

Die Use Cases wurden auf dem Kick-Off-Meeting der AG LTA 17.-18.08.2022 vorgestellt.
Die Dateien sind hier zu finden:

<https://drive.google.com/drive/folders/1bT2Y4jMPIZP-YrldDodyLnGwhcpkySQQ>.

An alle: gerne ergänzen, berichtigen etc.; bitte dabei die Strukturierung beibehalten.

14.11.2022: Diese Auswertung ist ein erster Entwurf und enthält an einigen Stellen noch Fragezeichen. Einzelne Use Cases fehlen noch.

NFDI4Memory (BSB/BVB)

Ceynowa
(Anlage 04, Anlage 05)

- Lokalen Ebene
 - Datenlieferant/-Quellen: BSB (vmtl.)
 - Objekte: ausgewählte Digitalisate, Webseiten, amtliche Veröffentlichungen
 - Archivsystem 1: Rosetta (Hosting: LRZ, Administration bei BVB-Verbundzentrale), BSB Rosetta-Nutzer
 - Light Archive, außer bei fehlenden Rechten (Nutzungsrechte, Schutzrechte)
 - Rechteklärung, Auszeichnung von Rechten in MD
 - Archivsystem 2: Tape + IPS-Software zur automatisierten Verwaltung (Hosting LRZ)
 - Bitstream Preservation + „Datenerhalt via Migration“?
 - 1.150 TB: hauptsächlich Digitalisate
 - Bedarf: hoch-leistungsfähige Storage Systeme
- Regionale Ebene, Rosetta als Dienst für HS, die selber Zugriff auf das System haben
 - Datenlieferant: Hochschulen in Bayern, bzw. deren Datenlieferanten/-Quellen (vmtl. Forschende, Datenproduzierende)
 - Objekte: aus dem Gesamtbereich der jeweiligen Bibliotheken, u. a. Forschungsdaten
 - Archivsystem: Rosetta (Hosting: LRZ, Administration und Unterstützung bei/von BVB-Verbundzentrale), Hochschulen in Bayern Rosetta-Nutzer
 - Noch offen: Standardisierung der Rechteeinholung und Umsetzung an den Hochschulen

- Noch offen: Standardisierung der Übermittlung/Dokumentation der Rechte
- Ziel: Erfahrungsaustausch mit ähnlichen konsortialen (?) Ansätzen
- Überregionale Ebene
 - DFG-Projekt OstData (Kontext FID Osteuropa, Ergebnisse für NFDI4Memory)
 - Datenlieferant/-Quellen: GWZO, Metadaten im Datenrepositorium OstData der BSB
 - Objekte: digitalisierte DIA-Sammlung
 - Archivsystem: Rosetta an BSB?
 - Noch offen: MD in anderem Format als Excel archivieren, präferiert Text-Format
 - Noch offen: LZA auch der Digitalisate
 - Noch offen: Dokumentation der Sperrfristen und entsprechende automatische Freischaltung

NFDI4Biodiversity

Puchta, Triebel
(Anlage 01, Anlage 02)

- SNSB, Triebel
 - Datenlieferant/-Quellen: DBMS DWB der SNSB (siehe <https://diversityworkbench.net>), Datenbestände in relationalen Datenbanken mit einheitlichem, dokumentiertem und publiziertem ER-Datenmodell (DWB)
 - Objekte: Sammlungsdaten zu Objekten der SNSB (auch vertrauliche Informationsbereiche), relationale Datenbanken mit Informationen zu “digital specimens” sensu <https://dissco.tech/2020/03/31/what-is-a-digital-specimen/>
 - Archivsystem: Digitales Archiv der Generaldirektion Staatliche Archive Bayerns (GDA) (GDA XML-Archivierungsclient (zu erstellen) mit AFIS zur Erschließung)
 - Ziel: Möglichkeit des Wiederaufbau einer neuen RDBMS-Produktivumgebung, zufriedenstellende Wiederherstellung der Verknüpfungen der relationalen DB

- Noch offen: Umsetzung des Archivformats (z. B. SIARD siehe <https://de.wikipedia.org/wiki/SIARD>), ausreichende MD, Festlegung Datenstruktur (SIP, AIPs)
- Altdaten landwirtsch. Förderung, Puchta (Staatl. Archive Bayern)
 - Datenlieferant/-Quellen: StMELF Bayern
 - Objekte: Landwirtschaftsdaten in veralteten proprietären Formaten (Hard- und Software bei StMELF nicht mehr vorhanden)
 - Archivsystem: Portal GFBio, ggf. Plattformen NFDI4Earth, FAIRagro?
 - Noch offen: Migration in langzeitarchivierungswürdige Formate in Kooperation mit DatArena ie. Computermuseum
 - Noch offen: Hinzufügen des Entstehungskontexts
 - Noch offen: Anreicherung der MD (Normdaten)
 - Noch offen: Definition AIP

NFDI4Earth, Staatl. Archive Bayern

Puchta, Schuermann
(Anlage 02)

- Datenlieferant/-Quellen: Datacubes
- Objekte: Geobasis- und Geofachdaten
- Archivsystem: ?, ist offen (Measure im NFDI4Earth-Antrag)
 - Erschließung noch offen
 - Bedingungen zur Bereitstellung noch offen
- Strategische oder rechtliche Grundlagen/Bedingungen:
 - Leitlinien, die von einer Bund-Länder-Arbeitsgruppe der Archiv- und Vermessungsverwaltungen erarbeitet wurden
 - Richtlinien für FDM, Richtlinien Einrichtung/Repo, Richtlinien FD-Bewertung
 - LTA-Measure im NFDI4Earth-Antrag

NFDI4Chem, 4Culture (FIZ-Karlsruhe)

(Anlage 12)

- NFDI4Chem
 - Datenlieferant/-Quellen: Datenproduzenten an Hochschulen, Forschungseinrichtungen etc.
 - Offen: Reduzierung der Nutzung proprietärer Formate bei Hersteller von Forschungsinstrumenten/-Software
 - Objekte: FD in binären, proprietären Formaten
 - Archivsystem: Systeme an Institutionen in NFDI4Chem
 - Offen: Emulation hat Lizenz-Probleme, aber die Interoperabilität ist auch eingeschränkt, Daten sollen ja kombinierbar sein und für machine learning verwendbar sein
- NFDI generell?
 - Datenlieferant/-Quellen: Datenproduzenten an Hochschulen, Forschungseinrichtungen etc.
 - Objekte: Multimediale Publikationen, Publikationen, die aus verschiedenen Komponenten bestehen, die nicht alle eigene Identifier haben
 - Archivsystem: Systeme an Institutionen in NFDI4Chem
 - Zu klären: Was ist das API, das Objekt? Verknüpfung von einzelnen Objekten über jeweilige Identifier?
- NFDI4Culture
 - Datenlieferant/-Quellen: Datenproduzenten an Hochschulen, Forschungseinrichtungen etc.
 - Objekte: 3D-Objekte
 - Archivsystem: Systeme an Institutionen in NFDI4Chem
 - Zu klären: archivwürdiges Dateiformat für 3D-Objekte

NFDI_xCS (Computer Science)

Wieder
(Anlage 06)

- Datenlieferanten/-Quellen: prinzipiell alle innerhalb NFDI, die mit Data Science Objekten arbeiten und einen entsprechenden Bedarf haben
- Objekte: Data Science Objekte (data, metadata, context information incl. software components)
- Archivsystem: Infrastruktur für Reusability von Data Science-Objekten, RDMC Instanzen und entsprechende

Infrastruktur dahinter (könnte kombinierbar sein mit klassischen Archivsystemen?)

- Container, sind in einer RDMC Instanz enthalten
 - Container: Research Data Management Container, z. B. Docker
 - Können auch Hardware-Emulations-Informationen enthalten
 - Container soll durch Forschende publiziert werden
 - Versionierung zur Dokumentation von Veränderungen/Nachnutzung
- Geplant: Curation/semantic preservation
 - RDMC Instanz-templates (e. g. per scientific sub-field), beinhalten Normen der Community
 - MD, Rechte-Verwaltung, Regeln, access control: per RDMC Instanz, reguliert durch Instanz-Eigentümer (Forschende) + evtl. abhängig von RD life-cycle step, evtl. in RDMC Instanz-template enthalten
 - review z. B. durch publication chair (PC chair, editor in chief)
 - Zuständigkeiten („support“) für RDMC Instanz entsprechend RD life-cycle step (und hinterlegten Rollen) von Erstellung bis Reuse
 - [Wer legt Rollen und Personen fest?]
 - [Aktiver Erhalt von Reusability über 10, 100 Jahre?]
 - „Transformation“: vmtl. der Daten und MD bei Zugriff/Ausgabe, [auch entsprechender neuer Ansprüche von Technologien nach 10, 100 Jahren?]
- Noch offen: Umsetzung der Archivierung der Container (im Projekt? Langfristig?), Erhalt der Container und der Container-Software
 - Container = Archiv-Paket?
 - Mit-Archivierung/-Erhalt von Abhängigkeiten von externen Infos (zB. Betriebssystem?) (evtl. im Container mit-archivieren? Alles?)

- Daten duplizieren? Vielleicht zu evaluieren, bei welchen Daten das sinnvoll ist
- Vs: FIZ Karlsruhe reißt den Container auseinander, sodass einzelne Teile geupdated werden können

DataPLANT/nfdi4plants

Gieschke
(Anlage 07, Anlage 08)

- Datenlieferant/-Quellen: Institutionen innerhalb DataPLANT/nfdi4plants?
- Objekte: Docker-Container-Images, enthält in Sprachen wie CWL beschriebene Workflows zur Orchestrierung von Ein- und Ausgaben wissenschaftlicher Software auf Hochleistungsrechnern, Container Image Format: layered tar.gz
- Archivsystem: Emulations-Infrastruktur
 - Geplant: Layered Emulations-Infrastruktur: Emulator-Schicht, stabile techn. Umgebung, stabile Anwendung (Schichten bleiben möglichst getrennt)
 - Container Runtime, Runtime Specification
 - Kernel (Linux) backwards compatibility wird verfolgt, aber keine Garantie
 - Hardware Platform abstracted, ISA/CPU nicht standardisiert
 - Anspruch: wenn ein Container einmal emuliert ist, ist er auch langfristig nachnutzbar
 - Noch ungeklärt
 - Einzelaufwand vs Massengeschäft, außerdem wer übernimmt den Aufwand? Datenlieferant oder Archiv?
 - man muss den Emulator selbst langfristig erhalten, evtl. muss man portieren
 - hebis: gibt es standardisierte Sprache zu technischen Risiken? -> Zeitschiene: ca. 30 Jahre? Demnächst kommen ARM (Active Risk Management?). Oder eine offene Infrastruktur z.B. RISC-basiert (RISC: Resilient Infrastructure & Sustainable Communities)

Uni Freiburg, HS Kehl, SFB 948/SFB 1015

Gieschke

(Anlage 07, Anlage 08)

- Datenlieferanten/-Quellen: SFB 948/SFB 1015 (?), ganz NFDI?
- Objekte: hoch-dynamische, interaktive Webseiten
- Archivsystem: Emulationsumgebung?
 - Umsetzung: LZA von Disk Image der Webserver
 - Noch offen: Sicherheits-Issues:
 - für den emulierten Webserver gibt es keine Sicherheits-Fix mehr
 - Sicherheits-bezogene Informationen: Abspeicherung in mit Sicherheits-Hürden versehenen Bereichen
 - Noch offen: Umgang mit integrierten Scripten, multimediale Anwendungen (zB. Flash) -> mit-emulieren?
 - Orchestrierung: Redirect?

MPCDF

Zastrow

(Anlage 11)

- Datenlieferanten/-Quellen: alle Institute der Max Planck Gesellschaft
- Objekte: von den Max Planck Instituten erzeugten wissenschaftlichen Daten, ca. 250 PB wissenschaftliche Daten
- Archivsystem: wird aufgebaut
 - MPCDF Technikpartner
- Bisher nur Bitstream Preservation: digitale Speichersysteme der MPCDF
- Curation/semantische Preservation
 - Support und Beratung durch MPCD über den gesamten Life Cycle der wissenschaftlichen Daten hinweg, Wahrnehmung in Verantwortung der Wissenschaftler vmtl.

DKFZ Heidelberg Krebsinstitut

Buchhalter

(Anlage 13 (fehlt noch), evtl. Anlage 14)

- Datenlieferanten/-Quellen: Ärzte, Unikliniken, Forschungsinstitute, HS?
- Objekte
 - Rohdaten müssen im Land verbleiben
 - Zunehmend Genom-Daten (evtl. jeder Krebsfall geht mit einer Komplett-Genom-Analyse einher) -> hexabite-Bereich

- Genome-Referenzen im Graph vernetzt?
- Bedarf: zT. auch Datensätze mit wenigen MD, die nur wenig angefragt werden
- Archivsystem: ?

LABW

Fähle, Naumann
(Anlage 17)

- Datenlieferanten/-Quellen: Hochschulen und Forschungseinrichtungen
- Objekte: E-Mails verschiedener Systemkontexte (Outlook, Webmail, Lotus Notes)
- Noch offen: Archivsystem/Infrastruktur
 - System, das einheitliche Speicherstrukturen und Containerformate erlaubt
 - System, das nach Schlagwörtern durchsuchbar ist
 - System, das Access Control anhand spezifischer Schlagwörter erlaubt
 - Transformation/Migration: E-Mails in für das Data Mining geeignete abgeleitete Formate zu bringen
- Noch offen: Curation/semantic Preservation
 - E-Mails mit Konto, Verfasser, Datum und sonstigen Schlagwörtern versehen/aus den Daten extrahieren (oder übernehmen vom Datenlieferanten?)
 - Rechtinformationen als Schlagwörter hinterlegen

NFDI4Microbiota, (FAIRagro)

Markus
(Anlage 9, Anlage 10)
NFDI4Microbiota

- Datenlieferanten/-Quellen: Datenbanken der Mikrobiologie, Fachrepositorium Lebenswissenschaften
- Objekte: Daten der Mikrobiologie (alle Formate möglich)
- Archivsystem
 - de.NBI (bitstream preservation)
 - EMBL-EBI (semantic preservation, Omics data in standardized formats)
 - ZB MED Digital Archive (Rosetta, semantic preservation für long-tail of research data in various formats, publiziert im Fachrepositorium Lebenswissenschaften)
- Noch offen:
 - LZA von Research Software (Bearbeitung eher über AG RSE)
 - LZA-Empfehlungen für Forschende (während der Forschung, bei Publikation)
 - Weiterentwicklung von LZA-Quality Control tools und -Registries
 - Evtl. eine retention policy, Microbiota-preservation policy

FAIRagro

- Datenlieferanten/-Quellen: u. a. BonaRes
- Objekte: agrarwissenschaftliche Daten
- Archivsystem: z. T. ZB MED Digital Archive (Rosetta, semantic preservation von BonaRes-Repo)
- Noch offen
 - LZA-Empfehlungen für Forschende (während der Forschung, bei Publikation)
 - Weiterentwicklung von LZA-Quality Control tools und -Registries
 - Evtl. eine retention policy

Text+

Leinen

(Anlage 15)

- Datenlieferanten/-Quellen: Verlage, publizierende Stellen in Deutschland, publizierende Stellen außerhalb Deutschlands
- Objekte: in Deutschland veröffentlichte Medienwerke, im Ausland veröffentlichte deutschsprachige Medienwerke und Übersetzungen deutschsprachiger Medienwerke in andere Sprachen, im Ausland veröffentlichte fremdsprachige Medienwerke über Deutschland, sogenannte Germanica, die zwischen 1933 und 1945 von deutschsprachigen Emigranten verfassten oder veröffentlichten Druckwerke (E-Books, E-Journals, E-Paper, digitale Hörbücher, Musikveröffentlichungen oder Websites)
- Archivsystem: kopal, [koLibRI](#) - kopal Library for Retrieval and Ingest
- noch offen
 - Volltextsuche
 - Überführung in XML-Formate, Textformate für spezifische Objektgruppen (z. B. Zeitungen)
 - Hinzufügen von ausreichenden MD (auf welcher Ebene, Registry)
 - Verwendbarkeit für Text Mining (z. B. standardisierte Formate)