical work of data centers (bitstream preservation). The greatest challenge of LTA is not so much the bit stream preservation but the preservation of information on a logical and semantic level. This kind of information preservation can be achieved generally by two methods, migration and emulation. While the former is used to transform data and content to current standards and formats, the latter attempts to emulate environments allowing to keep e.g. outdated software running. Both strategies lean on technical metadata that are stored along with the data. Accordingly, semantic preservation can only succeed with appropriate content-related, event-driven metadata, which is ideally already available when the data is created.

Most solutions addressing LTA draw on the OAIS model (ISO 14721:2012) describing the ingestion, archival proper and dissemination process of documents and data (SIP, AIP and DIP packages).[5] There are a variety of software solutions available adopting the OAIS model such as Rosetta, Archivematica or DIMAG that integrate other LTA services such as PRONOM, DROID or JHOVE for detecting and validating formats. In addition, there are communities, in Germany especially Nestor (participants of the group are members of Nestor) that have already successfully developed LTA strategies in an cooperative approach and policies the working group can draw upon.

Against this background, one objective of the working group is to identify suitable standards for archival purposes such as PREMIS (to be aligned with other NFDI sections such as metadata and connected to international initiatives such as OPF[6] or PLANETS[7]) and typical use cases for the application of migration or emulation procedures. By making use of the already existing infrastructure

[5] The next version 3.0 is available since 2019 and is in the standardization process, cf. https://cwe.ccsds.org/moims/_layouts/15/WopiFrame.aspx?sourcedoc={61C755A7-2C54-4D0D-A8F0-7B6A4228D74C}&file=OAIS%20final%20v3%20draft%20with%20changes%20wrt%20OAISv2%2020190924-rl.docx&action=default

[6] https://openpreservation.org/

[7] https://planets-project.eu/

it aims to design scenarios by means of which NFDI repositories can function as distributed but joint LTA archives, based on common rules, for preserving data at both the bit stream and the logical and semantic level (information preservation). It seeks to establish a common understanding of how various kinds of data is created, collected or selected by NFDI consortia and how it is stored and documented by appropriate metadata and according to the OAIS principles in drawing on work already done by stakeholders like Nestor and established discipline-specific solutions. In addition, the working group will address issues of cost calculation, scaling of services and implementation to provide relevant information for setting-up appropriate base service infrastructures for the entire NFDI. In particular it is planned to describe and conceive particular services suitable for NFDI LTA, to develop a financial and organizational model for ensuring sustainability and long-term operation of the resources and services brought into the NFDI and to explore how in view of a sustainable multi-cloud-based infrastructure IT systems can be adapted to interoperate on an application-oriented and technical level, e.g. by establishing workflows and appropriate interfaces. It is hoped that a basic service for LTA can be established enabling uniform access to data, software, and compute resources as well as sovereign data exchange being dedicated to the idea of collaborative work by interconnecting and building on already existing services.

Work Plan

(Work plan with milestone descriptions; initial runtime 2 years; approx. ½-1 page)

● Document analysis of NFDI proposals for collating requirements and use cases ● Collecting practical applications from various NFDI consortia ● Defining the scope: significant properties of data that have to be preserved	6 Months
● Identify appropriate Metadata standards for LTA ● Describing scenarios for migration and emulation ● Conceiving of suitable services	6 Months
● Exploring financial and organizational models of relevant services and infrastructures	3 Months
● Consolidation and federation of services to make them appropriate for the NFDI	9 Months
● IT-systems implementation plan	3 Months

If necessary, individual steps must be repeated for the consortia that are approved in 2022.

Collaboration Plan

(Interlinking of working group and different NFDI consortia; possible collaboration with initiatives outside the NFDI; approx. ¼-½ page)

- Exchange with NFDI consortia represented in the WG
  - FAIRagro [noch im Antragsverfahren]
  - DataPLANT (NFDI4PLANTS.ORG)
  - NFDI4Earth
  - NFDI4Biodiversity
  - NFDI4Chem
  - NFDI4Culture
  - NFDI4DataScience
  - NFDI4Ing
  - NFDI4Memory [noch im Antragsverfahren]
  - NFDI4Microbiota
  - NFDI4Objects [noch im Antragsverfahren]
  - NFDI4RSE [noch im Antragsverfahren]
  - PUNCH4NFDI
  - Text+

- Exchange with NFDI consortia not represented in the WG as well as WGs of other NFDI sections

- Collaboration with
  - DINI/Nestor
  - EOSC and other EU initiatives
  - RDA regarding standards

- Collaboration with larger institutions of the research infrastructure sector relevant for LTA

Initial Membership List

(Members from at least 6 institutions and at least 6 consortia)

Universitätsbibliothek der Humboldt-Universität zu Berlin
Prof. Dr. Andreas Degkwitz, andreas.degkwitz@ub.hu-berlin.de

Deutsche Nationalbibliothek (Text+, Nestor)
Dr. Peter Leinen, p.leinen@dnb.de
Frank Scholze, f.scholze@dnb.de

Universitäts- und Landesbibliothek Darmstadt (NFDI4Ing, Text+)
Prof. Dr. Thomas Stäcker, thomas.staecker@tu-darmstadt.de

FIZ Karlsruhe (NFDI4Chem, NFDI4Culture, MaRDI)
Felix Bach, felix.bach@fiz-karlsruhe.de
Matthias Razum, matthias.razum@fiz-karlsruhe.de
Moritz Schubotz, moritz.schubotz@fiz-karlsruhe.de

Universität Heidelberg (PUNCH4NFDI)
Prof. Dr. Stefan Wagner, s.wagner@lsw.uni-heidelberg.de

Niedersächsische Staats- und Universitätsbibliothek (NFDI4Biodiversity, NFDI4Culture, NFDI4Earth, Text+)
Regine Stein, regine.stein@sub.uni-goettingen.de

Prof. Dr. Wolfram Horstmann, horstmann@sub.uni-goettingen.de

GESIS – Leibniz-Institut für Sozialwissenschaften in Mannheim (NFDI4DataScience)
Prof. Dr. Stefan Dietze, stefan.dietze@gesis.org

Leibniz-Zentrum für Agrarlandschaftsforschung (FAIRagro [noch im Antragsverfahren])
Dr. Nikolai Svoboda, nikolai.svoboda@zalf.de

Deutsche Zentralbibliothek für Medizin (NFDI4MIcrobiota)
Dr. Katharina Markus, markus@zbmed.de

Albert-Ludwigs-Universität Freiburg, Rechenzentrum (NFDI4PLANTS)
Dirk von Suchodoletz, dirk.von.suchodoletz@rz.uni-freiburg.de
Klaus Rechert, klaus.rechert@rz.uni-freiburg.de

Karlsruher Institut für Technologie
Doris Ressmann, doris.ressmann@kit.edu

Friedrich-Schiller-Universität Jena, zedif: Kompetenzzentrum Digitale Forschung (NFDI4RSE [noch im Antragsverfahren])
Dr. Frank Löffler, frank.loeffler@uni-jena.de

Deutsches Elektronen-Synchrotron DESY
Dr. Thomas Schoerner-Sadenius, thomas.schoerner@desy.de

Geschäftsstelle NFDI4Ing
Thorsten Schwetje, geschaeftsstelle@nfdi4ing.de

Bayerische Staatsbibliothek
Dr. Klaus Ceynowa (NFDI4Memory [noch im Antragsverfahren]), ceynowa@bsb-muenchen.de

Generaldirektion der Staatliche Archive Bayerns
Dr. Markus Schmalzl (NFDI4Biodiversity, NFDI4Earth, NFDI4Objects [noch im Antragsverfahren], NFDI4Memory [noch im Antragsverfahren], FAIRago [noch im Antragsverfahren]), markus.schmalzl@gda.bayern.de
Dr. Michael Puchta (NFDI4Biodiversity, NFDI4Earth, NFDI4Objects [noch im Antragsverfahren], NFDI4Memory [noch im Antragsverfahren], FAIRago [noch im Antragsverfahren]), michael.puchta@gda.bayern.de

Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
Dr. Achim Bonte, achim.bonte@sbb.spk-berlin.de
Reinhard Altenhöner, reinhard.altenhoener@sbb.spk-berlin.de

Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden
Dr. Jens Bove, jens.bove@slub-dresden.de
Dr. Julia Meyer, julia.meyer@slub-dresden.de

Universitätsbibliothek Stuttgart
Dr. Helge Steenweg, helge.steenweg@ub.uni-stuttgart.de

GESIS - Leibniz-Institut für Sozialwissenschaften (KonsortSWD)
Dr. Claus-Peter Klas, claus-peter.klas@gesis.org

Staatliche Naturwissenschaftliche Sammlungen Bayerns (NFDI4Biodiversity)
Dr. Dagmar Triebel, triebel@snsb.de