



**ANTIFASCHISTISCHES
SCHUTZ-INSTITUT**
für reflexive Gesellschaftsethik

KI-LAB-26

Artikel und Methodenpapier zur Ordnung digitaler Archive



Autor: Andre(a) Schmidt – www.as-institut.de – **Stand:** 2026_06_06 – **Version:** 1.1

Inhaltsverzeichnis:

1. Publizistischer Artikel: Gegen die Verflachung

- 1.1 Einleitung: Das stille Problem der Gegenwart
- 1.2 Ausgangslage: Warum große Archive nicht von selbst intelligenter werden
- 1.3 Der Kern der Antwort: Vom Archiv zur Wissensinfrastruktur
- 1.4 Die epistemische Schutzfunktion der Architektur
- 1.5. Warum dies mehr ist als Dateimanagement
- 1.6 Veröffentlichung, Öffentlichkeit und der Ort der Publikation
- 1.7 Die größere Tragweite: Asymmetrische Befähigung
- 1.8 Emanzipation nicht als Pathos, sondern als methodischer Zugewinn
- 1.9 Möglichkeiten der Weiterentwicklung
- 1.10 Schluss: Eine Infrastruktur gegen das Vergessen der Formen

2. Ki-Lab-26 Archive Struktura v1.

- 2.0.0.1 Dokumenttyp
- 2.0.0.2 Lizenzhinweis
- 2.0.1 Abstract
- 2.0.2 Schlagwörter

2.1 Vorbemerkung

2.2 Problemstellung und Ausgangslage

- 2.2.1 Das Grundproblem großer persönlicher Wissensarchive
- 2.2.2 Begrenzte Kontextfenster als epistemische und technische Engstelle
- 2.2.3 Vom Archiv zur strukturierten Wissensinfrastruktur
- 2.2.4 Das Problem des Flattenings

2.3 Benennung des Themenfeldes auf Mikro-, Meso-, Makro- und Metaebene

- 2.3.1 Mikroebene: Quellennahe Einheitenerschließung, Segmentierung und Metadatenanreicherung
- 2.3.2 Mesoebene: Wissensorganisation, Korpusstrukturierung und taxonomisch kontrollierte Dossierbildung
- 2.3.3 Makroebene: Mehrschichtige Archiv- und Retrieval-Architektur
- 2.3.4 Metaebene: Provenienzbewusste Epistemologie digitaler Wissensinfrastrukturen

2.4 Projektdefinition

- 2.4.1 Arbeitsbezeichnung und Status

2.4.2 Gegenstandsbereich

2.4.3 Zielperspektive

2.5 Disziplinäre Einordnung des Forschungsbereichs

2.5.1 Archivwissenschaft und digitale Langzeitarchivierung

2.5.2 Informationswissenschaft und Knowledge Organization

2.5.3 Wissensrepräsentation und Ontologie-Engineering

2.5.4 Retrieval-Architekturen, graphbasierte und hybride Suchverfahren

2.5.5 Dokumentations- und Lehrforschung

2.6 Leitprinzipien des Gesamtmodells

2.6.1 Nicht-Destruktivität

2.6.2 Provenienzschutz

2.6.3 Ebenentrennung

2.6.4 Kontrollierte Taxonomie

2.6.5 Konservative Rekonstruktion

2.6.6 Mehrstufige Verdichtung

2.6.7 Querverschließung statt bloßer Baumlogik

2.6.8 Unsicherheit als positive Information

2.7 Architektur des Projektökosystems

2.7.1 Grundform: Staffeldächtnis

2.7.2 Die Ebenen des Modells

2.7.2.1 Roharchiv

2.7.2.2 Inventar- und Manifestebene

2.7.2.3 Segmentebene

2.7.2.4 Cluster- und Dossier-Ebene

2.7.2.5 Community- und Cluster-Reports

2.7.2.6 Oberkarten- und Bereichsebene

2.7.2.7 Root-Map

2.7.3 Vektorraumschicht

2.7.4 Wissensgraph-Schicht

2.7.5 Archivgraph und Inferenzgraph

2.7.6 Hierarchie und Query-Pfade

2.8 Die bislang ausgebildeten Teilprojekte

2.8.1 Screenshot-Archivierung und juristisch vorsichtige Vorstrukturierung

2.8.1.1 Zweck

2.8.1.2 Strukturprinzipien

2.8.1.3 Methodische Besonderheit

2.8.2 Telegram-Backups als Artikelarchiv

2.8.2.1 Ausgangslage

2.8.2.2 Zweck

2.8.2.3 Strukturprinzipien

2.8.2.4 Bedeutung für das Gesamtprojekt

2.8.3 Perspektive eines ChatGPT-Archivprojekts

2.8.3.1 Erwartete Struktur

2.8.3.2 Besonderheit

2.9 Dokumenttypen und Standardartefakte

2.9.1 Maschinenlesbare Inventare

2.9.2 Menschlich lesbare Review-Dateien

2.9.3 Ordnerzusammenfassungen

2.9.4 Sidecar-Dateien

2.9.5 Community- und Cluster-Reports

2.9.6 Dokumentausgaben

2.10 Taxonomische Ordnung und Versionierung

2.10.1 Funktion der Taxonomie

2.10.2 Typische Klassifikationsfelder

2.10.3 Methodische Bedeutung

2.10.4 Taxonomische Drift

2.10.5 Taxonomische und ontologische Versionierung

2.11 Provenienz, Unsicherheit und epistemische Disziplin

2.11.1 Warum Provenienz zentral ist

2.11.2 Transformationsprovenienz

2.11.3 Unsicherheit als positive Information

2.11.4 Dreiteilung als Standard

2.11.5 Relationsbezogene Unsicherheitslogik

2.12 Von der Hierarchie zum hybriden Modell

2.12.1 Die Grenze reiner Baumstrukturen

2.12.2 Die Notwendigkeit einer relationalen Zusatzschicht

2.12.2.1 Mögliche Knotentypen

2.12.2.2 Mögliche Beziehungstypen

2.12.3 Die ergänzende Vektorraumschicht

2.12.4 Baum, Graph und Vektor als komplementäre Logiken

2.12.5 GraphRAG als Bezug, nicht als erschöpfende Definition

2.13 Die Rolle der KI-gestützten Bearbeitung

2.13.1 KI nicht als Gedächtnis, sondern als Strukturierungswerkzeug

2.13.2 Arbeitsmodi

2.13.3 Grenzen

2.13.4 Maschinelle Verdichtung als dokumentationspflichtiger Akt

2.14 Rebuild- und Evolutionslogik

2.14.1 Versionierbarkeit

2.14.2 Rebuild-Regeln

2.14.3 Langfristigkeit

2.15 Theoretische Verdichtung für Dokumentation und Lehre

2.15.1 Forschungsfeldkarte

2.15.2 Architekturkarte

2.15.3 Methodenkarte

2.15.4 Epistemische Schutzkarte

2.15.5 Lehrmodule

2.16 Wissenschaftliche und praktische Tragweite

2.16.1 Wissenschaftliche Bedeutung

2.16.2 Praktische Bedeutung

2.17 Grenzen und offene Probleme

2.17.1 Taxonomische Drift

2.17.2 Überverdichtung

2.17.3 Segmentierungsprobleme

2.17.4 Graphische Komplexität

2.17.5 Vektorielle Intransparenz

2.17.6 Transformationsblindheit

2.18 Vorläufige Gesamtbilanz

2.19. Arbeitsdefinition für die weitere Entwicklung

2.20. Arbeitsauftrag für eine mögliche v1.2

1. Publizistischer Artikel: Gegen die Verflachung

Warum die Ordnung digitaler Archive zu einer Frage geistiger Souveränität wird

1.1 Einleitung: Das stille Problem der Gegenwart

Die digitale Gegenwart leidet nicht in erster Linie an Informationsmangel. Sie leidet an einer anderen, subtileren Störung: an der Unfähigkeit, wachsende Wissensbestände so zu organisieren, dass ihre innere Geschichte erhalten bleibt. Gespeichert wird heute beinahe alles. Erschlossen wird davon nur ein Bruchteil. Noch seltener wird es in einer Form erschlossen, die Herkunft, Kontext, zeitliche Schichtung, begriffliche Entwicklung, innere Widersprüche und unterschiedliche Geltungsgrade nicht einebnen, sondern sichtbar hält.

Genau an diesem Punkt setzt die im Projekt **Ki-Lab-26-Archive-Struktur-v1_1** formulierte Architektur an. Ihr Gegenstand ist auf den ersten Blick technisch: Archive, Metadaten, Segmentierung, Cluster, Vektorräume, Wissensgraphen, Retrieval-Pfade. In Wahrheit handelt es sich jedoch um weit mehr als eine technische Organisationsfrage. Es geht um die Bedingungen, unter denen Denken, Dokumentieren, Forschen und Erinnern im digitalen Raum souverän bleiben können. Es geht darum, ob ein wachsendes Archiv bloß zu einer Ansammlung unverbundener Dateien wird oder zu einer Wissensinfrastruktur, die spätere Erkenntnis nicht behindert, sondern ermöglicht.

Der Sinn dieses Vorhabens ist daher nicht bloß administrativ. Es soll nicht einfach „Ordnung schaffen“, sondern die Verflachung verhindern, die entsteht, wenn heterogene Materialien in glatte, gleichförmige Datenobjekte übersetzt werden. Der Begriff, der diese Gefahr präzise benennt, lautet hier **Flattening**: die Tendenz, unterschiedliche Dokumenttypen, Entstehungskontexte, Bearbeitungsstufen und epistemische Statuslagen in ein und dieselbe operative Oberfläche zu pressen. Das Projekt antwortet darauf mit einer Gegenbewegung: Nicht Reduktion auf das Einfachste, sondern eine geordnete Mehrschichtigkeit, in der Verdichtung möglich wird, ohne Herkunft und Pfadabhängigkeit zu zerstören.

1.2 Ausgangslage: Warum große Archive nicht von selbst intelligenter werden

Wer über Jahre hinweg forscht, publiziert, kommentiert, dokumentiert, chatbasiert arbeitet, Zwischenstände sichert, Screenshots ablegt, Exportdateien archiviert, Projekttexte fortschreibt und digitale Gesprächsverläufe sammelt, erzeugt kein geordnetes Werkverzeichnis, sondern ein

vielgestaltiges Gelände. In einem solchen Gelände liegen Essays neben Rohnotizen, Telegram-Exporte neben Dokumentfassungen, Chatverläufe neben publizierten Texten, Bilder neben strukturellen Analysen, technische Skizzen neben theoretischen Verdichtungen. Diese Heterogenität ist kein Unfall, sondern Ausdruck realer Arbeitsweise. Das Problem beginnt dort, wo diese reale Arbeitsweise später nicht mehr zugänglich ist, weil das Archiv zwar vollständig, aber operativ stumm geworden ist.

Mit jeder Vergrößerung des Bestands nimmt die Schwierigkeit zu, relevante Materialien wiederzufinden, Zusammenhänge quer über unterschiedliche Formate hinweg zu erkennen, ältere Bearbeitungsstufen von späteren Verdichtungen zu unterscheiden und zu klären, was an einem Datum, einer Zuschreibung oder einer Relation tatsächlich sicher ist. Diese Schwierigkeit verschärft sich weiter, sobald große Sprachmodelle hinzutreten. Denn solche Modelle verfügen zwar über bemerkenswerte Verdichtungs- und Rekombinationsfähigkeiten, aber nicht über unbegrenzte Kontextfenster. Sie „erinnern“ kein großes Archiv als Ganzes, sondern können immer nur einen Ausschnitt aktiv halten. Daraus folgt eine entscheidende Einsicht: Ein großes digitales Archiv wird nicht schon dadurch für KI nutzbar, dass es digital vorliegt. Es muss in einer Form vorbereitet werden, die sinnvolle Pfadwahl, schrittweise Verdichtung, gezieltes Nachladen und kontrollierte Rückbindung an Quellen erlaubt.

1.3 Der Kern der Antwort: Vom Archiv zur Wissensinfrastruktur

Die in Ki-Lab-26 ausgearbeitete Struktur schlägt deshalb vor, Archive nicht länger als flache Dateimengen zu behandeln, sondern als **mehrschichtige Wissensinfrastrukturen**. Der entscheidende Gedanke lautet: Zwischen Rohmaterial und späterer Anfrage muss eine Architektur treten, die weder das Material unberührt und damit unzugänglich lässt, noch es so stark glättet, dass seine innere Geschichte verlorengeht.

Diese Architektur arbeitet gestuft. Am unteren Ende steht das Roharchiv, das unangetastet bleibt. Darüber liegt eine Inventar- und Manifestebene, auf der Dateien, Exportordner, Nachrichtensequenzen oder Bildserien technisch und kontextuell beschrieben werden. Daran schließt sich eine Segmentebene an, auf der operative Einheiten gebildet werden: nicht zu klein, um sinnlos zu fragmentieren, nicht zu groß, um für spätere Auswertung unhandlich zu bleiben. Darüber entstehen Cluster und Dossiers, also thematisch kontrollierte Zusammenhänge. Noch darüber folgen Oberkarten, Bereichsübersichten und Root-Maps, die nicht den Gesamtinhalt wiederholen, sondern Zugänge ordnen. In der überarbeiteten v1.1 treten ergänzend Community- und Cluster-Reports hinzu, ferner eine Vektorraumschicht für latente semantische Nähe und perspektivisch eine Wissensgraph-Schicht für explizite relationale Quererschließung.

Damit wird aus einem Archiv nicht einfach ein schnelleres Suchsystem, sondern ein **Staffelgedächtnis**. Dieser Begriff ist zentral. Gemeint ist eine Gedächtnisform, in der nicht alles gleichzeitig präsent gehalten werden muss. Stattdessen trifft eine Anfrage zunächst auf verdichtete Oberstrukturen, von dort auf logisch einschlägige Unterbereiche und erst danach – falls nötig – auf tiefere Quellschichten. Das ist nicht bloß effizient. Es ist erkenntnistheoretisch bedeutsam. Denn es erlaubt Verdichtung, ohne den Anspruch auf Rückverfolgbarkeit preiszugeben.

1.4 Die epistemische Schutzfunktion der Architektur

Die Stärke des Modells liegt nicht allein in seiner Schichtung, sondern in seinen Schutzmechanismen. Diese Schutzmechanismen reagieren auf ein Problem, das in vielen digitalen

Systemen unterschätzt wird: Daten können nicht nur fehlen, sie können auch auf problematische Weise „zu glatt“ werden. Gerade dort, wo KI zur Verdichtung eingesetzt wird, entsteht leicht der Eindruck, eine Zusammenfassung oder ein Cluster sei einfach die neutrale Essenz dessen, was im Material ohnehin enthalten war. Das ist ein Irrtum. Jede Verdichtung ist ein Eingriff. Jeder Eingriff hat Voraussetzungen, Entscheidungen, Auslassungen, Hervorhebungen und blinde Flecken.

Darum insistiert die Architektur auf mehreren Trennungen. Erstens auf der Trennung von **Quelle**, **Normalisierung** und **Deutung**. Was direkt extrahiert oder sichtbar ist, darf nicht unbemerkt mit dem verschmelzen, was formell bereinigt oder interpretativ eingeordnet wurde. Zweitens auf der Trennung zwischen gesichertem Befund und unsicherer Relation. Ein Dokument kann eindeutig vorliegen, während die Beziehung, in der es zu einem anderen Dokument steht, nur eine Hypothese ist. Drittens auf der Ausweitung des Provenienzbegriffs: Nicht nur das Ausgangsmaterial, sondern auch maschinelle Verdichtungen müssen ihre eigene Entstehungsgeschichte mitführen. Modelltyp, Promptfamilie, Parametrisierung, Laufkontext und etwaige Nachbearbeitung gehören daher nicht an den Rand, sondern in die Dokumentation selbst.

Diese Architektur ist in diesem Sinn nicht nur leistungsorientiert, sondern auch defensiv. Sie schützt das Archiv gegen Verwechslungen, gegen stille Überschreibungen, gegen taxonomische Verwilderung, gegen Überverdichtung, gegen die rhetorische Unsichtbarkeit maschineller Filter und gegen jenen Verlust der inneren Geschichte, der viele gegenwärtige KI-Anwendungen epistemisch fragwürdig macht.

1.5. Warum dies mehr ist als Dateimanagement

Es wäre verfehlt, dieses Projekt als gehobene Ordnerlogik zu beschreiben. Sein eigentlicher Gegenstand ist nicht Dateiordnung, sondern **Wissensorganisation unter Bedingungen digitaler Überkomplexität**. Darum lässt es sich auch nicht treffend in bloß einem Forschungsbereich verorten. Archivwissenschaftliche Prinzipien wie Provenienz und Nicht-Destruktivität spielen ebenso hinein wie informationswissenschaftliche Fragen der Klassifikation, wissensrepräsentative Probleme von Knotentypen und Relationen, retrieval-architektonische Überlegungen zu Suchpfaden und Kontextfenstern sowie dokumentations- und lehrbezogene Fragen der Vermittelbarkeit. Die in v1.1 vorgenommene Präzisierung ist gerade darin wichtig, dass sie zu grobe Etiketten zurückweist. Es geht weder einfach um „Personal Knowledge Management“ noch schlicht um „Enterprise Content Management“. Der Gegenstand ist anspruchsvoller: eine hybride, provenancebewusste und mehrstufige Wissensinfrastruktur, die mit Hierarchie, Graph und Vektorraum komplementär arbeitet.

Darin liegt auch der sachliche Kern der Nähe zu Ansätzen wie GraphRAG. Nicht weil das Projekt einfach unter einen Trendbegriff subsumiert werden sollte, sondern weil es dieselbe strukturelle Einsicht teilt: Große Bestände werden nicht dadurch klüger nutzbar, dass man bloß mehr Rechenleistung auf sie wirft. Entscheidend ist die Qualität ihrer Vorstrukturierung. Wer Entitäten, Beziehungen, Segmentgrenzen, Cluster, Verdichtungen und Pfade kontrolliert organisiert, ersetzt rohe Rechengewalt zu einem erheblichen Teil durch **architektonische Intelligenz**.

1.6 Veröffentlichung, Öffentlichkeit und der Ort der Publikation

An diesem Punkt tritt eine naheliegende Frage auf, die im Zuge der Projektentwicklung ausdrücklich diskutiert wurde: Wie verhält sich eine solche Architektur zur Öffentlichkeit? Sollte

ein derartiges Papier in einem öffentlichen Kanal publiziert werden? Könnte es sicherheitsrelevant sein? Könnte es unerwünschte Wirkungen entfalten?

Diese Frage verdient eine nüchterne Behandlung. Im engeren sicherheitstechnischen Sinn ist eine Architektur dieser Art keine Anleitung zu Exploits, keine kryptographische Durchbruchstechnik, keine militärische Handlungsanweisung und keine operative Sabotageform. Sie beschreibt keine Waffentechnik, sondern eine Methode der Wissensorganisation. Insofern liegt ihre Brisanz nicht im Sinn klassischer Staatssicherheitsbedrohung. Ihre eigentliche Sprengkraft liegt woanders: in der Möglichkeit, Abhängigkeiten von zentralisierten, glättenden und vorstrukturierten Erkenntnisumgebungen zu reduzieren.

Gerade deshalb ist aber der **Publikationsort** nicht nebensächlich. Das Risiko liegt weniger im Papier selbst als in der Umgebung seiner Rezeption. Plattformen, die auf schnelle Reaktion, Affektverdichtung, ideologische Verkürzung und kontextarme Weiterverbreitung angelegt sind, können komplexe, methodisch strenge Texte in genau jener Weise verflachen, gegen die sich die Architektur wendet. Ein solcher Text kann dort aus seinem methodischen Zusammenhang gerissen, auf Schlagworte reduziert und von sehr verschiedenen Milieus als bloße Bestätigung bereits vorhandener Affekte vereinnahmt werden.

Die klügere Lösung besteht deshalb nicht in Geheimhaltung, sondern in **strategischer Provenienzicherung der Veröffentlichung selbst**. Eine Primärveröffentlichung in einem kontrollierten, dokumentierbaren und eigens verantworteten Raum – etwa auf einer eigenen Seite, in einem sauber referenzierbaren Repository oder in einer publizistisch kontrollierten Umgebung – wahrt die Integrität des Textes besser als eine unmittelbare Primärplatzierung in einer Plattformöffentlichkeit, deren Binnenlogik selbst flattening-begünstigend wirkt. Eine nachgelagerte Verlinkung in öffentliche Kanäle kann dann erfolgen, ohne die Herkunft des Textes zu entkoppeln.

1.7 Die größere Tragweite: Asymmetrische Befähigung

Noch interessanter als die Frage der Plattform ist eine zweite, strategisch weiter reichende Überlegung, die sich im Anschluss an die Diskussion aufdrängt: Wenn eine solche Architektur Einzelakteure, kleine Forschungszusammenhänge oder zivilgesellschaftliche Projekte emanzipieren kann, liegt dann nicht ebenso nahe, dass sie auch institutionelle Akteure mit knapperen Ressourcen stärkt?

Diese Überlegung ist ernst zu nehmen. Der entscheidende Punkt liegt darin, dass das Projekt einen anderen Weg der Leistungssteigerung anbietet als die gegenwärtig dominanten Modelle großer Technologiekonzerne. Dort wird das Problem großer Wissensmengen oft durch maximale Skalierung beantwortet: größere Modelle, größere Rechenzentren, größere Kontextfenster, größere Energiebudgets. Ki-Lab-26 setzt dem keine bloße Gegentechnik entgegen, sondern eine andere Rationalität. Nicht rohe Kapazität steht im Zentrum, sondern **strukturelle Vorintelligenz**. Wenn ein Archiv diszipliniert segmentiert, taxonomisch stabilisiert, provenancebewusst beschrieben, in Dossiers verdichtet und durch Graph- und Vektorraumschichten quer erschlossen ist, kann ein wesentlich kleineres Modell mit deutlich weniger Ressourcen Aufgaben übernehmen, die andernfalls massive Hardware- und Lizenzbudgets erfordern würden.

Hier liegt der Punkt dessen, was man als **asymmetrische Befähigung** bezeichnen kann. Gemeint ist die Möglichkeit, dass methodische und architektonische Qualität ein Defizit an roher Rechenmacht partiell kompensiert. Das ist für Einzelne bedeutsam, die ihre eigenen Archive emanzipiert

verwalten wollen. Es ist ebenso bedeutsam für kleinere Institute, Initiativen, Archive, NGOs oder auch Staaten und Verwaltungen, die nicht über die Finanz- und Infrastrukturbasis verfügen, um den großskaligen Plattformmodellen einfach zu folgen.

Diese Form der Befähigung ist nicht mit staatlicher Destabilisierung zu verwechseln. Sie bedroht nicht die Sicherheit von Staaten im engeren Sinn, wohl aber die Selbstverständlichkeit technologischer Hegemonie. Sie zeigt, dass Erkenntnisfähigkeit nicht notwendig identisch ist mit maximalem Budget. Wer die Architektur seiner Daten ernst nimmt, gewinnt einen Spielraum zurück, der in der gegenwärtigen Debatte oft unterbelichtet bleibt: **digitale Souveränität durch Methode.**

1.8 Emanzipation nicht als Pathos, sondern als methodischer Zugewinn

Der Begriff des Empowerments ist in vielen Diskursen abgegriffen. Im Zusammenhang dieser Architektur erhält er jedoch einen präziseren Sinn. Emanzipation meint hier nicht bloß das subjektive Gefühl von Kontrolle, sondern eine real veränderte Fähigkeit, mit eigenen Beständen anders zu arbeiten als unter Bedingungen zentralisierter Abhängigkeit.

Diese Befähigung hat mehrere Ebenen. Auf der persönlichen Ebene bedeutet sie, dass jahrelange Arbeits- und Denkverläufe nicht im digitalen Sediment verschwinden, sondern wieder anschlussfähig werden. Auf der projektbezogenen Ebene bedeutet sie, dass Publikationslinien, Konfliktchronologien, Theorieentwicklungen, Dokumentationsbestände und reflexive Schleifen nicht nur konserviert, sondern analysierbar bleiben. Auf der institutionellen Ebene bedeutet sie, dass Wissensarbeit nicht zwangsläufig an proprietäre Cloudfilter, externe Sicherheitsarithmetiken und kulturell fremde Bias-Regime gebunden bleibt. Auf der didaktischen Ebene bedeutet sie, dass man nicht nur Ergebnisse, sondern Methoden und Schutzmechanismen selbst vermitteln kann.

Gerade hierin liegt ein tieferer Sinn der in v1.1 neu eingeführten Lehr- und Dokumentationsdimension. Architekturkarten, Methodenkarten, Forschungsfeldkarten und epistemische Schutzkarten sind keine schmückenden Begleitmaterialien. Sie machen den Unterschied zwischen einem bloß funktionierenden System und einem System, dessen Rationalität nachvollziehbar, lehrbar und kollektiv weiterentwickelbar ist. Ein Archiv, das nur seinem Erbauer verständlich bleibt, ist operativ nützlich, aber kulturell fragil. Eine Wissensinfrastruktur, die ihre eigene Struktur mit dokumentiert, erzeugt dagegen Anschlussfähigkeit.

1.9 Möglichkeiten der Weiterentwicklung

Aus der vorliegenden Fassung ergeben sich mehrere klare Entwicklungspfade. Erstens die Formalisierung der Transformationsprovenienz, damit maschinelle Verdichtungen dauerhaft überprüfbar bleiben. Zweitens die Versionierung von Taxonomie und Ontologie, um begriffliche Drift nicht nur zu beklagen, sondern historisch kontrollierbar zu machen. Drittens der prototypische Aufbau einer Vektorraumschicht, die latente thematische Nähe jenseits expliziter Begriffsordnungen erschließt. Viertens die Entwicklung eines dualen Archiv- und Inferenzgraphen, damit gesicherte Provenienzrelationen nicht mit heuristischen Relationen verschwimmen. Fünftens die Ausbildung von Community- und Cluster-Reports als eigener Klasse globaler Verdichtung. Und schließlich die systematische Ausarbeitung didaktischer Begleitformen, die das Projekt über seine operative Funktion hinaus als Forschungs- und Lehrgegenstand verfügbar machen.

Diese Weiterentwicklungen sind nicht lediglich technische Ausbaustufen. Sie verschieben das Projekt von der Grundartikulation zu einer methodisch kontrollierten Instrumentierung. Das bedeutet: Aus der Theorie der Struktur wird Schritt für Schritt eine belastbare Praxis der Archiv- und Erkenntnissoveränität.

1.10 Schluss: Eine Infrastruktur gegen das Vergessen der Formen

Die eigentliche Pointe von Ki-Lab-26 liegt nicht darin, dass hier ein weiteres Ordnungssystem für Dateien vorgeschlagen wird. Die eigentliche Pointe liegt darin, dass das Projekt eine Form des Erinnerns im digitalen Raum verteidigt, die mehr ist als Speicherung. Es verteidigt die Erinnerung an Pfade, an Entstehungsweisen, an Zwischenstufen, an Unsicherheiten, an begriffliche Verschiebungen und an die Differenz zwischen dokumentiertem Befund und späterer Verdichtung.

In einer Zeit, in der immer mehr Systeme auf Glättung, Reibungsreduktion und schnelle Ausgabe hin optimiert werden, ist diese Architektur eine Gegenfigur. Sie behauptet, dass geistige Souveränität nicht im Besitz möglichst großer Modelle besteht, sondern in der Fähigkeit, die eigene Wissensumgebung so zu strukturieren, dass sie weder stumm noch verflacht wird. Sie zeigt, dass methodische Strenge, Provenienzbewusstsein und geschichtete Verdichtung nicht der Preis für Langsamkeit sind, sondern die Bedingung dafür, dass digitale Archive überhaupt wieder zu Orten der Erkenntnis werden können.

Gerade darum ist ein solches Projekt nicht nur technisch oder organisatorisch bedeutsam. Es ist kulturell, wissenschaftlich und politisch relevant. Nicht weil es im engeren Sinn eine Waffe wäre, sondern weil es eine Form der Souveränität vorbereitet, die in den kommenden Jahren an Bedeutung gewinnen dürfte: die Fähigkeit, unter Bedingungen begrenzter Mittel und begrenzter Kontextfenster dennoch strukturiert, quellennah, reflektiert und unabhängig denken zu können.

Wenn man den Sinn von Ki-Lab-26 in einem Satz bündeln wollte, könnte man sagen: Es ist der Versuch, dem digitalen Archiv seine Form, seiner Form die Geschichte und seiner Geschichte die zukünftige Nutzbarkeit zurückzugeben. Die Tragweite eines solchen Versuchs liegt nicht nur im Archiv selbst, sondern in der Möglichkeit, aus verstreuten, bedrohten oder verflachten Beständen wieder einen Raum zu machen, in dem Denken mehr sein kann als bloße Reaktion.

2. Methodenpapier: Ki-Lab-26 Archive Structure v1.1

Ki-Lab-26-Archive-Struktur-v1.1

Entwurf einer mehrschichtigen Archiv-, Verdichtungs- und Retrieval-Architektur für heterogene digitale Arbeitsbestände unter Bedingungen begrenzter Kontextfenster

****Dokumenttyp:**** Arbeitsgrundlage / wissenschaftlich orientierter Strukturartikel

****Version:**** v1.1

****Projektkontext:**** Ki-Lab-26

****Status:**** Konsolidierte Revisionsfassung
****Autor:**** Andre Schmidt - www.as-institut.de
****Stand:**** 2026 - 06 - 06
****Lizenzhinweis****

Dieses Werk ist lizenziert unter der
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International
(CC BY-NC-SA 4.0).

Sie dürfen dieses Werk unter folgenden Bedingungen vervielfältigen,
verbreiten, öffentlich zugänglich machen sowie bearbeiten:

- Namensnennung der Urheberperson in angemessener Form,
- Verlinkung bzw. Angabe der Lizenz,
- Kennzeichnung vorgenommener Änderungen,
- keine Nutzung zu primär kommerziellen Zwecken,
- Weitergabe bearbeiteter Fassungen unter derselben Lizenz.

Ein Eindruck offizieller Billigung durch die Urheberperson oder
durch das AS-Institut / European Processing darf durch die Nutzung
nicht erzeugt werden.

Rechte an gesondert gekennzeichnetem Fremdmaterial bleiben unberührt.

Empfohlene Quellenangabe:

Schmidt, Andre: Ki-Lab-26: Archive Strukture v1.1 (konsolidiert).

AS-Institut / European Processing, 2026.

Lizenz: CC BY-NC-SA 4.0 International.

Hinweis:

Maßgeblich ist der offizielle Lizenztext von Creative Commons.

Die freie nicht-kommerzielle Weitergabe dieser Publikation ist
ausdrücklich erwünscht.

Abstract

Der vorliegende Beitrag formuliert mit ****Ki-Lab-26-Archive-Strukture-v1.1**** eine konsolidierte und erweiterte Fassung einer mehrschichtigen Archiv-, Verdichtungs- und Retrieval-Architektur für heterogene digitale Arbeitsbestände. Ausgangspunkt ist die Beobachtung, dass große digitale Wissensarchive unter Bedingungen begrenzter Kontextfenster nicht mehr durch bloße Speicherung oder flache Suchbarkeit operativ nutzbar bleiben. Der Artikel entwickelt daher ein Modell, das nicht-destruktive Quellensicherung, Provenienzschutz, taxonomische Kontrolle, mehrstufige Verdichtung, segmentierende Zwischenebenen, thematische Clusterbildung, Oberkarten, eine ergänzende Vektorraumschicht sowie eine perspektivische Wissensgraph-Schicht miteinander

verbindet. Zugleich werden die disziplinären Bezüge des Projekts – insbesondere zu Archivwissenschaft, Informationswissenschaft, Wissensrepräsentation, Retrieval-Architekturen, Dokumentationsforschung und epistemischer KI-Forschung – präzisiert. Der Text entfaltet sowohl die methodischen Leitprinzipien als auch die bereits konkretisierten Teilprojekte, insbesondere die Screenshot-Archivierung und das Telegram-Artikelarchiv, und ordnet diese in ein übergreifendes Staffeldgedächtnis-Modell ein. Ziel ist der Aufbau einer langfristig erweiterbaren Wissensinfrastruktur, die pfadförmige, relationale, vektorielle und quellennah rückgebundene KI-Nutzung ermöglicht, ohne die epistemische Trennung von Quelle, Normalisierung und Deutung aufzugeben.

Schlagwörter

Digitale Wissensarchive; Archivarchitektur; Staffeldgedächtnis; Retrieval; Verdichtungslogik; Provenienzschutz; Taxonomie; Segmentierung; Clusterbildung; Wissensgraph; GraphRAG; Vektorraumschicht; Ontologie-Versionierung; Telegram-Backups; Screenshot-Archivierung; KI-gestützte Wissensorganisation; Kontextfenster

1. Vorbemerkung

Der folgende Text formuliert das bislang entstandene Projektökosystem als wissenschaftlich orientierten Grundsatzartikel aus. Er dient nicht der Selbstdarstellung eines bereits abgeschlossenen Systems, sondern der präzisen Explikation eines in Ausarbeitung befindlichen Strukturmodells. Sein Zweck ist doppelt: erstens die begriffliche und methodische Stabilisierung des bereits entworfenen Projektzusammenhangs; zweitens die Schaffung einer belastbaren Diskussions- und Arbeitsgrundlage, die in späteren Entwicklungsstufen erweitert, korrigiert oder partiell umgebaut werden kann. Der Text ist deshalb so angelegt, dass er auch für Leserinnen und Leser nachvollziehbar bleibt, die weder Einblick in den zugrunde liegenden Arbeitsdialog noch in die thematische Vorgeschichte des Projekts besitzen.

Die vorliegende Fassung v1.1 ist ausdrücklich als ****konsolidierte Vollrevision**** konzipiert. Sie ersetzt die Grundfassung v1 nicht durch ein bloßes Ergänzungspapier, sondern integriert deren Grundbestand in eine erweiterte, begrifflich präzisierte und architektonisch fortgeschriebene Gesamtfassung. Anlass dieser Revision war insbesondere die Frage, wie das Projekt auf unterschiedlichen wissenschaftlichen Skalen zu benennen, disziplinär zu verorten und um zusätzliche Komponenten – etwa eine Transformationsprovenienz maschineller Verdichtungen, eine Vektorraumschicht oder eine beziehungsbezogene Unsicherheitslogik – zu erweitern sei. V1.1 nimmt diese Impulse auf, ohne den Grundcharakter der ursprünglichen Architektur preiszugeben.

2. Problemstellung und Ausgangslage

2.1 Das Grundproblem großer persönlicher Wissensarchive

Digitale Langzeitprojekte erzeugen typischerweise keine homogene Bibliothek, sondern ein Gefüge aus heterogenen, partiell überlappenden, unterschiedlich stark strukturierten und verschieden granularen Beständen. Dazu gehören etwa Gesprächsarchive, Messenger-Exporte, Screenshot-Sammlungen, Essay- und Artikelbestände, Entwürfe, Metadateien, Seitenversionen, Notizfragmente, systematische Projekttexte, technische Dokumentationen und reflexive Zwischenstufen. Solche Archive wachsen meist nicht entlang einer einzigen Logik. Sie entstehen arbeitshistorisch, situativ, konfliktbedingt, publizistisch, dokumentarisch, juristisch, analytisch und bisweilen auch bloß aus dem Impuls, Material vorsorglich zu sichern.

Das resultierende Problem ist nicht bloß eines der Speichermenge. Entscheidender ist der Verlust an operativer Nutzbarkeit. Je größer ein Archiv wird, desto schwieriger wird es, relevante Materialien gezielt wiederzufinden, Zusammenhänge zwischen räumlich getrennten Materialien zu erkennen, unterschiedliche Dokumenttypen begrifflich zusammenzuführen und für spätere Auswertung oder Veröffentlichung belastbar vorzustrukturieren. Das gilt umso mehr, wenn die spätere Bearbeitung nicht ausschließlich durch menschliche Lektüre, sondern in erheblichem Umfang mit Hilfe sprachverarbeitender Modelle erfolgen soll, deren Kontextfenster stets begrenzt ist.

2.2 Begrenzte Kontextfenster als epistemische und technische Engstelle

Sprachmodelle können große Bestände nicht einfach „als Ganzes erinnern“. Sie müssen Material selektiv erhalten, nachladen, verdichten, rekombinieren und auf bestimmte Anfragen hin fokussiert erschließen. Daraus ergibt sich eine fundamentale Anforderung: Ein großes Archiv muss so organisiert werden, dass nicht immer das Ganze vorgelegt werden muss, sondern jeweils nur jene Ausschnitte, Verdichtungen, Verweise und Rückbindungen, die für die konkrete Anfrage wahrscheinlich relevant sind. Das Archiv darf daher nicht nur gespeichert, sondern muss in eine Form überführt werden, die pfadförmiges, hierarchisches, relationales und gegebenenfalls vektoriell ergänztes Retrieval erlaubt.

2.3 Vom Archiv zur strukturierten Wissensinfrastruktur

Das Projekt ****Ki-Lab-26-Archive-Strukture-v1.1**** setzt an genau dieser Stelle an. Es geht nicht um eine bloße Dateisortierung, auch nicht um eine simple Sammlung von Zusammenfassungen, sondern um den Entwurf einer mehrschichtigen Infrastruktur, die mehrere Anforderungen zugleich erfüllen soll:

Erstens muss sie der Heterogenität der Ausgangsbestände gerecht werden.

Zweitens muss sie Provenienz, Rückverfolgbarkeit und Unsicherheitsmarkierung sichern.

Drittens muss sie in mehreren Verdichtungsstufen arbeiten, damit große Bestände für Modelle mit begrenztem Kontextfenster nutzbar werden.

Viertens muss sie neben einer hierarchischen Ordnungslogik auch relationale Querschließung zulassen, damit nicht nur vertikale Pfade, sondern auch begriffliche und sachliche Nachbarschaften nutzbar werden.

Fünftens soll sie eine ergänzende Vektorraumschicht aufnehmen können, um auch solche semantischen Nachbarschaften erschließbar zu machen, die weder taxonomisch explizit benannt noch graphisch bereits modelliert sind.

Sechstens muss sie die Entstehung maschineller Verdichtungen selbst als provenancepflichtigen

Vorgang behandeln.

2.4 Das Problem des Flattenings

Eine zentrale Gefährdung großer digitaler Archive liegt im **Flattening**, also in der Tendenz, heterogene Materialien in flache, scheinbar gleichartige Datenobjekte zu überführen und dabei die innere Geschichte der Bestände, ihre Pfadabhängigkeiten, ihre Transformationsschritte und ihre Herkunftsbezüge zu verlieren. Flattening kann technisch, rhetorisch oder epistemisch auftreten: technisch, wenn Daten in uniforme Chunks zerlegt werden; rhetorisch, wenn Verdichtungen ihre Herstellungsbedingungen verschweigen; epistemisch, wenn Quelle, Normalisierung und Deutung ineinanderfallen.

Das vorliegende Projekt ist in seinem Kern als Gegenarchitektur gegen diese Verflachung zu verstehen. Es will nicht bloß Daten verfügbar machen, sondern ihre operative Nutzbarkeit erhöhen, ohne ihre innere Struktur zu tilgen.

3. Benennung des Themenfeldes auf Mikro-, Meso-, Makro- und Metaebene

Die Einordnung des Projekts hängt wesentlich davon ab, auf welcher analytischen Skala man es betrachtet. Eine singuläre Begriffszuweisung wäre deshalb irreführend.

3.1 Mikroebene: Quellennahe Einheitenerschließung, Segmentierung und Metadatenanreicherung

Auf der Mikroebene ist das Projekt am treffendsten als **quellennahe Einheitenerschließung, Segmentierung und Metadatenanreicherung heterogener digitaler Materialien** zu bezeichnen. Hierzu gehören:

- Parsing und Erfassung einzelner Dokumente, Dateien, Exporte, Nachrichten oder Bildartefakte,
- Segmentierung in operativ sinnvolle Einheiten,
-

- Provenienz- und Kontextanreicherung,
- Unsicherheitsmarkierung,
- sidecarbasierte oder manifeste Beschreibung.

Begriffe wie *Document Parsing* oder *Metadata Extraction* erfassen davon nur Teilaspekte. Ebenso ist *Semantic Chunking* nur begrenzt passend, weil das Projekt nicht bloß semantische Textstücke erzeugt, sondern je nach Materialtyp sehr unterschiedliche operative Einheiten: Screenshots, Nachrichtensequenzen, Export-Sets, Fallsegmente oder Dossierbausteine.

3.2 Mesoebene: Wissensorganisation, Korpusstrukturierung und taxonomisch kontrollierte Dossierbildung

Auf der Mesoebene handelt es sich um ****wissensorganisatorische Korpusstrukturierung**** beziehungsweise um ****taxonomisch kontrollierte Dossier- und Clusterbildung****. Hier werden Einzelmanifest-Einheiten, Segmente oder Fallteile in geordnete thematische Zusammenhänge überführt.

Der Begriff ***Personal Knowledge Management*** ist hierfür nur eingeschränkt tauglich. Er ist zu stark auf individuelle Selbstorganisation und Produktivitätsoptimierung zugeschnitten. Das vorliegende Projekt geht darüber hinaus, weil es archivische, dokumentarische, analytische, gegebenenfalls juristische und perspektivisch wissenschaftliche Anforderungen miteinander verbindet.

3.3 Makroebene: Mehrschichtige Archiv- und Retrieval-Architektur

Auf der Makroebene lässt sich das Projekt als ****mehrschichtige Archiv- und Retrieval-Architektur für heterogene Wissensbestände unter Bedingungen begrenzter Kontextfenster**** beschreiben. Hier sind Bezüge zu ***Retrieval-Augmented Generation***, modularen Retrieval-Verfahren oder hybriden Sucharchitekturen sachlich legitim, aber einzuhegen.

Das Projekt ist nicht selbst schon ein fertiges RAG-System, sondern eine vorbereitende Wissensinfrastruktur, aus der unterschiedliche Retrieval-Modi hervorgehen können: klassische Volltextsuche, Vektorraumsuche, graphbasierte Traversierung, pfadförmiges Nachladen verdichteter Oberstrukturen und hybride Suchformen. Der RAG-Bezug ist daher eine mögliche Betriebsform, nicht die erschöpfende Definition des Modells.

3.4 Metaebene: Provenienzbewusste Epistemologie digitaler Wissensinfrastrukturen

Auf der Metaebene bewegt sich das Projekt im Feld einer ****provenienzbewussten Epistemologie digitaler Wissensinfrastrukturen****. Hier treten Fragen auf wie:

- Unter welchen Bedingungen bleibt eine Verdichtung erkenntnistheoretisch verantwortbar?
- Wie lässt sich Herkunft sichern, ohne operative Nutzbarkeit zu opfern?
- Wie wird Unsicherheit nicht als Defizit, sondern als positive Information modelliert?
- Wie können KI-gestützte Rekonstruktions- und Verdichtungsleistungen dokumentiert werden, ohne stillschweigend Objektivitätsstatus zu erhalten?

Diese Metaebene schützt das Projekt davor, als rein technische Organisationslösung missverstanden zu werden.

4. Projektdefinition

4.1 Arbeitsbezeichnung und Status

****Ki-Lab-26-Archive-Strukture-v1.1**** ist als vorläufige, versionierte Arbeitsarchitektur zu verstehen. Die Bezeichnung markiert bewusst keinen fertigen Softwarestand, sondern eine

systematische Konzeption eines übergreifenden Archiv- und Retrieval-Ökosystems. Das Kürzel „v1.1“ bezeichnet daher keine kosmetische Fortschreibung, sondern eine inhaltlich konsolidierte und erweiterte Revisionsstufe.

4.2 Gegenstandsbereich

Das Projekt adressiert insbesondere folgende Bestandsformen:

- große Screenshot-Sammlungen, darunter Messenger-Screenshots, Screenshots von KI-Dialogen und sonstige dokumentarisch oder analytisch relevante Bildsicherungen,
- Telegram-Kanal-Backups mit publizierten oder publikationsnahen Texten, insbesondere HTML-Exporte mit ein- oder mehrteiligen `messages\*.html`-Strukturen,
- perspektivisch auch große Archive aus ChatGPT-Exporten oder vergleichbaren Konversationsarchiven,
- allgemein heterogene digitale Wissens-, Arbeits- und Dokumentationsbestände, die nicht als fertige Bibliothek, sondern als vielschichtige Werkstatt- und Archivlandschaft vorliegen.

4.3 Zielperspektive

Ziel ist die Überführung solcher Bestände in ein System, das mehrere Betriebsweisen zugleich ermöglicht:

- ****archivarische Erschließung****, also konservative, provenance-sensible Dokumentation dessen, was vorliegt;
- ****thematische Verdichtung****, also die Bildung von Segmenten, Dossiers, Clustern und Oberkarten;
- ****anfrageorientiertes Retrieval****, also eine pfadförmige, relationale und vektorielle Suchfähigkeit, die eine spätere KI-gestützte Auswertung entlang logisch plausibler Pfade ermöglicht;
- ****dokumentarische und didaktische Verdichtung****, also die Bereitstellung von Karten, Überblicken und Methodenmodulen für Vermittlung, Dokumentation und mögliche Lehre.

5. Disziplinäre Einordnung des Forschungsbereichs

Das Projekt liegt an einer Schnittstelle mehrerer Forschungsfelder, ohne vollständig in einem einzigen aufzugehen.

5.1 Archivwissenschaft und digitale Langzeitarchivierung

Sobald Nicht-Destruktivität, Provenienzschutz, Rückführbarkeit und quellennahe Ordnungsbildung leitend werden, ist der Bezug zur Archivwissenschaft grundlegend. Die Einsicht, dass Herkunftsbeziehungen und Kontextwahrung nicht nachträgliche Metadatenverzerrungen, sondern konstitutive Ordnungskriterien sind, verankert das Modell in archivwissenschaftlichen Denkformen.

5.2 Informationswissenschaft und Knowledge Organization

Sobald kontrollierte Vokabulare, taxonomische Drift, Themencluster, Relevanzstufen und strukturierte Dossierbildung ins Zentrum rücken, bewegt sich das Projekt im Bereich der Informationswissenschaft, insbesondere der Wissensorganisation, der Klassifikationstheorie und der strukturierenden Korpuserschließung.

5.3 Wissensrepräsentation und Ontologie-Engineering

Dort, wo das Modell Begriffe, Knotentypen, Beziehungstypen, Versionierungsfragen und maschinenlesbare Schemata systematisiert, berührt es das Feld der Wissensrepräsentation. Zu unterscheiden ist dabei strikt zwischen einer lockeren Begriffsordnung, einer kontrollierten Taxonomie, einer formalen Ontologie und einem tatsächlich operationalisierten Wissensgraphen.

5.4 Retrieval-Architekturen, graphbasierte und hybride Suchverfahren

Sobald hierarchische Verdichtung, thematische Oberkarten, Vektorraumschichten und Wissensgraphen anfragebezogen genutzt werden, ergibt sich der Bezug zu modernen Retrieval-Architekturen, einschließlich graphbasierter und hybrider RAG-Verfahren. Dieser Bezug ist heuristisch und technologisch wichtig, darf aber das Projekt nicht begrifflich vollständig kolonisieren.

5.5 Dokumentations- und Lehrforschung

Weil das Projekt zugleich dokumentiert, strukturiert, verdichtet, reflektiert und vermittelbar gemacht werden soll, gehört es auch in einen Bereich, der als **methodische Dokumentations- und Lehrforschung zu digitalen Wissensinfrastrukturen** beschrieben werden kann. Dies ist deshalb wichtig, weil v1.1 ausdrücklich nicht nur operative Architektur, sondern auch didaktische Verdichtung mitführen soll.

6. Leitprinzipien des Gesamtmodells

6.1 Nicht-Destruktivität

Das System arbeitet grundsätzlich nicht-destruktiv. Originalbestände werden nicht überschrieben, stillschweigend bereinigt, umbenannt, gelöscht oder in ihrem Quellkontext unkenntlich gemacht. Alle Arbeitsprodukte entstehen in abgeleiteten Zielstrukturen. Damit wird verhindert, dass spätere Rekonstruktion, juristische Bezugnahme oder wissenschaftliche Rückprüfung an einer veränderten Quellenbasis scheitern.

6.2 Provenienzschutz

Jede bearbeitete Einheit muss auf ihre Herkunft zurückführbar bleiben. Dies betrifft Pfade, Exportordner, Dateiteile, Nachrichtenbereiche, Quellformate, Zeitstempel, gegebenenfalls Hashes

und verwandte Fundstellen. In v1.1 wird dieser Provenienzbegriff erweitert: Auch **Transformationen** – etwa maschinell erzeugte Verdichtungen, Clusterberichte oder Community-Summaries – sind provenancepflichtig. Nicht nur das Ausgangsmaterial, sondern auch der Transformationsakt muss dokumentierbar sein.

6.3 Ebenentrennung

Ein Grundsatz des Modells lautet, dass zwischen mindestens drei Ebenen streng unterschieden werden muss:

- dem **direkt extrahierten oder direkt sichtbaren Material**,
- der **redaktionellen Normalisierung oder formalen Verdichtung**,
- der **interpretativen oder hypothesenhaften Einordnung**.

V1.1 ergänzt diese Trennung um eine vierte relevante Dimension: die **Transformationsprovenienz**, also die Frage, mit welchem Modell, Prompt, Parameterraum und Bearbeitungskontext eine maschinelle oder halbmaschinelle Verdichtung zustande kam.

6.4 Kontrollierte Taxonomie

Die Struktur darf nicht über unkontrolliert wuchernde Kategorien verfügen. Deshalb wird mit kontrollierten Wortschätzen gearbeitet: für Quellenarten, Dokumenttypen, Hauptthemen, Unterthemen, Relevanzstufen, Confidence-Werte, Manual-Review-Markierungen und verwandte Klassifikationsfelder. Neue Kategorien dürfen nicht beiläufig entstehen, sondern nur kontrolliert und dokumentiert eingeführt werden.

6.5 Konservative Rekonstruktion

Wo Unsicherheit besteht, hat konservative Nachvollziehbarkeit Vorrang vor eleganter, aber spekulativer Rekonstruktion. Das System soll nicht Lücken „schönschreiben“, sondern Grenzen sichtbar machen.

6.6 Mehrstufige Verdichtung

Ein Großarchiv wird nicht dadurch für eine KI nutzbar, dass man es schlicht in kleinere Dateien zerlegt. Es bedarf einer abgestuften Verdichtungslogik: Rohmaterial, Einzelmanifest, Segment, Cluster, Oberkarte, Root-Map. V1.1 ergänzt diese Staffelung um die Möglichkeit von Community- und Cluster-Reports sowie um vektorielle und graphische Zusatzschichten.

6.7 Querschließung statt bloßer Baumlogik

Die Struktur ist nicht als starrer Baum zu denken. Viele Themen sind multipel verankert. Ein gutes System benötigt deshalb neben Hierarchie auch Querregister, Schlagwortachsen, Verweiskfelder, eine perspektivische Wissensgraph-Schicht sowie eine ergänzende Vektorraumschicht.

###

6.8 Unsicherheit als positive Information

Unsicherheit ist keine lästige Restgröße, sondern ein informationshaltiger Bestandteil der Architektur. Sie ist nicht nur an Dokumente oder Rekonstruktionen zu binden, sondern in v1.1 auch an Relationen, Hypothesen und Ableitungen.

7. Architektur des Projektökosystems

7.1 Grundform: Staffelgedächtnis

Die leitende Grundidee des Systems lässt sich als ****Staffelgedächtnis**** beschreiben. Gemeint ist damit eine gestufte Gedächtnisarchitektur, in der nicht jede Anfrage auf das Roharchiv zugreifen muss, sondern zunächst auf verdichtete Oberstrukturen trifft, von dort aus logisch weiterverzweigt und erst im Bedarfsfall tiefere Ebenen aktiviert.

Diese Staffelung erfüllt mehrere Funktionen. Sie reduziert die Kontextlast für begrenzte Modelle, diszipliniert die Wissensorganisation und macht Such- und Denkpfade explizit. Nicht jedes Thema steht als flacher Datensatz neben jedem anderen, sondern ist in einem mehrstufigen Verhältnis aus Überblick, Teilstruktur und Detailfundierung verortet.

7.2 Die Ebenen des Modells

7.2.1 Roharchiv

Die unterste Ebene enthält das unbearbeitete Ausgangsmaterial. Dazu gehören Exportdateien, Screenshot-Sammlungen, Chat-Archive, HTML-Dateien, JSON-Dateien und andere Quellbestände. Diese Ebene bleibt unangetastet.

7.2.2 Inventar- und Manifestebene

Auf dieser Ebene wird pro Einheit oder Quellcontainer ein standardisiertes Inventar erzeugt. Beispiele sind:

- technische Gesamtinventare im CSV- oder JSON-Format,
- pro Datei oder pro Exportordner erzeugte Metadaten,
- Listen technischer Eigenschaften, Pfade, Zeitstempel, Größen, Quellteile,
- Kurzbeschreibungen und erste Klassifikationen.

Diese Ebene dient der maschinenlesbaren und menschlich überprüfbaren Grunderschließung.

7.2.3 Segmentebene

Hier werden thematisch oder funktional zusammenhängende Einheiten gebildet. Im Chat- oder Messenger-Kontext kann ein Segment eine Folge inhaltlich kohärenter Nachrichten sein; im

Dokumentkontext ein klar abgrenzbarer Teilzusammenhang; im Screenshot-Kontext eine Serie verwandter Screenshots. Die Segmentebene ist für spätere thematische Bündelung entscheidend, weil weder der gesamte Ursprungschat noch die einzelne Bilddatei immer die geeignete operative Einheit ist.

7.2.4 Cluster- und Dossier-Ebene

Auf dieser Ebene werden mehrere Segmente oder Einzelmanifest-Einheiten zu thematischen Clustern zusammengeführt. Ein Cluster ist keine lose Tag-Sammlung, sondern ein kontrolliert gebildetes Dossier, das ein Thema, einen Fall, eine Konfliktlage, eine Publikationslinie oder eine Projektdynamik zusammenfasst.

7.2.5 Community- und Cluster-Reports

V1.1 führt eine zusätzliche Artefaktklasse ein: **Community- und Cluster-Reports**. Diese Berichte stehen zwischen Dossier- und Oberkartenebene. Sie dienen dazu, größere thematische Felder global zu verdichten, ohne den Rückbezug auf Segmente, Dossiers und Quellen zu verlieren. Sie eignen sich besonders für globale Themenabfragen, Dokumentation, didaktische Aufbereitung und query-fokussierte Orientierung.

7.2.6 Oberkarten- und Bereichsebene

Hier entstehen verdichtete Übersichtsdateien über größere Bereiche: etwa Projektfelder, Themenfamilien, Konfliktachsen, institutionelle Zusammenhänge, wiederkehrende Motive oder Publikationsstränge. Diese Ebene dient der schnellen Orientierung und der Pfadwahl.

7.2.7 Root-Map

An der Spitze steht eine Root-Map oder Wurzelkarte. Sie enthält nicht den Gesamthalt des Archivs, sondern die primären Einstiegspunkte. Sie beantwortet nicht jede Sachfrage, sondern die vorgelagerte Frage, wo eine sachgerecht gestellte Folgefrage sinnvollerweise weitergeführt werden sollte. Eine Root-Map ist funktional eine oberste Navigationsstruktur; sie ist nicht schon deshalb eine formale Top-Level-Ontologie.

7.3 Vektorraumschicht

V1.1 ergänzt die hierarchische Grundarchitektur um eine **Vektorraumschicht**. In ihr können Segmente, Dossiers oder andere operative Einheiten als Embeddings repräsentiert werden. Diese Schicht ersetzt weder Taxonomie noch Hierarchie noch Wissensgraph, sondern ergänzt sie. Ihre Funktion liegt darin, auch solche semantischen Nachbarschaften auffindbar zu machen, die nicht bereits explizit taxonomisch benannt oder graphisch modelliert sind.

7.4 Wissensgraph-Schicht

Perspektivisch gehört zum Ausbau des Systems eine relationale Schicht in Form eines Wissensgraphen. Diese ergänzt die Hierarchie um explizite Entitäten und Beziehungen. Ihre

Einführung ist an zwei Bedingungen gebunden:

- die kontrollierte Modellierung von Knotentypen und Beziehungstypen,
- die differenzierte Geltungsmarkierung von Kanten.

7.5 Archivgraph und Inferenzgraph

V1.1 präzisiert die Graphperspektive weiter, indem zwischen zwei Graphformen unterschieden wird:

- dem ****Archivgraphen****, der gesicherte Provenienz-, Struktur- und Zugehörigkeitsrelationen abbildet,
- dem ****Inferenzgraphen****, der interpretative, heuristische oder hypothesenhafte Beziehungen abbildet.

Diese Trennung ist wichtig, um Deutung nicht unbemerkt in den Status archivischer Faktizität zu überführen.

7.6 Hierarchie und Query-Pfade

Das System ist so zu verstehen, dass eine spätere anfrageorientierte KI-Anwendung zunächst auf einer oberen Verdichtungsebene ansetzt. Von dort wird nicht blind Volltextsuche auf alles angewandt, sondern entlang plausibler Pfade entschieden, welche Unterbereiche, Subcluster oder Segmentdateien nachgeladen werden sollen. In einem ausgebauten System kann diese Pfadlogik durch graphische Traversierung und vektorielle Ähnlichkeitssuche ergänzt werden.

8. Die bislang ausgebildeten Teilprojekte

8.1 Screenshot-Archivierung und juristisch vorsichtige Vorstrukturierung

Ein bereits ausgearbeitetes Teilprojekt betrifft die Ordnung großer Screenshot-Bestände. Sein Kern ist eine nicht-destruktive, provenienzbewusste und juristisch vorsichtige Erschließung von etwaigen Messenger-Screenshots, Chat-Sicherungen und sonstigen Bilddokumenten.

8.1.1 Zweck

- Ordnung großer Screenshot-Mengen nach Jahr, Quelle und Thema,
- Erzeugung technischer Inventare,
- Anlegen von Sidecar-Dateien mit Metadaten, Kurzbeschreibung und Unsicherheitsmarkierung,
- Vorbereitung einer späteren juristischen oder analytischen Auswertung.

8.1.2 Strukturprinzipien

- Originaldateien werden nicht verändert.

- Die Zielstruktur folgt einem Muster wie `organized\_screenshots/YYYY/source/topic/`.
- Zu jeder Datei können beschreibende Metadateien erzeugt werden.
- Für unsichere oder gemischt deutbare Fälle existiert ein `manual\_review`-Bereich.

8.1.3 Methodische Besonderheit

Dieses Teilprojekt ist besonders stark auf Provenienz, epistemische Disziplin und die Trennung von sichtbarem Inhalt, inferierter Einordnung und Unsicherheit angelegt. Es ist damit eher archivisch-forensisch als bloß organisatorisch motiviert.

8.2 Telegram-Backups als Artikelarchiv

Ein zweites, deutlich anders strukturiertes Teilprojekt betrifft Telegram-Kanal-Backups und die Rekonstruktion artikelförmiger Beiträge.

8.2.1 Ausgangslage

Telegram-Exporte liegen nicht notwendig als eine einzelne HTML-Datei vor, sondern häufig als Exportordner mit mehreren zusammengehörigen Dateien wie `messages.html`, `messages2.html`, `messages3.html` usw. Diese Dateien dürfen nicht automatisch als getrennte Exporte behandelt werden, sondern bilden innerhalb eines Ordners meist ein gemeinsames Export-Set.

8.2.2 Zweck

- Erkennung artikelförmiger oder essayistischer Beiträge in Telegram-Backups,
- konservative Rekonstruktion einzelner Texte oder Sequenzen,
- Erzeugung einer eigenständigen Dokumentdatei pro Artikel,
- Begleitung jeder Dokumentdatei durch eine Metadaten-Markdown-Datei,
- Ordnung nach Jahr und Thema,
- Dublettenkontrolle über mehrere Backups hinweg.

8.2.3 Strukturprinzipien

- Primäre Quelleinheit ist nicht die einzelne Datei, sondern das Export-Set.
- Partgrenzen von HTML-Dateien sind nicht automatisch Artikelgrenzen.
- Zwischen Einzelnachricht und zusammengehöriger Nachrichtensequenz wird methodisch unterschieden.
- Es existieren Regeln für editorische Titelvergabe, Variantenbehandlung und `manual\_review`.

8.2.4 Bedeutung für das Gesamtprojekt

Dieses Teilprojekt bildet ein Muster dafür, wie aus halbstrukturierten Exporten publizistische Arbeitsprodukte rekonstruiert werden können, ohne die Quellenlage zu verfälschen.

8.3 Perspektive eines ChatGPT-Archivprojekts

Ein drittes, bislang noch konzeptionell vorbereitetes Teilprojekt betrifft große ChatGPT-Archive. Anders als beim Telegram-Projekt stehen dort nicht primär fertige Artikel im Vordergrund, sondern Gesprächsverläufe, Arbeitssegmente, Reflexionsschleifen, Entwürfe und thematisch mehrfach verzweigte Dialoge.

8.3.1 Erwartete Struktur

- Rohinventarisierung des Exportbestands,
- Segmentierung einzelner Chats in inhaltlich kohärente Einheiten,
- thematische Klassifikation und Clusterbildung,
- später Erzeugung von Clusterzusammenfassungen, Registern und gegebenenfalls publizistischen Destillaten.

8.3.2 Besonderheit

Hier ist die Segmentebene noch wichtiger als beim Telegram-Projekt, weil ein einzelner Chat häufig mehrere Themenfelder, Arbeitsmodi und Funktionsschichten enthält.

9. Dokumenttypen und Standardartefakte

Das Projektökosystem arbeitet mit einer Vielzahl standardisierter Artefakte. Diese sind nicht bloß technische Ausgaben, sondern konstitutive Teile der Architektur.

9.1 Maschinenlesbare Inventare

Beispiele:

- `inventory\_master.csv`
- `inventory\_master.json`
- `telegram\_exports\_inventory.csv`
- `telegram\_exports\_inventory.json`
- `articles\_master\_index.csv`
- `articles\_master\_index.json`

Diese Dateien dienen der systematischen Erfassung, der automatisierten Weiterverarbeitung und der Reproduzierbarkeit.

9.2 Menschlich lesbare Review-Dateien

Beispiele:

- `classification\_review.md`
-

`article\_extraction\_review.md`

Sie dokumentieren Bestandszahlen, Verteilungen, Unsicherheiten, Dublettenlagen, Grenzfälle und angewandte Regeln.

9.3 Ordnerzusammenfassungen

Für Themen- oder Bereichsordner werden `\_folder\_summary.md`-Dateien vorgesehen. Diese enthalten typischerweise:

- Ordnungszweck,
- Anzahl der enthaltenen Einheiten,
- thematische Kurzcharakteristik,
- Unsicherheiten,
- interne Cluster,
- Hinweise auf manuellen Prüffallbedarf.

9.4 Sidecar-Dateien

Pro Datei oder pro exportiertem Dokument können Metadaten-Dateien mit gleichem Basisnamen entstehen. Sie bündeln Identifikation, Provenienz, Klassifikation, Kurzbeschreibung, Relevanz, Unsicherheiten und Verweise. V1.1 erweitert diese Dateart um Felder zur Transformationsprovenienz, wo maschinelle Verdichtungen beteiligt sind.

9.5 Community- und Cluster-Reports

Als neue Artefaktklasse treten Community- und Cluster-Reports hinzu. Sie verdichten größere thematische Zusammenhänge, ohne zu bloßen Folder-Summaries zu werden. Sie eignen sich für globale Themenabfragen, Dokumentation, Methodendarstellung und didaktische Vermittlung.

9.6 Dokumentausgaben

Insbesondere im Telegram-Projekt ist die Ausgabe als eigenständige Dokumentdatei, vorzugsweise im ODT-Format, vorgesehen. Dabei gilt: Ist die direkte ODT-Erzeugung nicht zuverlässig, muss ein dokumentierter Zwischenweg genutzt werden. Eine intransparente Formatabweichung ist methodisch unerwünscht.

10. Taxonomische Ordnung und Versionierung

10.1 Funktion der Taxonomie

Die Taxonomie ist das Instrument, mit dem das System begriffliche Verwilderung verhindert. Sie standardisiert Kernfelder und schafft damit Konsistenz zwischen Teilprojekten.

10.2 Typische Klassifikationsfelder

- Jahr
- Quelle / Herkunft
- Dokumenttyp
- Hauptthema
- Unterthema
- Schlagwörter
- Relevanz
- Confidence
- Manual-Review-Status
- Titelstatus
- Extraktionstyp
- Quellstruktur / Part-Status

10.3 Methodische Bedeutung

Die Taxonomie ist nicht bloß ein Ordernamenkatalog. Sie hat epistemische Funktion. Sie zwingt dazu, Entscheidungen explizit zu treffen: Ist etwas ein Artikel oder nur ein Fragment? Ist ein Thema tatsächlich klar, oder müsste es in `manual\_review`? Ist ein Titel original oder redaktionell vergeben? Ist ein Zusammenhang direkt belegbar oder nur inferiert?

10.4 Taxonomische Drift

Je größer das System wird, desto größer wird die Gefahr begrifflicher Ausfransung. Ohne strikte Pflege der Taxonomie könnte das System in redundante oder konkurrierende Kategoriebildung zerfallen.

10.5 Taxonomische und ontologische Versionierung

V1.1 erweitert die ursprüngliche Driftkritik zu einer positiven Architekturkomponente: Taxonomie und Ontologie müssen versionierbar sein. Das bedeutet:

- Begriffsstände erhalten Versionsmarkierungen,
- Umbauten von Kategorien werden dokumentiert,
- Aufspaltungen, Zusammenführungen oder Umbenennungen bleiben historisch nachvollziehbar,
- Rebuilds höherer Ebenen können auf definierte Schemaänderungen bezogen werden.

Diese Forderung ist pragmatisch zu staffeln. Nicht jede frühe Projektphase erfordert sofort volle formale Ontologie-Niveaus. Wohl aber braucht das System von Anfang an maschinen- und menschenlesbare Versionsdisziplin.

11. Provenienz, Unsicherheit und epistemische Disziplin

11.1 Warum Provenienz zentral ist

In großen Langzeitarchiven droht der Verlust der Herkunftsbeziehung. Verdichtete Dateien wirken dann zwar operativ hilfreich, können aber nicht mehr verlässlich auf ihre Basis zurückgeführt werden. Das System setzt dem eine radikale Provenienzorientierung entgegen. Jede Verdichtung muss rückwärts lesbar bleiben.

11.2 Transformationsprovenienz

V1.1 verschiebt die Provenienzfrage über die Quellbasis hinaus auf die Ebene der Transformationen. Wenn ein Segment, Dossier, Cluster oder Community-Report maschinell erzeugt oder maschinell mitverdichtet wird, ist diese Entstehungsgeschichte selbst dokumentationspflichtig.

Zu erfassen sind dabei mindestens:

- verwendetes Modell,
- Modellversion oder identifizierbare Modellsignatur,
- relevante Laufparameter,
- Prompt oder Promptfamilie,
- Datum und Kontext der Erzeugung,
- gegebenenfalls menschliche Nachbearbeitung.

Dies verhindert, dass maschinell erzeugte Verdichtungen als neutraler Objektivitätsersatz erscheinen.

11.3 Unsicherheit als positive Information

Unsicherheit gilt im Modell nicht als Mangel, der nach Möglichkeit sprachlich geglättet werden sollte, sondern als eigenständige Information. Es ist ein Qualitätsmerkmal, wenn ein System sichtbar machen kann, dass eine Titelvergabe nur editorisch erfolgte, eine Datierung unsicher ist, eine Sequenzgrenze fraglich bleibt oder eine Dublettenzuordnung nicht abschließend entschieden wurde.

11.4 Dreiteilung als Standard

Die Felder `Direkt sichtbar` oder `Direkt extrahiert`, `Redaktionell normalisiert` und `Unsicher/prüfbedürftig` sind Ausdruck dieser Haltung. Sie erzwingen die methodische Trennung von Quelle, Bearbeitung und Deutung.

11.5 Relationsbezogene Unsicherheitslogik

V1.1 erweitert diese Logik auf Relationen. Ein Dokument kann sicher erschlossen sein, während die Aussage, es stehe in einer bestimmten Beziehung zu einem anderen Dokument, nur eine Hypothese sein kann. Für die spätere Wissensgraph-Schicht folgt daraus:

- Jede Kante benötigt einen Herkunftsnachweis.

- Jede Kante benötigt eine Geltungs- oder Confidence-Markierung.
- Zwischen direkter Quellrelation, abgeleiteter Relation und interpretativer Hypothesenrelation ist zu unterscheiden.

Diese Erweiterung verhindert, dass der Graph zu einem epistemisch flachen Behauptungsraum wird.

12. Von der Hierarchie zum hybriden Modell

12.1 Die Grenze reiner Baumstrukturen

Eine rein hierarchische Archivstruktur genügt für große, quervernetzte Bestände nur begrenzt. Viele Einheiten gehören zugleich mehreren thematischen Räumen an. Konfliktlagen sind sowohl rechtlich als auch sozialdynamisch, publizistisch und biografisch relevant. Texte sind oft zugleich Projektmaterial, Theoriebaustein, dokumentarischer Beleg und Reflexionsprodukt.

12.2 Die Notwendigkeit einer relationalen Zusatzschicht

Daher gehört zum langfristigen Ausbau des Systems eine zusätzliche relationale Schicht: ein kontrolliert aufgebauter Wissensgraph. Dieser ergänzt die Hierarchie um explizite Entitäten und Beziehungen.

12.2.1 Mögliche Knotentypen

- Personen
- Gruppen
- Projekte
- Publikationen
- Konfliktfälle
- Dokumente
- Konzepte
- Orte
- Zeiträume
- Institutionen
- Plattformen
- Verfahren

12.2.2 Mögliche Beziehungstypen

- gehört zu
- ist Vorstufe von
- ist Variante von
- steht im Konflikt mit
- kritisiert

- bezieht sich auf
- setzt voraus
- ist Gegenbegriff zu
- verwendet synonym
- dokumentiert
- reagiert auf
- gehört zur gleichen Chronologie

12.3 Die ergänzende Vektorraumschicht

Neben Hierarchie und Graph tritt in v1.1 ausdrücklich eine dritte Querlogik hinzu: die Vektorraumschicht. Sie ist dort produktiv, wo semantische Nähe besteht, ohne dass sie bereits taxonomisch oder graphisch expliziert wäre. Damit erweitert sich die Architektur von einer Kombination aus Baum und Graph zu einer Kombination aus **Baum, Graph und Vektorraum**.

12.4 Baum, Graph und Vektor als komplementäre Logiken

Die Hierarchie bleibt für Provenienz, Verdichtung und Pfadwahl zuständig. Der Graph ergänzt sie durch explizite relationale Quererschließung. Der Vektorraum ergänzt beide um latente semantische Nachbarschaftssuche. Erst aus der Verbindung dieser drei Logiken entsteht ein leistungsfähiges Staffeledächtnis.

12.5 GraphRAG als Bezug, nicht als erschöpfende Definition

Die Nähe des Projekts zu graphbasierten Retrieval-Verfahren ist sachlich gegeben. Gleichwohl sollte das Modell nicht schlicht mit einem Trendbegriff identifiziert werden. Der produktive Kern der GraphRAG-Assoziation liegt darin, die bereits angelegte relationale Querschicht systematisch auszubauen, nicht darin, die Gesamtarchitektur begrifflich auf einen einzelnen aktuellen Ansatz zu reduzieren.

13. Die Rolle der KI-gestützten Bearbeitung

13.1 KI nicht als Gedächtnis, sondern als Strukturierungswerkzeug

Das System geht nicht davon aus, dass eine KI das gesamte Archiv „im Kopf“ behalten soll. Vielmehr dient KI dazu, Archive in Formen zu überführen, die später von KI besser nutzbar sind. Es handelt sich also um eine reflexive Infrastruktur: KI baut die Bedingungen ihrer späteren besseren Einsetzbarkeit mit auf.

13.2 Arbeitsmodi

Daraus ergeben sich mehrere Einsatzweisen:

- Inventarisierung und technische Erschließung,

- inhaltliche Segmentierung,
- taxonomische Zuordnung,
- Dubletten- und Variantenanalyse,
- Clusterbildung,
- Erzeugung standardisierter Metadateien,
- Aufbau von Oberkarten und Community-Reports,
- perspektivische Anlage von Graphstrukturen,
- vektorielle Anreicherung operativer Einheiten.

13.3 Grenzen

Die Verantwortung für die epistemische Qualität der Struktur liegt nicht bei der KI allein, sondern in den Schemata, Taxonomien, Rebuild-Regeln und Prüfverfahren, die ihr vorgegeben werden. Ein schlecht definiertes Ordnungsschema erzeugt auch mit leistungsfähigen Modellen nur ein schlecht organisiertes Gedächtnis.

13.4 Maschinelle Verdichtung als dokumentationspflichtiger Akt

Maschinelle Verdichtung darf nicht als unsichtbarer Zwischenschritt auftreten. Sie muss in der Architektur als eigenständiger, dokumentierbarer und überprüfbarer Akt verankert werden. Dies gilt insbesondere dort, wo Cluster-Reports, Community-Summaries oder relationale Vorschläge erzeugt werden.

14. Rebuild- und Evolutionslogik

14.1 Versionierbarkeit

Das System ist von Anfang an auf Revision angelegt. Kategorien können nachgeschärft, Cluster umgebaut, Teilprojekte ergänzt und Graphschichten nachgerüstet werden. Deshalb ist die Versionierung des Gesamtmodells selbst

ein wichtiger Bestandteil der Architektur.

14.2 Rebuild-Regeln

Ein belastbares System benötigt implizit oder explizit Regeln dafür, wann höhere Ebenen neu erzeugt werden müssen. Beispiele:

- Wenn neue Telegram-Backups hinzukommen, müssen Inventar, Dublettenlogik und Master-Index aktualisiert werden.
- Wenn neue Screenshot-Bestände eingehen, muss die Jahr-Quelle-Thema-Struktur erweitert werden.
- Wenn die Taxonomie verändert wird, müssen betroffene Sidecar-Dateien, Ordnerzusammenfassungen, Community-Reports und Oberkarten gegebenenfalls neu normalisiert

werden.

- Wenn eine Graphschicht eingeführt oder verändert wird, müssen Knotentypen, Relationstypen und Kanten-Geltungslogiken auf bereits vorhandene Metadaten rückbezogen werden.
- Wenn Vektorraummodelle oder Embedding-Strategien geändert werden, müssen auch semantische Nachbarschaftsrelationen neu bewertet werden.

14.3 Langfristigkeit

Die Architektur ist daher nicht nur auf einmalige Aufräumarbeiten, sondern auf fortlaufende Archivpflege angelegt. Sie ist ein Betriebskonzept, kein Einzelwerkzeug.

15. Theoretische Verdichtung für Dokumentation und Lehre

Weil das Projekt nicht nur gebaut, sondern auch vermittelbar gemacht werden soll, benötigt es ausdrücklich didaktische Eigenformen. V1.1 ergänzt die Architektur daher um thematische Verdichtungen zweiter Ordnung, die nicht primär operative Artefakte, sondern dokumentarisch-didaktische Formate darstellen.

15.1 Forschungsfeldkarte

Eine Forschungsfeldkarte soll die beteiligten Disziplinen, Problemstellungen und Begriffsachsen übersichtlich darstellen. Sie dient dazu, das Projekt gegenüber Dritten nicht nur als technische Speziallösung, sondern als interdisziplinäre Forschungsarchitektur darzustellen.

15.2 Architekturkarte

Eine Architekturkarte dient der Darstellung der Ebenen:

- Roharchiv,
- Inventar/Manifest,
- Segment,
- Cluster/Dossier,
- Community-/Cluster-Report,
- Oberkarte,
- Root-Map,
- Vektorraumschicht,
- Wissensgraph-Schicht,
- Rebuild- und Provenienzlogik.

Sie ist für Lehre und Teamkommunikation unverzichtbar, weil sie die operative und die epistemische Architektur gleichzeitig sichtbar macht.

15.3 Methodenkarte

Eine Methodenkarte soll die unterschiedlichen Arbeitsoperationen voneinander unterscheiden:

- Parsing,
- Segmentierung,
- Metadatenanreicherung,
- taxonomische Zuordnung,
- Verdichtung,
- Relationenbildung,
- Vektorisierung,
- Versionierung,
- Rebuild,
- Review.

Sie verhindert, dass einzelne Projektbausteine als unscharfer Gesamtprozess missverstanden werden.

15.4 Epistemische Schutzkarte

Eine epistemische Schutzkarte bündelt jene Sicherungen, die das Projekt gegen Glättung und semantische Entgrenzung absichern:

- Nicht-Destruktivität,
- Provenienzpflcht,
- Trennung von Quelle, Normalisierung und Deutung,
- Manual-Review-Bereiche,
- Confidence-Logik,
- relationale Geltungsmarkierungen,
- Transformationsprovenienz maschineller Verdichtungen.

Diese Karte eignet sich besonders für Dokumentation, Lehre und methodische Selbstvergewisserung.

15.5 Lehrmodule

Für eine mögliche spätere Lehre lassen sich aus dem Projekt mindestens fünf Module ableiten:

1. ****Archivische Grundprinzipien im digitalen Raum****
2. ****Segmentierung und operative Einheitenbildung****
3. ****Taxonomie, Ontologie und Driftkontrolle****
4. ****Hierarchie, Graph und Vektorraum als komplementäre Retrieval-Logiken****
5. ****Epistemische Verantwortung in KI-gestützter Verdichtung****

Diese Module erlauben es, das Projekt nicht nur technisch umzusetzen, sondern auch als Forschungs- und Lehrgegenstand explizit auszuarbeiten.

16. Wissenschaftliche und praktische Tragweite

16.1 Wissenschaftliche Bedeutung

Das Projekt ist nicht nur technisch interessant, sondern berührt grundlegende Fragen digitaler Wissensorganisation: Wie lassen sich heterogene Arbeitsarchive so strukturieren, dass Verdichtung nicht zur Entkontextualisierung wird? Wie kann man unter Bedingungen begrenzter Kontextfenster dennoch mit großen Korpora arbeiten? Wie werden Provenienz, Unsicherheit und konzeptuelle Querbezüge zugleich gesichert? Wie lässt sich persönliche oder projektbezogene Wissensarbeit so aufbereiten, dass sie nicht in bloße Suchbarkeit zerfällt, sondern strukturell anschlussfähig bleibt?

16.2 Praktische Bedeutung

Praktisch eröffnet das Modell mehrere Nutzungen:

- juristisch vorsichtige Dokumentationsvorbereitung,
- Aufbau publizistischer Werkarchive,
- thematische Projektsteuerung,
- schnellere Wiederauffindbarkeit komplexer Denk- und Arbeitsverläufe,
- Vorbereitung späterer Publikationen,
- schrittweiser Ausbau zu einem KI-nutzbaren Langzeitarchiv,
- Erstellung methodischer und didaktischer Verdichtungsformen für Dokumentation und Lehre.

17. Grenzen und offene Probleme

17.1 Taxonomische Drift

Je größer das System wird, desto größer wird die Gefahr begrifflicher Ausfransung. Ohne strikte Pflege der Taxonomie könnte das System in redundante oder konkurrierende Kategoriebildung zerfallen.

17.2 Überverdichtung

Zu starke Verdichtung kann Quellenarmut erzeugen. Wenn Oberkarten, Clusterdateien oder Community-Reports ihre Rückbindung verlieren, wird das System oberflächlich effizient, aber epistemisch schwach.

17.3 Segmentierungsprobleme

Gerade bei Chats, Messenger-Verläufen und mehrteiligen Exporten bleibt die Frage nach den richtigen Grenzen schwierig. Hier ist `manual\_review` nicht Ausnahme, sondern ein notwendiger Bestandteil der Qualitätssicherung.

17.4 Graphische Komplexität

Die Einführung einer Wissensgraph-Schicht verspricht höheren Querzugriff, birgt aber auch das Risiko begrifflicher Inflation. Ohne kontrollierte Knotentypen, Relationstypen und Verifikationslogik würde die Graphschicht mehr Unordnung als Erkenntnis produzieren.

17.5 Vektorielle Intransparenz

Die Vektorraumschicht erweitert die Suchfähigkeit, kann aber semantische Nähe auf nicht unmittelbar einsichtige Weise modellieren. Dadurch entsteht eine neue Form epistemischer Intransparenz, die durch klare Dokumentation, Reproduzierbarkeit und nicht-absolut gesetzte Ähnlichkeitswerte abgefedert werden muss.

17.6 Transformationsblindheit

Wenn maschinelle Verdichtungen nicht mitsamt ihrer Entstehungsparameter dokumentiert werden, entsteht eine zusätzliche Blindstelle. Dann erscheinen Zusammenfassungen oder Cluster als fertige Objekte, obwohl sie in Wahrheit durch spezifische Modell- und Promptsentscheidungen geprägt sind.

18. Vorläufige Gesamtbilanz

****Ki-Lab-26-Archive-Strukture-v1.1**** bezeichnet den Entwurf eines mehrschichtigen Archivökosystems, das aus der Einsicht hervorgeht, dass große digitale Arbeitsbestände weder durch rohe Speicherung noch durch flache Suchbarkeit angemessen beherrschbar werden. Das Modell verbindet nicht-destruktive Quellensicherung, standardisierte Metadaten, kontrollierte Taxonomie, segmentierende Zwischenebenen, thematische Dossiers, Community-Reports, oberste Pfadkarten, eine ergänzende Vektorraumschicht und perspektivisch eine relationale Wissensgraph-Schicht. Seine bereits konkretisierten Teilprojekte – Screenshot-Archivierung und Telegram-Artikelarchiv – fungieren als operative Testfelder für allgemeine Prinzipien, die später auf weitere Bestandsformen, insbesondere Chat-Archive, ausgedehnt werden können.

Die Stärke des Modells liegt nicht in einer einzelnen technischen Operation, sondern in der Kombination aus Quellenbewahrung, Verdichtung, Pfadlogik, Querserschließung und Transformationsbewusstsein. Es ist weder bloß Dateimanagement noch bloß Retrieval, weder nur Archiv noch nur Wissensgraph, sondern der Versuch, ein wachsendes persönliches und projektbezogenes Archiv in eine differenzierte, langfristig erweiterbare Wissensinfrastruktur zu überführen.

Gerade deshalb ist der Charakter von v1.1 ernst zu nehmen. Die vorliegende Fassung beansprucht nicht Endgültigkeit, wohl aber methodische Strenge. Sie versteht sich als Fundament, auf dem spätere Versionen aufbauen können: präziser in der Segmentlogik, breiter in der Bestandsintegration, differenzierter in der Taxonomie, robuster in der Rebuild-Organisation, transparenter in der Transformationsprovenienz und komplexer in der Verbindung von Hierarchie,

Graph und Vektorraum.

19. Arbeitsdefinition für die weitere Entwicklung

Für die weitere Diskussion lässt sich das Projekt in einer konzentrierten Formel so bestimmen:

****Ki-Lab-26-Archive-Strukture-v1.1 ist eine versionierte, nicht-destruktive, provenienzbewusste, taxonomisch kontrollierte und mehrstufig verdichtende Wissensinfrastruktur für heterogene digitale Archivbestände, die operative Einheiten quellenlogisch erschließt, in Dossiers, Community-Reports und Oberkarten verdichtet, durch relationale und vektorielle Zusatzschichten erweitert und dadurch eine pfadförmige, epistemisch rückgebundene und perspektivisch graphisch ausbaubare KI-Nutzung unter Bedingungen begrenzter Kontextfenster ermöglicht.****

20. Arbeitsauftrag für eine mögliche v1.2

Aus v1.1 ergeben sich für eine nächste Ausbaustufe mehrere naheliegende Aufgaben:

- Formalisierung der Transformationsprovenienz,
- Einführung einer versionierten Taxonomieverwaltung,
- prototypische Vektorraumschicht,
- Entwurf eines dualen Archiv- und Inferenzgraphen,
- Entwicklung von Community- und Cluster-Reports,
- Ausarbeitung der Forschungsfeld-, Architektur-, Methoden- und epistemischen Schutzkarten,
- Aufbau eines lehrgeeigneten Begleitformats.

Damit verschiebt sich der Schwerpunkt von der bloßen Grundartikulation des Systems zur methodisch kontrollierten Instrumentierung seiner Zusatzschichten.