Diskriminierungssensible Metadatenpraxis

Ein Handbuch zur ethischen Beschreibung historischer Quellen und Forschungsdaten

AUTOR:INNEN

Moritz Mähr

Noëlle Schnegg

ZUGEHÖRIGKEITEN

University of Bern

ETH Zurich

Universität Basel

ANDERE FORMATE

MS Word

PDF

VERÖFFENTLICHUNGSDATUM

10. Februar 2026

GEÄNDERT

30. Oktober 2026

ZUSAMMENFASSUNG

Dieses Handbuch bietet eine praxisorientierte Anleitung für die diskriminierungssensible Beschreibung von Metadaten zu historischen Quellen und Forschungsdaten. Es richtet sich an Forschende Historiker*innen, Archivar*innen, Bibliothekar*innen und Daten-Kurator*innen in der Geschichtswissenschaft und in Gedächtnisinstitutionen (GLAM). Das Handbuch verbindet theoretische Reflexionen zu Normativität, Bias und Oppression mit konkreten Handlungsempfehlungen für den gesamten Forschungsdatenlebenszyklus. Es behandelt sowohl technische Aspekte der Metadatenmodellierung als auch ethische Fragen der Repräsentation marginalisierter Gruppen. Durch die Kombination von Theorie und Praxis unterstützt es Institutionen und Projekte dabei, Diskriminierung in der Metadatenpraxis zu erkennen und diskriminierungssensible Alternativen umzusetzen.

SCHLÜSSELWÖRTER

Diskriminierungssensible Metadaten, Ethische Metadatenpraxis, Historische Quellen, Digitale Geschichtswissenschaft, FAIR-Prinzipien, CARE-Prinzipien, Bias in Daten, Metadatenstandards, Digital Humanities

CODE-LINKS

Diskussion Living Document, Quellcode

Aktuelle Fassung – Vorabversion 2.0.0-alpha.0

Dies ist die aktuell veröffentlichte Fassung des Handbuchs *Diskriminierungssensible Metadatenpraxis*. Sie ist eine Vorabversion der überarbeiteten zweiten Auflage und noch nicht redaktionell abgeschlossen. Feedback, Korrekturen und Ergänzungen sind willkommen via E-Mail oder GitHub.

Frühere Fassungen: Erste Auflage, Version 1.0.1 (archiviert unter der DOI 10.5281/zenodo.17073511) und Version 1.0.0 (archiviert unter 10.5281/zenodo.11124720).

⚠ Content-Note

Dieses Dokument enthält Abbildungen von historischen Quellen, die diskriminierende Sprache, Bilder oder Darstellungen enthalten. Sie sind Ausdruck von Vorurteilen, Stereotypen oder Gewalt gegen bestimmte Gruppen in der Vergangenheit.

Vorwort zur zweiten Auflage

Diskriminierungssensible Metadatenpraxis. Ein Handbuch zur ethischen Beschreibung historischer Quellen und Forschungsdaten liegt nun als Vorabversion der überarbeiteten zweiten Auflage vor. Dieses Handbuch ist neu in zwei Teile gegliedert. Der Praxisteil ist vorangestellt und führt entlang des Forschungsdatenlebenszyklus durch Entscheidungen, Checklisten und Beispiele für die Erstellung und Pflege von Metadaten. Der theoretische Teil bündelt die begrifflichen und konzeptionellen Überlegungen zu Sprache, Klassifikation, Repräsentation und Macht. Leser*innen können dennoch wahlweise mit der konzeptionellen Rahmung oder mit der direkten Anwendung beginnen: Beide Teile verweisen aufeinander und können im Zusammenspiel genutzt werden.

Die erste Fassung des Handbuchs entstand 2023 im Rahmen des Forschungsprojekts Stadt.Geschichte.Basel als Hilfestellung zur Beschreibung von Objekten auf der Forschungsdatenplattform und erschien am 3. Juni 2024. Diskussionen, Workshops und Rückmeldungen aus der Community machten jedoch deutlich, dass der ursprüngliche Anspruch, zu diskriminierungsfreien Metadaten anzuleiten, nicht einlösbar war. In der zweiten Auflage korrigieren wir diesen Anspruch: Im Zentrum steht nun die Anleitung zu einem diskriminierungssensiblen Umgang mit Machtverhältnissen und Kontexten.

Zugleich haben wir den Praxisteil deutlich ausgebaut und klar vom theoretischen Teil getrennt. Stadt.Geschichte.Basel bleibt ein wichtiges Fallbeispiel, steht jedoch nicht mehr im Zentrum des Handbuchs. Unser Zielpublikum sind Forschende Historiker*innen, Archivar*innen, Bibliothekar*innen und Daten-Kurator*innen in Projekten aus Geschichtswissenschaft und GLAM im deutschsprachigen und europäischen Raum: von universitären Editionsvorhaben bis hin zu digital zugänglichen Sammlungen.

Wir laden alle Leser*innen ein, dieses “Living Document” mit uns gemeinsam weiterzuentwickeln und durch Feedback, Ergänzungen oder Fallbeispiele zu bereichern.

Basel, 10. Februar 2026

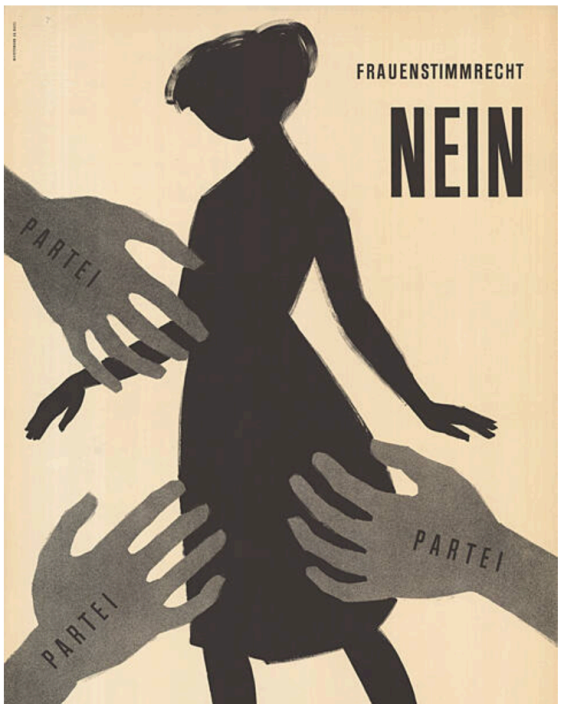
Moritz Mähr & Noëlle Sarah Schnegg

Einleitung

Dieses Handbuch ist eine praxisorientierte Anleitung für die diskriminierungssensible Beschreibung von Metadaten zu historischen Quellen und Forschungsdaten. Es richtet sich an Forschende Historiker*innen, Archivar*innen, Bibliothekar*innen und Daten-Kurator*innen in der Geschichtswissenschaft und in GLAM-Institutionen (Galleries, Libraries, Archives, Museums). Das Handbuch richtet sich sowohl an Einsteiger*innen als auch an erfahrene Fachleute und stellt Best Practices bereit.

Warum Metadaten? In einer zunehmend digital vermittelten Welt machen Metadaten historische Bestände auffindbar, zugänglich, interoperabel und nachnutzbar (im Sinne der FAIR-Prinzipien). Frei zugängliche, maschinenlesbare Metadaten ermöglichen die Integration in Suchmaschinen, Datenportale und virtuelle Forschungsumgebungen. Dadurch verändert sich, wie Historiker*innen Quellen erforschen, interpretieren und verstehen.

Start › Sammlungsobjekte



SAMMLUNG: Plakate

Frauenstimmrecht, Nein

BESCHREIBUNG: Eidg. Abstimmung 1. Feb. 1959

ENTWURF: Werner Nänny (CH, 1916 - 1992)

AUFTRAG: Basler Frauenkomitee gegen das Frauenstimmrecht (CH)

DRUCK: Wassermann AG, Basel (CH, gegründet 1897)

DATIERUNG: [1959]

TECHNIK: Lithografie

VERMASSUNG: 127 × 90 cm

SIGNATUR: CH-000957-X:11785

SAMMLUNGSGRUPPEN: Thema: Politik, Technik: Lithografie

Screenshot aus der Plakatsammlung Basel: Abstimmungsplakat “Frauenstimmrecht, Nein” zur eidgenössischen Abstimmung vom 1. Februar 1959, entworfen von Werner Nänny im Auftrag des Basler Frauenkomitees gegen das Frauenstimmrecht, gedruckt von Wassermann AG, Basel.

Ein Blick in die deutschsprachigen Archive zeigt, dass vielerorts bereits auf die FAIR-Prinzipien bei Metadaten geachtet wird. Eine diskriminierungssensible Metadatenpraxis bleibt jedoch meist aus. Oft fehlen kontextualisierende Beschreibungen, in denen Diskriminierungsformen explizit benannt und in ihren jeweiligen historischen Kontext eingeordnet werden. Im Online-Katalog der “Plakatsammlung der Schule für Gestaltung” beschränkt sich die Beschreibung des Plakats “Frauenstimmrecht, Nein” auf sachliche Angaben: “Eidgenössische Abstimmung, 1. Februar 1959”. Hinweise auf den historischen, politischen und insbesondere den sexistischen Kontext fehlen.

Mit diesem Handbuch setzen wir uns für eine diskriminierungssensible Metadatenpraxis ein. Wir erkennen an, dass Diskriminierung tief in gesellschaftlichen und institutionellen Strukturen verankert ist und sich nicht allein durch eine wohlüberlegte Begriffswahl vermeiden lässt.

Diskriminierungssensibilität bedeutet, aufmerksam zu bleiben für Veränderungen von gesellschaftlichen Normen, die Vielfalt von Diskriminierungserfahrungen und die Unabgeschlossenheit von Dekolonisierungsprozessen. Sie fordert uns auf, unsere eigene Positionierung und bestehende Machtverhältnisse kritisch zu hinterfragen und zu reflektieren.

Im Handbuch verwenden wir bewusst eine breite Definition von Diskriminierung, um möglichst viele Anwendungsfälle abzudecken. Unter Diskriminierung verstehen wir schwerwiegende Formen der Benachteiligung. Eine Benachteiligung wird zu einer Diskriminierung, wenn sie in einem unmittelbaren Zusammenhang mit der tatsächlichen oder zugeschriebenen Zugehörigkeit zu einer bestimmten Gruppe oder einem Merkmal steht. Zu diesen Gruppenzugehörigkeiten oder Merkmalen zählen die soziale Stellung, das biologische und soziale Geschlecht sowie die Geschlechtsidentität, die ethnische Herkunft, diskriminierende Fremdzuschreibungen (zum Beispiel rassistische, antisemitische oder koloniale Kategorisierungen), die Religionszugehörigkeit, die Weltanschauung und politische Überzeugungen, die Sprache, eine Behinderung oder chronische Erkrankung, eine genetische Disposition, das Lebensalter, die sexuelle Orientierung, das Körpergewicht und die Lebensform (zum Beispiel Fahrende). Diskriminierungen entstehen laufend, weil gesellschaftliche Werte und Normen bestimmte Gruppen stigmatisieren.

Der Band gliedert sich in zwei Teile. Er beginnt mit der Praxis. Im Praxisteil behandeln wir den diskriminierungssensiblen Umgang mit Metadaten entlang der Phasen des Forschungsdatenlebenszyklus (Higgins 2008):

1. **Planung und Konzeption:** Konzeptionelle Grundentscheidungen, Auswahl von Standards, Rollen, ethische und rechtliche Aspekte, Fokus auf Einwilligung und Schutz vulnerabler Gruppen. Kapitel 3.1
2. **Datensammlung und Quellenkritik:** Datenerzeugung und -sammlung, Nutzung offener Formate und normierter Begrifflichkeiten zur Interoperabilität sowie Kontextualisierung der Quellen. Kapitel 3.2
3. **Datenverarbeitung und Anreicherung:** Datenaufbereitung, Metadatenanreicherung, Dokumentation, frühzeitige Standardisierung. Kapitel 3.3
4. **Speicherung und Verwaltung:** Strukturiertes, sicheres Speichern, Zugriffskontrolle, Pflege, Versionierung, diskriminierungssensible Zugangsregelungen. Kapitel 3.4
5. **Veröffentlichung und Zugang:** Zugänglichmachung über Repositorien, persistente Identifikatoren (PIDs), Lizenzen, FAIR/CARE-Prinzipien. Kapitel 3.5
6. **Nachnutzung und Wiederverwendung:** Recherchierbarkeit, Interoperabilität, Kontextinformation zur Vermeidung von Fehlinterpretationen. Kapitel 3.6
7. **Archivierung und Löschung:** Auswahl von Archivierungsstandards und Speicherorten, rechtliche und ethische Vorgaben. Kapitel 3.7

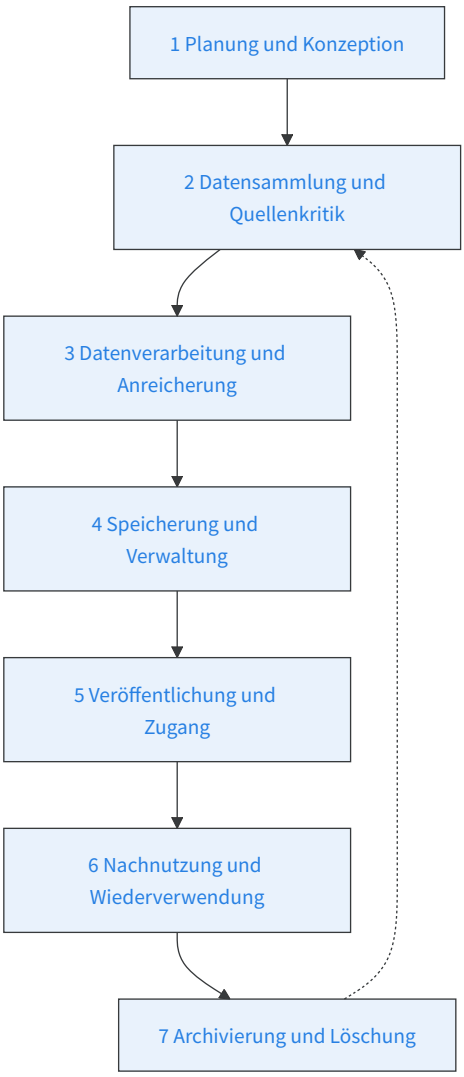


Abbildung 1: Darstellung des Datenlebenszyklus nach dem Curation Lifecycle Model (DCC), adaptiert in sieben linearen Phasen von der Planung bis zur Archivierung beziehungsweise Löschung.

Der zweite Teil bündelt die theoretischen und technischen Schlüsselbegriffe und verknüpft sozial- und informationswissenschaftliche Perspektiven. Er dient als konzeptioneller Rahmen und als Nachschlagebereich (Normativität, Formen der Diskriminierung, Bias und Oppression sowie Grundlagen zu Forschungs- und Metadaten, Metadatenstandards und FAIR / CARE / LOUD).

Das Ziel dieses Handbuchs ist es, anhand konkreter Beispiele, Methoden und Strategien Hilfestellungen zu bieten, die es der Leser*in erlauben sollen, Diskriminierung in der Metadatenpraxis zu erkennen und Entscheidungen in Bezug auf den eigenen Forschungskontext und auf die zur Verfügung stehenden Ressourcen fällen zu können. Zudem greifen wir konkrete Beispiele aus der deutschsprachigen, geschichtswissenschaftlichen Forschungspraxis auf, um auf wiederkehrende Stolperfallen hinzuweisen. Zur Orientierung im Handbuch dient folgende Entscheidungshilfe:

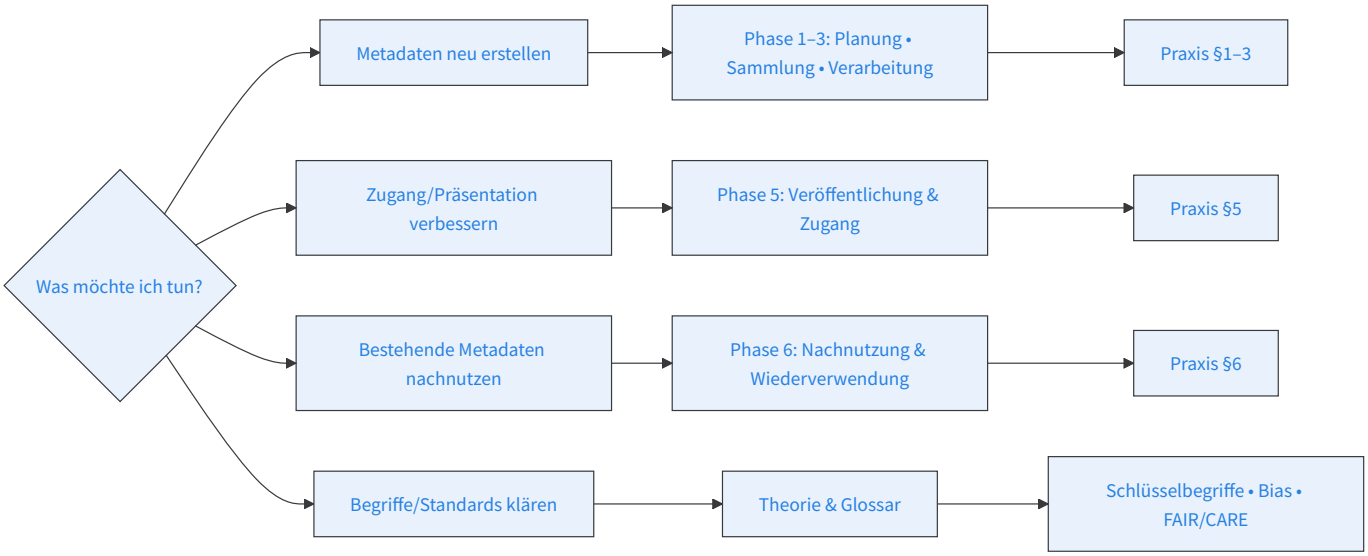


Abbildung 2: Entscheidungshilfe zur Orientierung im Handbuch. Die Startfrage ‘Was möchte ich tun?’ verzweigt in vier Pfade—Metadaten neu erstellen; Zugang/Präsentation verbessern; Bestehende Metadaten nachnutzen; Begriffe/Standards klären—mit Verweisen auf zugehörige Phasen, Praxisabschnitte und Checklisten.

Die Autor*innen bringen unterschiedliche fachliche und persönliche Hintergründe in das Handbuch ein. Moritz Mähr (weisser cis Mann) ist promovierter Historiker. Noëlle Schnegg (weisse cis Frau) studiert Geschichte und Nahoststudien. Beide sind in der Schweiz aufgewachsen und verfügen über privilegierte gesellschaftliche Rahmenbedingungen, wobei sich individuelle Erfahrungen, beispielsweise hinsichtlich Sexismus, unterscheiden. Diese Offenlegung dient der Transparenz und Einordnung der Perspektiven im Handbuch.

Wir orientieren uns am Contributor Covenant, einem verbreiteten Code of Conduct für offene Communities, und verpflichten uns, diskriminierende Inhalte klar zu kennzeichnen und kontextualisiert aufzuarbeiten. Reproduktion problematischer Inhalte erfolgt ausschliesslich zu notwendigen Analysezwecken.

Als Living Document lebt dieses Handbuch von der Community. Antidiskriminierungsarbeit ist ein nie endender gesellschaftlicher Prozess, weshalb auch dieses Handbuch niemals abgeschlossen sein wird. Vielmehr erfordert es kontinuierliche, kritische Reflexion, Überarbeitung und Anpassung. Verbesserungsvorschläge können via E-Mail oder als Kommentar auf GitHub eingereicht werden.

Praxis: Diskriminierungssensibler Umgang mit Metadaten

Metadaten sind weit mehr als neutrale Beschreibungen von Forschungsdaten oder kulturellen Objekten – sie prägen fundamental, wie Inhalte gefunden, verstanden und interpretiert werden. In digitalen Archiven, Repositorien und Forschungsprojekten entscheiden Metadaten darüber, welche Geschichten erzählt werden und welche im Verborgenen bleiben. Dabei reproduzieren sie oft unbewusst historische Machtstrukturen, diskriminierende Begriffe oder ausschliessende Kategorisierungen.

Ein diskriminierungssensibler Umgang mit Metadaten bedeutet, diese Wirkmacht bewusst zu reflektieren und verantwortungsvoll zu gestalten. Es geht darum, marginalisierte Stimmen sichtbar zu machen, respektvolle Sprache zu verwenden und transparente Entscheidungsprozesse zu dokumentieren. Gleichzeitig müssen historische Kontexte gewahrt und problematische Inhalte nicht beschönigt, sondern klar benannt und kontextualisiert werden.

Der Praxisteil orientiert sich am DCC Curation Lifecycle nach Higgins (2008) und führt durch alle Phasen der Metadatenerstellung – von der ersten Planung bis zur Langzeitarchivierung. Die sieben Hauptkapitel behandeln jeweils spezifische Herausforderungen und bieten konkrete Handlungsempfehlungen, Checklisten und Praxisbeispiele. Dabei werden sowohl die etablierten FAIR-Prinzipien als auch die CARE-Prinzipien berücksichtigt, die besonders bei kulturell sensiblen Inhalten von Bedeutung sind.

Methodisch stützt sich der Praxisteil zusätzlich auf deutschsprachige und europäische Infrastruktur- und Standardisierungsdebatten, insbesondere zu Metadatenstandards, Dokumentation und Interoperabilität (Schopper 2024; forschungsdaten.info 2024; «Europeana Data Model. EDM Documentation», o. J.). Diese Perspektive ergänzt den internationalen Forschungsstand um regional etablierte Arbeitsweisen in Archiven, Bibliotheken, Museen und universitären Datenprojekten.

Die Phasen sind iterativ angelegt – Entscheidungen wirken rückkoppelnd auf frühere Schritte, und neue Erkenntnisse erfordern oft Anpassungen in der Herangehensweise. Ziel ist eine belastbare Nachvollziehbarkeit, verbesserte Auffindbarkeit und angemessene Kontextualisierung sensibler Inhalte, die sowohl wissenschaftlichen Standards als auch ethischen Anforderungen genügen.

1. Planung und Konzeption

1.1 Zielsetzung klären

Warum beschreibe ich? Die Herkunft, Erhebungsbedingungen und Weitergabe der Metadaten müssen nachvollziehbar sein. Zudem müssen Metadaten in Archiven und Repositorien auffindbar sein. Ihre Entstehung, ihr Erwerb und ihre Rezeption sollten klar kontextualisiert werden, damit sie richtig verstanden und eingeordnet werden können, denn die Qualität von Metadaten steuert Interoperabilität und Nachhaltigkeit.

Für wen beschreibe ich? Um die Zielgruppen definieren zu können, müssen Informationsbedürfnisse und Arbeitskontexte erhoben werden. Auch potenziell diskriminierungserfahrene Nutzer*innen müssen (insbesondere bei der Veröffentlichung) berücksichtigt werden.

Was beschreibe ich? Dazu muss der Umfang der zu erfassenden Forschungsdaten, die zur Verfügung stehenden Ressourcen und Prioritäten festgelegt werden. Damit einher geht auch die Bestimmung, wie einheitlich die Metadaten aussehen sollen. Dazu kann ein Pilot mit zufällig gezogenen Quellen durchgeführt werden, um Aufwand, Tiefe und Lücken realistisch zu schätzen.

Checkliste Zielsetzung

- [ ] Ziel, Zielgruppen, Nutzungsszenarien dokumentiert
- [ ] Prioritäten und Mindeststandards definiert
- [ ] Pilotkorpus erfasst und Aufwand gemessen
- [ ] Sensible Inhalte antizipiert und gekennzeichnet

Praxisbeispiel: Stadt.Geschichte.Basel

Eine frühe Klärung von Objekttypen, Zielgruppen und Kontexten war zentral. Das Schema blieb bewusst simpel sowie flexibel und wurde iterativ angepasst. Neben technischen Angaben waren kontextualisierende Informationen nötig. Diskriminierende Inhalte wurden historisch, sozial und politisch eingebettet. Fehlende Urheber*innen erforderten eigene Recherche. Dabei halfen verlinkte Nachschlagewerke (zum Beispiel Basler Stadtbuch). Der Überblick über diskriminierende Themen entstand jedoch erst nach vielen aufwendigen Annotationen und war nicht vorgängig absehbar (Mähr 2022).

1.2 Ethische und rechtliche Rahmenbedingungen

Frühzeitige Klärung ist entscheidend: Urheber- und Leistungsschutzrechte, Datenschutz und Persönlichkeitsrechte, Rechte Dritter sowie der Schutz vulnerabler Gruppen müssen berücksichtigt werden. Grundlage bildet ein Rechteinventar, das pro Objekt Herkunft, Urheber*in, Rechteketten und Personenbezug dokumentiert (Weitzmann und Klimpel 2016).

Der Umgang mit personenbezogenen Daten verlangt eine klare Rechtsgrundlage (Einwilligung, Vertrag, gesetzliche Grundlage, berechtigtes Interesse) und gegebenenfalls Schutzmassnahmen wie Anonymisierung, Zugriffsstufen oder Content-Notes (siehe Kapitel 3.5.2).

Für die Veröffentlichung sind Lizenzangaben und Rechteaussagen (zum Beispiel CC-Lizenzen, RightsStatements.org) maschinenlesbar zu dokumentieren. Interne Review- und Eskalationswege sichern die Nachvollziehbarkeit und Rechtssicherheit. Bei komplexen Fällen ist juristische Expertise beizuziehen.

Checkliste ethische und rechtliche Rahmenbedingungen

- [ ] Rechtslage pro Inhaltstyp dokumentiert und geklärt
- [ ] Einwilligungen und Einschränkungen erfasst
- [ ] Personenbezug und Schutzmassnahmen geprüft
- [ ] Lizenz/ rights statement festgelegt
- [ ] Review- und Eskalationswege definiert

1.3 Standards und Infrastruktur festlegen

Als Ausgangspunkt sollten etablierte Schemata wie das DCMES, MODS oder Schema.org geprüft werden. Ihre Verbreitung in der Fachcommunity und die Anschlussfähigkeit an bestehende Systeme sichern Interoperabilität und Nachhaltigkeit. Je nach Projekt kann es sinnvoll sein, Schemata zu erweitern oder zu kombinieren, etwa EAD oder MARC auf Sammlungsebene und DCMES, MODS, LIDO oder VRA Core auf Objektebene (Schopper 2024).

Darüber hinaus ist die frühzeitige Planung von Normdaten, kontrollierten Vokabularen, technischen Plattformen und Datenflüssen entscheidend. Bei der Definition der Metadatenfelder empfiehlt es sich, zunächst alle relevanten Informationen zu sammeln (zum Beispiel Zeitangaben, Geokoordinaten, Lizenzhinweise, Versionsgeschichte). Ob Angaben in eigenen Feldern oder in Kommentarfeldern erfasst werden, hängt von Automatisierbarkeit, Standardisierung und Aggregationsanforderungen ab.

Ein weiterer wichtiger Aspekt in diesem Schritt ist die Sensibilität gegenüber impliziten Annahmen, die mit Metadatenschemata und Ontologien verbunden sind. Jede Klassifikation trifft bestimmte Aussagen über die Welt und ist daher nie neutral. Durch ihre Ordnungsstrukturen suggerieren Metadatenschemata häufig Allgemeingültigkeit und Objektivität, blenden jedoch Widersprüche, Ein- und Ausschlüsse aus. Beispiele hierfür sind die implizite Annahme eines Kernfamilienmodells in standardisierten Erhebungsbögen oder die westlich geprägten Vorstellungen von Besitz und Urheberschaft, die sich etwa in den Dublin-Core-Elementen “creator” und”rights” niederschlagen.

Bei der Festlegung von Metadatenfeldern stellt sich häufig die Frage, ob bestimmte Angaben in eigenständigen Feldern oder in einem allgemeinen Kommentarfeld erfasst werden sollten. Die Entscheidung hängt dabei von Kriterien wie Automatisierbarkeit, Häufigkeit der Nutzung, Standardisierung, Anforderungen durch Aggregatoren sowie der gewünschten Flexibilität ab. In der Praxis erfolgt die Modellierung der Felder meist iterativ und parallel zur Quellenannotation. Dabei kann es vorkommen, dass ursprünglich geplante Felder wieder verworfen oder angepasst werden müssen, weil sich herausstellt, dass sie nur für einen Teil der Objekte relevant sind.

Schliesslich ist es wichtig, sämtliche Entscheidungen und Abwägungen bei der Schemaauswahl und Felddefinition nachvollziehbar zu dokumentieren. So wird nicht nur die interne Konsistenz gewährleistet, sondern auch eine spätere Nachvollziehbarkeit durch andere Forscher*innen und Projekte sichergestellt.

Ein weiterer zentraler Punkt bei der Auswahl und Planung von Infrastrukturen ist die Abwägung zwischen proprietären Lösungen und offenen, quelloffenen Systemen. Proprietäre Software und kommerzielle Plattformen können kurzfristig Vorteile durch Benutzer*innenfreundlichkeit, Support und Marktverbreitung bieten. Gleichzeitig bergen sie das Risiko des Vendor Lock-in: Daten und Workflows sind an ein spezifisches System gebunden, wodurch langfristige Migrationen, Interoperabilität und Kostenkontrolle erschwert werden. Gerade bei Forschungsdaten widerspricht dies den Prinzipien von Nachhaltigkeit, Offenheit und FAIR / CARE.

Checkliste Standards und Infrastruktur

- [ ] Zentrales Schema + Erweiterungen dokumentiert
- [ ] Normdaten/Vokabulare ausgewählt
- [ ] Plattform, Exportformate und persistente Identifikatoren (PIDs; zum Beispiel DOI) definiert

Praxisbeispiel: Stadt.Geschichte.Basel

Im Projekt Stadt.Geschichte.Basel wurde ein flexibles Metadatenmodell entwickelt, das Metadaten von Objekten und dazugehörigen Medien klar voneinander trennt. Zeitangaben werden nach dem EDTF Standard erfasst, damit auch unsichere oder mehrdeutige Daten präzise dokumentiert werden können (Stadt.Geschichte.Basel 2025). Für die Arbeit wird ein Mix aus etablierten Open Source Werkzeugen (etwa Omeka, QGIS, R und Python) und kommerziellen Plattformen (GitHub, ArcGIS) eingesetzt. Die Daten werden vor Veröffentlichung geprüft, angereichert und zur Nachnutzung bereitgestellt. Für die langfristige Sicherung werden sie im DaSCH und auf Zenodo archiviert. Begleitend sorgen Schulungen, Data Stewards und die Einbindung in Lehre und Praktika für nachhaltige Nutzung. Leitlinien wie FAIR, Open Access und digitale Nachhaltigkeit prägen die Infrastruktur. Proprietäre Dienste werden durch offene Standards und Repositorien abgesichert. Zudem gibt es klare Regelungen zu Datenschutz und Barrierefreiheit. Die Dokumentation wird innerhalb der Projeklaufzeit regelmässig aktualisiert (Mähr 2022).

2. Datensammlung und Quellenkritik

2.1 Primärerschliessung vs. Nachnutzung

In dieser Phase werden die eigentlichen Forschungsdaten erzeugt oder gesammelt. Dies kann durch empirische Erhebungen, Messungen, Archivarbeit oder Datenübernahmen aus anderen Quellen erfolgen. Die Nutzung von strukturierten Erhebungsinstrumenten, offenen Formaten und normierten Begrifflichkeiten trägt hier zur künftigen Interoperabilität bei.

In der Praxis findet oftmals eine Kombination aus Primärerschliesung und Re-Use statt. So kann es von Vorteil sein, jeweils bei den Gedächtnisinstitutionen und Portalen nachzuschauen. Beispielsweise liegen bei Europeana oftmals schon Metadaten bereit, die ganz oder in Teilen übernommen werden können. Im Idealfall sind diese Datensätze bereits miteinander kompatibel (beziehungsweise interoperabel) und können reibungs- und verlustarm ineinander überführt werden. In der Realität muss jedoch beim Zusammenführen, Anreichern, Ausdünnen und Korrigieren bestehender Metadatensätze meist eine Reihe von Entscheidungen, Abwägungen und Anpassungen getroffen werden. Ausserdem ist im Blick auf eine diskriminierungssensible Metadatenpraxis zu beachten, dass Archive stets einen selektiven Einblick auf Geschichte geben, dem man sich bei deren Verwendung stets bewusst sein sollte. Ein Rückgriff auf bestehende Metadaten birgt deshalb stets auch das Risiko einer unbewussten Reproduktion potenziell diskriminierender Inhalte.

Genozid-Denkmal

Dieses Objekt wird von Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg bereitgestellt und gepflegt

Auf der Website des Anbieters ansehen

Gut zu wissen | Alle Metadaten

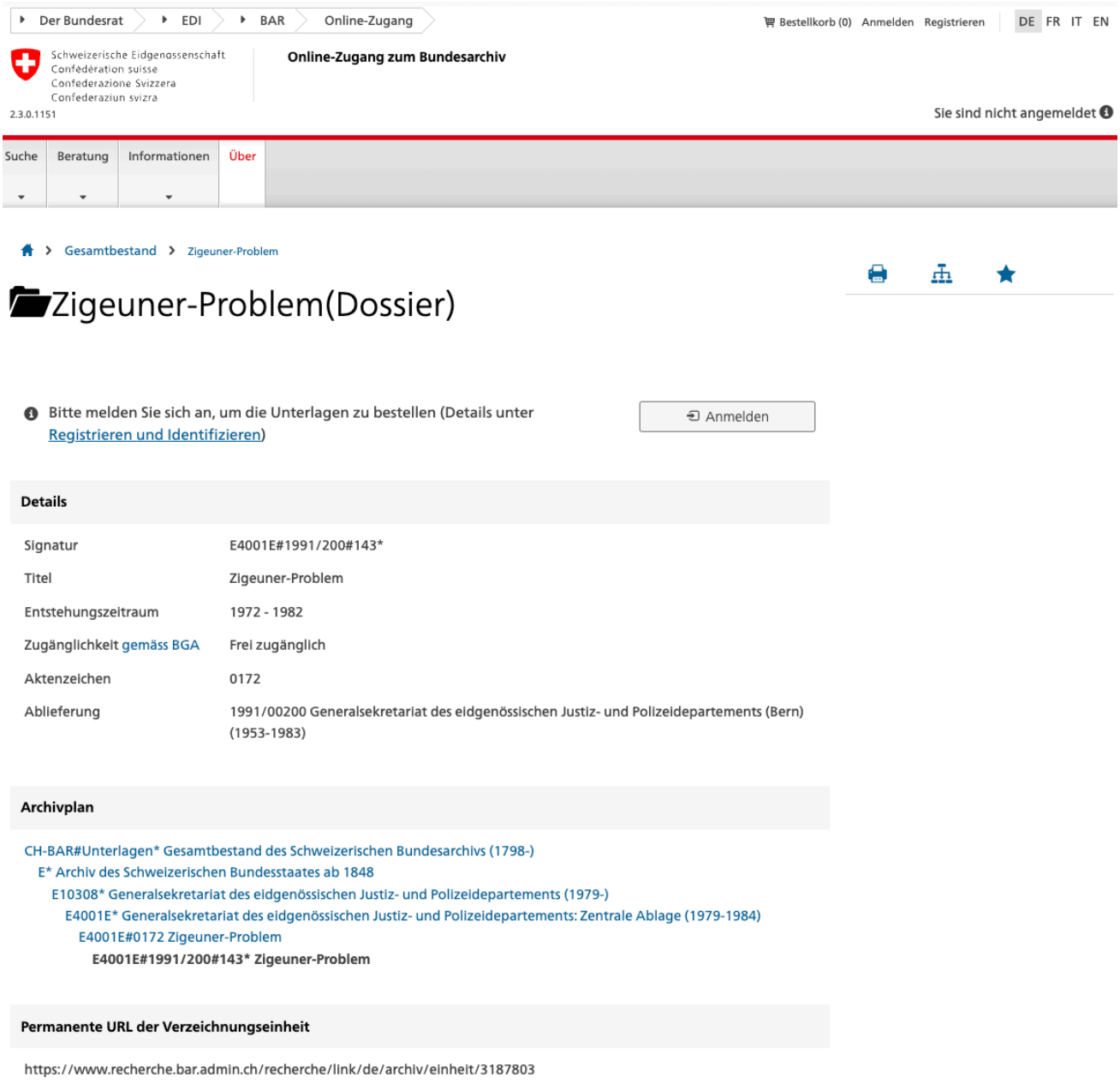
Betreff	**Skulptur**
Art des Objekts	**memorials (monuments)**
Datenpartner	**Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg**
Aggregator	**Deutsche Digitale Bibliothek**
Rechtehinweise der Medien in diesem Datensatz (sofern nicht anders angegeben)	**http://rightsstatements.org/vocab/InC/1.0/**
Erstellungsdatum	**1965**
Orte	**Ejmiacin (Standort)**
Kennung	**http://www.deutsche-digitale-bibliothek.de/item/EIXXGS4W5AVJLCA556YUS4IRIS2CRMF5**
Bereitstellendes Land	**Germany**
Name der Sammlung	**199_DDB_BildarchivFotoMarburg**
Erstmals auf Europeana veröffentlicht	**2022-01-31T11:04:15.728Z**
Zuletzt aktualisiert vom Datenpartner	**2022-01-31T11:04:15.728Z**

Screenshot aus Europeana: Metadaten zum “Genozid-Denkmal” in Ejmiacin, Armenien.

Zum Objekt “Genozid-Denkmal” aus Europeana liegen schon einige Metadaten bereit. Erfasst sind der **Betreff** (Skulptur) und die **Objektart** (memorials/monuments). Als **Datenpartner** wird das Deutsche Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg angegeben, während die **Deutsche Digitale Bibliothek** als Aggregator fungiert und die Metadaten zusammenzog. Weitere Angaben betreffen das **Erstellungsdatum** (1965), den **Ort** (Ejmiacin), die **Sammlung** (199 DDB BildarchivFotoMarburg) sowie die **Rechteinformation** nach rightsstatements.org.

Im Beispiel wird deutlich, dass die Metadaten zwar formale Metadaten enthalten, jedoch wesentliche Aspekte einer diskriminierungssensiblen Metadatenpraxis fehlen. So bleibt die Beschreibung auf einer technischen, oberflächlichen Ebene, ohne den historischen Kontext, also den Genozid an den Armenier*innen, zu benennen. Auch die Betroffenenperspektive findet nirgends Ausdruck: Weder werden anerkannte Selbstbezeichnungen verwendet, noch gibt es Verweise auf Ressourcen der armenischen Community. Stattdessen stehen ausschliesslich die beteiligten Institutionen als Datenpartner im Vordergrund. Zudem beschränken sich die Metadaten auf Deutsch und Englisch; eine Mehrsprachigkeit unter Einbezug der armenischen Sprache wäre jedoch zentral, um kulturelle Kontexte sichtbar zu machen. Schliesslich behandelt der Rechtehinweis lediglich die Nachnutzung des Fotos, nicht aber die kulturelle Sensibilität des Denkmals selbst. Eine diskriminierungssensible Metadatenpraxis sollte im Gegensatz über die rein formale Erfassung hinausgehen, indem sie präzisere Benennungen wählt, historische Einordnungen in einem Beschreibungsfeld ergänzt, Mehrsprachigkeit ermöglicht und vor allem die Perspektiven und Stimmen der betroffenen Communities einbezieht. Dies müsste bei der Erstellung alles berücksichtigt und neu erfasst werden.

Beim Zusammenführen von Metadaten sollten in einem letzten Schritt alle Fehlstellen, Unstimmigkeiten und möglichen Informationsverluste transparent gemacht werden. Zudem ist es wichtig, die getroffenen Entscheidungen nachvollziehbar zu dokumentieren, um den Prozess für andere überprüfbar und weiterführbar zu halten. Dazu gehört auch, kontaminiertes Archivgut (Diskriminierungen, Lücken, Biases oder ethische Konflikte) zu dokumentieren und in den neuen Metadaten oder Paradaten auszuweisen.



Screenshot aus dem Online-Zugang des Schweizerischen Bundesarchivs zur Akte “Zigeuner-Problem” (Signatur E4001E#1991/200#143*), erstellt 1972–1982, abgelegt beim Generalsekretariat des Eidgenössischen Justiz- und Polizeidepartements.

Das Dossier “Zigeuner-Problem” im Schweizer Bundesarchiv macht deutlich, wie selbst Archivtitel Diskriminierung reproduzieren können. Diese stigmatisierende Fremdbezeichnung entstand in der damaligen Behördenpraxis und sollte zwar für heutige Nutzer*innen zur Sicherung der Authentizität erhalten bleiben, gleichzeitig jedoch in den Metadaten oder Paradaten klar als diskriminierende Bezeichnung markiert und kontextualisiert werden. Konkret könnte beispielsweise in den Metadaten ein Hinweis auf die Verfolgung von Sinti und Roma sowie alternative Suchbegriffe respektive Selbstbezeichnungen stehen.

Kritische Fragen an bestehende Metadaten

- Was ist archiviert und wie strukturiert?
- Welche Elemente werden explizit erfasst, wie präsentiert?
- Nach welcher Logik ist die Sammlung geordnet?
- Wo liegen Lücken, Biases, Unsicherheiten in Metadaten und Sammlung?
- In welchem Entstehungs- und Institutionskontext steht die Sammlung?
- Welche Rückfragen an die Institution sind nötig?

Checkliste Re-Use

- [ ] Re-Use-Quellen bewertet und zitiert
- [ ] Übernahmen, Streichungen, Korrekturen protokolliert
- [ ] Bias-Befunde und Lücken sichtbar gemacht

2.2 Kontextualisierung der Quellen

Wenn wie in den Beispielen “Genozid-Denkmal” oder “Zigeuner-Problem” eine Kontextualisierung fehlt, muss in diesem Schritt eine Analyse des Entstehungs- sowie des Verwendungskontextes vorgenommen werden. Mit Blick auf einen diskriminierungssensiblen Umgang mit historischen Quellen gilt es zu beachten, dass sich diskriminierende Ideologien im Lauf der Zeit verändert haben. Eine fundierte Analyse der historischen Überlieferungsgeschichte erlaubt es, die Wirkungsmacht dieser Ideologien zum Zeitpunkt der Objekterstellung zu rekonstruieren und damit kritisch zu reflektieren. Im Zentrum steht dabei die Frage “Unter welchen Umständen und mit welcher Absicht wurde die Quelle verfasst?” (Hitz und Stüdli 2016).

Zudem helfen bei der Kontextualisierung die acht W-Fragen: Wer? Wann? Wo? Welche Quelle? Warum? An wen? Wie/wer überlieferte? Wovon zeugt sie, wovon schweigt sie?

Durch das Durcharbeiten der W-Fragen können die sozialen und politischen Kontexte (**Entstehungskontext**) erfasst werden. Diskriminierungsformen wie beispielsweise Rassismus sind eng mit sozialen, politischen, wissenschaftlichen, institutionellen, ökonomischen und kulturellen Strukturen verbunden und sind in Hinblick auf Erstellung, Rezeption und Reproduktion der Quelle gleichermassen relevant.

Neben dem Entstehungskontext muss auch der **Verwendungskontext der Quelle** ermittelt und kritisch reflektiert werden. Wichtig hier sind die Fragen, wie das Objekt beziehungsweise die Quelle aktuell zur Verfügung gestellt wird. In welchem archivarischen Kontext (Ordnungssystem, Schlagwörter, Fachbereich) ist es zu finden? Was ist der Umfang und Zustand der zu beschreibenden Metadaten? Wer waren die betreuenden Archivar*innen und was ist über deren Arbeitskontexte bekannt? Dabei ist auch ein Blick in die Rezeption zu werfen. Wie wird die Quelle dort interpretiert und eingeordnet? Und welche sozialen, wissenschaftlichen und politischen Kontexten waren dabei wiederum wirksam?

Checkliste Kontextualisierung

- [ ] Entstehungs- und Nutzungskontext beschrieben
- [ ] Relevante Akteur*innen und Machtverhältnisse benannt
- [ ] Rezeption und Deutungskonflikte referenziert

3. Datenverarbeitung und Anreicherung

3.1 Technische Standards implementieren

Zentrale Grundlage einer nachhaltigen und nachvollziehbaren Datenverarbeitung ist die Festlegung von Feldstrukturen, Datentypen, Formaten und Kardinalitäten. Diese Elemente sollten in einem iterativen Prozess modelliert werden, der eng mit der fortlaufenden Annotation verknüpft ist. Ergänzend sind Metafelder vorzusehen, die Auskunft geben über Quellenlage, Zuverlässigkeit, Präzision, Bearbeitungsstand, Veröffentlichungsstatus, Versionierung, Lizenz sowie Zitiervorschläge. Für sensible Inhalte sind Sichtbarkeits- und Zugriffskontrollen (Flags) einzuplanen, um eine differenzierte Steuerung zwischen interner Verarbeitung und externer Veröffentlichung zu ermöglichen.

Die Automatisierung von Verarbeitungsschritten setzt atomar strukturierte Felder voraus. Freitextangaben, die nur in seltenen Fällen vorkommen, sollten hingegen möglichst zentral gebündelt werden, um Redundanzen und Validierungsprobleme zu vermeiden.

Ontologien sind nicht neutral Schemata treffen Annahmen über Welt und Personen. Jede Festlegung von Datenmodellen, Schemata und Kontrollstrukturen impliziert Annahmen über Weltbilder, soziale Kategorien und rechtliche Rahmenbedingungen. Beispiele hierfür sind das binäre Geschlechtsmodell, rigide Namensfelder ohne kulturelle Differenzierung oder die in westlich geprägten Standards vorherrschende Urheberrechtslogik, die sich in Feldern wie `creator` oder `rights` niederschlägt. Solche Modellierungen sind niemals selbstverständlich, sondern müssen als bewusste Entscheidungen transparent dokumentiert werden. Eine präzise Aufzeichnung dieser Abwägungen ist daher verpflichtend, um die Nachvollziehbarkeit und kritische Überprüfbarkeit der Datenarchitektur sicherzustellen.

Feldstruktur, Datentypen, Formate und Kardinalitäten festlegen. Iterativ modellieren und parallel annotieren. Ergänzende Felder für Quellenlage, Zuverlässigkeit, Präzision, Bearbeitungsstand, Veröffentlichungsstatus, Version, Lizenz, Zitiervorschlag vorsehen. Sichtbarkeitsflags für sensible Inhalte einplanen. Automatisierung erfordert atomare Felder; seltene, frei formulierte Angaben gegebenenfalls zentral bündeln.

Beispiele: binäres Geschlechtsmodell, starre Namensfelder, westliche Urheberrechtslogik in creator/rights. Dokumentation aller Abwägungen ist Pflicht.

Checkliste Technische Standards

- [ ] Feldkatalog mit Definitionen erstellt
- [ ] Maschinenlesbare Datentypen und Validierung gesetzt
- [ ] Sichtbarkeits- und Sensitivitätslogik modelliert

3.2 Beschreibung und Verschlagwortung

Nach der Wahl der Metadatenstandards (Kapitel 7.4), müssen die Felder befüllt werden. Dazu kann für jedes Feld auf eine Reihe sogenannter **kontrollierter Vokabulare** beziehungsweise **Normdatensätze** (auch Normdateien genannt) zurückgegriffen werden. Hier kann zwischen Normdatensätzen (etwa der GND) sowie kontrollierten Vokabularen und spezifischen Schlagwortindizes (etwa thematischen Indizes wie GenderOpen) unterschieden werden. Für Ortsverzeichnisse wird oft auch Gazetteer verwendet. Auch DCTerms stellt mit den **DCMI Metadata Terms** übersichtliche Datenwertstandards mit einigen Dutzend Einträgen bereit. Diese sollten nicht mit dem DCMES verwechselt werden.

Für eine diskriminierungssensible Metadatenpraxis ist es wichtig, Normdaten und Vokabulare auch inhaltlich kritisch zu hinterfragen. So ist beispielsweise auf die Grenzen von GND hinzuweisen: Viele Begriffe fehlen, und häufig liegt eine binäre Logik zugrunde. Über lokale GND-Redaktionen können jedoch Ergänzungsanträge eingereicht werden, um diskriminierende Lücken zu schliessen oder problematische Einträge zu verändern. Als Beispiel ist der Begriff *Gender* zu nennen, der 2024 offiziell in GND aufgenommen wurde.

Darüber hinaus sollte Mehrsprachigkeit systematisch berücksichtigt werden, nicht als nachträgliche Ergänzung, sondern als grundlegende Systemfrage. Begriffe sind stets kulturell kodiert und lassen sich nicht ohne Weiteres in andere Sprachen übertragen, wie das Beispiel *Race* ≠ *Rasse* verdeutlicht. Auch beim Einsatz von Mappings und Crosswalks empfiehlt es sich, mit Äquivalenzklassen zu arbeiten, deren Definitionen jedoch kritisch zu prüfen. Widersprüche oder Ausschlüsse müssen sichtbar gemacht und, falls nötig, in Anträgen an die zuständigen Redaktionen thematisiert werden.

Es kann zwischen **intrinsischen und extrinsischen Schlagwörtern** unterschieden werden. Dabei geht es darum, ob die verwendeten Begriffe der Quelle selbst entstammen (intrinsisch), oder ob es sich um äussere Zuschreibungen oder Kontextinformationen handelt (extrinsisch). Schemata wie Dublin-Core erlauben es jedoch, intrinsische und extrinsische Schlagwörter demselben Element zuzuordnen (zum Beispiel `dc:subject`), ohne diese explizit als solche auszuzeichnen, wobei extrinsisch häufiger ist. Bei der extrinsischen Verschlagwortung sowie bei Objekt- und Bildbeschreibungen ist es nicht nur relevant, welche Begriffe gewählt werden, sondern auch, wer dabei Mitsprache erhält. Diese Entscheide sollten dokumentiert werden.

Checkliste Beschreibung

- [ ] Vokabulare und Normdateien pro Feld definiert
- [ ] Regeln für Intrinsisch/Extrinsisch festgelegt
- [ ] Unsicherheit formalisiert und sichtbar gemacht

Lücken, Widersprüche, Unsicherheiten und häufige Fehler

- **Lücken:** Fehlende Personenidentitäten explizit kennzeichnen. Keine Annahmen zu Nationalität, Geschlecht oder anderen Merkmalen treffen.
- **Widersprüche:** Abweichende Datierungen oder Ortsangaben parallel erfassen und Quellenlage dokumentieren.
- **Unsicherheit:** Unscharfe Angaben formal kennzeichnen (zum Beispiel EDTF "~1905").
- **Häufige Fehler:**
  - Nacktheit und Sexualität ohne Kontext sexualisieren
  - Personen über Diskriminierungsmarker essenzialisieren
  - Koloniale Namensformen unkritisch übernehmen; stattdessen Synonyme und Alternativnamen pflegen

Freitext-Beschreibung und bewusste Begriffsentscheidungen

- Respektvolle, auf Care ausgerichtete Sprache statt vermeintlicher Neutralität
- Aktive und präzise Ausdrucksweise ("X tötete Y" statt Passivformen)
- Machtverhältnisse benennen, wenn kontextrelevant
- Keine heroisierenden Darstellungen von Bestandsbildner*innen
- Problematische Ereignisse klar benennen: Lynching, Vergewaltigung, Mord
- Sprachspezifika berücksichtigen (zum Beispiel *Race* ≠ *Rasse*)
- Begriffe kontextualisieren ("Zwerg" vs. "kleinwüchsige Person")
- Person-first-Formulierungen verwenden ("Person mit …")
- Selbstbezeichnungen respektieren (zum Beispiel "Crip" nur als Selbstbeschreibung)
- Problematische LCSH- oder GND-Heading beibehalten, aber mit Bemerkung versehen

3.3 KI-Unterstützung und Automatisierung

Die Integration von Large Language Models (LLMs) in die Metadatenerstellung verändert die Arbeit in Gedächtnisinstitutionen und Forschungsprojekten grundlegend. Entscheidend ist nicht Effizienz, sondern der kontrollierte Umgang mit Risiken wie Verzerrung, Halluzinationen und Intransparenz. KI-generierte Inhalte müssen klar gekennzeichnet werden, ergänzt durch Risiko- und Bias-Assessments sowie eine dokumentierte Entscheidungsgrundlage (Impact-Assessment). Transparenz und Nachvollziehbarkeit von Modellen, Prompts und Versionen sind zwingend.

Künstliche Intelligenz bei der automatischen Verschlagwortung bei der DNB

Die Deutsche Nationalbibliothek setzt seit 2012 KI für die automatische Verschlagwortung ein. Während grosse Datenmengen effizient erschlossen werden, zeigen sich Grenzen an den Rändern der statistischen Verteilung (seltene Begriffe) und bei komplexen Themen. Bei einer sozialwissenschaftlichen Reihe musste die Automatisierung wegen hoher Fehlerquoten eingestellt werden (Junger und Scholze 2021).

LLMs reproduzieren kulturelle und historische Prägungen, verstärken Stereotypen und verschleiern Unsicherheiten durch Halluzinationen. Ohne redaktionelle Kontrolle drohen fehlerhafte oder spekulative Inhalte, die sich in Katalogen verfestigen.

Multimodale KI für automatische Alt-Texte bei Stadt.Geschichte.Basel

KI-Modelle generieren im Projekt automatische Alt-Texte und verbessern so die Barrierefreiheit. Ohne systematische Qualitätskontrolle bleiben jedoch Risiken inkonsistenter oder diskriminierender Beschreibungen (Mähr und Twente 2025).

"Human-in-the-loop" ist kein Garant für Sicherheit: Fachpersonen benötigen Zeit, Schulung und klare Workflows. Definierte Rollen (zum Beispiel Kuratierung, Technik, Ethik-Board), nachvollziehbare Versionshistorien und Eskalationsverfahren sind notwendig, ebenso wie gezielte Trainings zu Bias-Erkennung und Prompt-Design.

Bias-Erkennung in Metadaten mit dem DE-BIAS-Tool

Das europäische Projekt DE-BIAS entwickelte Werkzeuge zur automatisierten Erkennung diskriminierender oder historisch belasteter Begriffe in Metadaten von Kulturerbeinstitutionen. Kernstück ist ein KI-gestütztes Webtool, das Sammlungsdaten analysiert, problematische Ausdrücke markiert und kontextualisierte Alternativvorschläge anbietet – gestützt auf ein kontrolliertes, mehrsprachiges Vokabular (derzeit in fünf Sprachen verfügbar).
Besonders relevant ist DE-BIAS für Gedächtnisinstitutionen, die historische Metadatenbestände pflegen, deren Begriffe aus kolonialen, rassistischen oder sexistischen Diskursen stammen können. Durch transparente Annotation und Exportfunktionen lassen sich Korrekturläufe effizient planen. Doch auch hier gilt: Die maschinelle Erkennung ersetzt keine kritische Auseinandersetzung, wie auch die verhältnismässig tiefe Erkennungsrate von DE-BIAS zeigt. Es müssen konkrete Einsatzszenarien definiert werden (zum Beispiel automatische Transkription, Übersetzungen, Vorschläge für Schlagwörter) und es ist klarzustellen, wann ausschliesslich menschliche Expertise erforderlich ist (zum Beispiel Kontextualisierung oder Bewertung kontroverser Inhalte). Ergänzend können Ansätze der Explainable AI (xAI) und Metriken wie Fairness, Precision und Recall helfen, die Qualität systematisch zu evaluieren.

Checkliste KI-Einsatz

- [ ] KI-Einsatzbereiche, Modelle, Trainings-/Prompt-Kontexte dokumentiert
- [ ] Stufen der Qualitätskontrollen, Freigaben, Rückverfolgbarkeit definiert
- [ ] KI-Anteile in Feldern gekennzeichnet
- [ ] Risiko- und Bias-Assessments durchgeführt und dokumentiert
- [ ] Datenschutz, Urheberrecht und FAIR / CARE-Prinzipien geprüft
- [ ] Kontrollschwellen (zum Beispiel Confidence-Scores) für menschliche Überprüfung festgelegt
- [ ] Rollen, Eskalationsverfahren und Verantwortlichkeiten definiert
- [ ] Modelle, Prompts und Updates versioniert und veröffentlicht
- [ ] Mitarbeitende geschult und fortlaufend sensibilisiert

4. Speicherung und Verwaltung

4.1 Repositorien und Plattformen

Im Zentrum dieser Phase steht die dauerhafte, sichere und strukturierte Ablage der Forschungsdaten. Dazu werden Speichersysteme, Zugriffsrechte und Backup-Strategien festgelegt. Relevanz haben dabei sowohl technische Standards und Vorgaben zur Datensicherheit als auch diskriminierungssensible Zugangsregelungen. Ebenso zentral sind die fortlaufende Pflege und Versionierung der Daten, um Transparenz und Nachvollziehbarkeit zu gewährleisten.

Die Auswahl geeigneter Plattformen sollte sich nach Kriterien wie Dauerhaftigkeit, Persistenten Identifikatoren (PIDs) und Exportpfaden richten. In Betracht kommen insbesondere Zenodo, DaSCH und SWISSUbase oder institutionelle Repositorien mit langfristiger Ausrichtung. Barrierefreiheit ist mitzudenken. Ein Data Dictionary ist bereitzustellen. Es sollte Felddefinitionen, Datentypen, zulässige Werte, Beispiele und Annotationsregeln enthalten. Zugriffsstufen und sensible Felder sind technisch verpflichtend einzurichten.

Offene Publikationsinfrastrukturen wie Zenodo bieten insbesondere kleineren Teams und Projekten ohne institutionelle Anbindung einen niederschwelligen Weg zur langfristigen Sicherung und Sichtbarkeit von Forschungsdaten. Zugleich ermöglichen abgestufte Zugriffskontrollen oder kontextspezifische Lizenzen den Schutz sensibler Daten, ohne deren Existenz oder Forschungskontext zu verschleiern.

4.2 Versionierung und Historisierung

Metadaten sollten nicht gelöscht, sondern in ihrer Entwicklung dokumentiert werden. Jede Änderung ist mit den Angaben *Wer–Wann–Warum* zu versehen. Frühere Forschungsstände sind als Referenzen zu bewahren, um Transparenz über die eigene Institutions- und Forschungsgeschichte sicherzustellen. Dafür gilt es, geeignete Workflows in der eingesetzten Software zu nutzen. Bei kleineren Projekten bieten sich versionierte Releases in Zenodo an; auch der Einsatz von *Semantic Versioning* kann hier sinnvoll sein. Ein Beispiel für eine strukturierte Dokumentation findet sich unter «Europeana Data Model. EDM Documentation» (o. J.).

Die Wahl des Versionierungsmodus richtet sich nach Komplexität und Kollaborationsgrad. Für sehr kleine und kurzlebige Artefakte genügt eine konsistente Dateibenennung, ergänzt um ein zentrales **Änderungsprotokoll**, das jeweils Änderungen, Begründungen und betroffene Dateien aufführt. Für kleine Teams mit laufender Zusammenarbeit liefern Plattformen wie Nextcloud oder ownCloud automatische Dateiversionen; dennoch sollten die Entscheidungsgründe zusätzlich im **Änderungsprotokoll** festgehalten werden, da Dateiversionshistorien selten aussagekräftige Provenienzen abbilden. Für kollaborative Projekte mit Verzweigungen, Peer-Review und hohen Anforderungen an Reproduzierbarkeit ist der Einsatz von **Git** vorzuziehen.

Checkliste Speicherung und Verwaltung

- [ ] Repositorium und persistente Identifikatoren (PIDs) gewählt
- [ ] Data Dictionary publiziert
- [ ] Versionierungsstrategie definiert und technisch verankert

5. Veröffentlichung und Zugang

In der Phase der Veröffentlichung werden die Forschungsdaten, sofern rechtlich und ethisch vertretbar, für Dritte zugänglich gemacht. Üblicherweise erfolgt dies über Repositorien, verbunden mit der Vergabe persistenter Identifikatoren wie DOIs, der Auswahl geeigneter Lizenzen sowie der Bereitstellung umfassender Metadaten. Ziel ist es, Offenheit und Nachnutzbarkeit im Sinne von Open Data und den FAIR-Prinzipien zu ermöglichen, zugleich aber die CARE-Prinzipien zu berücksichtigen, die gegebenenfalls Einschränkungen erfordern.

5.1 Zielgruppe definieren

Metadaten erfüllen die Funktion von Findhilfen und Kontextträgern. Im Rahmen der Publikation müssen Ort, Vermittlungsstrategien und Mechanismen zur Langzeitkontrolle geklärt werden. Dazu gehört auch die Definition von Zugangsebenen, die Planung verschiedener Zielgruppen und ihrer Einstiegspfade sowie die Festlegung des Grades an Kontextualisierung auf Bestands-, Dossier- und Objektebene.

Von Beginn des Projekts an sollte der dauerhaft lauffähige Endzustand definiert werden. Dadurch wird eine Übernahme, Versionierung und Zitierbarkeit durch Gedächtnisinstitutionen erleichtert. Eine solche Rückwärtsplanung reduziert Betriebsrisiken und erleichtert die Übergabe (The Endings Project Team 2023).

Adressierbarkeit und Persistenz erfordern die Vergabe kanonischer, parameterloser URLs für alle Entitäten. Editionen sollen versioniert veröffentlicht werden, Umzüge müssen durch Weiterleitungen abgesichert sein, und Zitierempfehlungen sind sichtbar zu machen.

Checkliste Zielgruppe

- [ ] Zielgruppen, Einstiegspfade und Zugangsebenen definiert
- [ ] Statischer Endzustand mit Rückwärtsplanung festgelegt
- [ ] Kanonische, parameterlose URLs oder DOIs vergeben
- [ ] Editionen versioniert, Weiterleitungen bei Umzügen vorgesehen
- [ ] Zitierempfehlungen sichtbar integriert

5.2 Umgang mit sensiblen und diskriminierenden Inhalten

Sensible Daten wie Informationen zu Personen, Kulturgut oder Fundorten erfordern besonderen Schutz. Ebenso gilt es, diskriminierende Inhalte (Kapitel 4.1) klar zu benennen, kontextualisiert zu präsentieren und nicht zu beschönigen. Content-Notes bieten dabei eine Möglichkeit, Zugänge zu moderieren und gleichzeitig Transparenz herzustellen.

Bei sensiblen Daten können Geodaten archäologischer Fundorte, personenbezogene Abbildungen lebender oder identifizierbarer Menschen sowie Kontaktdaten von Bearbeiter*innen problematisch sein. Bei diskriminierenden Inhalten sollte stets die Wirkung geprüft und die Reproduktion schädlicher Anschauungen vermieden werden.

Einwilligungen sind insbesondere bei historischen Bildern häufig nicht vorhanden. Dies erfordert eine bewusste Reflexion des Veröffentlichungswegs sowie gegebenenfalls Alternativen wie Verpixelungen, Kontextlayer oder eingeschränkten Zugang.

Entscheidungen zu Sichtbarkeit, Content-Notes und alternativen Darstellungsweisen sind pro Edition zu dokumentieren. Eingriffe in Daten oder Medien sollen reversibel sein, und Änderungen sind mit Begründungen im Änderungsprotokoll auszuweisen.

Checkliste sensible und diskriminierende Inhalte

- [ ] Sensible Daten identifiziert und Schutzmassnahmen getroffen
- [ ] Diskriminierende Inhalte benannt und kontextualisiert
- [ ] Content-Notes formuliert
- [ ] Einwilligungsfragen reflektiert und Alternativen geprüft
- [ ] Governance-Regeln dokumentiert und Änderungen nachvollziehbar protokolliert

5.3 Technische Strategien

Das Zielbild einer Veröffentlichung ist ein statischer Webauftritt, realisiert mit HTML, CSS und minimalem JavaScript. Er soll ohne Datenbank und ohne proprietäre Dritt-Dienste auskommen. Redundanz wird bewusst in Kauf genommen, um Resilienz zu erhöhen, während statische Suchlösungen zu bevorzugen sind.

Der Build- und Validierungsprozess steht unter Versionskontrolle. Quellen, Skripte und Konfigurationen müssen nachvollziehbar versioniert sein. Jeder Build validiert Ein- und Ausgaben strikt, Fehler stoppen den Prozess. Änderungen führen deterministisch zu einem Neu-Build. Die Qualitätssicherung umfasst Schematests, Linkchecks, Validierungen von HTML, CSS und XML sowie Prüfsummen.

Sichtbarkeitssteuerung geschieht vor dem Export oder in Ableitungsschritten, beispielsweise durch serverseitiges Entfernen sensibler Felder, die Bereitstellung maskierter Bildvarianten, alternative Einstiegspfade mit Content-Notes oder gestufte Editionen mit klarer Kennzeichnung. Die Referenzausgabe selbst bleibt statisch.

Offene, verlustfreie und nachnutzbare Formate sind zu verwenden, sofern keine Schutzgründe dagegensprechen. Dazu gehören PDF/A für formatierten Text, XML / HTML / MD / TXT für strukturierten Text, reines TXT für Plaintext, TIFF oder PNG für Bilder, SVG für Vektoren und CSV für Tabellen. Durchgängig wird UTF-8 als Zeichenkodierung empfohlen.

Eine README.md-Datei im Wurzelverzeichnis erläutert Zweck, Geltungsbereich, Struktur, Rollen, Benennungsregeln, Lizenzen, Rechte und Restriktionen, Versionierung und Releaseschema. Zusätzlich sind Datenmodell, kontrollierte Vokabulare und zulässige Werte als statische Referenzdokumente zu hinterlegen.

Exportschnittstellen sind durch API-Keys und Rate-Limits abzusichern, und sie sollen ausschliesslich freigegebene Felder enthalten. Zur Reproduktion wird eine Datenbeilage im Paket geliefert und versioniert. Falls Plattformfunktionen wie LIDO erforderlich sind, bleiben diese auf die vorgelagerte Produktionsschicht beschränkt; das Endprodukt wird stets statisch ausgespielt. Nutzer*innen müssen die Wahl haben, sensible Inhalte durch ein bewusstes Opt-in mit klaren Hinweisen zu aktivieren.

Checkliste technische Strategien

- [ ] Statischer Webauftritt ohne serverseitige Abhängigkeiten definiert
- [ ] Build- und Validierungsprozess unter Versionskontrolle implementiert
- [ ] Sichtbarkeitssteuerung in Export- oder Ableitungsschritten umgesetzt
- [ ] Offene, verlustfreie Formate mit UTF-8-Kodierung verwendet
- [ ] README.md und Referenzdokumente bereitgestellt und versioniert
- [ ] API-Exporte abgesichert und reproduzierbare Datenbeilage integriert
- [ ] Opt-in-Mechanismus für sensible Inhalte verfügbar

5.4 Transparente Dokumentation

Die Dokumentation bleibt öffentlich zugänglich und macht Entscheidungsprozesse, Mitbestimmungen, Konflikte, Lücken und Versionen sichtbar. Beispielhaft dient die Dokumentation der Stadtgeschichte Basel. Jede Edition sollte statt eines “rolling release” fixiert vorliegen, ein Build-Datum und eine Versionsbezeichnung tragen und eine Zitierempfehlung enthalten. Persistente URIs müssen stabil bleiben, Umzüge sind durch Weiterleitungen abzusichern. Jede Edition wird als fixiertes Paket archiviert, das Webartefakte, Datenkopien, Dokumentationen, Prüfsummen, Lizenzdateien und optional eine CITATION.cff-Datei umfasst.

Offene Prinzipien werden auf der Projektseite explizit sichtbar gemacht. Dies schliesst eine Verlinkung auf die *Endings Principles* sowie die Veröffentlichung von Compliance-Checklisten und Diagnosetools ein, welche die Qualitätssicherung und die Abschlussreife dokumentieren.

Checkliste Dokumentation

- [ ] Publikationsdokumentation online und öffentlich zugänglich
- [ ] Editionen mit Datum, Version und Zitierempfehlung publiziert
- [ ] Persistente URIs stabil, Umzüge mit Weiterleitungen gesichert
- [ ] Jede Edition als Paket mit Prüfsummen, Daten und Dokumentation archiviert

6. Nachnutzung und Wiederverwendung

Die veröffentlichten Daten treten nach ihrer Bereitstellung in einen neuen Lebenszyklus ein. Sie können von der wissenschaftlichen Community, Gedächtnisinstitutionen oder zivilgesellschaftlichen Akteur*innen recherchiert, zitiert, kombiniert und für neue Fragestellungen genutzt werden. Damit Nachnutzung nicht zur unkontrollierten Reproduktion von Diskriminierungen oder Fehlinterpretationen führt, sind klare Leitprinzipien zu formulieren. Die FAIR-Prinzipien bilden dabei die Grundlage, müssen jedoch im Sinne der CARE-Prinzipien ergänzt werden. Es gilt zu beachten, dass Nachnutzung nie neutral erfolgt: Sie schafft neue Kontexte und Bedeutungen. Ein expliziter Governance-Rahmen, der Verantwortlichkeiten und Grenzen definiert, reduziert das Risiko von schädlichen Re-Use-Szenarien.

6.1 Interoperabilität erweitern

Damit Daten dauerhaft anschlussfähig bleiben, ist ihre strukturelle und semantische Interoperabilität sicherzustellen. Praktisch bedeutet dies, dass Daten nicht nur in offenen Formaten vorliegen, sondern auch durch klar beschriebene Schnittstellen zugänglich sind. Neben einfachen Exporten in standardisierten Formaten (zum Beispiel

CSV oder XML) können Linked-Data-Ansätze mit RDF oder JSON-LD den Zugang erweitern. Austauschprotokolle wie OAI-PMH oder IIIF erleichtern den Bezug durch Drittinstitutionen. Persistente Identifikatoren – etwa DOIs für Datensätze oder ORCID-IDs für Autor*innen – gewährleisten Nachvollziehbarkeit. Mappings zwischen Standards (zum Beispiel Dublin Core, LIDO oder EDM) sollten versioniert und dokumentiert vorliegen, um die Transformation transparent zu machen.

6.2 Zitation und Provenienz

Für die Nachnutzung ist ein klarer Zitierleitfaden erforderlich. Neben der Angabe von DOIs oder anderen Identifikatoren empfiehlt sich die Bereitstellung maschinenlesbarer Zitationsinformationen, etwa in einer CITATION.cff-Datei. Provenienzangaben müssen nicht nur die Herkunft der Daten dokumentieren, sondern auch Unsicherheiten sichtbar machen – beispielsweise durch Zeitangaben im EDTF-Format oder Vertrauenswerte bei automatisierten Prozessen. So bleibt nachvollziehbar, welche Transformationen ein Datensatz durchlaufen hat.

6.3 Nutzungsbedingungen und Lizenzen

Eine transparente Lizenzpolitik ist unabdingbar. Neben den gängigen Creative-Commons-Lizenzen sollten bei kulturell sensiblen Inhalten CARE-orientierte Einschränkungen vorgesehen werden. Hier bieten sich etwa TK-Labels an, die kulturelle Rechte und Einschränkungen sichtbar machen. Insbesondere bei historischen oder diskriminierenden Inhalten muss reflektiert werden, welche Formen der Nachnutzung ausgeschlossen oder eingeschränkt werden sollen, ohne den wissenschaftlichen Diskurs zu behindern.

6.4 Maschinenlesbarer Zugang

Die Bereitstellung von Schnittstellen (APIs) oder statischen Datenpaketen erleichtert die Weiterverwendung. Dabei sind klare Nutzungsregeln, Fair-Use-Begrenzungen und versionierte Snapshots wichtig, um Konsistenz zu gewährleisten. Nutzer*innen müssen erkennen können, ob sie mit einer stabilen Version oder einem fortlaufend veränderlichen Datensatz arbeiten.

6.5 Feedback und Korrekturschleifen

Nachnutzung sollte nicht als einseitiger Prozess verstanden werden. Rückmeldungen von Nutzer*innen müssen systematisch erfasst, geprüft und dokumentiert werden. Issue-Tracker, Feedback-Formulare oder moderierte Mailinglisten können hier als Kanäle dienen. Eine dokumentierte Änderungspolitik (Changelog, Deprecation-Policy) erhöht die Transparenz. Insbesondere Communities, die von diskriminierenden Inhalten betroffen sind, sollten in die Nachbesserung und Ko-Kuration eingebunden werden.

Checkliste Nachnutzung und Wiederverwendung

- [ ] Linked-Data-Exporte, PIDs und Crosswalks verfügbar
- [ ] Zitations- und Lizenzrichtlinien veröffentlicht
- [ ] API oder Dumps versioniert zugänglich
- [ ] Feedback-Prozess etabliert, Nutzung dokumentiert

7. Archivierung und Löschung

Nach Abschluss des Forschungsdatenzyklus folgt entweder die dauerhafte Archivierung oder – wo rechtlich oder ethisch geboten – die Löschung von Daten. Beide Prozesse erfordern institutionalisierte Strategien, die auf internationalen Standards aufbauen, rechtliche Rahmenbedingungen einhalten und eine transparente Governance etablieren. Das OAIS-Referenzmodell bietet einen Orientierungsrahmen für die Organisation der digitalen Langzeitarchivierung, während Leitlinien wie die NDSA Preservation Levels konkrete Schwellenwerte und Anforderungen formulieren.

7.1 Langzeitarchivierung konkret

Eine nachhaltige Langzeitarchivierung setzt auf offene, standardisierte Verpackungs- und Dokumentationsformate. Containerformate erlauben es, Datenpakete mit Metadaten, Prüfsummen und Kontextinformationen zu bündeln. Standards wie METS oder PREMIS sichern die Nachvollziehbarkeit von Provenienzen und Veränderungen. Technisch sind Integritätsprüfungen mit Prüfsummen und Redundanzstrategien (zum Beispiel die 3-2-1-Regel: drei Kopien, zwei Medientypen, ein Standort ausserhalb) zentral. Neben Formaten wie PDF/A, TIFF, CSV oder XML sind auch Migrationen in neue Standards einzuplanen, um die Lesbarkeit langfristig zu gewährleisten. Ergänzend müssen Build-Informationen, Software-Umgebungen oder Containerabbilder dokumentiert werden, falls Daten nur mit spezifischen Werkzeugen reproduzierbar bleiben.

7.2 Löschung, Rückgabe und Takedown

Nicht alle Daten können oder dürfen unbegrenzt aufbewahrt werden. Rechtliche Löschfristen, datenschutzrechtliche Ansprüche oder Rückgaberechte von Datensubjekten erfordern kontrollierte Verfahren. Löschungen sollten nicht stillschweigend erfolgen, sondern durch sogenannte Tombstone-Seiten sichtbar bleiben, die auf die frühere Existenz des Datensatzes hinweisen. Wo möglich, sollte Löschung durch selektive Redaktion (zum Beispiel Anonymisierung oder Einschränkung einzelner Felder) ersetzt werden, um die Kontextintegrität zu wahren. Ein formelles Deletions-Audit sowie nachvollziehbare Änderungsprotokolle sichern die Transparenz. Für sensible Kulturdaten kann auch eine Rückgabe an betroffene Communities (Repatriierung) erforderlich sein.

7.3 Zugriff und Schutz

Archivierte Daten müssen auch in Zukunft geschützt und verantwortungsvoll zugänglich bleiben. Zugriffsstufen und Embargoregelungen steuern, wer Daten unter welchen Bedingungen einsehen kann. Verschlüsselung während der Speicherung und Übertragung ist ebenso obligatorisch wie eine sichere Schlüsselverwaltung. Audit-Logs machen Zugriffe nachvollziehbar, während Notfallpläne sicherstellen, dass Daten auch bei Systemausfällen oder institutionellen Veränderungen erhalten bleiben.

Checkliste Archivierung und Löschung

- [ ] Archivpakete mit dokumentierten Provenienzen und Prüfsummen erstellt
- [ ] Lösch- und Rückgabeprozesse rechtlich und organisatorisch festgelegt
- [ ] Tombstones und Änderungsprotokolle veröffentlicht

- [ ] Zugriffskontrollen, Verschlüsselung und Logs aktiv
- [ ] Redundanz- und Notfallpläne etabliert, Kosten langfristig kalkuliert

Handlungsleitende Prinzipien

1. Transparenz als Infrastrukturprinzip

Jeder Arbeitsablauf einer Pipeline erhält einen öffentlich zugänglichen Entscheidungs- und Änderungslog. Dokumentiert werden: Erhebungsdesign, Auswahlrationale, Fehlerraten, Mapping-Regeln sowie abgelehnte Alternativen (versioniert).

Für Datensätze und Modelle gelten strukturierte Begleitdokumente als Standard: Datasheets und Model Cards (Steckbriefe für Daten/Modelle). Sie enthalten gruppenspezifische Qualitätsmetriken, bekannte Lücken und Nutzungseinschränkungen. Diese Dokumente sind zitierfähig und mit persistenten Identifikatoren in Repositorien verankert.

2. Varianten- und Mehrsprachigkeitsfähigkeit im Metadatenmodell

Konzepte und Namen werden als mehrsprachige, relationale Entitäten modelliert (zum Beispiel SKOS):

- `prefLabel` = bevorzugte Bezeichnung
- `altLabel` = alternative Bezeichnung
- `hiddenLabel` = versteckte Schreibweisen
- `exactMatch/closeMatch` = präzise oder nahe Entsprechungen
- Provenienz = dokumentierte Herkunft

Identitäten und Rollen erhalten kontrolliert-offene Felder statt binärer Pflichtangaben. Selbstbezeichnungen und Schreibvarianten gelten als gleichwertige, abfragbare Identifikatoren. Unsicherheiten werden in eigenen Feldern mit standardisierten, maschinenlesbaren Qualifikatoren vermerkt.

3. Partizipative Kuratierung und Governance

Kuratierung erfolgt gemeinsam mit betroffenen Communities. Dies umfasst:

- Verfahren für Einspruch, Korrektur und Takedown
- klare Zuständigkeiten und angemessene Vergütung
- Vereinbarungen zur Wissenssouveränität (FAIR / CARE)

Für sensible Bestände werden abgestufte Zugangsmodelle (differenzierte Rechtevergabe) und Schutzkennzeichnungen eingesetzt. Alle Entscheidungen sind befristet und revidierbar.

4. Ausgleichende Technik entlang der gesamten Prozesskette

Die Auswahl folgt stratifizierten Plänen mit *Equity-Buckets* (Fairness-Schichten), um unterrepräsentierte Gruppen systematisch einzubeziehen.

- **Digitalisierung und Erkennung:** gezielte Feinjustierung für schwach performende Segmente; Fehlerbilanzen werden nach Schrift, Sprache, Medium und Gruppe veröffentlicht.
- **Retrieval-Systeme:** Einsatz von Query-Expansion (Anfrageerweiterung), Cross-Lingual IR (mehrsprachiges Informationsretrieval) und Re-Ranking (Neuanordnung der Treffer), um Recall-Differenzen zwischen Gruppen zu reduzieren. Ranking-Entscheide werden erklärt.

5. Analytische Robustheit als Standardpraxis

Studien verpflichten sich zu:

- Vorabdefinition von Annahmen
- Sensitivitätsanalysen (Reaktion auf kleine Änderungen)
- Tests alternativer Operationalisierungen
- Gewichtung bekannter Erhebungs- und Selektionsfehler

Fehlende Daten werden transparent behandelt, zum Beispiel durch Mehrfachimputation mit Diagnose der MAR/MNAR-Annahmen (zufälliges/nicht-zufälliges Fehlen). Für kausale Aussagen sind kontrafaktische Szenarien, Instrumentvariablen oder natürliche Experimente zu prüfen. Alle Limitierungen sind offen zu deklarieren.

6. Gerechtigkeitsmetriken und Monitoring

Erfolg wird anhand equity-orientierter Kennzahlen gemessen:

- gruppenspezifischer Recall / Precision (Trefferquote/Vollständigkeit und Genauigkeit)
- Fehlerratenparität (Vergleichbarkeit der Fehlerquoten)
- Exposure-Anteile im Ranking (Sichtbarkeit in Trefferlisten)
- Bearbeitungszeiten für Korrekturen
- Community-Zufriedenheit

Diese Metriken fliessen in kontinuierliche Audits ein. Abweichungen lösen definierte Korrekturpfade aus. Ein öffentliches Issue-Tracking ermöglicht externe Prüfung und Nachsteuerung.

7. Reproduzierbarkeit und digitale Nachhaltigkeit

Workflows sind:

- containerisiert (standardisierte, portable Software-Umgebungen)
- daten- und modellseitig versioniert
- in offenen, langfristigen Formaten und Schnittstellen zugänglich

Abhängigkeiten werden minimiert. Rechenaufwände werden bilanziert und, wo möglich, durch Minimal Computing (ressourcenschonende Verfahren) ersetzt. Langzeitarchivierung und Wiederverwendbarkeit sind von Beginn an Designkriterien.

Zusammenfassung

Handlungsleitend ist ein zyklisches Vorgehen:

1. **Planen** (Ziele, Risiken, Metriken)
2. **Umsetzen** (bei der Annotation mit einer diversen Auswahl an Beispielen beginnen)
3. **Prüfen** (Audits, Robustheit, Community-Feedback)
4. **Anpassen** (Governance-Entscheide revidieren, restliche Objekte auszeichnen)

So werden Diskriminierungen nicht nur benannt, sondern in überprüfbaren Schritten reduziert. Zugleich wird die Nachvollziehbarkeit und Integrität der Forschung gestärkt.

Methodische Grenzen einer diskriminierungssensiblen Praxis

Historische Daten sind in ihrer Entstehung, Überlieferung und Digitalisierung selektiv. Retrodigitalisierte Bestände sind geprägt von den Normen ihrer Entstehungszeit, den Kriterien archivarischer Auswahl sowie den technischen Entscheidungen heutiger Digitalisierungsprozesse. Eine vollständige Überwindung dieser Vorprägungen ist nicht erreichbar; realistisch ist nur, sie sichtbar zu machen und mögliche und tatsächliche Folgen aufzuzeigen. Damit verschiebt sich der Anspruch von vermeintlicher Neutralität zu expliziter Reflexivität: Provenienzangaben, Auswahlkriterien und Erschliessungsentscheidungen werden systematisch dokumentiert, versioniert und in Analysen berücksichtigt.

Messbarkeit bleibt begrenzt, weil zentrale Grössen nur über Proxy-Variablen zugänglich sind. OCR / HTR -Fehler, Normalisierungen und Kategorienschnitte erzeugen Verzerrungen, die nicht homogen über Sprachen, Schriften und Gruppen wirken. Operationalisierungen sollten deshalb mit Fehlermodellen verknüpft werden, die Unsicherheiten quantifizieren; Zum Beispiel Konfidenzintervalle für CER / WER und gruppenspezifische Recall / Precision . Fairnessbegriffe sind zudem konkurrierend: Parität in der Trefferquote, Gleichheit der Fehlerraten und Nutzenmaximierung lassen sich oft nicht gleichzeitig erreichen. Solche Zielkonflikte sind offen zu benennen und als Governance-Entscheidungen zu verantworten.

Die verschiedenen Formen von Verzerrung, direkt, indirekt, strukturell oder institutionell, wirken nicht isoliert, sondern greifen ineinander. Entscheidungen im Digitalisierungsprozess, etwa eine stratifizierte Auswahl, prägen dadurch unmittelbar spätere Suchergebnisse und deren Interpretation. Hinzu treten klassische Repräsentativitätsprobleme: Korpusgrenzen, Überlieferungs- und Auswahlverzerrungen sowie zeitliche Verschiebungen (“dataset shift”) mindern die Übertragbarkeit von Befunden. Kausale Schlussfolgerungen aus solchen Beobachtungsdaten sind deshalb nur unter starken Zusatzannahmen belastbar. Potenzielle Störfaktoren wie Confounding, Selektions- oder Messfehler sind als zentrale Hypothesen zu behandeln – nicht als nachträgliche Randbemerkung.

Schliesslich sind rechtlich-ethische und ökologische Grenzen mitzudenken. Re-Identifikationsrisiken steigen mit Verknüpfbarkeit; CARE -Prinzipien und Wissenssouveränität kollidieren mit radikal offener Nachnutzung. Digitale Nachhaltigkeit verlangt formatarme, langlebige Lösungen und eine Energie- und Speicherökonomie, die den wissenschaftlichen Nutzen gegen ökologische Kosten abwägt. Der methodische Mindeststandard ist deshalb “interpretative Bescheidenheit”: Ergebnisse werden als bedingt, kontextualisiert und replizierbar ausgewiesen.

Theorie: Schlüsselbegriffe und Konzepte

Die Entwicklung und Anwendung einer diskriminierungssensiblen Metadatenpraxis setzt ein gemeinsames Verständnis zentraler Begriffe und Konzepte voraus. Dieses Kapitel definiert und kontextualisiert die theoretischen und technischen Schlüsselbegriffe, die den Rahmen des Handbuchs bilden. Im Fokus stehen Begriffe wie **Normativität**, **Diskriminierung**, **Bias** und **Oppression**, die sowohl aus geistes- und kulturwissenschaftlicher als auch aus informations-, bibliotheks- und archivwissenschaftlicher Perspektive diskutiert werden. Neben diesen kritischen Grundbegriffen werden auch technische Termini wie **Metadatenstandard**, **Normdaten** und **Datenwertstandard** eingeführt, um die Brücke zwischen inhaltlicher Reflexion und technischer Implementierung zu schlagen.

Dieses Kapitel bestimmt die Begriffe so, dass sie sowohl analytisch tragfähig als auch praktisch operationalisierbar sind. Die Auswahl und Definition der Begriffe orientiert sich an internationalen Menschenrechtsstandards (humanrights.ch, o. J.) und interdisziplinären Ansätzen wie Data Feminism, Data Justice und Critical Data Studies (Mehrabi u. a. 2021; Loukissas 2019; D'Ignazio und Klein 2020). Zugleich wird die Debatte gezielt in deutschsprachige und europäische Fachkontexte rückgebunden, etwa durch queerfeministische und archivkritische Beiträge (Sparber 2016; Gruber 2022; Rottmann 2023) sowie durch institutionsbezogene Analysen aus dem europäischen GLAM-Bereich (Kaiser u. a. 2023; Baroncini u. a. 2025). Damit bildet dieses Kapitel die konzeptionelle und terminologische Grundlage für alle weiteren Analysen und Empfehlungen im Handbuch.

Diskriminierung in und durch Daten

Daten sind nicht neutral (Normativität)

WIKIDATA Search Wikidata Search English

Crimea (Q7835)

Item Discussion Read View history Tools

Eastern European peninsula, in the Black Sea and Sea of Azov, disputed between Ukraine (de jure) and Russia (de facto)
Kirim | Qırım | Kimmería | Taurikê | Crimean Peninsula | Crim. | Crimean P. | Taurike | Taurica | Kimmeria | Cimmeria

In more languages

Language	Label	Description	Also known as
default for all languages	No label defined	–	
English	Crimea	Eastern European peninsula, in the Black Sea and Sea of Azov, disputed between Ukraine (de jure) and Russia (de facto)	Kirim Qırım Kimmería Taurikê Crimean Peninsula Crim. Crimean P. Taurike Taurica Kimmeria Cimmeria
German	Krim	Halbinsel im nördlichen Schwarzen Meer	Tauris
Alemannic	Krim	No description defined	
French	Crimée	péninsule d'Europe de l'Est, entre la mer Noire et la mer d'Azov, disputée entre l'Ukraine (de jure) et la Russie (de facto)	

All entered languages

Statements

instance of	peninsula 1 reference

Screenshot des Wikidata-Eintrags zu Crimea (Q7835), der die Halbinsel Krim beschreibt. Angezeigt werden Sprachvarianten (unter anderem Englisch, Deutsch, Französisch), alternative Bezeichnungen sowie die Beschreibung als "Eastern European peninsula, in the Black Sea and Sea of Azov, disputed between Ukraine (de jure) and Russia (de facto)".

Daten und Metadaten erscheinen auf den ersten Blick als objektive Repräsentationen der Wirklichkeit, doch sind sie stets in historisch gewachsene Machtverhältnisse und normative Ordnungen eingebettet. Das Beispiel der Krim in Wikidata Q7835 und die beigefügte Karte der Halbinsel illustrieren anschaulich, wie scheinbar neutrale Darstellungen in Metadaten inhärent Stellung beziehen und politische Konfliktlagen spiegeln.

Politische Karte der Krim mit Städten, Verkehrswegen und geografischen Bezeichnungen. Die Karte zeigt die Halbinsel im Schwarzen Meer, angrenzend an die Ukraine und Russland.

Die Karte präsentiert die Krim als “Autonome Republik Krim” und zeigt damit explizit eine völkerrechtlich orientierte, pro ukrainische Perspektive: Die Benennung und farbliche Abgrenzung der Krim als Teil der Ukraine (“Oblast”, “Autonome Republik”) betont den de-jure-Status nach ukrainischem Recht und den internationalen Menschenrechtsstandards. Es werden zentrale Orte wie Sewastopol oder Kertsch mit ihrer ukrainischen Transkription aufgeführt. Die Grenze zu Russland ist klar als Staatsgrenze markiert. In der Detaildarstellung (oben rechts) wird die Krim deutlich der Ukraine zugeordnet, wodurch der umstrittene völkerrechtliche Status als Teil Russlands nicht gleichwertig visualisiert wird.

Dieses kartografische Beispiel steht paradigmatisch für die normativen Setzungen, die auch digital kodierten Datenstrukturen wie Wikidata zugrunde liegen. So wird die Krim dort sowohl als Teil der Ukraine (“de jure”) als auch Russlands (“de facto”) geführt, jedoch übernehmen viele verknüpfte Objekte wie Städte oder administrative Einheiten diese Mehrdeutigkeit nicht konsistent oder bilden einzig die ukrainische Perspektive ab. Die Karte operiert dabei mit einem klaren Framing zugunsten der ukrainischen Souveränität und blendet alternative Klassifikationsoptionen, wie etwa die Bezeichnung “Republik Krim” als Teil Russlands, weitgehend aus.

Dies verdeutlicht, dass jede Form von Daten- und Metadatenmodellierung – sei sie visuell, textuell oder strukturell – auf normativen Entscheidungen beruht, die oft unsichtbar bleiben. Die Auswahl von Bezeichnungen, die Sichtbarkeit von Grenzen oder die Hierarchisierung von Souveränitätsansprüchen spiegeln und stabilisieren bestehende Machtasymmetrien und wirken als “invisible hand of classification” (Bowker und Star 1999). Für Nutzer*innen entsteht der Eindruck technischer Neutralität, obwohl sowohl die Karte als auch digitale Datenbanken politische Aushandlungsprozesse und Interessen materialisieren. Mit technischer Neutralität ist hier gemeint, dass Entscheidungen in Interface, Datenmodell und Standardisierung als rein technische Notwendigkeiten erscheinen, obwohl sie normative Setzungen enthalten.

Eine diskriminierungssensible Metadatenpraxis setzt daher an der Schnittstelle von Technik und Gesellschaft an: Sie verlangt die konsequente Offenlegung und Reflexion der eigenen Klassifikationsschemata, Entscheidungslogiken und Datenstrukturen. In der Umsetzung heisst das: transparente Dokumentation von Modellierungsentscheidungen (Schema, Version, Geltungsbereich), parallele Felder für Mehrperspektivität (*de jure*/*de facto*, Selbst- und Fremdbezeichnungen), überprüfbare Quellen- und Provenienzangaben sowie die explizite, maschinenlesbare Markierung von Konflikten und Unsicherheiten in den Metadaten (Status, Zeitraum, räumliche Gültigkeit). Die Beschreibung durch Metadaten und ihre Repräsentation in digitalen Datenbanken sind bewusst gestaltete, nachvollziehbare Praktiken, die nie lediglich “abbilden”, sondern normativ wirken.

Direkte Diskriminierung

Die Begriffsbestimmung von Diskriminierung erfolgt in der Einleitung (Kapitel 2). In diesem Kapitel operationalisieren wir direkte Diskriminierung für die Metadatenpraxis: Eine Regel, Entscheidung oder Handlung ist direkt diskriminierend, wenn sie Personen aufgrund eines geschützten Merkmals ungleich behandelt und diese Ungleichbehandlung nicht durch einen legitimen, verhältnismässigen Zweck gedeckt ist. Im Fokus stehen Mechanismen, Indikatoren und Gegenmassnahmen.

Kernelemente

- *Unmittelbare Ungleichbehandlung:* Eine Regel oder Praxis behandelt Personen explizit unterschiedlich.
- *Merkmalsbezug:* Die Ungleichbehandlung knüpft an ein tatsächliches oder zugeschriebenes (geschütztes) Merkmal an.
- *Fehlende Rechtfertigung:* Es gibt keinen legitimen, verhältnismässigen Zweck, der die Ungleichbehandlung trägt.

Beispiele

- **Zugangsbeschränkungen in Archiven**
  - *Mechanismus:* Normative Ausschlussregeln nach Geschlecht, Konfession oder Stand.
  - *Effekt:* Systematische Tilgung von Stimmen aus der historischen Überlieferung.

    - *Indikatoren:* Explizite Zutrittsordnungen, fehlende Nutzungsprotokolle für ausgeschlossene Gruppen.
    - *Gegenmassnahmen:* Retrospektive Dokumentation von Ausschlüssen, Priorisierung von Erschliessungslücken, inklusionsorientierte Benutzungsordnungen.
- **Staatliche Register**
    - *Mechanismus:* Kategorisierung nach “Rasse”, “Stamm”, “Religion” zur Hierarchisierung.
    - *Effekt:* Ungleichbehandlung durch Verwaltung und Recht.
    - *Indikatoren:* Kategorien mit sanktions- oder leistungsrelevanter Wirkung.
    - *Gegenmassnahmen:* Historisierung und Kontextualisierung problematischer Kategorien, Schutzkennzeichnungen, restriktive Nachnutzungsbedingungen.
- **Berufsrollen in amtlichen Dokumenten**
    - *Mechanismus:* Erfassung von Frauen nur relational (“Frau des Schmieds”) statt als eigenständige Akteurinnen.
    - *Effekt:* Unsichtbarmachung ökonomischer Tätigkeit.
    - *Indikatoren:* Geringer Anteil eigenständiger Berufsangaben bei Frauen.
    - *Gegenmassnahmen:* Nachträgliche Normalisierung, alternatives Namens- und Rollenmodell, Varianten als gleichwertige Identifikatoren pflegen.
- **Diskriminierende Beschreibungen und Zuschreibungen**
    - *Mechanismus:* Abwertende oder stereotypisierende Formulierungen in Katalogtexten, Abstracts und Schlagwörtern sowie unkommentierte Übernahme diskriminierender Fremdbezeichnungen aus Quellen oder Altmetadaten.
    - *Effekt:* Stigmatisierung und Verletzung; Reproduktion historischer Hierarchien; Verzerrung von Deutung und Auffindbarkeit.
    - *Indikatoren:* Wertende Adjektive als Default, Täterperspektive ohne Kontext, problematische Begriffe als bevorzugte Benennung, fehlende Kontext- und Provenienzfelder.
    - *Gegenmassnahmen:* Trennung von Quellzitat und Beschreibungsebene; Kontextualisierung über Hinweise/ Content-Notes ; dokumentierte Redaktionsregeln, Versionierung und Feedbackwege (siehe Kapitel 4.1.8).

Indirekte Diskriminierung

Indirekte Diskriminierung liegt vor, wenn formal neutrale Kriterien, Methoden oder Regelungen zur Erhebung, Auswahl, Beschreibung oder Interpretation historischer Daten in ihrer Wirkung systematisch bestimmte Gruppen benachteiligen. Das betrifft zum Beispiel unterdokumentierte Gruppen oder Praktiken, die durch etablierte Routineprozesse weiter marginalisiert werden.

Kernelemente

- *Formale Neutralität:* Kriterien oder Prozesse wirken auf den ersten Blick allgemeingültig.
- *Ungleichwirkung:* In der Anwendung entstehen systematische Nachteile für bestimmte Gruppen (Disparate Impact).
- *Alternative Gestaltung:* Es gäbe weniger benachteiligende, fachlich gleichwertige Verfahren oder Ausnahmen.

Beispiele

- **Alphabetische Namensregister**
    - *Mechanismus:* Ordnung nach Familiennamen mit Relationen über Haushaltsvorstände; Witwen unter Namen des Ehemannes.
    - *Effekt:* Systematische Unauffindbarkeit von Frauen, Kindern und Menschen aus Kulturen ohne Familiennamen.
    - *Indikatoren:* Hoher Anteil von “siehe”-Verweisen statt eigener Einträge.
    - *Gegenmassnahmen:* Sekundärregister nach Vornamen/Rollen, relationale Verknüpfungen, Namensvarianten als Primärschlüssel.
- **Sprachliche Dokumentation**
    - *Mechanismus:* Dominanz von Amtssprachen (Latein, Französisch, Hochdeutsch) in der Überlieferung.
    - *Effekt:* Geringe Sichtbarkeit von Minderheitensprachen und -praktiken.
    - *Indikatoren:* Anteil nicht erfasster beziehungsweise nicht indexierter Sprachvarietäten.
    - *Gegenmassnahmen:* Mehrsprachige Erschliessung, Community-basierte Übersetzungen, SKOS -Labels pro Sprachvarietät.
- **Digitale Volltextsuche**
    - *Mechanismus:* Auswahl- und Digitalisierungsbias; OCR / HTR bevorzugt standardisierte Drucke.
    - *Effekt:* Unterrepräsentation von Frauen, Arbeiter*innen, Handschriften.
    - *Indikatoren:* CER / WER nach Schrift/Medium; Recall -Differenzen in Korpora.
    - *Gegenmassnahmen:* Stratifizierte Digitalisierung, gezieltes Fine-Tuning , ausgleichendes Ranking , Query-Expansion .
- **Zensuskategorien**
    - *Mechanismus:* Erfassung nur von Haushaltsvorständen.
    - *Effekt:* Statistische Unsichtbarkeit von Frauen, Kindern und anderen Haushaltsangehörigen.
    - *Indikatoren:* Fehlende Individualdatensätze für Nicht-Vorstände.
    - *Gegenmassnahmen:* Rekonstruktion von Haushalten, Metadaten zu Erhebungsdesigns, methodische Gewichtung in Analysen.
- **Unzureichende Findmittel und Zugangsbarrieren**
    - *Mechanismus:* Formal “neutrale” Find- und Zugriffspfade (Fachsprache, fehlende Barrierefreiheit, technische Hürden, unvollständige Erschliessung) begünstigen geübte Nutzergruppen und etablierte Forschungspraktiken.
    - *Effekt:* Bestände und Perspektiven marginalisierter Gruppen bleiben praktisch unzugänglich; Nachnutzung konzentriert sich auf gut erschlossene, kanonisierte Quellen.
    - *Indikatoren:* Hoher Anteil von 0-Treffer-Suchen, geringe Nutzung bestimmter Bestände, fehlende barrierefreie Metadaten/Interfaces.
    - *Gegenmassnahmen:* Barrierearme Sprache, mehrere Einstiegspfade, bessere Findmittel und Dokumentation von Lücken, nutzerzentrierte Tests mit diversen Gruppen.

Strukturelle Diskriminierung

Strukturelle Diskriminierung bezeichnet Benachteiligungen, die in etablierten Praktiken der Sammlung, Dokumentation, Bewahrung und Zugänglichmachung verankert sind. Ordnungen des Archivierens, Katalogisierens und Kuratierens reproduzieren häufig patriarchale, koloniale und heteronormative Sichtweisen. Diese wirken sich auf Auswahlprozesse, die Sprache der Erschliessung und institutionelle Routinen aus. Zahlreiche digitale Infrastrukturen orientieren Sprache, Usability und Standards primär an westlichen Forschungstraditionen; indigene und nicht-westliche Perspektiven bleiben dadurch marginalisiert.

Kernelemente

- *Einbettung in Routinen und Infrastrukturen:* Benachteiligungen entstehen nicht primär durch einzelne Akte, sondern durch gewachsene Standards, Prozesse und Klassifikationen.
- *Kumulativität und Pfadabhängigkeit:* Viele kleine Entscheidungen verstärken sich über Zeit und werden schwer revidierbar.
- *Unsichtbarkeit als Normalform:* Ungleichheiten erscheinen als "Datenlage" oder "Sachzwang" und werden dadurch stabilisiert.

Beispiele

- **Überlieferungs- und Erschliessungslücken**
  - *Mechanismus:* Selektive Sammlung, Archivierung und Beschreibung; marginalisierte Gruppen werden als "nicht relevant" klassifiziert, nur am Rand dokumentiert oder in Sammelkategorien aufgelöst.
  - *Effekt:* Aktive oder fortgeschriebene Unsichtbarkeit in Archiven und Geschichtsschreibung; die "Datenlage" erscheint neutral, ist aber Ergebnis von Machtverhältnissen.
  - *Indikatoren:* Leere oder stark ungleich verteilte Facetten, Sammelrubriken ("miscellaneous"), geringe Granularität bei marginalisierten Akteur*innen, systematische Lücken nach Region/Status.
  - *Gegenmassnahmen:* Gap-Analysen, partizipative Sammlungs- und Erschliessungsstrategien, dokumentierte Erschliessungslücken, gezielte Erschliessungs- und Redescription-Projekte.
- **Eurozentrische Normdaten und kontrollierte Vokabulare**
  - *Mechanismus:* Normdaten privilegieren westliche Taxonomien; indigene Konzepte werden "gemappt".
  - *Effekt:* Semantische Asymmetrien, Fehlklassifikation, schlechteres Retrieval.
  - *Indikatoren:* Hoher Anteil unscharfer Schlagwörter; geringe Mehrsprachigkeit in `skos:prefLabel/altLabel`.
  - *Gegenmassnahmen:* Community-kuratiertes Vokabular, Mehrsprachigkeit, präzise `skos:exactMatch/closeMatch`, dokumentierte Provenienz.
- **Binäre Personen- und Geschlechtermodelle in Metadatenstandards**
  - *Mechanismus:* Pflichtfelder erzwingen binäre Geschlechter, patronymische Hauptformen.
  - *Effekt:* Unsichtbarkeit nicht-binärer Identitäten und eigenständiger Rollen.
  - *Indikatoren:* Anteil Datensätze ohne Felder für Selbstbezeichnung; Normalisierungszwang in Normdaten.
  - *Gegenmassnahmen:* Kontrolliert-offene Felder (`genderIdentity`, `role`), Varianten als gleichwertige Identifikatoren, versionsgeführte Entscheidungsprotokolle.
- **Digitale Pipeline-Bias: Auswahl, OCR/HTR und Ranking**
  - *Mechanismus:* Kanonzentrierte Auswahl; Trainingsdaten für dominante Schriften/Sprachen; Indexgewichte bevorzugen gut erkannten Text.
  - *Effekt:* Höhere Fehlerraten und schlechtere Auffindbarkeit für Minderheitensprachen und Handschriften.
  - *Indikatoren:* CER/WER nach Schrift/Varietät; Recall-Differenzen; Abdeckung pro Segment.
  - *Gegenmassnahmen:* Stratifizierte Auswahlpläne, publizierte Fehlerbilanzen, Fine-Tuning für unterrepräsentierte Schriften, re-ranking, CLIR.
- **Kanon- und Metrikgetriebene implizite Priorisierung**
  - *Mechanismus:* Mittelvergabe nach Nutzung und Zitation.
  - *Effekt:* Sichtbarkeitsspirale zugunsten von Eliten und Zentren.
  - *Indikatoren:* Budget- und Seitenzahlen pro Gruppe/Region; "Anfragen vs. unerschlossene Bestände".
  - *Gegenmassnahmen:* Equity-Buckets, Social-Impact-KPIs, kooperative Digitalisierung, transparente Trade-offs.

Institutionelle Diskriminierung

Institutionelle Diskriminierung entsteht, wenn interne Ordnungen und Routinen von einer spezifischen Gedächtnisinstitution systematisch bestimmte Gruppen benachteiligen. Sie verschränkt sich oft mit struktureller Diskriminierung.

Kernelemente

- *Organisationsspezifische Regeln:* Benachteiligung wird durch interne Policies, Zuständigkeiten, Budgets oder Systementscheidungen erzeugt.
- *Reproduzierbarkeit im Betrieb:* Effekte sind wiederkehrend (nicht zufällig), weil sie in Prozessen, Rollen und Steuerung verankert sind.
- *Verantwortung und Governance:* Gegenmassnahmen erfordern klare Zuständigkeiten, Dokumentation und überprüfbare Entscheidungswege.

Beispiele

- **Digitalisierungspriorisierung nach Kanon**
  - *Mechanismus:* Auswahlkriterien privilegieren stark nachgefragte Bestände.
  - *Effekt:* Quellen marginalisierter Gruppen bleiben analog.
  - *Indikatoren:* Disparitäten zwischen Anfragevolumen und Erschliessungsgrad pro Community.
  - *Gegenmassnahmen:* Quoten für unterrepräsentierte Bestände, partizipative Auswahlprozesse.
- **Katalogisierungsrichtlinien mit normativen Kategorien**
  - *Mechanismus:* obligatorische Felder wie "männlich/weiblich"; beschränkte Auswahl von kolonial geprägten Schlagwörter.
  - *Effekt:* Fehlklassifikation, Reproduktion diskriminierender Vokabulare.
  - *Indikatoren:* Anteil problematischer Pflichtfelder, Korrekturhistorien.
  - *Gegenmassnahmen:* Revision von Richtlinien, optionale Felder, Governance für Vokabularänderungen.
- **Lizenz- und Gebührenpolitik**
  - *Mechanismus:* Hohe Reproduktionskosten, restriktive Lizenzen.
  - *Effekt:* Erschwerte Nachnutzung für Forscher*innen ohne starke institutionelle Einbindung, besonders in ressourcenärmeren Ländern und Regionen.
  - *Indikatoren:* Anteil Open-Access-Objekte; Gebühren pro Nutzungstyp.
  - *Gegenmassnahmen:* OA-First-Policy, gestaffelte Gebühren, CARE-kompatible Zugangsmodelle.
- **Metadatenstandards ohne Mehrsprachigkeit**
  - *Mechanismus:* bewusste Zulassung nur dominanter Amtssprachen oder standardisierter Umschriften.
  - *Effekt:* Unsichtbarkeit von Namen, Toponymen und Konzepten in Minderheitensprachen.
  - *Indikatoren:* Sprachenabdeckung in Feldern und Indizes.
  - *Gegenmassnahmen:* Mehrsprachige Felder, lokale Namensformen als gleichberechtigte Labels, Such-Unterstützung für Varianten.
- **Ausschluss aus Entscheidungsprozessen**

    - *Mechanismus:* Metadaten-Governance (Vokabularpflege, Redaktionsregeln, Sperrentscheidungen, Priorisierung) wird ohne marginalisierte Gruppen gestaltet oder ihre Beteiligung bleibt symbolisch.
    - *Effekt:* Kategorien, Prioritäten und Korrekturen reproduzieren Dominanz; Konflikte um Begriffe und Darstellung eskalieren, Vertrauen sinkt.
    - *Indikatoren:* Keine dokumentierten Beteiligungsformate, fehlende Beschwerde-/Feedbackwege, wiederkehrende Auseinandersetzungen um "legacy metadata".
    - *Gegenmassnahmen:* Partizipative Gremien/Advisory Boards, transparente Entscheidungsprotokolle, klare Änderungs- und Eskalationsprozesse.
- **Präsentation sensibler Inhalte ohne Schutzkonzept**
    - *Mechanismus:* Veröffentlichung und Interface-Design ohne sichtbare Kontextualisierung, Hinweise oder abgestufte Zugänge für diskriminierende und personenbezogene Inhalte.
    - *Effekt:* (Re-)Traumatisierung, Stigmatisierung und Missbrauchsrisiken; Abschreckung bestimmter Nutzergruppen.
    - *Indikatoren:* Fehlende Content-Notes /Hinweise, keine Zugriffsstufen, keine dokumentierten Risikoabwägungen.
    - *Gegenmassnahmen:* Hinweis- und Kontextlayer, abgestufte Zugänge/Opt-in, dokumentierte Review-Prozesse (siehe Praxisteil, Abschnitt 5.2).

Statistische Diskriminierung

Statistische Diskriminierung bezeichnet die Benachteiligung von Individuen, wenn unter unvollständiger Information Entscheidungen auf gruppenbezogenen Durchschnittswerten beruhen.

Kernelemente

- *Informationsasymmetrie:* Über Einzelne liegen weniger, über Gruppen mehr Information vor. Entscheidungen werden auf Gruppenstatistiken gestützt.
- *Gruppenzuschreibung:* Wahrscheinlichkeiten oder Durchschnitte werden auf einzelne Personen übertragen.
- *Effekt:* Benachteiligung von Personen, die nicht dem Gruppenprofil entsprechen.

Beispiele

- **Automatisches Geschlechter-Mapping in Normdaten**
    - *Mechanismus:* Imputation (das heisst Schätzung fehlender Werte) von `gender` aus Namensstatistiken oder Sprachmodellen.
    - *Effekt:* Systematische Fehlzuordnung bei nicht-westlichen, historischen oder trans Namen.
    - *Indikatoren:* Häufung von Zuweisungen nahe dem Klassifikationsschwellenwert; überproportionale Korrekturen nach Herkunftsregion.
    - *Gegenmassnahmen:* Enthaltungsregel bei Unsicherheit; separate Felder für Selbstangaben und Quellen; kalibrierte Schwellen; aktive Nachannotation.
- **Entity-Resolution nach Produktivität**
    - *Mechanismus:* Zusammenführung häufiger Namen auf den produktivsten Normdatensatz.
    - *Effekt:* False Merges löschen weniger sichtbare Personen; Zitationen und Werke werden fehlgeleitet.
    - *Indikatoren:* Unplausible Sprünge in Grad- und Zitationsverteilungen; Cluster mit hoher Namensähnlichkeit und heterogener Provenienz.
    - *Gegenmassnahmen:* Konservative Blocking-Regeln (Vorabfilter in Record-Linkage-Verfahren, um Vergleichsmenge einzuschränken); zeit- und ortsgebundene harte Constraints; gezieltes Review kleiner oder minorisierter Cluster; vollständige Provenienzspeicherung.
- **Historische Geokodierung mit modernen Gazetteers**
    - *Mechanismus:* Ambige Toponyme werden auf heutige Mehrheitsorte gemappt. Gazetteers sind strukturierte Ortsnamendatenbanken, die meist gegenwärtige Ortsinformationen priorisieren.
    - *Effekt:* Verdrängung historischer Minderheitensiedlungen; Fehlkontexte in Karten.
    - *Indikatoren:* Hoher Anteil Default-Zuordnungen ohne Jahrgang; Dominanz grosser Orte bei kurzen Ortsnamen.
    - *Gegenmassnahmen:* Zeitgeschichtliche Gazetteers (das heisst nach historischen Epochen getrennt); Unsicherheitsgeometrien; "unresolved" statt Zwangszuordnung; expliziter Quellenhinweis im Datensatz.
- **Digitalisierungspolitik nach Zitationsmetriken**
    - *Mechanismus:* Auswahl via globalem Impact-Score als Proxy für "Wert".
    - *Effekt:* Periphere Stimmen bleiben analog; selbstverstärkende Sichtbarkeit etablierter Kanons.
    - *Indikatoren:* Schiefe Verteilungen zugunsten kanonisierter Autorinnen und Autoren sowie Orte; geringe Diversität an den Sample-Rändern.
    - *Gegenmassnahmen:* Stratifizierte Sampling-Pläne; Equity-Buckets (gezielt definierte Auswahlkategorien, um Diversität sicherzustellen); Offenlegung der Auswahlfunktion; Simulation und Vergleich alternativer Auswahlregeln.

Intersektionalität: Verflechtung von Diskriminierungsformen

Begriff und Herkunft

Intersektionalität bezeichnet einen Analyseansatz, der Diskriminierung nicht als isolierte Einzelphänomene (zum Beispiel nur Sexismus oder nur Rassismus) versteht, sondern als **verflochtene, sich wechselseitig konstituierende** Macht- und Ungleichheitsverhältnisse. Der Begriff wurde in der juristischen und feministischen Theorie prominent, um genau die Fälle zu beschreiben, in denen Personen durch das Raster einer Single-Axis-Logik (einachsige Perspektive) fallen: Wer nur nach „Geschlecht" oder nur nach „Rasse" fragt, übersieht die spezifischen Erfahrungen an der Kreuzung mehrerer Kategorien (Crenshaw 1989; Collins 2000).

Für die Daten- und Metadatenpraxis ist daran besonders wichtig, dass Intersektionalität nicht einfach „Mehrfachzugehörigkeit" meint. Es geht nicht darum, Diskriminierungen additiv zu summieren, sondern darum, dass Kategorienwahl, Modellierung und Beschreibung oft so angelegt sind, dass bestimmte Gruppen **weder in der Kategorie A noch in der Kategorie B** sichtbar werden, sondern zwischen beiden verschwinden. Diese Problemlage ist auch in datengetriebenen Systemen empirisch gut belegt: Fehlerraten und Sichtbarkeit unterscheiden sich nicht nur zwischen Einzelkategorien, sondern besonders stark an deren Schnittstellen (Buolamwini und Gebru 2018; Noble 2018).

Relevanz für Metadatenpraxis

Metadaten entstehen durch Entscheidungen darüber, welche Merkmale erfasst werden, wie sie benannt werden und welche Beziehungen zwischen Entitäten modelliert werden. Wenn diese Entscheidungen nur auf einer Achse (einachsige Perspektive) getroffen werden, sind intersektionale Effekte naheliegend:

- **Klassifikationslogik:** Facetten oder Normdaten sind so strukturiert, dass „Frauen" implizit als „weiss" und „Schwarze Menschen" implizit als „männlich" gelesen werden (Crenshaw 1989).
- **Retrieval-Logik:** Such- und Filterfunktionen behandeln Kategorien als unabhängig. Eine Suche nach „Frauen" liefert dann vor allem Datensätze, die in dominanten Kontexten bereits gut beschrieben wurden.

- **Quellenlage und Erschliessung:** Historische Überlieferung und institutionelle Sammelpraktiken sind ungleich verteilt. Wenn diese Asymmetrien nicht explizit dokumentiert werden, erscheinen sie als „Datenlage“ statt als Ergebnis von Machtverhältnissen.

Beispiel: Historische Volkszählungsdaten erfassen „Geschlecht“ und „Ethnie“ als getrennte Variablen. Eine einachsige Auswertung kann nahelegen, dass „Frauen“ im Datensatz ausreichend sichtbar sind, weil viele weibliche Einträge existieren, oder dass „Schwarze Personen“ sichtbar sind, weil viele Einträge als „Black“ markiert sind. Eine intersektionale Perspektive fragt dagegen gezielt nach der Schnittmenge und kann zeigen, dass Schwarze Frauen in bestimmten Erhebungs- oder Auswertungslogiken systematisch unterrepräsentiert sind (zum Beispiel durch Haushaltsvorstandslogik, Namensnormalisierung, fehlende Berufsangaben, oder durch „unbestimmte“ Restkategorien).

Praktische Implikationen

Intersektionalität ist in der Praxis kein „zusätzlicher Tag“, sondern eine Anforderung an Datenmodell, Arbeitsabläufe und Auswertung:

- **Mehrfachkategorisierung ermöglichen:** Wo Kategorien erfasst werden, sollten Kombinationen nicht technisch ausgeschlossen werden (zum Beispiel „nicht-binär“ und „trans“ und „historische Fremdbezeichnung“ als unterschiedliche, kontextgebundene Felder statt als exklusive Auswahl).
- **Schnittmengen prüfen:** Qualitätskontrollen sollten nicht nur pro Kategorie („Frauen“, „Minderheitensprache“) erfolgen, sondern auch für Kreuzungen (zum Beispiel „Frauen + Minderheitensprache“, „Migrantisch + Arbeiterklasse“, „indigen + Archivzugang“).
- **Granularität vs. Schutz:** Je granularer Merkmale kombiniert werden, desto leichter kann Re-Identifikation werden. Intersektionale Sichtbarkeit muss daher mit Datenschutz- und Schutzkonzepten zusammen gedacht werden (Kapitel 3.5).

Diskriminierungsformen sollten nicht isoliert betrachtet werden. Jede Analyse sollte prüfen, ob bestimmte Gruppen an der Kreuzung mehrerer Merkmale besonders betroffen sind. Praktisch heisst das: Intersektionalität als Prüffrage in Planung, Datensammlung, Beschreibung und Veröffentlichung verankern (Kapitel 3.1 und Kapitel 3.5).

Historische Begriffe: Bewahren oder Korrigieren?

Das Dilemma

Historische Quellen enthalten häufig diskriminierende Begriffe und Kategorisierungen. Für eine diskriminierungssensible Metadatenpraxis entsteht daraus ein Spannungsfeld zwischen zwei Ansprüchen:

1. **Quellentreue und historischer Kontext:** Begriffe sind Teil der Überlieferung, sie können für Forschung, Provenienz und Rekonstruktion historischer Semantiken relevant sein.
2. **Schutz vor Verletzung und Perpetuierung:** Begriffe können beim Wiedergeben, Indexieren und Sichtbarmachen erneut verletzen und diskriminierende Ordnungen stabilisieren.

Diese Spannung ist nicht rein sprachlich. Sie betrifft Arbeitsabläufe, Suchbarkeit, Normdatenpflege, Zugangsentscheidungen und die Frage, wer über Begriffe mitentscheiden darf.

Archivische Provenienz (Bewahrungsprinzip)

Archivische Provenienz betont Bewahrung und dokumentarische Integrität: Beschreibung solle die Überlieferung nicht “korrigieren”, um keine Geschichtsfälschung zu betreiben. Problematische Begriffe werden als historische Dokumente verstanden, deren Veränderung Kontext zerstören kann. Zusätzlich wird argumentiert, dass Forschende Originalbegriffe benötigen, um Quellen zu finden und historische Sprachgebrauche nachzuvollziehen (Jenkinson 1922).

Für Metadatenpraxis kann daraus eine Strategie folgen, die auf minimale Eingriffe setzt: Originaltermini werden beibehalten, allenfalls ergänzt um Erläuterungen.

Reparative Beschreibung

Reparative Beschreibung argumentiert, dass Archive, Kataloge und Metadaten nie neutral sind: Sie sind eine Zugriffsschicht, die aktiv steuert, was sichtbar, auffindbar und legitim erscheint. Problematische Begriffe unkommentiert fortzuführen, ist daher selbst eine Entscheidung mit Folgen. Reparative beziehungsweise kritische Beschreibung zielt darauf, diskriminierende Beschreibungspraktiken zu erkennen, zu kontextualisieren und zu verändern. Leitfäden wie die Anti-Racist Description Resources und Rassismuskritische Dokumentation betonen dabei Transparenz, Kontext, Community-Einbindung und den bewussten Umgang mit „legacy metadata“ (Archives for Black Lives in Philadelphia. Anti-Racist Description Working Group 2020; Schweizerisches Nationalmuseum u. a. 2025). Die dekoloniale Archivdebatte formuliert dazu explizit, dass Beschreibungspraxis nicht nur Dokumentation, sondern politische Praxis ist und daher revidierbare Verantwortungsentscheidungen braucht (Ghaddar und Caswell 2019; Caswell 2019).

In deutschsprachigen Archivdebatten wird diese Spannung unter anderem im Kontext queerfeministischer Kritik an Katalog- und Ordnungspraktiken diskutiert (Sparber 2016; Gruber 2022; Rottmann 2023).

Kompromissmodelle in der Praxis

In der Praxis existieren Ansätze, die Quellentreue und Gegenwartssensibilität kombinieren:

1. **Dual Cataloging:** Historische und aktuelle Begriffe parallel führen (zum Beispiel in getrennten Feldern oder über Varianten), sodass sowohl Recherche nach historischen Termini als auch respektvolle Darstellung möglich ist.
2. **Kontextualisierung:** Problematische Begriffe markieren und erklären (zum Beispiel über Content Notes, Hinweise zur Terminologie und kurze Einordnungen).
3. **Community Involvement:** Betroffene Communities in Entscheidungen einbeziehen (zum Beispiel bei Vokabularen, Korrekturen, Sperrfristen oder Kontexttexten).
4. **Versionierte Beschreibung:** Änderungen nachvollziehbar dokumentieren (Wer hat wann warum geändert? Welche Varianten bleiben recherchierbar?).
5. **Community Archives:** Selbstbestimmte Dokumentationspraktiken respektieren und nicht durch institutionelle Normdatenlogik „glattziehen“ (Flinn 2007).

Wenn diese Ansätze aus Ressourcen- oder Governance-Gründen nicht vollständig umgesetzt werden können, empfiehlt sich als Minimalstrategie ein kontextsensitives Vorgehen, das klar zwischen Überlieferungs- und Beschreibungsebene unterscheidet:

- Für **Metadaten** (Beschreibungsebene): aktuelle, respektvolle Begriffe verwenden und historische Varianten dokumentieren (zum Beispiel als Varianten/Altlabels) statt sie als bevorzugte Benennung auszugeben.

- Für **Transkriptionen/Editionen**: Originalwortlaut beibehalten, aber kontextualisieren (zum Beispiel mit Content Notes und eindeutigen Hinweisen auf diskriminierende Sprache).
- Für **Normdaten**: Änderungen versionieren und begründen; die Provenienz von Begriffen und Korrekturen dokumentieren.

Zentral ist dabei die Frage der Verantwortung: Wer entscheidet über Begriffe, wer trägt Risiken, und wie werden Betroffene beteiligt? (Kapitel 3.5.4, Kapitel 3.8.3 und Kapitel 3.5).

Verzerrungen und Fehler (Bias)

Unter einem **Bias** wird eine Verzerrung – eine systematische Abweichung einer objektiven Darstellung – verstanden. Technisch kann ein Bias in Forschungsdaten etwa durch eine unausgewogene Datenauswahl, eine stereotypisierende Begriffsauswahl oder durch algorithmische Vorannahmen entstehen. Inhaltlich äussert sich ein Bias in verschiedenen Formen wie in der Auswahl dessen, was überhaupt als erzählenswert gilt, in der Art und Weise der Beschreibung historischer Ereignisse oder in den moralischen oder interpretativen Bewertungen, die Historiker*innen vornehmen. Diese Voreingenommenheit ist nicht nur ein Fehler, sondern ergibt sich aus der grundlegenden Tatsache, dass jede historische Darstellung selektiv und perspektivisch ist. In der Metadatenerstellung ist oftmals ein Bias in den Beschreibungstexten zu finden. So werden beispielsweise Frauen in historischen Quellen oftmals auf ihr äusseres Erscheinungsbild reduziert, während bei Männern häufig zuerst ihre beruflichen Tätigkeiten erwähnt werden. Diese Reduktion der Frau auf ihr äusseres Erscheinungsbild steht in einer langen Tradition der Objektivierung weiblicher Körper. Wenn im Kontext der Überarbeitung der Metadaten nicht nur der Bias kritisch reflektiert wird, sondern der Fehler als Reproduktion von Geschlechterstereotypen dargestellt wird, werden diskriminierende Strukturen sichtbar und können so schrittweise abgebaut werden.

Verhältnis von Bias und Diskriminierung

Das Problem der Begriffsverwirrung

In der Forschung wie auch in praktischen Debatten wird **Bias** häufig als Sammelbegriff verwendet: mal als neutral-technische Bezeichnung für Verzerrungen, mal als moralische Kurzformel für „Ungerechtigkeit“ oder „Diskriminierung“. Diese Doppelverwendung führt zu zwei wiederkehrenden Problemen. Erstens werden Phänomene vermischt, die sich analytisch trennen lassen: Verzerrungen in Daten, Modellen und Interfaces einerseits und Diskriminierung als soziale Praxis beziehungsweise als Ergebnis ungleicher Macht- und Teilhabeverhältnisse andererseits. Zweitens wird dadurch die Verantwortung verschoben: Was als „Bias“ benannt wird, erscheint oft als technisches Qualitätsproblem, selbst wenn die Ursachen (und Folgen) in Kategorienwahl, institutionellen Routinen, Normdaten, Ausschlüssen und historischen Machtverhältnissen liegen.

Für eine diskriminierungssensible Metadatenpraxis ist diese Unschärfe besonders relevant, weil Metadaten sowohl **Beschreibung** (eine technische Zugriffsschicht) als auch **Deutung** (eine historisch-soziale Perspektive) enthalten. Ein und derselbe Befund – etwa stereotype oder auslassende Beschreibungen – kann daher als „Bias“ (Verzerrung) beschrieben werden, aber zugleich als Form von Diskriminierung wirken, wenn er bestimmte Gruppen systematisch unsichtbar macht, herabsetzt oder ihre Handlungsmacht verzerrt.

Bias als technischer Fehler

Eine einflussreiche Position – prominent in Fairness- und Bias-Forschung zu maschinellem Lernen – versteht Bias primär als **mess- und modellierbare Abweichung**, die sich diagnostizieren und (zumindest teilweise) korrigieren lässt. In diesem Zugriff ist Bias ein Problem der Datenqualität, der Messung, der Modellannahmen oder der Evaluation: Man sucht nach Quellen der Verzerrung (zum Beispiel nicht-repräsentative Datensätze, ungeeignete Benchmarks, fehlerhafte Proxy-Variablen) und nach Verfahren zur Reduktion (zum Beispiel Re-Sampling, Re-Weighting, Regularisierung, Post-Processing). Surveys wie Mehrabi et al. ordnen solche Bias-Quellen systematisch und machen sichtbar, dass „Bias“ je nach Pipeline-Stufe sehr unterschiedliche technische Ursachen haben kann (Mehrabi u. a. 2021).

Wichtig ist: Selbst innerhalb dieses technisch orientierten Zugriffs ist „Fairness“ kein einfacher Zielwert. Verschiedene Fairness-Definitionen (zum Beispiel Demographic Parity, Equalized Odds, Individual Fairness) können sich gegenseitig ausschliessen, sodass Korrekturen immer auch **Trade-offs** erzeugen (Chouldechova 2017; Kleinberg u. a. 2017). Für die Praxis heisst das: Bias-Reduktion ist ein sinnvolles Instrument, aber sie löst nicht automatisch die Frage, welche Norm (Gleichheit, Chancengleichheit, Schadensminimierung, Rechenschaft) im jeweiligen Kontext gilt.

Gerade hier liegt der Unterschied zur Diskriminierung: Bias beschreibt in erster Linie eine diagnostizierbare Verzerrung in einem technischen oder methodischen Prozess; Diskriminierung bezeichnet die sozial wirksame Benachteiligung von Personen und Gruppen in ungleichen Machtverhältnissen. Ein System kann damit statistisch “fairer” werden und dennoch diskriminierend wirken, wenn es historisch belastete Kategorien fortschreibt, Ausschlüsse reproduziert oder nur dominante Nutzergruppen als implizite Norm behandelt (Buolamwini und Gebru 2018; Noble 2018).

Bias als Symptom struktureller Diskriminierung

Eine zweite Position – anschlussfähig an Critical Data Studies und feministische/antirassistische Technik- und Wissenskritik – versteht Bias nicht primär als “Fehler im System”, sondern als **Symptom**: Daten und Modelle reproduzieren gesellschaftliche Ungleichheiten, weil sie in Institutionen, Klassifikationen, Quellenlagen und epistemischen Routinen eingebettet sind. Aus dieser Perspektive ist es irreführend, Diskriminierung als Ausnahme zu behandeln, die „aus den Daten entfernt“ werden kann. Vielmehr sind Kategorien, Trainingsdaten, Normvokabulare, Such- und Sortierlogiken und auch das, was als „gute“ Daten gilt, historisch geprägt und sozial umkämpft (Noble 2018; Buolamwini und Gebru 2018).

Data Feminism betont in diesem Sinn, dass Datenarbeit immer auch Machtarbeit ist: Wer wird gezählt, wer wird beschrieben, welche Klassifikationen werden als neutral ausgegeben, und wessen Perspektiven fehlen in der Infrastruktur? (D’Ignazio und Klein 2020) Für Metadaten heisst das: Selbst wenn einzelne Verzerrungen technisch reduziert werden, kann die Praxis weiterhin diskriminierend wirken – etwa wenn die zugrunde liegenden Kategorien Gruppen unsichtbar machen, wenn historische Fremdbezeichnungen unkritisch fortgeschrieben werden oder wenn institutionelle Routinen bestimmte Quellen systematisch ausschliessen.

Bias, Diskriminierung und Unterdrückung: Ein integrativer Blick

Dieses Handbuch unterscheidet Bias und Diskriminierung analytisch, hält sie aber praktisch für eng verschränkt.

Erstens: Die Bias-Taxonomie in diesem Kapitel ist als **Diagnose- und Handlungsinstrument** gedacht. Sie hilft, Verzerrungen in Daten (zum Beispiel Auswahl- und Messfehler), in Algorithmen (zum Beispiel Evaluations- oder Optimierungsartefakte) und in Nutzerinteraktionen (zum Beispiel Ranking- und Popularitätslogiken) zu lokalisieren und gezielt anzugehen.

Zweitens: Bias wird hier zugleich als **operationalisierte, analysierbare Form** von Diskriminierung verstanden – im Sinne von „sichtbaren Spuren“ ungleicher Macht- und Teilhabeverhältnisse in Daten- und Metadatenpraktiken. Bias-Korrekturen sind daher oft notwendig (zum Beispiel um unmittelbare Schäden zu reduzieren), aber sie sind selten hinreichend, wenn Kategorien, institutionelle Ziele und Verantwortlichkeiten unberührt bleiben.

Hier setzt das Konzept der **Oppression** (Unterdrückung) an. Oppression ist nicht bloss eine einzelne Handlung der Benachteiligung, sondern das Zusammenspiel von Praktiken, Diskursen und Institutionen, die Handlungsspielräume ganzer Gruppen systematisch einengen. In der feministischen Machttheorie wird Unterdrückung als **strukturelles “power-over”** verstanden: eine dauerhafte, nicht-zufällige Konstellation, in der Institutionen, Normen und symbolische Ordnungen bestimmte Gruppen in ihren Optionen einschränken, während andere privilegiert werden (Young 1990). Unterdrückung wirkt **materiell** (physisch, ökonomisch), **symbolisch** (Stigmatisierung, Unsichtbarmachung) und **epistemisch** (Festlegung dessen, was als Wissen gilt).

Historische Daten- und Metadatenpraktiken können Unterdrückung auf mindestens drei Ebenen reproduzieren: ontologisch, epistemisch und infrastrukturell. Diese Formen wirken kumulativ: Jede einzelne Drahtstrebe, beispielsweise ein kontrolliertes Vokabular, scheint zunächst harmlos. Doch im Zusammenspiel entsteht ein Käfig, der die Bewegungsfreiheit bestimmter Gruppen systematisch einschränkt (Frye 1983).

Ontologische Gewalt

Wirkmechanismus: Erzwingt Kategorien, die der Selbstbeschreibung der Betroffenen widersprechen oder sie auf defizitäre Merkmale reduziert (Bowker und Star 1999).

Beispiele:

- Binärer `gndo:gender`-Wert (“male”/“female”) löscht nicht-binäre Identitäten.
- Koloniale Ethnonyme in Normdaten (“Hottentotte”) perpetuieren rassistische Klassifikationen.

Epistemische Gewalt

Wirkmechanismus: Verunmöglicht Wissen, indem es bestimmte Perspektiven ausschliesst oder als “Rauschen” markiert (Spivak 1988).

Beispiele:

- Aggregations-Metadaten, die Briefe von Dienstbotinnen als “miscellaneous” ablegen.
- Diplomatenschriftverkehr wird dagegen fein granular erschlossen.

Infrastrukturelle Gewalt

Wirkmechanismus: Fixiert Benachteiligungen durch technische Standards, die schwer veränderbar sind (lock-in) (Rodgers und O’Neill 2012).

Beispiele:

- Vorgabefelder in Sammlungssoftware ohne Mehrsprachigkeit erzwingen englischsprachige Schlagwörter und verdrängen indigene Begriffe.
- Unveränderbare Feldlängen lassen traditionelle Namen abschneiden.

Verzerrung und Fehler in Daten (Data Bias)

Messfehler (Measurement Bias)

Bias, der durch fehlerhafte, unvollständige oder inadäquate Messung von Variablen entsteht.

Beispiel 1: In digitalen Editionen historischer Texte wird “Bedeutung” oft über die Häufigkeit bestimmter Begriffe erfasst, doch fehlerhafte OCR-Erkennung - etwa wenn das historische “ſ” nicht als “s” erkannt wird – oder uneinheitliches Tagging führen leicht zu systematischen Verzerrungen.

Beispiel 2: Erfassung von Geschlecht in historischen Volkszählungen: “Beruf: Haushaltsvorstand” wird in digitalen Datenbanken oft als “männlich” codiert, was weibliche Haushaltsvorstände systematisch ausschliesst.

Auslassungsfehler (Omitted Variable Bias)

Entsteht, wenn relevante Variablen im Modell fehlen, was zu verzerrten Ergebnissen führt.

Beispiel: Eine digitale Netzwerkanalyse historischer Korrespondenz lässt informelle Kommunikationswege (zum Beispiel persönliche Treffen, mündliche Überlieferung) aus, was zu verzerrten Interpretationen von Kommunikationsnetzwerken führt.

Repräsentationsfehler (Representation Bias)

Bias durch nicht-repräsentative Stichproben, zum Beispiel geografische oder demografische Unterrepräsentation im Datensatz.

Beispiel: Digitalisierte Zeitungsarchive decken meist nur bestimmte (oft bürgerliche oder urbane) Presse ab; Arbeiterzeitungen, marginalisierte Gruppen oder nicht-deutschsprachige Publikationen fehlen und werden in der Forschung unsichtbar.

Aggregationsfehler (Aggregation Bias)

Fehlerhafte Verallgemeinerung von Gruppenergebnissen auf Individuen oder Subgruppen.

Simpson’s Paradox: Aggregierte Trends können täuschen, weil sich Zusammenhänge auf Subgruppenebene ins Gegenteil verkehren.

Beispiel: Eine Methode scheint insgesamt erfolgreicher, ist aber in allen Teilgruppen weniger erfolgreich, die Aggregation verschleiert dies.

Modifiable Areal Unit Problem (MAUP): Ergebnisse hängen von der gewählten räumlichen Aggregation ab.

Beispiel: Zusammenfassung von unterschiedlichen Sozialstrukturen (zum Beispiel alle “Arbeiter” im 19. Jh.) überdeckt regionale Unterschiede – etwa zwischen Textilarbeiterinnen im Ruhrgebiet und Landarbeitern in Ostpreussen.

Stichprobenfehler (Sampling Bias)

Bias durch nicht-zufällige Auswahl von Stichproben führt zu mangelnder Generalisierbarkeit.

Beispiel: Oral History-Projekte, die ausschliesslich mit Zeitzeugen arbeiten, die aktiv Kontakt zu Forscher*innen aufnehmen, erfassen tendenziell eher politisch engagierte oder bildungsnahe Akteur*innen.

Längsschnittfehler (Longitudinal Data Fallacy)

Fehlschluss durch Vermischung von Kohorten in Querschnittsdaten, anstatt echte zeitliche Entwicklung zu betrachten.

Beispiel: Analyse von Wikidata-Einträgen zu historischen Persönlichkeiten über Jahrzehnte hinweg, ohne zu berücksichtigen, dass sich die Erfassungsregeln oder Community-Praktiken im Zeitverlauf ändern.

Historische Verzerrung (Historical Bias)

Bias, der bereits in der gesellschaftlichen Realität existiert und sich in den Daten widerspiegelt, auch bei perfekter Stichprobe.

Beispiel: Digitale Repositorien, die historische Demografie abbilden, spiegeln patriarchale Strukturen wider: Die geringe Zahl von “Frauen in Führungspositionen” ist kein Datenfehler, sondern gesellschaftliche Realität.

Populationsfehler (Population Bias)

Unterschiede zwischen Nutzenden der Plattform und der Zielpopulation, zum Beispiel durch Demografie.

Beispiel: Wikidata-Einträge zu Historiker*innen stammen überproportional von männlichen, westlichen Beitragenden, was sich in der Sichtbarkeit und Kategorisierung niederschlägt.

Verzerrungen in und durch Algorithmen (Bias in Algorithms)

Algorithmischer Fehler (Algorithmic Bias)

Bias, der durch algorithmische Designentscheidungen entsteht, unabhängig von Bias in den Daten (zum Beispiel durch Auswahl der Optimierungsfunktion).

Beispiel: Topic Modeling von philosophischen Texten ergibt “Themen”, die Resultat von Wortlisten und Stoppwortdefinitionen sind, aber von Nicht-Expert*innen als inhaltlich signifikante Topoi interpretiert werden.

Evaluationsfehler (Evaluation Bias)

Verzerrung durch ungeeignete oder unausgewogene Benchmarks bei der Modellbewertung.

Beispiel: Trainings- und Testsets für Named-Entity-Recognition-Modelle im Bereich Geschichte verwenden hauptsächlich Quellen des 20. Jahrhunderts. Modelle performen deshalb schlecht bei mittelalterlichen oder frühneuzeitlichen Texten.

Verzerrungen durch Nutzerinteraktion (User Interaction Bias)

Darstellungsfehler (Presentation Bias)

Ungleichgewicht, das durch visuelle, typografische oder layoutbezogene Hervorhebungen entsteht. Interface-Entscheide lenken Aufmerksamkeit und Interpretationsrahmen, bevor inhaltliche Qualität bewertet wird.

Beispiel: Quellen, die ohne Einschränkungen zugänglich sind, werden farblich hervorgehoben im Archivkatalog. Damit werden sie häufiger angeklickt und verdrängen, die weniger zugänglichen Quellen überdurchschnittlich.

Rangfolgenfehler (Ranking Bias)

Systematische Verzerrung durch Sortierlogiken, die Klicks, Zitationszahlen oder Metadatenfülle belohnen. Höhere Position führt zu mehr Aufmerksamkeit, was die ursprüngliche Rangordnung verstärkt, unabhängig von Relevanz.

Beispiel: Museumsportal sortiert nach “Meist betrachtet”. Kolonialzeitliche Exponate mit früher Social-Media-Reichweite dominieren, während neu katalogisierte Objekte aus dem Globalen Süden kaum Sichtbarkeit erhalten.

Popularitätsfehler (Popularity Bias)

Beliebtere Objekte werden häufiger gezeigt und verstärken dadurch ihre Popularität, unabhängig von Qualität.

Beispiel: In Crowdsourcing-Projekten zu alten Handschriften dominieren wenige besonders aktive User, sodass ihre Lesarten überproportional häufig übernommen werden.

Emergenter Fehler (Emergent Bias)

Bias, der erst durch langfristige Interaktion mit Nutzenden oder gesellschaftlichen Wandel entsteht.

Beispiel: Ein digitales Editionsprojekt zu mittelalterlichen Urkunden wird ursprünglich als Forschungsinfrastruktur für Editionsphilologie konzipiert. Mit der Zeit beginnen jedoch genealogische Communities, die Daten für Familienforschung zu nutzen. Dadurch verschiebt sich die Nachfrage in Richtung Namens- und Ortsindexierung. Die Infrastrukturbetreiber passen ihre Metadatenstrukturen und Suchfunktionen an diese Nutzergruppen an, was wiederum philologische Tiefeninformationen (zum Beispiel Variantenkritik) systematisch marginalisiert.

Selbstselektionsverzerrung (Self-Selection Bias)

Bias durch selbstselektierende Teilnehmende, zum Beispiel in Umfragen.

Beispiel: Digitalisierungsprojekte zu Privatarchiven werden eher von Familien mit hohem kulturellem Kapital initiiert, während marginalisierte Familien seltener teilnehmen.

Soziale Verzerrung (Social Bias)

Individuelle Entscheidungen werden durch soziale Einflüsse verzerrt.

Beispiel: Bei kollaborativen Transkriptionsprojekten orientieren sich neue User*innen an den Lesegewohnheiten erfahrener Transkribierender und übernehmen deren Fehler.

Verhaltensverzerrung (Behavioral Bias)

Unterschiedliches Verhalten von Nutzenden je nach Plattform, Kontext oder Zeit.

Beispiel: Historiker*innen recherchieren systematischer, während Lai*innen häufig nach Familiennamen oder spektakulären Ereignissen suchen. Dies beeinflusst Zugriffszahlen.

Zeitliche Verzerrung (Temporal Bias)

Verzerrungen, die sich aus zeitlichen Veränderungen in Verhalten oder Population ergeben.

Beispiel: Die Häufigkeit von Suchbegriffen in Archiven schwankt mit Debatten (zum Beispiel "Pandemie" 2020/21), was langfristige Analysen verzerrt.

Inhaltsproduktionsfehler (Content Production Bias)

Verzerrungen, die auf Unterschieden in Struktur, Lexik, Semantik und Syntax nutzergenerierter Inhalte beruhen.

Beispiel: Digitale Foren zur Wissenschaftsgeschichte werden auf Englisch dominiert; andere Sprachen sind unterrepräsentiert.

Daten über Daten

Was sind Forschungsdaten?

Unter Forschungsdaten verstehen wir sämtliche digitale Repräsentationen von physischen und virtuellen Objekte, die Forscher*innen während ihrer Forschung verwenden und produzieren und die als digitale Daten repräsentiert werden können. Dazu gehören Quellen, Transkriptionen oder Reproduktionen, Exzerpte, Zeitreihen, Tabellen, Diagramme, Karten, Modelle, Bilder, Videos, Interviews, Artikel, Sekundärliteratur, Software, Quellcode, Workflows, Forschungsprotokolle, Datensätze und Ähnliches.

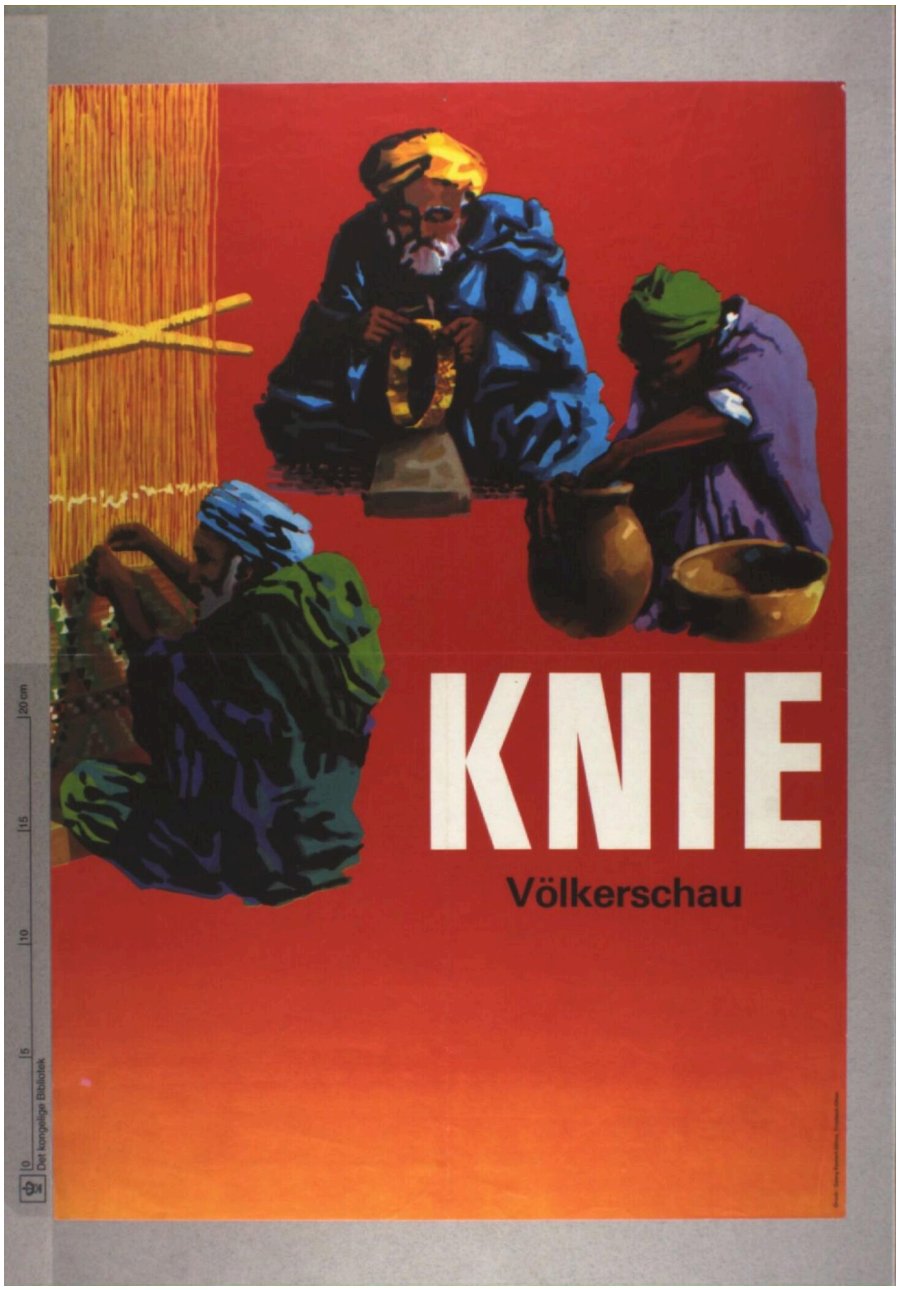
Bei historischen Forschungsprojekten stammt ein grosser Teil der Forschungsdaten oft aus den Beständen von Gedächtnisinstitutionen wie Archiven, Bibliotheken und Museen oder steht in publizierter Form in Büchern oder Artikeln zur Verfügung. In vielen Fällen sorgen diese Einrichtungen für die Langzeitarchivierung der Quellen. Dann kann über die DOI oder die Signatur direkt auf die Objekte (sowie auf deren Metadaten) verwiesen werden. Meistens werden in Forschungsprojekten jedoch zusätzliche Metadaten erhoben oder bestehende Metadaten korrigiert. Dabei kann es sich etwa um Quellenannotationen, erweiterte Beschreibungen oder korrigierte Angaben handeln. In diesen Fällen empfiehlt es sich, einen neuen, möglichst kompletten Metadatensatz zu erstellen und mit Verweis auf das Original auf einer geeigneten Plattform zur Verfügung zu stellen. Redundanz ist bei Forschungsdaten wünschenswert und erhöht ihre Verfügbarkeit und Auffindbarkeit.

Im Rahmen der Forschung werden oft Daten aus historischen Quellen abgeleitet und zusammengestellt. Dazu gehören etwa Textdaten (zum Beispiel Forschungsprotokolle), Zeitreihen (gegebenenfalls dargestellt als Diagramme oder Tabellen) oder georeferenzierte Karten und Netzwerkdarstellungen (zum Beispiel basierend auf Grabungsdaten oder Briefwechseln).

Da viele textuelle Forschungsdaten nur auf Papier oder in unstrukturierter digitaler Form vorliegen, ist die Extraktion strukturierter Daten aus diesen Materialien oft sehr aufwändig (scannen, bereinigen, annotieren und so weiter). Neben den für die Forschung relevanten Daten müssen auch die dazugehörigen Prozessinformationen und unterstützenden Daten (Software, Algorithmen, Protokolle und so weiter) dokumentiert, archiviert und zugänglich gemacht werden. Diese Informationen sind unerlässlich, um die Nachvollziehbarkeit und Reproduzierbarkeit der Forschungsergebnisse zu gewährleisten.

Was sind Metadaten?

Metadaten enthalten strukturierte Informationen über Objekte, insbesondere über deren Inhalt, Kontext und Struktur. Dabei ermöglichen beziehungsweise erleichtern sie deren Identifikation, Auffindbarkeit, Organisation, Verwaltung, Kontextualisierung und Nutzung. Metadaten sollten so strukturiert sein, dass sie die wichtigsten Attribute des beschriebenen Objekttyps modellieren. Ihre Speicherung erfolgt entweder unabhängig von oder auch zusammen mit den ihnen zugeordneten Daten.



Plakat des Circus Knie für eine “Völkerschau”, Darstellung von stereotypisierten Szenen mit Männern beim Handwerk, Schweiz, 20. Jahrhundert.

Zur Illustration Begrifflichkeiten rund um Metadaten greifen wir auf die Metadaten des Beispiels *Knie Völkerschau* zurück. Wir haben es ausgewählt, weil es exemplarisch für problematische, kolonial geprägte Darstellungsweisen in populärer Unterhaltungskultur des 20. Jahrhunderts steht und über reichhaltige Metadaten verfügt.

“Metadaten-Tabelle für: *Knie Völkerschau*”

Feld (Dublin Core / Europeana)	**Wert**
dc:title	Knie Völkerschau
dc:description	Litografi, flerfarvet tryk : mål: 500 x 350 mm flerfarvet tegning af tre håndværkere ved arbejdet
dc:date	1969?
dc:type	Billede, Todimensionalt billedmateriale Still image, poster
dc:subject	Knie Cirkus Håndværker
dc:identifier	http://www.kb.dk/images/billed/2010/okt/billeder/object488811/en/
dc:rights	Billedet er muligvis beskyttet af loven om ophavsret CC BY-NC-ND 4.0
edm:isShownBy	http://kb-images.kb.dk/DAMJP2/DAM/Samlingsbilleder/0000/488/811/PL000012/full/full/0/native.jpg
edm:isShownAt	http://www.kb.dk/images/billed/2010/okt/billeder/object488811/en/
edm:provider	DK-National Aggregation Service

Feld (/ Europeana)	Wert
edm:dataProvider	Det Kongelige Bibliotek, Nationalbibliotek og Kobenhavns Universitetsbibliotek
edm:country	Denmark
edm:language	da (Dänisch)
edm:preview	https://api.europeana.eu/thumbnail/v2/url.json?uri=http%3A%2F%2Fkb-images.kb.dk%2FDAMJP2%2FDAM%2FSamlingsbilleder%2F0000%2F488%2F811%2FPL000012%2Ffull%2Ffull%2F0%2Fnative.jpg\&type=IMAGE
Europeana-ID	92023/images_billed_2010_okt_billeder_object488811
Europeana-LandingPage	https://www.europeana.eu/en/item/92023/images_billed_2010_okt_billeder_object488811
IIIF Manifest	https://iiif.europeana.eu/presentation/92023/images_billed_2010_okt_billeder_object488811/manifest

Die Metadaten des Plakats *Knie Völkerschau* machen deutlich, dass Metadaten sowohl **intrinsische** (dem Objekt selbst inhärente) als auch **extrinsische** (dem Objekt zugeschriebene) **Informationen** enthalten können. So verweist etwa `dc:description` auf intrinsische Eigenschaften wie Format, Material und Gestaltung des Drucks, während Felder wie `dc:subject` oder `edm:provider` extrinsische Klassifikationen und institutionelle Zuschreibungen dokumentieren (forschungsdaten.info 2024).

Darüber hinaus lassen sich die im Beispiel vorliegenden Metadaten verschiedenen **Funktionskategorien** zuordnen:

- **Bibliographische Metadaten**: Titel (`dc:title`), Identifier (`dc:identifier`, *Europeana-ID*), Rechte (`dc:rights`).
- **Administrative Metadaten**: Angaben zu Datenprovidern, Aggregationsdiensten und Zugangs-URLs (`edm:provider`, `edm:isShownAt`).
- **Fachspezifisch-inhaltliche Metadaten**: thematische Schlagworte (`dc:subject`) oder die Beschreibung der dargestellten Handwerker-Szenen (`dc:description`).

Im Kontext der Digitalisierung ist zusätzlich von **Paradaten** beziehungsweise **Prozessmetadaten** zu sprechen, die im Europeana-Datensatz implizit mitschwingen, auch wenn sie nicht explizit aufgeführt sind. Dazu gehören etwa Kameraeinstellungen oder Farbprofile, die während der Erstellung des Digitalisats generiert wurden (Hart u. a. 2025).

Das Beispiel illustriert zudem die Differenz zwischen **physischen Objekten** und deren **digitalen Repräsentationen**:

- Das physische Plakat selbst besitzt Eigenschaften wie Grösse, Material und Zustand.
- Das Digitalisat ist als JPEG-Datei mit spezifischer Auflösung verfügbar (`edm:isShownBy`, `edm:preview`).
- Der Digitalisierungsprozess erzeugt weitere Informationen, etwa zu Perspektive, Bildausschnitt und Farbwiedergabe, die für eine präzise Kontextualisierung entscheidend sind.

Metadatenstandards

Zur Strukturierung von Metadaten existieren verschiedene **Standards und Verfahren**, die meistens durch eine Fachcommunity entwickelt und gepflegt werden. Sie dienen der Sicherung von **Qualität, Konsistenz und Interoperabilität**, bringen jedoch auch Herausforderungen wie **Standardisierungsdruck, mögliche Auslassungen** und **Semantisierungsverluste** mit sich. Dabei lassen sich sieben Ebenen oder Typen von Standards unterscheiden.

Eine detaillierte Aufschlüsselung dieser sieben Ebenen mit Beispieltabelle ist in den Anhang ausgelagert (siehe Kapitel 7.4).

Im Praxisteil dieses Handbuchs gehen wir insbesondere auf Schemas, Inhalts- und Wertstandards ein und beziehen uns dabei vorwiegend auf die Dublin Core Metadata Initiative (DCMI). Eine umfassende Übersicht weiterer verbreiteter Standards liefert etwa Gabay u. a. (2023).

FAIR, CARE und LOUD: Überblick und Zielsetzung

Dieses Kapitel bündelt drei komplementäre Orientierungen der Forschungsdatenpraxis. **FAIR** adressiert Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit als technische und organisatorische Leitlinien (Wilkinson u. a. 2016; Go FAIR Initiative, o. J.). **CARE** rückt kollektiven Nutzen, Kontrollhoheit, Verantwortung und Ethik in den Fokus, insbesondere bei indigenen Daten (Global Indigenous Data Alliance, o. J.; Carroll u. a. 2021). **LOUD** verschiebt den Blick von der reinen Bereitstellung hin zur tatsächlichen Nutzbarkeit in wissenschaftlichen Arbeitsabläufen. Ziel ist eine lesbare, praxisnahe Ordnung der Prinzipien für historische Forschung und Metadatenarbeit.

Begriffsklärungen: offen, öffentlich, Open Data

"**Offen**" bezeichnet rechtlich und technisch definierte Nachnutzbarkeit. Offenheit wird durch Lizenzierung, Standards und Dokumentation hergestellt. "**Öffentlich**" meint faktische Sichtbarkeit ohne Zugangsbeschränkungen. Offenheit ist möglich ohne Öffentlichkeit, etwa bei Zugang auf Antrag bei offener Lizenz und klaren Metadaten; umgekehrt können öffentlich sichtbare Daten nicht offen sein, wenn Lizenz oder Nutzungsbedingungen Nachnutzung verhindern. "**Open Data**" steht normativ für Daten, die allen zur Verfügung stehen, typischerweise unter Free-Culture-Lizenzen wie CC 0, CC BY oder CC BY-SA. Offene Daten senken Zugangshürden und fördern Kollaboration; sie sind jedoch kein Selbstzweck. Sensible Quellen, personenbezogene Metadaten und schutzwürdige Kontexte erfordern abgestufte Zugänge, transparente Bedingungen und sorgfältige Abwägungen. Für Gedächtnisinstitutionen liefern die OpenGLAM-Prinzipien, das 5-Sterne-Modell für Open Data und die Open-Data-Policy-Guidelines zusätzliche Orientierung; zum Gemeinfreiheitsregime in der Schweiz informiert ein Factsheet des IGE.

FAIR: Nachnutzbarkeit als Leitlinie

FAIR steht für Findable (auffindbar), Accessible (zugänglich), Interoperable (interoperabel) und Reusable (wiederverwendbar). Die Prinzipien wurden 2016 als Leitlinien für eine nachhaltige und maschinenlesbare Datenpraxis formuliert:

- **Auffindbar:** Daten und Personen erhalten persistente Identifikatoren (zum Beispiel DOI, ORCID). Metadaten sind strukturiert, suchbar und beschreibend; minimal gefordert sind ein stabiler Ort und eine Zitierempfehlung.
- **Zugänglich:** Zugänge und Bedingungen sind dokumentiert. Auch wenn Daten nicht öffentlich sind, bleiben Metadaten frei zugänglich und der Weg zum Zugang nachvollziehbar.
- **Interoperabel:** Formate und Schemata folgen Standards; kontrollierte Vokabulare, Normdaten und Ontologien ermöglichen Verknüpfung und maschinelle Weiterverarbeitung.
- **Wiederverwendbar:** Lizenztexte, Provenienz, Versionierung, Qualitätsangaben und methodische Kontexte ermöglichen kritische Prüfung und Weiterverwendung. FAIR verlangt keine vollständige Öffentlichkeit, sondern klare Bedingungen für nachhaltige Nachnutzung.

> ⓘ **Praxisbeispiel: Open Research Data Template**
>
> *Stadt.Geschichte.Basel publiziert und dokumentiert Forschungsdaten mit dem Open Research Data Template von Moritz Mähr, um Open- und FAIR-Prinzipien operationell umzusetzen. DOIs via Zenodo sichern Zitierbarkeit und Auffindbarkeit; GitHub und GitHub Pages stellen Repositorium und lesbare Dokumentation bereit. Klare Lizenzierung (zum Beispiel CC BY 4.0 für Daten, AGPL 3.0 für Code), standardisierte Ordner- und Dateistrukturen sowie Automationen erhöhen Interoperabilität. README, CHANGELOG und CITATION.cff, Versionierung und Issue-Vorlagen fördern Transparenz und Wiederverwendbarkeit.*

Datenethik und CARE

Konventionelle Datenpraktiken vernachlässigen oft Entstehungskontexte, Machtverhältnisse und Folgewirkungen. **CARE** steht für Collective Benefit (Kollektiver Nutzen), Authority to Control (Kontrolle über die Daten), Responsibility (Verantwortung), Ethics (Ethik) und schliesst diese Lücke und richtet Datenpraxis an kollektiven Rechten und Pflichten aus:

- **Kollektiver Nutzen** fordert, dass Datenpraxis dem kollektiven Nutzen der betroffenen Gemeinschaften dient und nicht nur externen Forschungsinteressen.
- **Kontrolle über die Daten** sichert Anspruchsgruppen die Hoheit über den gesamten Lebenszyklus von Daten. Ein Beispiel sind TK-Labels von Local Contexts, die Nutzungsbedingungen sichtbar machen.
- **Verantwortung** betont die Verantwortung von Forscher*innen und Institutionen, Risiken zu minimieren, Transparenz zu sichern und Rechenschaft abzulegen.
- **Ethik** verlangt einen Umgang, der über Rechtskonformität hinausgeht und auf Respekt, kulturelle Sensibilität und Schadensvermeidung zielt.

Die Prinzipien wurden 2018 im Rahmen der International Data Week entworfen und zwischen 2019 und 2020 von der Global Indigenous Data Alliance ausgearbeitet (Global Indigenous Data Alliance, o. J.; Carroll u. a. 2021). Gedächtnisinstitutionen agieren dabei zwangsläufig als Gatekeeper; Governance-Modelle sollen Entscheidungsrechte explizit abbilden. In Kanada fassen die OCAP®-Prinzipien Ownership, Control, Access und Possession als konkrete Form indigener Datenhoheit zusammen.

LOUD: Nutzungsorientierte Erweiterung

LOUD steht für Linked (verknüpft), Open (offen), Usable (nutzbar) Data (Daten) und schliesst die Kluft zwischen abstrakten Datenmodellen und Forschungspraxis:

- **Verknüpft** verlangt eindeutige Referenzen und semantische Verknüpfungen, damit Datensätze anschlussfähig in Wissensnetzen zirkulieren.
- **Offen** steht für rechtlich wie technisch barrierearme Bereitstellung, inklusive offener Formate und Schnittstellen.
- **Nutzbar** betont dokumentierte Provenienz, verständliche Zugänge, einfache Exporte und APIs sowie Qualitätssicherung, damit Forscher*innen ohne Spezialwerkzeuge arbeiten können.
- **Daten** rückt Inhalte in den Mittelpunkt und fordert wiederverwendbare Bereitstellung statt statischer Visualisierungen.

LOUD operationalisiert FAIR in Richtung Nutzung und bleibt mit CARE kompatibel, weil Nutzungsbedingungen und Rechte modelliert und kommuniziert werden.

Entscheiden, dokumentieren, umsetzen

Für offen lizenzierte Daten und breite Nachnutzung stehen FAIR und LOUD im Vordergrund; bei sensiblen Beständen bestimmen CARE und gegebenenfalls OCAP® die Zugangs- und Governance-Modelle. In Aggregations- und Vernetzungsprojekten zählen Interoperabilität und Verlinkung, ergänzt um klare Nutzungsregeln. In jedem Fall gilt: Zuständigkeiten, Bedingungen und Grenzen schriftlich fixieren, Identifier und Metadaten konsequent vergeben, Formate und Vokabulare standardisieren, Änderungen versionieren und ethische Anforderungen explizit adressieren.

Danksagung

Wir danken allen Personen und Institutionen, die zur Erarbeitung der ersten Auflage 2024 und zur laufenden Überarbeitung dieses Handbuchs beigetragen haben. Ihre Hinweise, Kritik und Praxisperspektiven haben Terminologie, Beispiele, Ontologie-Mapping, Versionierung und Leitlinien zu FAIR- und CARE-Prinzipien geschärft.

Besonderer Dank gilt Levyn Bürki für die substanziellen Beiträge zur Neustrukturierung und Weiterentwicklung des Handbuchs. Die Mitarbeit von September 2024 bis Juni 2025 umfasste zentrale konzeptionelle Entscheidungen und die massgebliche Ausarbeitung wesentlicher Inhalte.

Für die aufmerksame Lektüre und konstruktiven Rückmeldungen zur ersten Auflage 2024 danken die Autor*innen Eric Decker, Céline Hug, Lucie Kolb, Jonas Lendenmann, Noah Regenass und Stephanie Willi. Für inhaltliche Hinweise und Korrekturen nach der Veröffentlichung danken wir Esther Ernst-Mombelli (GND-Redaktion, Universitätsbibliothek Basel), Marc Bayard (GND-Redaktion, Universitätsbibliothek Bern), Philipp Messner (Plakatsammlung SfG Basel), Elias Zimmermann (Universität Zürich/Genf), Karin Lackner (Universitätsbibliothek Graz) sowie Roberta Flora Spano (ETH-Bibliothek Zürich, Sammlungen und Archive).

Das Handbuch profitierte von Diskussionen in folgenden Foren und Fachcommunitys: dem Roundtable “FAIR and CARE” der DARIAH-CH Study Days (22. November 2024, FHNW) mit Beiträgen von Iolanda Pensa, Elena Chestnova, Lucie Kolb und Linda Ludwig; dem Workshop “Grosse Anforderungen an kleine Textfelder – Ethische Fragen an Metadaten historischer Quellen” am Herder-Institut (21.–22. November 2024, Marburg), wo Noëlle Schnegg und Levyn Bürki Erfahrungen aus dem Handbuchprojekt vorstellten; dem Panel “Die unsichtbaren Anforderungen der digitalen Geschichtswissenschaft” an den 7. Schweizerischen Geschichtstagen (8.–11. Juli 2025, Universität Luzern) mit einem Impuls von Noëlle Schnegg und Beiträgen aus Gedächtnisinstitutionen wie Bundesarchiv, ETH-Bibliothek, histify, Personenportal SH, SAMARA, Nationalbibliothek, Wirtschaftsarchiv, SSRQ, swisstopo und Transcriptiones; der Einladung zum 9. Bibliothekskongress (24.–27. Juni 2025, Wien); sowie dem Panel “Diskriminierungssensible Metadaten für historische Sammlungen” an der Digital-Humanities-Konferenz 2025 (16. Juli 2025, Universidade NOVA de Lisboa) mit Beiträgen von Levyn Bürki, Joris Burla, Peggy Große, Mario Kliewer, Jonas Lendenmann, Lisa Quade, Moritz Mähr, Noëlle Schnegg und Elias Zimmermann. Für die Einladung zum SODa-Forum “Das Handbuch zur Erstellung diskriminierungsfreier Metadaten aus Perspektive der Universitätssammlungen” (11. September 2025), zum Workshop “Relativieren oder limitieren? Zum Umgang mit Dark Heritage in Sammlungen und Archiven” (13. und 14. November; Dresden) und zum Workshop “Metadaten in den Humanities” (4. und 5. Dezember, Zürich) danken wir den Organisator*innen; die dortigen Rückmeldungen fliessen in die fortlaufende Überarbeitung ein.

Das Handbuch ist ein “Living Document”. Wir laden die Communities der Digital Humanities, der Gedächtnisinstitutionen und der Forschung ein, weitere Hinweise, Praxisbeispiele und Korrekturen beizusteuern. Für verbleibende Unschärfen und Fehler tragen die Autor*innen die Verantwortung.

Literatur

/ *OpenGLAM Principles*. o. J. Zugegriffen 9. September 2025. https://openglam.org/how/.

Archives for Black Lives in Philadelphia. Anti-Racist Description Working Group. 2020. *Archives for Black Lives in Philadelphia: Anti-Racist Description Resources*. Archives for Black Lives in Philadelphia. https://archivesforblacklives.files.wordpress.com/2020/10/ardr_202010.pdf.

Ausgangslage: Inwieweit schützt das aktuelle Recht vor Diskriminierung? 2025. Antidiskriminierungsgesetz ADG - Dossier. Humanrights.ch. https://www.humanrights.ch/de/ipf/menschenrechte/diskriminierung/antidiskriminierungsgesetz/adg-ausgangslage/.

Baroncini, S., C. Buyken, J. Garzón Rodríguez, I. K. Marino, S. Oberbichler, und C. Petz. 2025. «From Access to Ethics: Data Discussions at the DH2025 Conference in Lisbon». Billet. *DH Lab*, August 1. https://doi.org/10.58079/14gdf.

Basta, Christine, Marta R. Costa-jussà, und Noe Casas. 2019. «Evaluating the Underlying Gender Bias in Contextualized Word Embeddings». arXiv:1904.08783. Preprint, arXiv, April 18. https://doi.org/10.48550/arXiv.1904.08783.

Berlin (Redaktion), Digital History. 2025. «Wie verändern große Sprachmodelle die Geschichtswissenschaften? Eine studentische Blogreihe zur Anwendung von LLMs in der historischen Forschung.» Billet. *Digital History Berlin*, Mai 9. https://doi.org/10.58079/13wcq.

Bowker, Geoffrey C., und Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. The MIT Press. https://doi.org/10.7551/mitpress/6352.001.0001.

Budde, Gunilla. 2008. «Quellen, Quellen, Quellen …». In *Geschichte.Studium - Wissenschaft - Beruf*, herausgegeben von Dagmar Freist und Hilke Günther-Arndt. Berlin. https://elearning.uni-oldenburg.de/sendfile.php?type=0&file_id=5435fe2c79f320a9d7465f0f61851ca5&file_name=budde_quellen-quellen-quellen.pdf&set_language=de_DE.

Buolamwini, Joy, und Timnit Gebru. 2018. «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification». *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 81: 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, und Shelley Stall. 2021. «Operationalizing the CARE and FAIR Principles for Indigenous Data Futures». *Scientific Data* 8 (1): 108. https://doi.org/10.1038/s41597-021-00892-0.

Caswell, Michelle. 2019. «Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal». *Archival Science*, Online-Vorab-Publikation. https://doi.org/10.1007/s10502-019-09316-w.

Chouldechova, Alexandra. 2017. «Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments». *Big Data* 5 (2): 153–63. https://doi.org/10.1089/big.2016.0047.

Collins, Patricia Hill. 2000. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. 2nd ed. Routledge.

Costa-Jussà, Marta R., Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, und Ksenia Kharitonova. 2022. «Interpreting gender bias in neural machine translation: Multilingual architecture matters». *Proceedings of the aaai conference on artificial intelligence* 36 (11): 11855–63. https://ojs.aaai.org/index.php/AAAI/article/view/21442.

Crenshaw, Kimberlé. 1989. «Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics». *University of Chicago Legal Forum* 1989 (1): 139–67. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

D'Ignazio, Catherine, und Lauren Klein. 2020. «Collect, Analyze, Imagine, Teach». In *Data Feminism*. The MIT Press. https://data-feminism.mitpress.mit.edu/pub/ei7cogfn.

Debora Nozza. o. J. «Debora Nozza». Zugegriffen 14. Juli 2025. https://www.deboranozza.com/.

Deutsche Nationalbibliothek. o. J. «Start der Erschließungsmaschine eMA». Zugegriffen 9. Juni 2025. https://jahresbericht.dnb.de/Webs/jahresbericht/DE/2022/Hoehepunkte/Erschliessungsmaschine/erschliessungsmaschine_node.html.

Erschließungsmaschine gestartet – blog.dnb.de. o. J. Zugegriffen 9. Juni 2025. https://blog.dnb.de/erschliessungsmaschine-gestartet/.

«Europeana Data Model. EDM Documentation». o. J. Zugegriffen 9. September 2025. https://pro.europeana.eu/page/edm-documentation.

Eynden, Veerle van den. 2011. *Managing and Sharing Data: Best Practice for Researchers*. 3rd ed., fully rev. UK Data Archive.

Flinn, Andrew. 2007. «Community Histories, Community Archives: Some Opportunities and Challenges». *Journal of the Society of Archivists* 28 (2): 151–76. https://doi.org/10.1080/00379810701611936.

forschungsdaten.info. 2024. «Metadaten und Metadatenstandards». April 25. https://forschungsdaten.info/themen/beschreiben-und-dokumentieren/metadaten-und-metadatenstandards/.

Frye, Marilyn. 1983. *The politics of reality: Essays in feminist theory*. Crossing Press.

Gabay, Simon, Tobias Hodel, Moritz Mähr, u. a. 2023. «Datenstandards für die historische Forschung – Ein White-Paper der Schweizerischen Gesellschaft für Geschichte». *Whitepaper*, Online-Vorab-Publikation, November 10. https://doi.org/10.5281/ZENODO.10122052.

Ghaddar, J. J., und Michelle Caswell. 2019. «‹To Go Beyond›: Towards a Decolonial Archival Praxis». *Archival Science* 19 (2): 71–85. https://doi.org/10.1007/s10502-019-09311-1.

Global Indigenous Data Alliance. o. J. «CARE Principles for Indigenous Data Governance». Zugegriffen 9. September 2025. https://www.gida-global.org/care.

Go FAIR Initiative. o. J. «FAIR Principles». Zugegriffen 22. September 2022. https://www.go-fair.org/wp-content/uploads/2022/01/FAIRPrinciples_overview.pdf.

Gooding, Paul, Melissa M. Terras, und Sarah Ames, Hrsg. 2025. *Library Catalogues as Data: Research, Practice and Usage*. Facet Publishing.

Gruber, Anastasia. 2022. «Vom Knüpfen feministischer Begriffsnetze: Ariadnes Faden & geschlechtersensible Normdaten». *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 75 (2). https://doi.org/10.31263/voebm.v75i2.7327.

Hart, Alexander, Graham Reid, Helena Hartmann, Charlotte R. Pennington, Marta Topor, und Flávio Azevedo. 2025. «Paradata (Paradaten)». Februar 7. https://forrt.org/glossary/german/paradata/.

Higgins, Sarah. 2008. «The DCC Curation Lifecycle Model». *International Journal of Digital Curation* 3 (1): 134–40. https://doi.org/10.2218/ijdc.v3i1.48.

Hitz, Benjamin, und Beat Stüdli. 2016. «Werkzeugkasten Geschichte. Online-Handbuch für Studium und wissenschaftliches Arbeiten & PDF-Ausgabe». https://dg.philhist.unibas.ch/de/studium/werkzeugkasten-geschichte/.

humanrights.ch. o. J. «Formen der Diskriminierung». Zugegriffen 8. September 2025. https://www.humanrights.ch/de/ipf/menschenrechte/diskriminierung/diskriminierungsverbot-dossier/juristisches-konzept/formen-der-diskriminierung/.

Jaillant, Lise, Claire Warwick, Paul Gooding, Katherine Aske, Glen Layne-Worthey, und J. Stephen Downie, Hrsg. 2025. *Navigating artificial intelligence for cultural heritage organisations*. UCL Press. https://doi.org/10.14324/111.9781800088375.

Jenkinson, Hilary. 1922. *A Manual of Archive Administration*. Clarendon Press.

Junger, Ulrike, und Frank Scholze. 2021. «Neue Wege und Qualitäten – Die Inhaltserschließungspolitik der Deutschen Nationalbibliothek». In *Qualität in der Inhaltserschließung*, herausgegeben von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl, und Hans Schürmann. De Gruyter Saur. https://doi.org/10.1515/9783110691597-004.

Kaiser, Katja, Ina Heumann, Tahani Nadim, u. a. 2023. *Promises of mass digitisation and the colonial realities of natural history collections*. 11 (Februar): 13–25.

Kleinberg, Jon, Sendhil Mullainathan, und Manish Raghavan. 2017. «Inherent Trade-Offs in the Fair Determination of Risk Scores». *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*. https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

König, Mareike. 2024. «ChatGPT und Co. in den Geschichtswissenschaften – Grundlagen, Prompts und Praxisbeispiele». Billet. *Digital Humanities am DHIP*, August 19. https://doi.org/10.58079/126eo.

Loukissas, Yanni Alexander. 2019. *All Data Are Local: Thinking Critically in a Data-Driven Society*. The MIT Press. https://doi.org/10.7551/mitpress/11543.001.0001.

Mähr, Moritz. 2022. «Konzept „Stadt.Geschichte.Basel digital“: Zusammenfassung des Konzepts (April 2022) und des Zwischenberichts (Dezember 2022)». April. https://dokumentation.stadtgeschichtebasel.ch/products/interna/digitalkonzept/.

Mähr, Moritz. (2025) 2025. *Dublin Core Metadata Enhancer*. Markdown. August 30, released August 31. https://github.com/Stadt-Geschichte-Basel/dublin-core-metadata-enhancer.

Mähr, Moritz, und Moritz Twente. 2025. «Seeing History Unseen: Evaluating Vision-Language Models for WCAG-Compliant Alt-Text in Digital Heritage Collections». *Anthology of Computers and the Humanities* 3 (November): 1148–68. https://doi.org/10.63744/njQVYcLndSPE.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, und Aram Galstyan. 2021. «A Survey on Bias and Fairness in Machine Learning». *ACM Computing Surveys* 54 (6): 115:1–35. https://doi.org/10.1145/3457607.

Metadata Resources | Steven Jack Miller. o. J. Zugegriffen 7. September 2025. https://sites.uwm.edu/mll/metadata-resources/.

Nayak, Shravan, Kanishk Jain, Rabiul Awal, u. a. 2024. «Benchmarking Vision Language Models for Cultural Understanding». arXiv:2407.10920. Preprint, arXiv, Oktober 14. https://doi.org/10.48550/arXiv.2407.10920.

Net, Francesc, Marc Folia, Pep Casals, Andrew D. Bagdanov, und Lluis Gomez. 2024. «EUFCC-340K: A Faceted Hierarchical Dataset for Metadata Annotation in GLAM Collections». arXiv:2406.02380. Preprint, arXiv, Juni 4. https://doi.org/10.48550/arXiv.2406.02380.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5.

Open a GLAM Lab. 2019. 1. Aufl. PubPub. https://doi.org/10.21428/16ac48ec.f54af6ae.

Research Data Alliance International Indigenous Data Sovereignty Interest Group. 2019. «CARE Principles for Indigenous Data Governance». September 6. https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5d79c383e904c741c9e9cd86/1568260995760/CARE+Principles+for+Indigenous+Data+Governa

«Rethinking Collections as Data». 2019. *Open a Glam Lab*, Oktober 30. https://glamlabs.pubpub.org/pub/urhm68yr/release/1.

Rodgers, Dennis, und Bruce O’Neill. 2012. «Infrastructural violence: Introduction to the special issue». *Ethnography* 13 (4): 401–12. https://doi.org/10.1177/1466138111435738.

Rottmann, Andrea. 2023. «Que/e/r zur Ordnung des Archivs? Queere Geschichte in staatlichen Archiven erforschen». *Archiv - Theorie & Praxis* 76 (4): 21–24. https://www.archive.nrw.de/sites/default/files/media/files/Archiv.theoriepraxis23-4-Internet.pdf.

Schopper, Daniel. 2024. «Einführung in Metadaten und Dokumentation». DARIAH Campus. https://campus.dariah.eu/resources/hosted/einfuehrung-metadaten.

Schweiz, Verband der Museen der. o. J. «Webdossier: KI im Museum». museums.ch. Zugegriffen 7. September 2025. https://www.museums.ch/de/unser-engagement/mitteilungen/aktuelles/webdossier-ki-im-museum-6966.html.

Schweizerisches Nationalmuseum, Chonja Lee, Monique Ligtenberg, u. a. 2025. *Rassismuskritische Dokumentation. Ein Handbuch für GLAM-Institutionen*. Zenodo. https://doi.org/10.5281/zenodo.17657486.

Sparber, Sandra. 2016. «What’s the frequency, Kenneth? – Eine (queer)feministische Kritik an Sexismen und Rassismen im Schlagwortkatalog». *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 69 (2): 216–35. https://doi.org/10.31263/voebm.v69i2.1629.

Spivak, Gayatri Chakravorty. 1988. «Can the subaltern speak?» In *Marxism and the interpretation of culture*, herausgegeben von Cary Nelson und Lawrence Grossberg. University of Illinois Press.

Stadt.Geschichte.Basel. 2025. «Datendokumentation. Forschungsdatenplattform Stadt.Geschichte.Basel». https://forschung.stadtgeschichtebasel.ch/data.html.

Stanczak, Karolina, und Isabelle Augenstein. 2021. «A Survey on Gender Bias in Natural Language Processing». arXiv:2112.14168. Preprint, arXiv, Dezember 28. https://doi.org/10.48550/arXiv.2112.14168.

The Endings Project Team. 2023. «Endings Principles for Digital Longevity». März 3. https://endings.uvic.ca/principles.html.

Tóth-Czifra, Erzsébet. 2019. «DARIAH Pathfinder to Data Management Best Practices in the Humanities». DARIAH-Campus, DARIAH-Campus, Mai 3. https://campus.dariah.eu/resources/pathfinders/dariah-pathfinder-to-data-management-best-practices-in-the-humanities.

Trognitz, Martina. 2024a. «Grundlagen des Datenmanagements». DARIAH-Campus, DARIAH-Campus, September 16. https://campus.dariah.eu/resources/hosted/grundlagen-datenmanagement.

Trognitz, Martina. 2024b. «Langzeitarchivierung». DARIAH-Campus, DARIAH-Campus, November 21. https://campus.dariah.eu/resources/hosted/langzeitarchivierung.

Weitzmann, John H., und Paul Klimpel. 2016. *Rechtliche Rahmenbedingungen für Digitalisierungsprojekte von Gedächtnisinstitutionen*. Manual. 3. Auflage, Mai 2016 – Nachdruck Dezember 2016. digiS – Servicestelle Digitalisierung Berlin, Zuse Institute Berlin / iRights.Law – Anwälte für die digitale Welt. https://doi.org/10.12752/2.0.002.2.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, u. a. 2016. «The FAIR Guiding Principles for Scientific Data Management and Stewardship». *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Young, Iris Marion. 1990. *Justice and the politics of difference*. Princeton University Press.

Anhang

Checkliste

1. Planung und Konzeption
   1.1 Zielsetzung klären

- [ ] Ziel, Zielgruppen, Nutzungsszenarien dokumentiert
- [ ] Prioritäten und Mindeststandards definiert
- [ ] Pilotkorpus erfasst und Aufwand gemessen
- [ ] Sensible Inhalte antizipiert und gekennzeichnet

1.2 Ethische und rechtliche Rahmenbedingungen

- [ ] Rechtslage pro Inhaltstyp dokumentiert und geklärt
- [ ] Einwilligungen und Einschränkungen erfasst
- [ ] Personenbezug und Schutzmassnahmen geprüft
- [ ] Lizenz/ rights statement festgelegt
- [ ] Review- und Eskalationswege definiert

1.3 Standards und Infrastruktur festlegen

- [ ] Zentrales Schema + Erweiterungen dokumentiert
- [ ] Normdaten/Vokabulare ausgewählt
- [ ] Plattform, Exportformate und persistente Identifikatoren (PIDs ; zum Beispiel DOI) definiert

2. Datensammlung und Quellenkritik
   2.1 Primärerschliessung vs. Nachnutzung

- [ ] Re-Use-Quellen bewertet und zitiert
- [ ] Übernahmen, Streichungen, Korrekturen protokolliert
- [ ] Bias-Befunde und Lücken sichtbar gemacht

2.2 Kontextualisierung der Quellen

- [ ] Entstehungs- und Nutzungskontext beschrieben
- [ ] Relevante Akteur*innen und Machtverhältnisse benannt
- [ ] Rezeption und Deutungskonflikte referenziert

3. Datenverarbeitung und Anreicherung
   3.1 Technische Standards implementieren

- [ ] Feldkatalog mit Definitionen erstellt
- [ ] Maschinenlesbare Datentypen und Validierung gesetzt
- [ ] Sichtbarkeits- und Sensitivitätslogik modelliert

3.2 Beschreibung und Verschlagwortung

- [ ] Vokabulare und Normdateien pro Feld definiert
- [ ] Regeln für Intrinsisch/Extrinsisch festgelegt
- [ ] Unsicherheit formalisiert und sichtbar gemacht

3.3 KI-Unterstützung und Automatisierung

- [ ] KI-Einsatzbereiche, Modelle, Trainings-/Prompt-Kontexte dokumentiert
- [ ] Stufen der Qualitätskontrollen, Freigaben, Rückverfolgbarkeit definiert
- [ ] KI-Anteile in Feldern gekennzeichnet
- [ ] Risiko- und Bias-Assessments durchgeführt und dokumentiert
- [ ] Datenschutz, Urheberrecht und FAIR / CARE -Prinzipien geprüft

- [ ] Kontrollschwellen (zum Beispiel Confidence-Scores) für menschliche Überprüfung festgelegt
- [ ] Rollen, Eskalationsverfahren und Verantwortlichkeiten definiert
- [ ] Modelle, Prompts und Updates versioniert und veröffentlicht
- [ ] Mitarbeitende geschult und fortlaufend sensibilisiert

4. Speicherung und Verwaltung
   - [ ] Repositorium und persistente Identifikatoren (PIDs) gewählt
   - [ ] Data Dictionary publiziert
   - [ ] Versionierungsstrategie definiert und technisch verankert
5. Veröffentlichung und Zugang

   5.1 Zielgruppe definieren
   - [ ] Zielgruppen, Einstiegspfade und Zugangsebenen definiert
   - [ ] Statischer Endzustand mit Rückwärtsplanung festgelegt
   - [ ] Kanonische, parameterlose URLs oder DOIs vergeben
   - [ ] Editionen versioniert, Weiterleitungen bei Umzügen vorgesehen
   - [ ] Zitierempfehlungen sichtbar integriert

   5.2 Umgang mit sensiblen und diskriminierenden Inhalten
   - [ ] Sensible Daten identifiziert und Schutzmassnahmen getroffen
   - [ ] Diskriminierende Inhalte benannt und kontextualisiert
   - [ ] Content-Notes formuliert
   - [ ] Einwilligungsfragen reflektiert und Alternativen geprüft
   - [ ] Governance-Regeln dokumentiert und Änderungen nachvollziehbar protokolliert

   5.3 Technische Strategien
   - [ ] Statischer Webauftritt ohne serverseitige Abhängigkeiten definiert
   - [ ] Build- und Validierungsprozess unter Versionskontrolle implementiert
   - [ ] Sichtbarkeitssteuerung in Export- oder Ableitungsschritten umgesetzt
   - [ ] Offene, verlustfreie Formate mit UTF-8-Kodierung verwendet
   - [ ] README.md und Referenzdokumente bereitgestellt und versioniert
   - [ ] API-Exporte abgesichert und reproduzierbare Datenbeilage integriert
   - [ ] Opt-in-Mechanismus für sensible Inhalte verfügbar

   5.4 Transparente Dokumentation
   - [ ] Publikationsdokumentation online und öffentlich zugänglich
   - [ ] Editionen mit Datum, Version und Zitierempfehlung publiziert
   - [ ] Persistente URIs stabil, Umzüge mit Weiterleitungen gesichert
   - [ ] Jede Edition als Paket mit Prüfsummen, Daten und Dokumentation archiviert
6. Nachnutzung und Wiederverwendung
   - [ ] Linked-Data-Exporte, PIDs und Crosswalks verfügbar
   - [ ] Zitations- und Lizenzrichtlinien veröffentlicht
   - [ ] API oder Dumps versioniert zugänglich
   - [ ] Feedback-Prozess etabliert, Nutzung dokumentiert
7. Archivierung und Löschung
   - [ ] Archivpakete mit dokumentierten Provenienzen und Prüfsummen erstellt
   - [ ] Lösch- und Rückgabeprozesse rechtlich und organisatorisch festgelegt
   - [ ] Tombstones und Änderungsprotokolle veröffentlicht
   - [ ] Zugriffskontrollen, Verschlüsselung und Logs aktiv
   - [ ] Redundanz- und Notfallpläne etabliert, Kosten langfristig kalkuliert

Glossar

Term	Definition

AAT	Art & Architecture Thesaurus ist ein strukturiertes Vokabular des Getty Research Institute für Kunst-, Architektur- und Kulturgeschichte. Bietet hierarchisch organisierte Begriffe und mehrsprachige Synonyme für die konsistente Beschreibung kultureller Objekte.
API	Application Programming Interfaces sind Schnittstellen zwischen verschiedenen Softwareanwendungen, die den strukturierten Datenaustausch ermöglichen. In der Metadatenpraxis wichtig für die automatisierte Übertragung und Synchronisation von Datenbeständen.
CARE	Ethische Grundsätze für Forschungsdaten indigener und marginalisierter Gemeinschaften. Collective Benefit (Kollektiver Nutzen), Authority to Control (Kontrollbefugnis), Responsibility (Verantwortung), Ethics (Ethik). Ergänzen die FAIR-Prinzipien um Aspekte der Datensouveränität.
CC	Creative Commons (CC) ist ein standardisiertes Lizenzsystem mit modularen Bedingungen (zum Beispiel BY, SA, NC, ND) zur rechtssicheren Freigabe und Nachnutzung von Inhalten und Daten.
CER	Character Error Rate ist die Fehlerrate bei der automatischen Texterkennung (OCR/HTR), gemessen als Anteil falsch erkannter Zeichen. Wichtige Metrik zur Bewertung der Qualität digitalisierter Texte.
CHANGELOG.md	Datei zur nachvollziehbaren Dokumentation von Änderungen zwischen Versionen eines Projekts.
CIDOC CRM	Conceptual Reference Model ist ein internationaler Standard (ISO 21127) für die konzeptuelle Modellierung von Kulturerbe-Informationen. Ermöglicht die semantische Verknüpfung heterogener Datenquellen aus Museen, Archiven und Bibliotheken.
CITATION.cff	Standardisierte Metadatendatei für Zitationshinweise in Software- und Datenprojekten.
Content-Note	Hinweisfeld zur ergänzenden Beschreibung und Kontextualisierung des Inhalts einer Ressource. Dient der Auffindbarkeit sowie der informierten Nutzung, insbesondere bei potenziell belastenden oder diskriminierenden Inhalten. In diesem Handbuch verwenden wir bevorzugt “Content Note” (Hinweis) statt “Content Warning”.
Content-Warning	Begriff für Hinweise auf sensible oder potenziell belastende Inhalte. In diesem Handbuch verwenden wir stattdessen “Content Note”.
Crosswalk	Verfahren zur Zuordnung von Datenelementen zwischen verschiedenen Metadatenstandards. Ermöglicht die Migration und Interoperabilität von Metadaten.
CSS	Cascading Style Sheets ist eine Stylesheet-Sprache zur Gestaltung von Webseiten.
CSV	Comma-Separated Values ist ein einfaches, textbasiertes Dateiformat zur strukturierten Speicherung tabellarischer Daten. Weit verbreitet für den Datenaustausch zwischen verschiedenen Systemen.
Data Dictionary	Strukturierte Dokumentation aller Datenfelder eines Datensatzes. Enthält Feldnamen, Definitionen, Datentypen, zulässige Werte und Hinweise zur Verwendung. Siehe https://library.ucmerced.edu/data-dictionaries
Data Sheets	Strukturierte Begleitdokumente für Datensätze mit Angaben zu Entstehung, Qualität, Grenzen und Nutzung.
DCMES	Dublin Core Metadata Element Set ist das Basisschema der Dublin Core Metadata Initiative mit 15 Kernelementen zur Beschreibung digitaler Ressourcen.
DCTerms	Dublin Core Terms ist die erweiterte Menge von Dublin-Core-Begriffen (Properties und Klassen), die über das DCMES hinaus differenzierte Metadatenmodellierung ermöglicht.
Deprecation-Policy	Richtlinie, wie Funktionen, Schnittstellen oder Datenstrukturen als veraltet markiert und schrittweise abgelöst werden.
DOI	Digital Object Identifier ist ein persistenter Identifikator für digitale Objekte wie Publikationen, Datensätze oder Software-Releases. Ein DOI verweist auf eine stabile Referenz (Resolver) und unterstützt Zitierbarkeit und Nachvollziehbarkeit.
Dublin Core	Weit verbreiteter Metadatenstandard mit 15 Kernelementen (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights) zur grundlegenden Beschreibung digitaler Ressourcen.
Dump	Vollständiger oder teilweiser Export eines Datenbestands in ein übertragbares Format.
EAD	Encoded Archival Description ist ein XML-Standard für die strukturierte Beschreibung von Archivbeständen. Ermöglicht hierarchische Darstellung von Findmitteln und unterstützt komplexe Provenienz- und Kontextinformationen.
EDM	Europeana Data Model ist das Datenmodell der Europeana für die Aggregation und Bereitstellung europäischer Kulturobjekte. Basiert auf Linked Data-Prinzipien und ermöglicht semantische Verknüpfungen zwischen verschiedenen Sammlungen.
EDTF	Extended Date/Time Format ist ein Standard für die Darstellung unsicherer, approximativer oder komplexer Datumsangaben in historischen Quellen. Unterstützt Zeiträume, Unsicherheitsmarkierungen und alternative Kalendersysteme.
Entity-Resolution	Automatisiertes Verfahren zur Identifikation und Zusammenführung von Datensätzen, die sich auf dieselbe Entität (Person, Ort, Organisation) beziehen. Zentral für die Erstellung konsistenter Normdaten.
FAIR	Grundsätze für Forschungsdatenmanagement. Findable (Auffindbar), Accessible (Zugänglich), Interoperable (Interoperabel), Reusable (Nachnutzbar). Fördern die offene und nachhaltige Nutzung wissenschaftlicher Daten.
Fine-Tuning	Anpassung vortrainierter Machine Learning-Modelle an spezifische Datensätze oder Aufgaben. In der Metadatenpraxis relevant für OCR/HTR-Verbesserung oder Named Entity Recognition.
Gazetteer	Kontrolliertes Ortsverzeichnis mit standardisierten Ortsnamen, Variantenformen, Georeferenzen und Identifikatoren.
GenderOpen	Fachspezifischer Schlagwortindex und Suchinstrument für Geschlechterforschung und Gleichstellungsfragen.
Git	Verteiltes Versionskontrollsystem zur Nachverfolgung von Änderungen an Dateien und Projekten.
GND	Gemeinsame Normdatei ist eine kooperativ gepflegte Normdatei deutschsprachiger Bibliotheken für Personen, Körperschaften, Kongresse, Geografika, Sachschlagwörter und Werktitel. Zentrale Infrastruktur für die Vernetzung bibliographischer Daten.
HTML	HyperText Markup Language ist die Standard-Auszeichnungssprache für Inhalte im Web.
HTR	Handwritten Text Recognition ist die automatische Erkennung handschriftlicher Texte mittels Machine Learning. Spezialisierte Weiterentwicklung der OCR-Technologie für historische Handschriften und Dokumente.
HTTP	Hypertext Transfer Protocol ist das grundlegende Protokoll für den Transfer von Webressourcen. HTTP(S)-auflösbare Identifikatoren sind zentral für verlinkbare und maschinenlesbare Metadaten.
IIIF	International Image Interoperability Framework sind Standards für die interoperable Bereitstellung von Bildern und audiovisuellen Medien. Ermöglicht einheitliche Darstellung, Annotation und Manipulation digitaler Objekte unabhängig vom Hosting-System.
ISO	International Organization for Standardization ist eine internationale Organisation für Normung, die technische Standards für verschiedene Bereiche entwickelt, einschliesslich Informationsmanagement und Dokumentation.
JavaScript	Programmiersprache für interaktive Webanwendungen und dynamische Inhalte.
JSON	JavaScript Object Notation ist ein leichtgewichtiges, textbasiertes Datenformat für den strukturierten Datenaustausch. Weit verbreitet in Web-APIs und zunehmend auch für Metadaten verwendet.

JSON-LD	JSON for Linking Data ist eine Erweiterung von JSON für Linked Data-Anwendungen. Ermöglicht die semantische Annotation von JSON-Strukturen durch Verweise auf kontrollierte Vokabulare.
LCSH	Library of Congress Subject Headings ist ein umfangreiches, hierarchisch strukturiertes Schlagwortsystem der Library of Congress. Internationaler Standard für die sachliche Erschliessung in Bibliotheken.
LIDO	Lightweight Information Describing Objects ist ein XML-Schema für die Beschreibung von Museumsobjekten. Ermöglicht detaillierte Erfassung von Provenienz, Materialien, Techniken und kulturhistorischen Kontexten.
LOD	Linked Open Data bezeichnet offen lizenzierte, über URIs verknüpfte Daten im Web. LOD baut auf RDF und standardisierten Vokabularen auf und fördert Interoperabilität, Nachnutzbarkeit und semantische Verknüpfung.
LOUD	Linked Open Usable Data ist eine Weiterentwicklung der Linked Data-Prinzipien mit Fokus auf praktische Nutzbarkeit. Betont entwicklerfreundliche APIs, JSON-LD-Format und ausführliche Dokumentation.
MAR/MNAR	Kategorien für fehlende Daten: MAR (Missing At Random) und MNAR (Missing Not At Random).
MARC	Machine-Readable Cataloging ist ein internationaler Standard für die maschinelle Verarbeitung bibliographischer Daten. Grundlage der meisten Bibliothekskataloge, zunehmend durch flexiblere RDF-basierte Formate ergänzt.
MD	Dateiendung für Markdown-Dateien. Markdown ist eine leichtgewichtige Auszeichnungssprache für textbasierte Dokumente.
METS	Metadata Encoding and Transmission Standard ist ein XML-Schema für die Strukturierung und Verknüpfung von Metadaten digitaler Objekte. Weit verbreitet in digitalen Bibliotheken und Archiven.
Minimal Computing	Ansatz für digitale Praxis mit möglichst geringem Ressourcenverbrauch.
Model Cards	Strukturierte Dokumente zur Beschreibung von KI-Modellen, einschliesslich Einsatzzweck, Leistungsgrenzen und Risiken.
MODS	Metadata Object Description Schema ist ein bibliothekarischer Metadatenstandard auf XML-Basis. Bietet reichhaltigere Beschreibungselemente als DCMES bei guter Interoperabilität mit bibliographischen Standards.
Named-Entity-Recognition	Automatisierte Identifikation und Klassifikation von Entitäten (Personen, Orte, Organisationen) in Texten. Wichtig für die Anreicherung von Metadaten und die Verknüpfung mit Normdaten.
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting ist ein Protokoll für die automatisierte Sammlung von Metadaten aus verteilten Repositorien. Ermöglicht die Aggregation von Metadaten ohne Duplikation der eigentlichen Inhalte.
OAIS	Open Archival Information System ist ein Referenzmodell für digitale Langzeitarchivierung (ISO 14721). Definiert Rollen, Funktionen und Informationsmodelle für vertrauenswürdige digitale Repositorien.
OCAP	Principles for Indigenous Data Governance: Ownership, Control, Access und Possession. Beschreibt Anspruchsrechte und Kontrollmöglichkeiten indigener Gemeinschaften über ihre Daten.
OCR	Optical Character Recognition ist die automatische Erkennung von Zeichen in digitalisierten Texten. Grundlage für die Volltextsuche in historischen Dokumenten, unterliegt aber systematischen Fehlern bei historischen Schriften.
Ontologie	Formale, explizite Spezifikation von Konzepten und Beziehungen in einem bestimmten Wissensbereich. In der Metadatenpraxis wichtig für die semantische Modellierung und Interoperabilität von Daten.
Open Data	Daten, die rechtlich und technisch so bereitgestellt werden, dass sie frei genutzt, weiterverarbeitet und geteilt werden können.
ORCID	Open Researcher and Contributor ID ist ein persistenter Identifikator für Forschende. ORCID-IDs helfen, Personen eindeutig zuzuordnen und Publikationen sowie Forschungsdaten mit Autor*innen zu verknüpfen.
PDF-A	ISO-standardisiertes Dateiformat für die Langzeitarchivierung von PDF-Dokumenten.
PID	Persistenter Identifikator für digitale Objekte, Personen oder Entitäten.
Pipeline	Abfolge automatisierter Verarbeitungsschritte zur Transformation, Anreicherung oder Analyse von Daten. In der Metadatenpraxis häufig genutzt für ETL-Prozesse (Extract, Transform, Load).
PNG	Portable Network Graphics ist ein verlustfreies Rasterbildformat.
Precision	Anteil relevanter Treffer unter allen gefundenen Treffern.
PREMIS	Preservation Metadata Implementation Strategies ist ein Datenmodell für Bewahrungsmetadaten in digitalen Archiven. Dokumentiert Herkunft, Authentizität, Bewahrungsaktivitäten und technische Eigenschaften digitaler Objekte.
PURL	Persistent URL ist ein dauerhafter Web-Identifikator, der über Weiterleitungen auf wechselnde Zieladressen verweist. PURLs unterstützen stabile Referenzen, auch wenn sich die konkrete Ressourcen-URL ändert.
Query-Expansion	Automatische Erweiterung von Suchanfragen durch semantisch verwandte Begriffe. Verbessert die Vollständigkeit von Suchergebnissen, besonders bei heterogenen Metadaten.
Ranking	Anordnung von Suchergebnissen nach einer Relevanzfunktion.
RDF	Resource Description Framework ist ein Standard des World Wide Web Consortium für die strukturierte Beschreibung von Ressourcen. Grundlage des Semantic Web und von Linked Data-Anwendungen.
README.md	Einführende Projektdokumentation im Wurzelverzeichnis eines Repositoriums.
Recall	Anteil gefundener relevanter Treffer an allen tatsächlich relevanten Treffern.
Retrieval	Prozess der Suche und des Abrufs relevanter Informationen aus Datenbeständen. In der Metadatenpraxis entscheidend für die Nutzerzufriedenheit und Auffindbarkeit von Ressourcen.
Retrieval-System	Technisches System zum Auffinden und Abrufen von Informationen aus Datenbeständen.
rights statement	Standardisierte Rechteaussage zur Kennzeichnung von Nutzungsstatus, Zugangsbedingungen oder rechtlichen Einschränkungen digitaler Objekte. Häufig umgesetzt über kontrollierte URI-Sets wie RightsStatements.org.
Schema.org	Vokabular für strukturierte Daten im Web. Erleichtert die maschinenlesbare Auszeichnung von Inhalten und deren Auffindbarkeit in Suchsystemen.
SKOS	Simple Knowledge Organization System ist ein Standard für die Repräsentation von Klassifikationssystemen, Thesauri und anderen kontrollierten Vokabularen als Linked Data. Ermöglicht die Verknüpfung verschiedener Begriffssysteme.
SVG	Scalable Vector Graphics ist ein XML-basiertes Vektorgrafikformat.
TEI	Text Encoding Initiative ist ein internationaler Standard für die digitale Kodierung literarischer und historischer Texte. Basiert auf XML und ermöglicht detaillierte strukturelle und inhaltliche Annotation.
TIFF	Tagged Image File Format ist ein häufig verlustfreies Rasterbildformat für hochwertige Digitalisierung und Langzeitarchivierung.

TK-Labels	Traditional Knowledge Labels sind maschinen- und menschenlesbare Hinweise auf kulturelle Rechte, Kontexte und Einschränkungen, insbesondere für indigenes Wissen. Sie werden unter anderem von Local Contexts bereitgestellt.
Topic Modeling	Computergestützte Methode zur automatischen Identifikation thematischer Muster in Textsammlungen. Hilfreich für die Analyse grosser Korpora, erfordert aber kritische Interpretation der Ergebnisse.
Turtle	Terse RDF Triple Language ist ein kompaktes, menschenlesbares Serialisierungsformat für RDF-Daten. Weit verbreitet in Linked-Data-Workflows und für den Austausch semantischer Metadaten.
URI	Uniform Resource Identifier ist ein standardisierter Bezeichner für Ressourcen im Web. URIs ermöglichen eindeutige Referenzen auf Objekte, Konzepte und Relationen in Metadaten und Linked-Data-Systemen.
UTF-8	Unicode-Transformation-Format mit 8-Bit-Einheiten; Standard für Zeichenkodierung im Web und in vielen Datenformaten.
VIAF	Virtual International Authority File ist eine internationale Normdatei, die Normdaten von Nationalbibliotheken weltweit verknüpft. Ermöglicht die eindeutige Identifikation von Personen und Körperschaften über Sprachgrenzen hinweg.
VRA Core	Visual Resources Association Core ist ein Metadatenstandard für die Beschreibung visueller Kulturobjekte. Speziell entwickelt für Bilder, Kunstwerke und architektonische Objekte.
WER	Word Error Rate ist die Fehlerrate bei der automatischen Texterkennung, gemessen als Anteil falsch erkannter Wörter. Ergänzt die Character Error Rate (CER) bei der Bewertung von OCR/HTR-Qualität.
Workflow	Systematische Abfolge von Arbeitsschritten in der Datenverarbeitung. In der Metadatenpraxis wichtig für Qualitätssicherung und Konsistenz.
XML	eXtensible Markup Language ist eine strukturierte Auszeichnungssprache für die plattformunabhängige Darstellung hierarchisch organisierter Daten. Grundlage vieler Metadatenstandards wie METS, EAD und TEI.

Handbücher und Leitfäden

Dieser Entscheidungsbaum hilft bei der Orientierung der Kapitel und verlinkt auf externe Ressourcen.

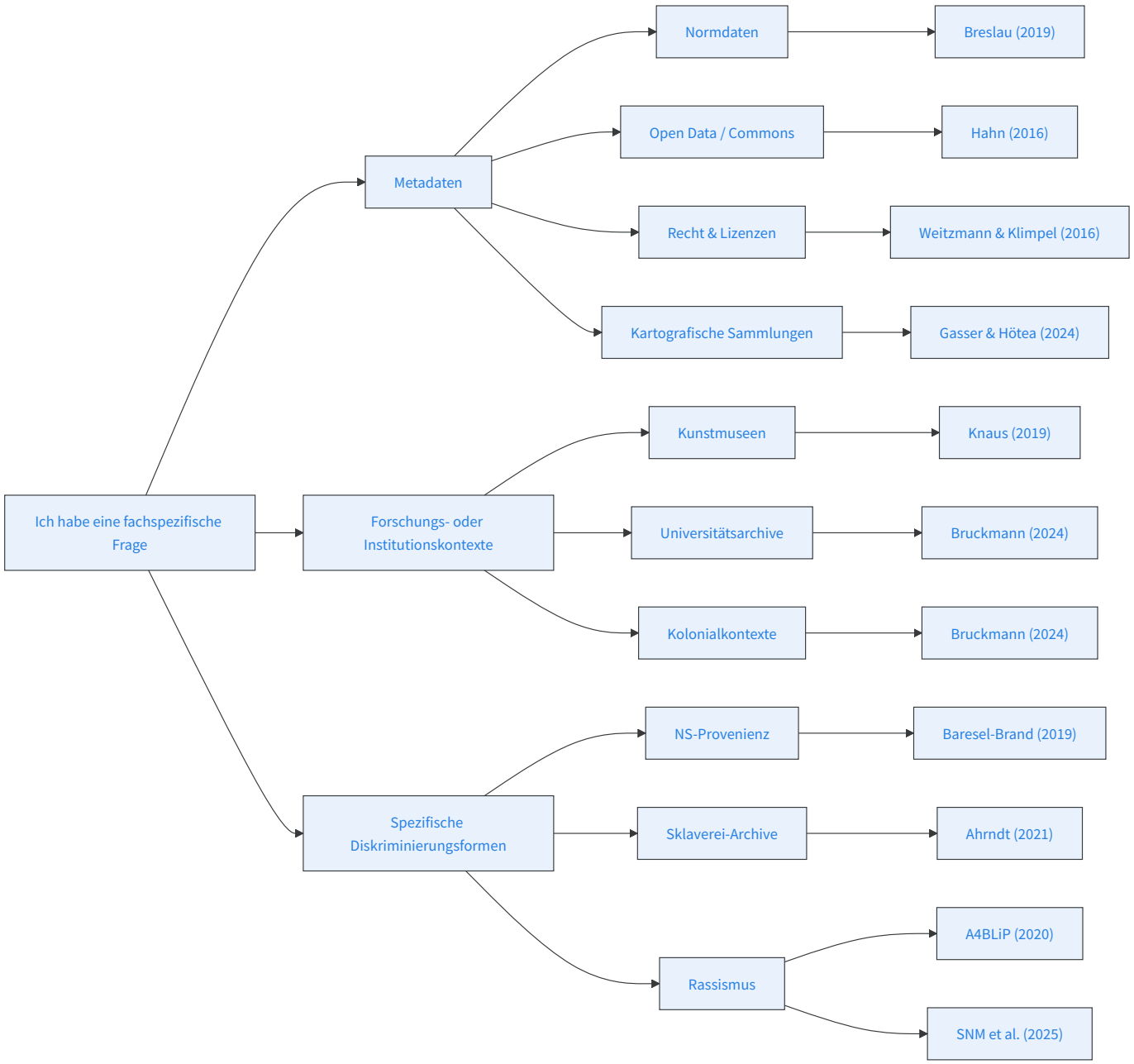


Abbildung 3: Hilfestellungen für weiterführende Fragen, die über die behandelten Themen des Handbuchs hinausgehen.

Metadatenstandards: sieben Ebenen

Dieser Anhang ergänzt Kapitel 4.3.3 und entfaltet die sieben Ebenen oder Typen von Metadatenstandards anhand des Beispiels *Knie Völkerschau*.

1. Datentypstandard *(Feld- oder Attributniveau)*

Legt fest, **in welchem Format einzelne Werte** codiert oder dargestellt sein müssen – zum Beispiel Zahlen, Datumsangaben, Zeichenketten, Booleans (Wahrheitswerte wie wahr oder falsch beziehungsweise 1 und 0).

Am Beispiel *Knie Völkerschau* zeigt sich, wie einzelne Werte formal typisiert werden:

- `dc:title` ist ein einfacher Textstring in Originalsprache → `rdf:langString` mit Sprachtag `de`
- `dc:date` enthält den unsicheren Jahreswert "1969?", typisiert als `xsd:string`, könnte aber in normierten Fällen auch `xsd:gYear` sein
- `edm:language` verwendet ISO 639-1 Codes ("da" für Dänisch)
- `dc:identifier` oder `edm:isShownBy` sind URIs → `anyURI`

2. Wertstandard *(Zulässige Werte für Felder)*

Enthält normierte, kontrollierte Begriffe oder Referenzen. Diese dienen der **Vergleichbarkeit, Suche und Aggregation**.

Das Beispiel *Knie Völkerschau* nutzt sowohl kontrollierte als auch freie Werte:

- `dc:type` enthält "Still image", "Poster" – konform mit dem **AAT** oder **EDM Type Vocabulary**
- `edm:country` ist `DK` (*Denmark*) → entspricht **ISO 3166-1 Alpha-2**
- `dc:rights` verweist auf die **CC-Lizenz** `http://creativecommons.org/licenses/by-nc-nd/4.0/`
- `dc:subject` ("Knie", "Cirkus", "Håndværker") könnte auf eine kontrollierte Vokabularquelle wie **GND** oder **Wikidata** gemappt werden

3. Inhaltsstandard *(Semantische Bedeutung von Feldern)*

Legt fest, **welche Informationen in welchen Feldern** erfasst werden sollen und was sie **semantisch bedeuten**. Dabei kann folgende Frage gestellt werden: "Wie befülle ich das Feld?"

Die Belegung der Felder des Beispiels *Knie Völkerschau* folgt dem **DCMES** und dem **Europeana Data Model (EDM)**:

- `dc:title` enthält den Titel des Plakats ("Knie Völkerschau")
- `dc:description` beschreibt Motiv, Technik und Masse der Lithografie
- `dc:subject` benennt thematische Schlagwörter
- `dc:rights` informiert über urheberrechtliche Bedingungen
- `edm:isShownBy` verweist auf das direkt eingebettete Digitalisat

4. Strukturstandard / Schema *(Datenmodellierung)*

Definiert **die erlaubten Felder und ihre Beziehungen**, oft in hierarchischen oder relationalen Strukturen. Ein Schema kann sowohl inhaltlich als auch formal spezifiziert sein.

Oft wird zwischen Strukturstandard und Schema nicht klar unterschieden. Während der Strukturstandard die allgemeine Anordnung und Bedeutung der Felder beschreibt, legt ein Schema konkret fest, wie diese umgesetzt werden; etwa welche Felder verpflichtend sind oder wie oft sie vorkommen dürfen. Ein Schema lässt sich als Bauplan verstehen, der auf einem übergeordneten, abstrakten Modell beruht.

Die Datenstruktur des Beispiels *Knie Völkerschau* entspricht dem **Europeana Data Model (EDM)**:

- `ProvidedCHO` (Cultural Heritage Object) enthält zum Beipsiel `dc:title`, `dc:date`, `dc:type`
- `Aggregation` verknüpft mit `edm:isShownBy`, `edm:preview`, `edm:provider`
- `Proxy` erlaubt Mehrsprachigkeit, etwa in `dc:description` (da) und `dc:title` (de)

5. Formatstandard *(maschinenlesbare Serialisierung & Kodierung)*

Legt fest, **wie der strukturierte Metadatensatz technisch gespeichert oder übertragen** wird. Es ist die konkrete **Kodierung und Serialisierung** der Metadatenstruktur, wie sie vom Computer ausgelesen wird. Der Fokus liegt auf der einfachen und effizienten Lesbarkeit für Maschinen, nicht für Menschen.

Anmerkung: XML oder RDF sind **Struktur- und Formatstandards zugleich**, abhängig davon, ob man ihre **logische Modellierung** oder die **Serialisierung** betont.

Im Fall von *Knie Völkerschau* liegen die Metadaten folgenden Formaten vor:

- Europeana-API liefert Metadaten im **JSON-LD** (zum Beispiel via `/record/v2/...json`)
- Exportierbar als **RDF**/**XML** oder **Turtle**
- URIs sind **HTTP-resolvable**, kodiert in **UTF-8**

6. Präsentationsstandard *(Visualisierung, Darstellung für Menschen)*

Definiert die **Gestaltung und Darstellung von Metadaten für Endnutzer*innen**, zum Beispiel in Web-Interfaces, Katalogsystemen oder PDFs. Diese Standards betreffen **Layout, Labels, Reihenfolge**, aber nicht die maschinenlesbare Struktur.

Die Europeana-Oberfläche stellt die Metadaten des Objekts in einer klaren, mehrsprachigen Ansicht dar:

- `dc:title` wird als **Titel** angezeigt

- `dc:rights` erscheint mit CC-Icons und verlinkter Lizenz
- `edm:preview` zeigt ein **Thumbnail**, während `edm:isShownBy` das hochauflösende Bild öffnet
- Schlagwörter (aus `dc:subject`) werden als **Filterfacetten** nutzbar gemacht

7. Interoperabilitäts- und Austauschstandards *(Systemvernetzung)*

Ermöglichen den **Austausch, Aggregation und Mapping** zwischen verschiedenen Standards, Datenmodellen oder Plattformen. Der Fokus liegt wie bereits bei den Formatstandards auf der Maschinenlesbarkeit.

Das Beispiel *Knie Völkerschau* ist vollständig eingebettet in ein interoperables Framework:

- Über die maschinenlesbare Schnittstelle OAI-PMH werden die Informationen aus den lokalen Repositorien gesammelt und an einen Aggreator (zum Beispiel Europeana) übermittelt. Die Herkunftsinformation kann im Feld `edm:dataProvider` eingebettet werden.
- IIIF Manifest verfügbar für DeepZoom oder Bildannotation
- Verlinkung auf Lizenzen, Orte, Konzepte via URI (LOUD-Prinzipien)
- Es sind Mappings in beide Richtungen möglich: von MARC zu Dublin Core und umgekehrt, von TEI zu EDM und umgekehrt sowie von CIDOC CRM zu EDM und umgekehrt.

Ein Feld, sieben Ebenen

"Metadaten-Tabelle für: *Knie Völkerschau* mit sieben Ebenen"

Metadatum (Feld)	① Datentyp	② Wertstandard	③ Inhaltsstandard	④ Struktur / Schema	⑤ Format	⑥ Präsentation	⑦ Interoperabilität
`dc:title`	`rdf:langString` (`de`)	Freitext	**Dublin Core: Title**	DCMES, EDM	JSON, RDF/XML	UI-Label: *Titel*	Europeana API, OAI-PMH
`dc:description`	`rdf:langString` (`da`)	Freitext, gegebenenfalls Getty AAT-Terms	**DC: Description**	DCMES	JSON	Accordion-Feld, Tooltip	OAI-PMH, JSON-LD
`dc:date`	`xsd:gYear` oder `xsd:string`	ISO 8601, "1969?" (unsicher)	**DC: Date**	EDM Timespan, DCMES	RDF/XML, JSON-LD	Formatierter Zeitstempel	`edm:hasMet` + IIIF Timeline
`dc:type`	URI / `rdf:langString` (`da`)	**AAT URI**, `poster`, `Still image`	**DC: Type**	DCMES, EDM + Europeana Vocabularies	RDF	Facette: "Medientyp"	Linked Open Data / EDM
`dc:subject`	`rdf:langString` (`da`)	**GND** / kontrollierte Schlagwörter	**DC: Subject**	DCMES	JSON	Schlagwortliste	VIAF, Wikidata, Linked Data
`dc:identifier`	`anyURI`	HTTP-URL	**DC: Identifier**	DCMES, EDM Aggregation	JSON-LD	als anklickbarer Link	persistente URI (PURL/DOI)
`dc:rights`	URI + `xsd:string`	**CC URI**, RightsStatements.org	**DC: Rights**	DCMES, EDM Rights	RDF, JSON	CC-Logo, Textlink	Interoperables Lizenzsystem
`edm:isShownBy`	`anyURI`	IIIF Image Service URI	**EDM: isShownBy**	EDM Aggregation	RDF	Bild-Embed	IIIF Image API
`edm:isShownAt`	`anyURI`	Originalquellenlink	**EDM: isShownAt**	EDM Aggregation	JSON-LD	Button "Zur Quelle"	OAI-PMH, Europeana Portal
`edm:preview`	IIIF Thumbnail URI	IIIF-compliant	**EDM: preview**	EDM Aggregation	JSON-LD	Thumbnail-Bild	IIIF Presentation API
`edm:dataProvider`	URI + Name (String)	Europeana Org-ID	**EDM: dataProvider**	EDM	RDF	Link zur Institution	Europeana Registry / LOD
`edm:language`	ISO 639-1 Code	`da`, `en`, `de`, …	**EDM: language**	EDM	JSON	Flaggensymbol + Sprachname	Multilingual indexing
`edm:country`	ISO 3166-1 Alpha-2	`DK` → *Denmark*	**EDM: country**	EDM	RDF/XML	Anzeige Herkunftsland	Europeana-Facettierung

Wiederverwendung

CC BY-SA 4.0

Zitat

Mit BibTeX zitieren:

```
@report{mähr2026,
  author = {Mähr, Moritz and Schnegg, Noëlle},
  title = {Diskriminierungssensible Metadatenpraxis},
  version = {2.0.0-alpha.0},
  date = {2026-02-10},
  url = {https://maehr.github.io/diskriminierungsfreie-metadaten/},
  langid = {de-CH},
  abstract = {Dieses Handbuch bietet eine praxisorientierte Anleitung
    für die diskriminierungssensible Beschreibung von Metadaten zu
    historischen Quellen und Forschungsdaten. Es richtet sich an
    Forschende Historiker\textbackslash*innen,
    Archivar\textbackslash*innen, Bibliothekar\textbackslash*innen und
    Daten-Kurator\textbackslash*innen in der Geschichtswissenschaft und
    in Gedächtnisinstitutionen (GLAM). Das Handbuch verbindet
    theoretische Reflexionen zu Normativität, Bias und Oppression mit
    konkreten Handlungsempfehlungen für den gesamten
    Forschungsdatenlebenszyklus. Es behandelt sowohl technische Aspekte
    der Metadatenmodellierung als auch ethische Fragen der
    Repräsentation marginalisierter Gruppen. Durch die Kombination von
    Theorie und Praxis unterstützt es Institutionen und Projekte dabei,
    Diskriminierung in der Metadatenpraxis zu erkennen und
    diskriminierungssensible Alternativen umzusetzen.}
}
```

Bitte zitieren Sie diese Arbeit als:

Mähr, Moritz, and Noëlle Schnegg. 2026. *Diskriminierungssensible Metadatenpraxis*. Versions 2.0.0-alpha.0. https://maehr.github.io/diskriminierungsfreie-metadaten/.