Langzeitarchivierung von Forschungsdaten

Mit NFDI-Fokus

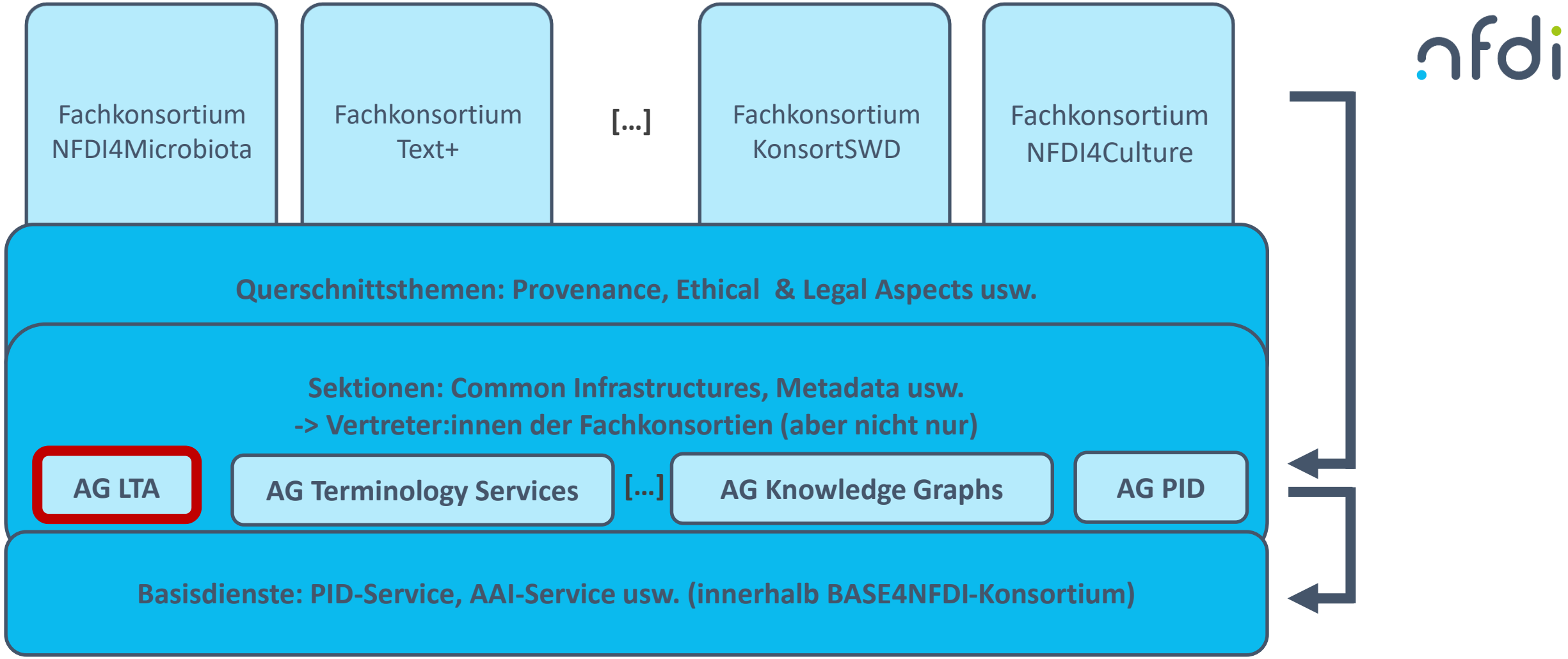
Katharina Markus
25.09.2025, LaVaH Workshop

Informationszentrum Lebenswissenschaften
ZB MED

Worum geht‘s genau – Definitionen und Erklärungen

- **Forschungsdaten:**
  - Daten, die im Zuge wissenschaftlicher Vorhaben entstehen (RfII (2018) Leistung aus Vielfalt)
  - Daten für die Nutzung in wissenschaftlichen Vorhaben
- **Nationale Forschungsdateninfrastruktur (NFDI):**
  - Bundesweit finanzierte Initiative und Verein
  - Ziel: Forschungsdaten in Deutschland besser auffindbar, zugänglich und nutzbar zu machen, Standard-Entwicklung, Ausbildung, usw.
  - Besteht v. a. aus Fachkonsortien
- **NFDI Arbeitsgruppe Long Term Access and Preservation (AG LTA):**
  - Arbeitsgruppe zur Bearbeitung des Themas Langzeitarchivierung in der NFDI, Konzept https://doi.org/10.5281/zenodo.14939132
  - Fachkonsortien-übergreifend

Worum geht‘s genau – Querschnittsthemen in der NFDI



Nicht gezeigt: Direktorat, Wissenschaftlicher Senat, Kuratorium usw.

Konzept der AG LTA: https://doi.org/10.5281/zenodo.14939132

Forschungsdaten: Basis von Wissenschaft und Forschung

- Wettbewerbsfähigkeit von Forschungsstandorten
- Kosteneinsparung durch Wiederverwendbarkeit
- Gute wissenschaftliche Praxis (Nachvollziehbarkeit, Plausibilität, Replizierbarkeit, Quellen-Verfügbarkeit)
- Datensouveränität, Daten- und Dokument-Integrität, Resilienz

- Bsp:
  - Altdaten für Klimaforschung
  - Alte Texte für Text und Data Mining
  - Langzeitversuche und -Studien in Agrarforschung und Gesundheitsforschung
  - …

Forschungsdaten: Basis von Wissenschaft und Forschung

Notwendig für neue Forschung:
Datennachnutzungs-Beispiele

Table 1:
Examples of dataset reuse for a novel purpose with the limitations/risks associated with each method.

Examples
Genome
Assembly of new genome sequences, for example organellar genome sequences, based on public datasets (Dierckxsens, Mardulyn & Smits, 2016)
Transcriptome
Co-expression analysis to find connected genes, for example identification of long non-coding RNAs associated with atherosclerosis progression (Wang et al., 2019); Co-expression networks, for example related to bamboo development using public RNA-Seq data (Ma et al., 2018) or related to cellulose synthesis using public microarray data (Persson et al., 2005); Construction of regulatory networks using co-expression data, for example co-expression network analysis to reveal genes in growth-defence trade-offs under JA signalling (Zhang et al., 2020)
Proteome
Identification of antimicrobial peptides (Porto, Pires & Franco, 2017)
Metabolome
Metabolic modelling (Brinkrolf et al., 2018)
Phenotype
Deep learning methods for image-based phenotyping, for example leaf counting (Ubbens et al., 2018) or root and shoot feature identification (Pound et al., 2017)
Ecology
Modelling and prediction of the variability of biodiversity to explain ecological and evolutionary mechanisms (Jetz, Fine & Mace, 2012)

Quelle: https://doi.org/10.7717/peerj.9954/table-1 in: Sielemann K, Hafner A, Pucker B. 2020.. PeerJ 8:e9954

Notwendig für gute wissenschaftliche Praxis:
Reproduzierbarkeit, verlässlich korrekte Daten

Retraction notice

Retraction notice to A model study into the effects of light and temperature on the degradation of fingerprint constituents [Science and Justice, 54 (2014) 346 - 350]

This article has been retracted at the request of the authors. The authors identified a inconsistency in the accepted paper and were unable to reproduce the average values that were used for the graphs and tables in the paper, due to the loss of the raw data. This, in turn, means that the authors cannot fulfil the demands of the Association of Dutch Universities and the Royal Dutch Academy of Science in respect to their ethical and research data standards.

Quelle: Amorós and Puit (2014) Retraction Notice https://doi.org/10.1016/j.scijus.2015.04.005

Digitale Informationen sind gefährdet

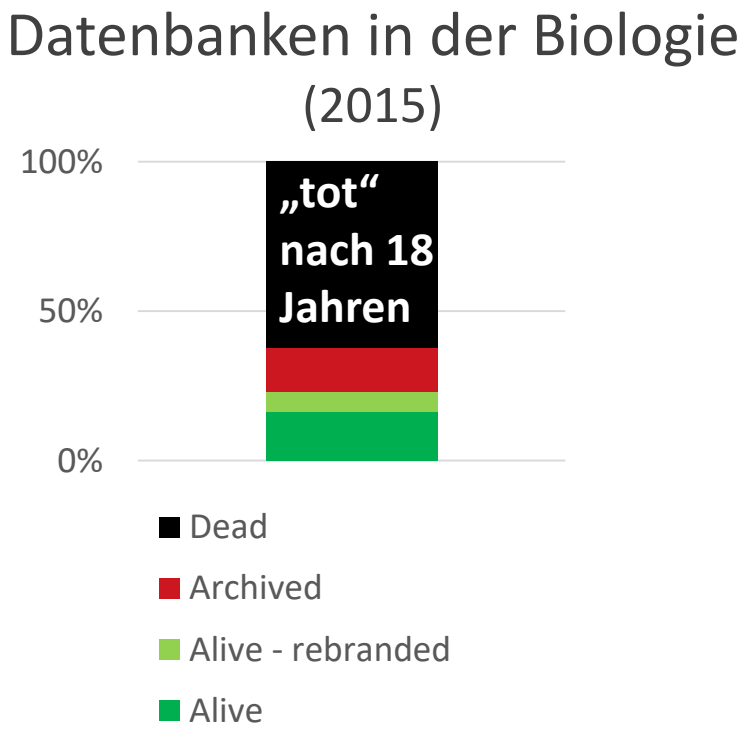
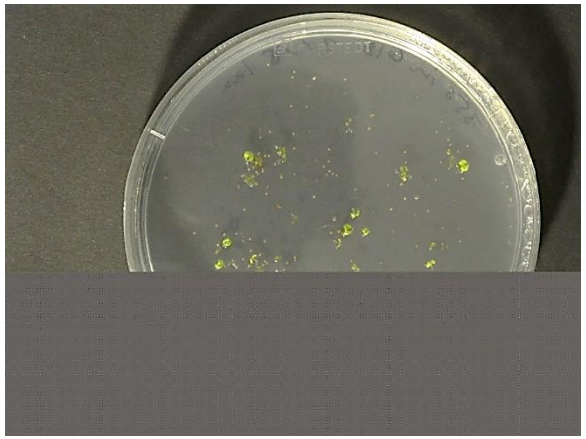
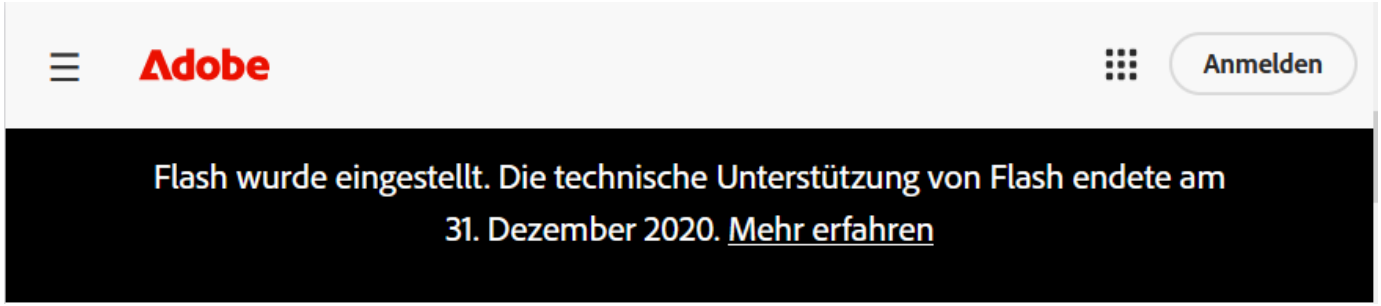


Fig. 1. Showing the proportions of databases that were alive, dead (or becoming so) after a period of 18 years. Quelle: Attwood et al. (2015) *EMBnet. journal*, *21*, 803 https://doi.org/10.14806/ej.21.0.803

Fig. 2. One of the 9-Track ½” Magnetic Tape which stored the original Landsat data. The bit sequence from one of the Landsat files rendered via a hex editor in 2020. Quelle: Doig (2020) https://www.dpconline.org/blog/james-doig-preserving-the-bits



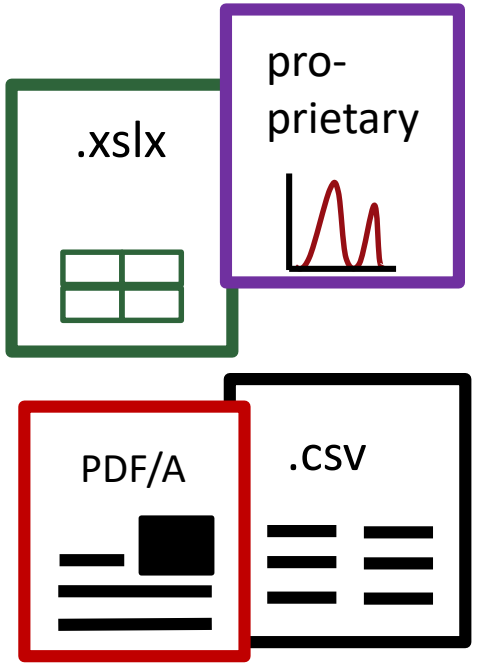
Quelle: Katharina Markus



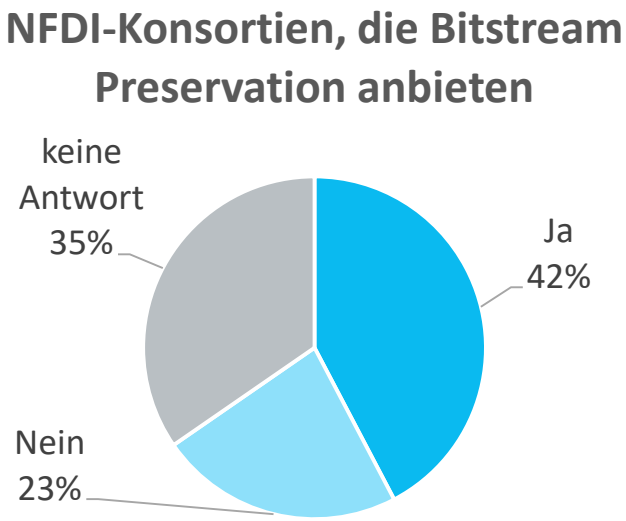
Quelle: https://www.adobe.com/de/products/flashplayer/end-of-life-alternative.html

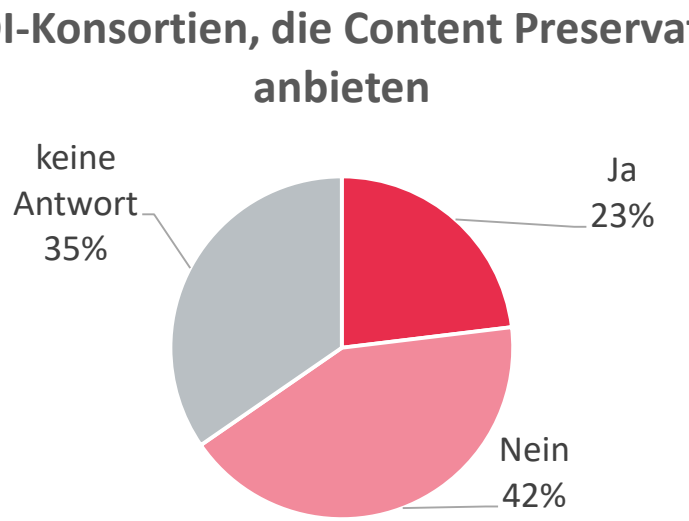
Herausforderungen: Datenvolumina und Komplexität

- **Datenmengen** und **Auswahl / Bewertung**
  - Datentransfer, Serverplatz, Skalierbarkeit, Ressourcen
  - Auswahl, was langzeitarchiviert werden soll (und ggf. wie lange): wer und wie?
- Komplexe **Erhaltungsmaßnahme** und **Qualitätskontrollen**
  - Formatvielfalt (Dateiformate, Datenbanken, Datenmodelle)
    - Z. B. .txt, .csv, .md, .bam.bai, .fa und viele mehr
- **Nachvollziehbarkeit** und benötigte **Kontextinformationen**
  - Z. B. erstellende Software, Computerumgebung, Datenanalysebedingungen und Methodenbeschreibung, verwendete Standards, Nutzungsrechte, usw.



Herausforderungen: Stand Langzeitarchivierung in der NFDI





Quelle: https://doi.org/10.5281/zenodo.10822613

- Umfrage 2023 innerhalb der NFDI (durch NFDI AG LTA)
- Insg. 26 NFDI-Konsortien, 17 Konsortien haben geantwortet
- 11 von 17 Konsortien gaben an, einen LZA-Dienst zu benötigen (2: nein, 4: Bedarf unklar)
- **Bitstream Preservation**: Datenspeicherung, Datenerhalt auf dem bit-Level
- **Content Preservation**: Informationserhalt über Technologie-Änderungen hinweg, Update von Metadaten, etc.

Herausforderungen: Langzeitarchivierung in NFDI-Konsortien

LZA in den Drittmittelanträgen der Fachkonsortien:

- Einbindung existierender Langzeitarchive mit CoreTrustSeal-Zertifizerung (Bsp. FAIRagro, NFDI4Culture)
- Einbindung existierender Fachdaten-Repositorien und –Datenbanken (Bsp. NFDI4Biodiversity)
- Community-Langzeitarchivierungs-Lösung muss erarbeitet werden (eigenes Measure im NFDI-Antrag) (Bsp. NFDIxCS, NFDI4Earth, Text+)
- Awareness, Training, Leitfäden (Bsp. NFDI4Culture, NFDI4Microbiota)

Disclaimer:

- nicht alle Konsortien, die an LZA arbeiten, sind als Beispiele gelistet
- Viele genannte Konsortien arbeiten in mehreren der vier Kategorien

Was braucht es: kooperative Langzeitarchivierung

Arbeitsteilung verschiedener Institutionen und Initiativen

1. **Zusammenarbeit** von Fachdatenbanken, Landesinitiativen und Unibibliotheken, außeruniversitären Forschungsinstitutionen usw.
2. Klare und verbindliche **Verantwortung** (Policies und Zertifizierung)
   - Welches Erhaltungslevel, welche Datenrettungsmaßnahmen
   - Welche Daten
   - Vertrauenswürdigkeit über CoreTrustSeal und nestor-Siegel

Was braucht es: kooperative Langzeitarchivierung

3. Wer hat was bereits langzeitarchiviert: **Dokumentation** über zentrale internationale Datenbank DataCite

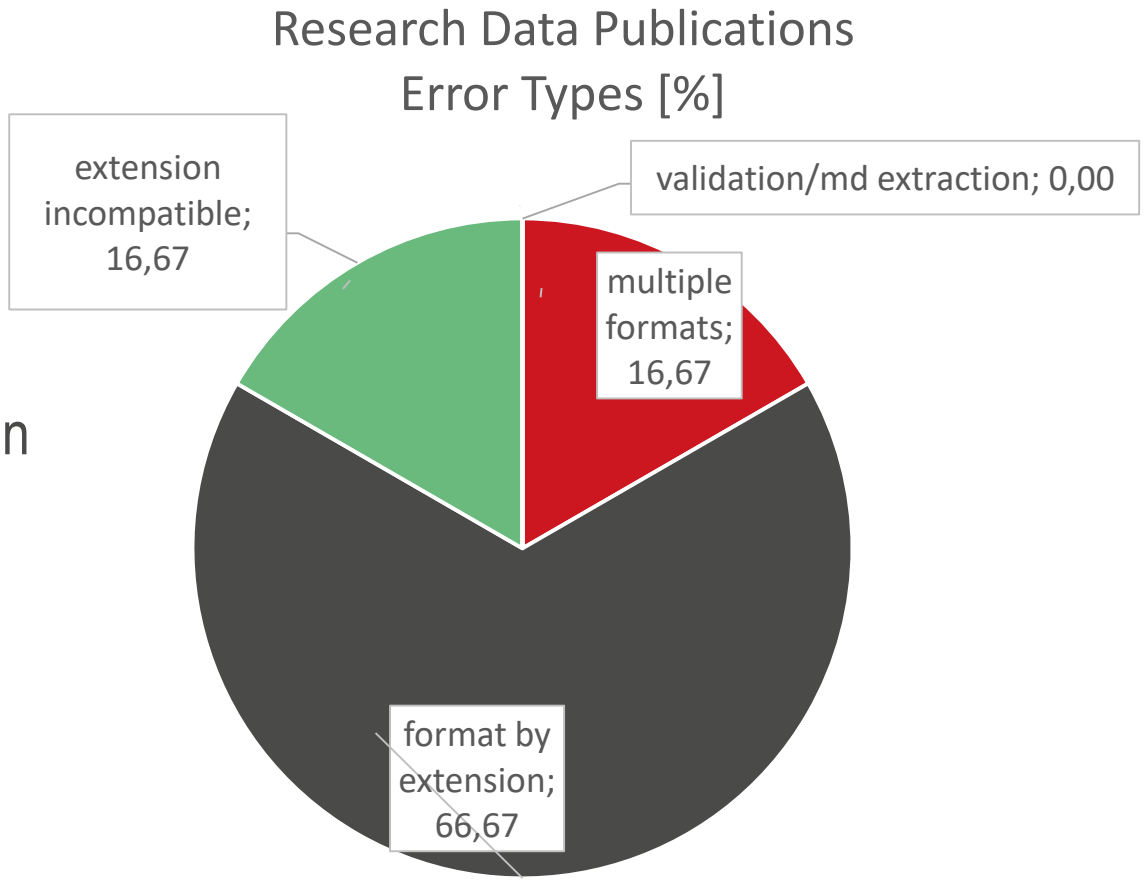
- Metadaten-Element vorgeschlagen (NFDI AG LTA)

Quelle: https://github.com/datacite/datacite-suggestions/discussions/149

- Ziel: Abfrage über DataCite nach Publikation mit Langzeitarchivierungs-Eintrag
- Verknüpfung mit Zertifizierung-Status des Langzeitarchivs/LZA-Service

Was braucht es: ausreichend Personal

- Aufgaben u. a.:
  - Entscheidungen zu sinnvollem **Ressourcen**-Einsatz, Einhaltung von LZA-**Standards**, Durchführung von **Zertifizierung**
  - Beobachtung von (Technologie-) **Entwicklungen** in der LZA-Community und in der Forschungscommunity
  - **Digital forensics/Technical Analyst**-Arbeit, Bewertung von Fehlermeldungen, Reparaturen, Erhaltungsmaßnahmen
- Fachliche Expertise von Vorteil



Publications may in be part of several error types, due to multiple files with different errors in one publications; RD publications n=12
Publications in PUBLISSO – Repository for Life Sciences (2023)
Quelle: https://doi.org/10.5281/zenodo.15197112

Was braucht es: LZA bereits im Forschungsprozess mitdenken

- Frühe **Qualitätskontrolle**
  - vermeidet später Ressourcen-intensive Korrekturen oder nicht-reparierbare Daten
  - Betrifft auch fehlende Informationen (z. B. Kontextinformationen, Rechte)
- LZA in **Data Management Plans (DMPs)**
  - Arbeit am NFDI-DMP-Template-Framework (DMP4NFDI und NFDI AG LTA) (Template V1: Schönau *et al.* 2025)
  - Steht Konsortien + entsprechenden Forschenden zur Verfügung
- Kriterien für die **Auswahl** von Daten für die Langzeitarchivierung
  - Andiskutiert im NFDI AG LTA-Workshop 2024 (Markus *et al.* 2025)
  - Aktivitäten in LaVaH, einzelnen NFDI-Konsortien, EOSC (EOSC-EDEN), etc.

Fazit

- Langfristiger Erhalt von Forschungsdaten: essentiell für gute Forschung und Wissenschaft
- Herausforderung:
  - Große Datenmengen, komplexe/vielfältige Daten, Formate und entspr. LZA-Maßnahmen
  - Nur wenige NFDI-Konsortien bieten z. Zt. Langzeitarchivierung *über* den Erhalt auf dem Bit-Level *hinaus* an (Content Preservation)
- Empfehlungen:
  - Bedarf an Zusammenarbeit von Institutionen und Initiativen, Verbindlichkeit und Zertifizierung
  - Öffentliche und zentrale Dokumentation: wer hat was langzeitarchiviert?
  - Fachpersonal in Langzeitarchiven
  - Qualität, Kontextinformation und Auswahl im Forschungsprozess, u. a. über Data Management Plans

Links, Literatur, Quellen

- Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen 2016, 160 S. urn:nbn:de:101:1-201606229098
- Markus, K., Leinen, P., & Stäcker, T. (2025). Concept for Setting up an LTA Working Group in the NFDI Section "Common Infrastructures". Zenodo. https://doi.org/10.5281/zenodo.14939132
- Beispiele Datennachnutzung und Retraction: mit Dank an die ZB MED-FDM- und Trainings-Teams
- Sielemann K, Hafner A, Pucker B. 2020. The reuse of public datasets in the life sciences: potential risks and rewards. PeerJ 8:e9954 https://doi.org/10.7717/peerj.9954
- Amorós and Puit (2014) Retraction notice to A model study into the effects of light and temperature on the degradation of fingerprint constituents [Science and Justice, 54 (2014) 346 - 350] https://doi.org/10.1016/j.scijus.2015.04.005
- Attwood, Teresa K.; Agit, Bora; Ellis, Lynda B.M.. Longevity of Biological Databases. EMBnet.journal, [S.l.], v. 21, p. e803, may 2015. ISSN 2226-6089. Available at: <https://journal.embnet.org/index.php/embnetjournal/article/view/803>. Date accessed: 29 aug. 2025. doi:https://doi.org/10.14806/ej.21.0.803.
- Doig (2020) Preserving the bits: a salutary tale from the National Archives of Australia https://www.dpconline.org/blog/james-doig-preserving-the-bits
- Markus, K., Naumann, K., Schmalzl, M., Watson, J., & Triebel, D. (2024). Langzeitarchivierung in der NFDI. Zenodo. https://doi.org/10.5281/zenodo.10822613
- kpletsch (2025) Metadata Element for Long-term Preservation https://github.com/datacite/datacite-suggestions/discussions/149
- Markus, K. (2024, September 4). Evaluation of Research Data File Errors - First Results. Zenodo. https://doi.org/10.5281/zenodo.15197112
- Markus, K., Leinen, P., Naumann, K., & Valena, P. (2025). Long Term Data Issues: Workshop about Preservation, Archiving and Access in the NFDI. Long Term Data Issues: Workshop about Preservation, Archiving and Access in the NFDI, online. Zenodo. https://doi.org/10.5281/zenodo.15106332
- Projekt-Webseite von EOSC EDEN und EOSC FIDELIS https://eden-fidelis.eu/
- Schönau, S., Windeck, J., Gonzalez Ocanto, M., Wallace, D., Castro, L. J., Diederichs, K., & Schmitz, D. (2025). The NFDI DMP Template Framework (1-0-0). Zenodo. https://doi.org/10.5281/zenodo.16737079

Für weitere Informationen stehe ich Ihnen gerne zur Verfügung

Dr. Katharina Markus

Leitung Digitale Langzeitarchivierung

ZB MED

Gleueler Straße 60
50931 Köln

markus@zbmed.de

Danke!

ZB MED Informationszentrum Lebenswissenschaften

www.zbmed.de