



~~Künstliche Intelligenz~~
~~Large Language Models~~

~~*Sprachmodelle und*~~ *Forschungsdatenmanagement*

institutionelle (und persönliche) Überlegungen zum Einsatz probabilistischer
Technologien im Forschungsdatenmanagement



Data Center for the Humanities (DCH)

geisteswissenschaftliches Datenzentrum an der Philosophischen Fakultät der
Universität zu Köln

360° FDM-Serviceportfolio

- Allgemeine FDM-Beratung
- Antragsunterstützung
- Begleitendes FDM
- Ressourcenbetreuung
- Archivierung
- Datenpublikationsstrategien
- Audiovisuelle Sprachdaten
- Awarenessmaterial
- Lehre
- On-/Off-Boarding



Disclaimer I (persönlich)

- (zu) große Herausforderung
- keine realistische Chance dem eigenen Auftrag gerecht zu werden
- Einsatz von KI und erzeugte Ergebnisse haben das Potenzial das Vertrauen in Forschung zu erschüttern
- mehr als skeptisch gegenüber dem Einsatz und verlässlichen Nutzen von KI im FDM



Disclaimer II

(institutionell)

- **keine** spezifischen Anforderungen bezüglich des Einsatzes von Sprachmodellen in der FDM-Beratung und darüber hinaus
- **keine** Implementierung von Sprachmodellen in der eigenen Arbeit



Versuch

- explorative Prüfung von Potenzialen für nachhaltiges Forschungsdatenmanagement
- institutionelle und persönliche Einordnung
- ...worüber wir nicht (genug) reden (wobei wir gestern schon damit angefangen haben)
- Dialog



Azeroual 2026

- Herausforderungen
 - Abhängigkeiten von kommerziellen Anbietern und Plattformen
 - Datenhoheit, Datenschutz und rechtliche Unsicherheiten
 - Intransparenz und mangelnde Nachvollziehbarkeit
 - Infrastruktur- und Ressourcenabhängigkeit
 - Fehlende Governance und institutionelle Strategien
 - Spannungsfeld zwischen Innovation und Kontrolle

Engel 2026

- Ansätze für Organisationsformen zum Umgang mit KI in Universitäten

La Sala et al. 2025

- FDM und FAIR als Grundlage für zuverlässige KI-Analysen
- Chancen
 - Automatisierte Datenbereinigung
 - Identifizierung von Mustern in umfangreichen Datensätzen
 - Entwicklung von Datenmanagementplänen



Anwendungsbeispiele

I: Entwicklung einer nachhaltigen Strategie für den Betrieb einer existierenden lebenden Ressource

II: Automatisierte Kodierung von FDM-Beratungsprotokollen zur weiteren Analyse von FDM-Bedarfslandschaften



Anwendungsbeispiel I

Digitale Datenbank

Aufnahmen von Sprecher:innen einer spezifischen Sprache

700 Interviews, jeweils 2 Stunden

300 GB Daten, hauptsächlich .wav

Zugänglichkeit über Website

MySQL, PHP5 + zusätzliche Software: Lame, MP3split, Ffmpeg
Hosting auf Windows-Server einer anderen universitären Bibliothek

Die Bibliothek informierte den
Forschenden, dass sie die Ressource
innerhalb eines Monats abschalten muss...



...und Panik brach aus



Anwendungsbeispiel I

- **Lobbyarbeit**
 - Überzeugung verschiedener Stakeholder
- **Migration**
 - Windows → Linux
- **Reduktion**
 - weniger dynamische Funktionalitäten
- **Implementation**
 - simple Suchfunktion
- **Upload und Testing**
 - Ablage von Ressource und Daten auf verschiedenen Infrastrukturkomponenten
- **Anpassungen**
 - work in progress



Es gibt eine Audiodatenbank, die ca. 300 GB Audiodaten (im .wav Format) enthält. Die Datenbank ist über eine Website zugänglich. Es wird eine MySQL-Datenbank sowie aktuell PHP 5 genutzt. Personen haben freien Zugang zu den Datenbankinhalten, können diese über einen in der Website eingebauten Player abspielen und herunterladen. Außerdem sind in der Website LAME, MP3split und FFmpeg implementiert. Datenbank und Website werden aktuell auf einem Windows-Server betrieben.

Welche Maßnahmen sollten organisatorisch und technisch ergriffen werden, damit Datenbank und Website langfristig online zur Verfügung gestellt werden können? Der Aufwand bei Betrieb von Datenbank und Website (Kuration, Updates, Anpassungen im Softwarecode etc.) sollte dabei so gering wie möglich gehalten werden. Die Funktionalitäten von Datenbank und Website sollten in ihrer aktuellen Form möglichst erhalten bleiben.

Entwickle mir bitte ein entsprechendes Nachhaltigkeitskonzept.



UzK:GPT | KI:Connect





Etwas ist schiefgelaufen!



Anfrage konnte nicht verarbeitet werden

Bitte versuchen Sie es erneut!

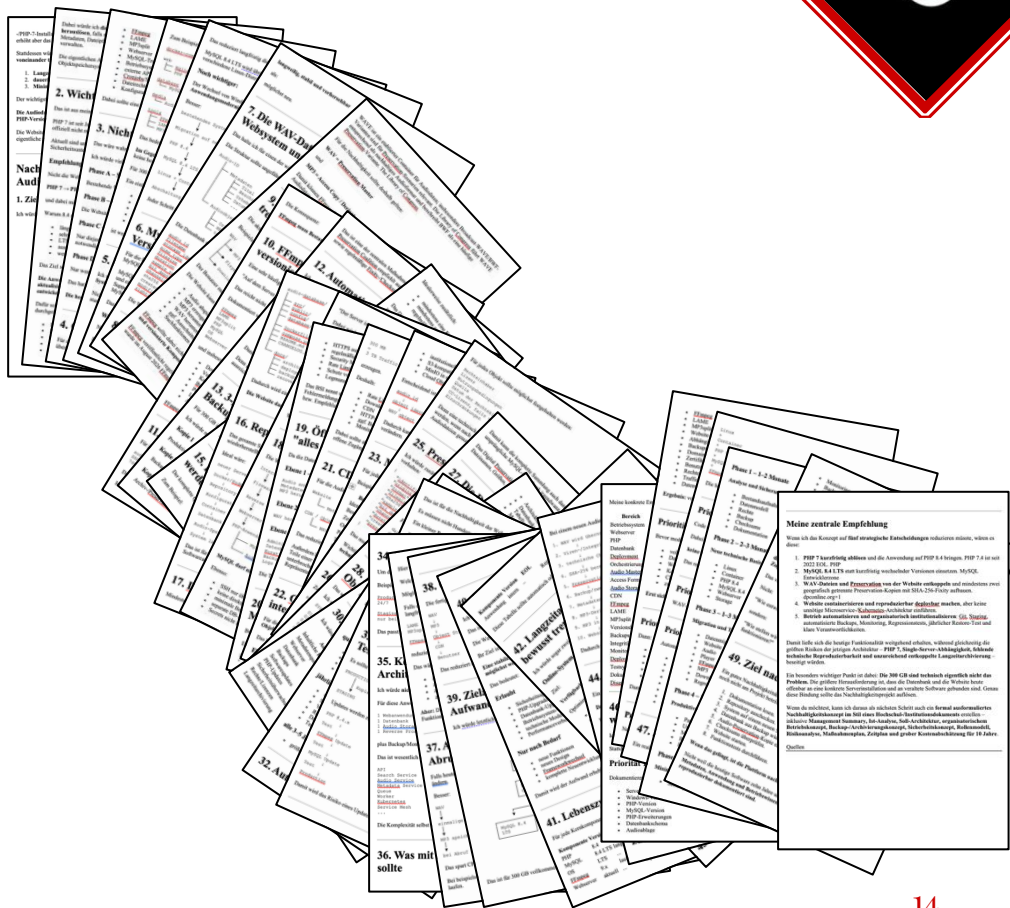
Verstanden

- Mistral Small 3.2 24B Instruct 2506
- OpenAI GPT OSS 120B



ChatGPT





- ChatGPT 5.6 Luna
- 44 Seiten
- 49 individuelle Schritte (mit detaillierten Unterschriften)

u. a.

- Zeitplan und Arbeitsablauf
- grafisch arrangerierte Architektur-Komponenten
- Backup-Strategien
- Disaster-Recovery-Tests
- „Bus Factor“ reduzieren
- Update-Prinzip: möglichst langweilig



Zusammenfassung

Empfehlungen

konkrete Maßnahmen

- veraltete Software ablösen

PHP 5 → PHP 8.4

MySQL → MySQL 8.4 LTS

Windows-Server → Linux LTS

FFmpeg, LAME, MP3Split → Überprüfung und ggf. Reduktion



- Container



- Trennung Audiodaten und Ressource



- Backups und Integritätsprüfungen



- Dokumentation



- automatisierte Updates und Testing



- Definition von Verantwortlichkeiten



- regelmäßige Notfalltests





Bitte überprüfe die folgende Website in Bezug auf ihre technische und organisatorische Nachhaltigkeit: [URL zur Ressource].

Wenn es Maßnahmen für einen noch nachhaltigeren Betrieb der Website gibt, liste diese bitte auf.



Zusammenfassung

Die Ressource ist inhaltlich und organisatorisch bereits vergleichsweise nachhaltig, insbesondere weil es sich um ein langfristig betreutes wissenschaftliches Forschungsdatenangebot handelt. Das DCH für die Ressource als eine der „betreuten Ressourcen“ und übernimmt damit grundsätzlich die dauerhafte Betreuung solcher lebenden Systeme.

Technisch wirkt die Website allerdings relativ alt und funktional stark auf die Bereitstellung der Forschungsdaten fokussiert. Das ist nicht automatisch schlecht für die ökologische Nachhaltigkeit – im Gegenteil: Eine einfache, schlanke Website kann sehr ressourcensparend sein. Es gibt aber einige Punkte, bei denen sich Energieverbrauch, Wartbarkeit, Barrierefreiheit, Sicherheit und langfristige Verfügbarkeit gleichzeitig verbessern ließen.



- langfristige Betreuung
- wissenschaftliche Nachhaltigkeit
- relativ schlanke Website
- gute Erschließung



- Audio optimieren
Komprimierung, Paginierung
- Langzeitarchivierung absichern
Trennung Daten – Ressource; Backups dokumentieren
- technische Wartung dokumentieren
- Performance verbessern
Komprimierung, optimierte Datenübertragung
- CO₂-Verbrauch erfassen
- Barrierefreiheit verbessern
- Daten langfristig nutzbar machen
Nutzung von PIDs; maschinenlesbare Metadaten

Erste Einschätzung

- Umfangreiche Evaluation einer Ressource, die nachhaltig aufgestellt werden soll
- Umfangreiche Evaluation einer Ressource, die vermeintlich bereits nachhaltig aufgestellt wurde
- Notwendigkeiten:
 - Finanzielle und technische Voraussetzungen
 - Kompetenzen
 - Personal
 - ...





Anwendungsbeispiel II

Formales Beschreibungsmodell für das Management von Forschungsdaten

465 individuelle Kategorien mit unterschiedlichen Beschreibungs- und Definitionslevel

Qualitative Inhaltsanalyse

basierend auf der Auswertung von 146 FDM-Beratungsprotokollen

Professionalisierung von FDM

gemeinsame Sprache; Protokollierung und Analyse von FDM-Beratungsprozessen; Formalisierung von FDM



Typ der Kategorie:	thematische Kategorie
Name der Kategorie:	Metadatensammlung (d. aF.)
Inhaltliche Beschreibung:	bezeichnet die Sammlung von erforderlichen Metadaten durch anfragende Forschende zu einem spezifischen Zweck.
Anwendung der Kategorie:	wird verwendet, wenn Informationen darüber identifiziert werden können, dass anfragende Forschende für die Bedienung eines FDM-Bedarfs erforderliche und geeignete Metadaten sammeln und/oder definieren.
Beispiele für Schlüsselbegriffe:	“Zusenden von Spreadsheet mit Metadaten”, “Sammlung von Metadaten für das lokale, fachspezifische Repositorium”, “Sammlung von Metadaten (orientiert an den Vorgaben des lokalen Rechenzentrums zur Beschreibung von Daten”
Weitere Anwendungen:	/
Abgrenzung zu anderen Kategorien:	/
Unterkategorie von:	Maßnahmen anfragende Forschende (d. aF.) – Datenvorbereitung (d. aF.)
Teil von Hauptkategorie:	Archivierung, Publikation
Oberkategorie von:	/





Es gibt ein Kategoriensystem, mit dessen Hilfe schriftliche Protokolle, die Beratungsgespräche zum Thema Forschungsdatenmanagement (FDM) dokumentieren, kodiert werden können. Das Modell verfügt über folgende Ober- und Unterkategorien.

Die Oberkategorie „Fachrichtung“ bezeichnet wissenschaftlich-inhaltliche Fachbereiche. Die Oberkategorie „Fachrichtung“ hat folgende Unterkategorien:

- ...

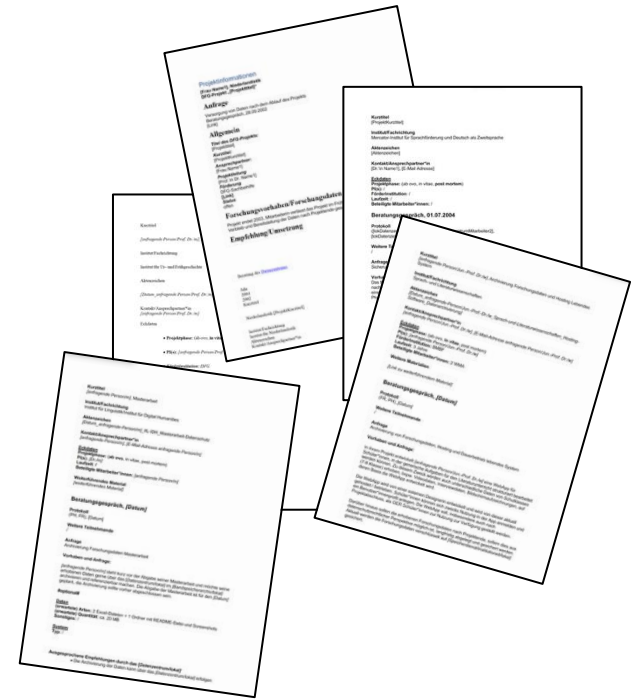


...Kodiere bitte das angehängte Protokoll mit den oben genannten und definierten Unterkategorien. Das bedeutet, dass Du Textstellen, die inhaltlich zu den genannten und definierten Unterkategorien passen, mit den entsprechend passenden Unterkategorien markiert und verknüpft werden. Bitte gib mir die von Dir verwendeten Unterkategorien sowie die entsprechend kodierten Textstellen als Tabelle aus.

Verschiedene Informationslevel

- a. inhaltliche Beschreibungen
- b. Anwendung
- c. Beispiele
- d. Anwendung + Beispiele
- e. inhaltliche Beschreibung + Anwendung
- f. inhaltliche Beschreibung + Beispiele
- g. inhaltliche Beschreibung + Anwendung + Beispiele

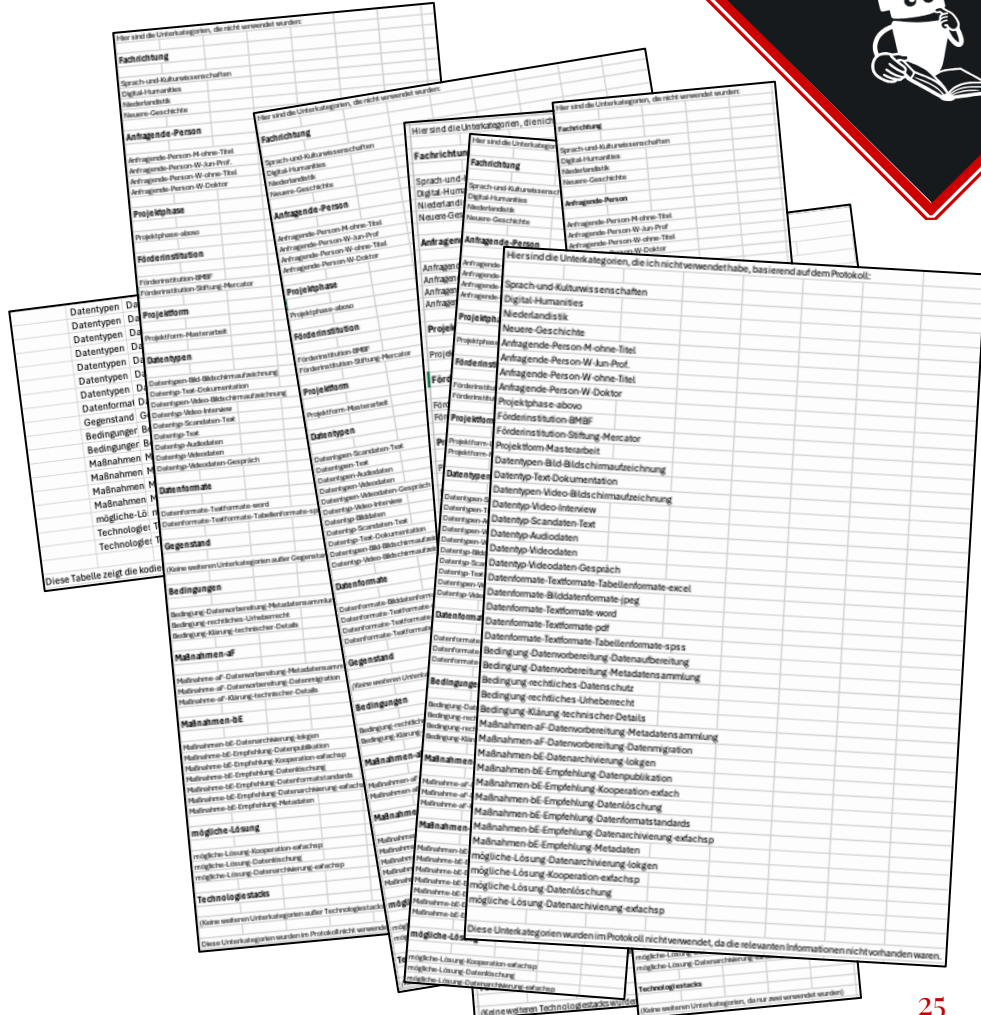
alle Kategorien, die in den untersuchten FDM-Beratungsprotokollen vorkommen + weitere Kategorien.





UzK:GPT | KI:Connect





- 25

[illegible]

- ☒ Matches
- ☐ Keine Kodierung
- ☒ Richtiger Code, alternative Textstelle
- ☒ Falscher Code, richtige Textstelle
- ☐ Kodierung neuer Textstellen

Erste Einschätzung

- grundsätzlich hohe Trefferquote 
- kein erkennbarer Einfluss der Informationseinheiten / -dichte auf Resultate
- keine dominante Häufigkeit von falschen Kodierungen – dennoch nicht zu ignorierende Fehler und Halluzinationen
- weiteres Vorgehen:
 - DFG-Projekt in Planung → Training eines RAG-Systems



Zusammenfassung

- vorhandene Szenarien für den Einsatz von KI-Agenten zur Unterstützung von nachhaltigem FDM
 - Entwicklung nachhaltiger Konzepte für den Betrieb lebender Ressourcen
 - Protokollierung und Analyse von FDM-Beratungsgesprächen und -landschaften
- weitere Potenziale
 - Entwicklung nachhaltiger lebender Ressourcen und Infrastrukturen
 - Front- und Backend Programmierarbeiten
 - User-oriented Interfaces
 - Datenmanagementplänen (bzw. grundsätzlich FDM-Workflows aus Forscher:innen-Perspektive)
 - Verarbeitung von Forschungsdaten und -ergebnissen (Analyse, Auszeichnung, Kodierung, Dokumentation etc.)

...aber ist es das wirklich wert?

...und geht es beim Thema FDM und KI nicht eigentlich um viel mehr?



...worüber wir nicht (genug) reden

- Angriffsrisiken für akademische IT-Infrastrukturen
- KI-, Daten- und Medien-Inkompetenzen
- digitaler Kolonialismus
- False Data / Fake Data (Vertrauen)
- fehlende Transparenz
- unkontrollierte und nicht verarbeitbare Informations- und Datenfluten
- politische, wirtschaftliche und ideologische Abhängigkeiten und Einflüsse



Forschungsdatenmanagement | Demokratie

- Forschungsdatenmanagement als **Grundbedingung für verlässliche, erkenntnisbasierte, überprüfbare und reproduzierbare Forschung** und die Verankerung von Vertrauen in Fakten
- **massenhafte Bereitstellung FAIRer Forschungsdaten und -ergebnisse**, nicht nur zum Selbstzweck, sondern v. a. zur Beförderung einer offenen, faktengeleiteten und pluralistischen, demokratischen Gesellschaft als **elementare Aufgabe des Forschungsdatenmanagements**





- Azeroual, Otmane. 2026. „Digitale Souveränität Im Zeitalter Generativer KI: Herausforderungen Und Handlungsoptionen Für Wissenschaftliche Informationsinfrastrukturen“. *Information – Wissenschaft & Praxis*, Online-Vorab-Publikation, Juni 20. <https://doi.org/10.1515/iwp-2026-3017>.
- Engel, Karsten. 2026. „Künstliche Intelligenz in Wissenschaftlichen Bibliotheken . Einige Organisationsformen Im Spannungsfeld Universitärer Governance“. *Bibliothek Forschung Und Praxis* 50 (2): 175–86. <https://doi.org/10.1515/bfp-2025-0061>.
- Imbusch, Peter, und Joris Steg. 2024. „Künstliche Intelligenz und gesellschaftlicher Wandel – eine Herausforderung für demokratische Macht- und Herrschaftsverhältnisse“. In *Künstliche Intelligenz, Mensch und Gesellschaft*, herausgegeben von Michael Heinlein und Norbert Huchler. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-43521-9_15.
- Kondratenko, Aleksandr, Anna Kondratenko, und Konstantin Kamberov. 2026. *May AI be sustainable? An outlook on sustainability and technology*. 5/2025 (Mai). <https://www.springerprofessional.de/en/may-ai-be-sustainable-an-outlook-on-sustainability-and-technolog/51841524>.
- Kreissl, Reinhard, und Roger Von Laufenberg. 2024. „Risiken und Gefahren der ‚Künstlichen‘ ‚Intelligenz‘“. In *Künstliche Intelligenz, Mensch und Gesellschaft*, herausgegeben von Michael Heinlein und Norbert Huchler. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-43521-9_10.
- Kurz, Susanne, und Øyvind Eide. 2024. *Vertrauen in die Wirklichkeit – AI, Trust und Reliability in den Digital Humanities*. Februar 21. <https://doi.org/10.5281/ZENODO.10698375>.
- La Sala, Beate Ulrike, Arnela Balic, und Robert Werth. 2025. *Künstliche Intelligenz und FDM: Synergien, Herausforderungen und Grenzen der Risikominderung*. <https://doi.org/10.11588/HEIDOK.00036411>.
- Luusua, Aale. 2023. „Katherine Crawford: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence: Yale University Press (April 6, 2021). Hardcover, 336 Pages, ISBN: 9780300209570“. *AI & SOCIETY* 38 (3): 1257–59. <https://doi.org/10.1007/s00146-022-01488-x>.
- Mühlhoff, Rainer. 2025. *Künstliche Intelligenz und der neue Faschismus*. Reclams Univesal-Bibliothek. Reclam.
- Raab, Heike. 2024. „Toys are us: KI, Diversität und soziale Ungleichheit. Oder: KI für Alle?“ In *Künstliche Intelligenz, Mensch und Gesellschaft*, herausgegeben von Michael Heinlein und Norbert Huchler. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-43521-9_11.



Dialog

Dr. Patrick Helling
Data Center for the Humanities (DCH)
Universität zu Köln
Universitätsstraße 22
50937 Köln

E-Mail: patrick.helling@uni-koeln.de
Tel.: +49 221 470 1307
ORCID: 0000-0003-4043-165X



The Simpsons, Season 13, Episode 13
(nicht KI-generiert)