



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

DIGITALE DIENSTE / FORSCHUNGSUNTERSTÜTZENDE DIENSTE



Metadaten & KI - Generierung, Extraktion und Bereitstellung

Andreas Frech
19.05.2025



- Was sind Metadaten?
- Was ist KI?
- Warum sind Metadaten für sie wichtig?
- Welche Anwendungsszenarien gibt es?
- Was sind die Voraussetzungen?



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

DIGITALE DIENSTE / FORSCHUNGSUNTERSTÜTZENDE DIENSTE



Metadaten



Metadaten oder Metainformationen sind strukturierte Daten, die Informationen über Merkmale anderer Daten enthalten.

Wikipedia (<https://de.wikipedia.org/wiki/Metadaten>)

Metadaten sind strukturierte Informationen, die beschreiben, erklären, lokalisieren, oder es sonst wie einfacher machen, eine Informationsquelle abzurufen, zu verwenden, oder zu verwalten. Metadaten werden oft Daten über Daten oder auch Informationen über Informationen genannt.

*Understanding Metadata /
National Information Standards Organization NISO*

(<https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf>)

Deskriptiv: Beschreibung einer Ressource oder Objekts
Auffinden, Identifizieren

Strukturell: Text-, Objekt-, Dokumentstruktur
Darstellen, Zusammenhänge herstellen

Markup: Auszeichnung von Inhalt (Strukturell oder semantisch)
Hervorheben, Navigieren

Administrativ: Verwaltung
Dokumentieren und nutzbar machen

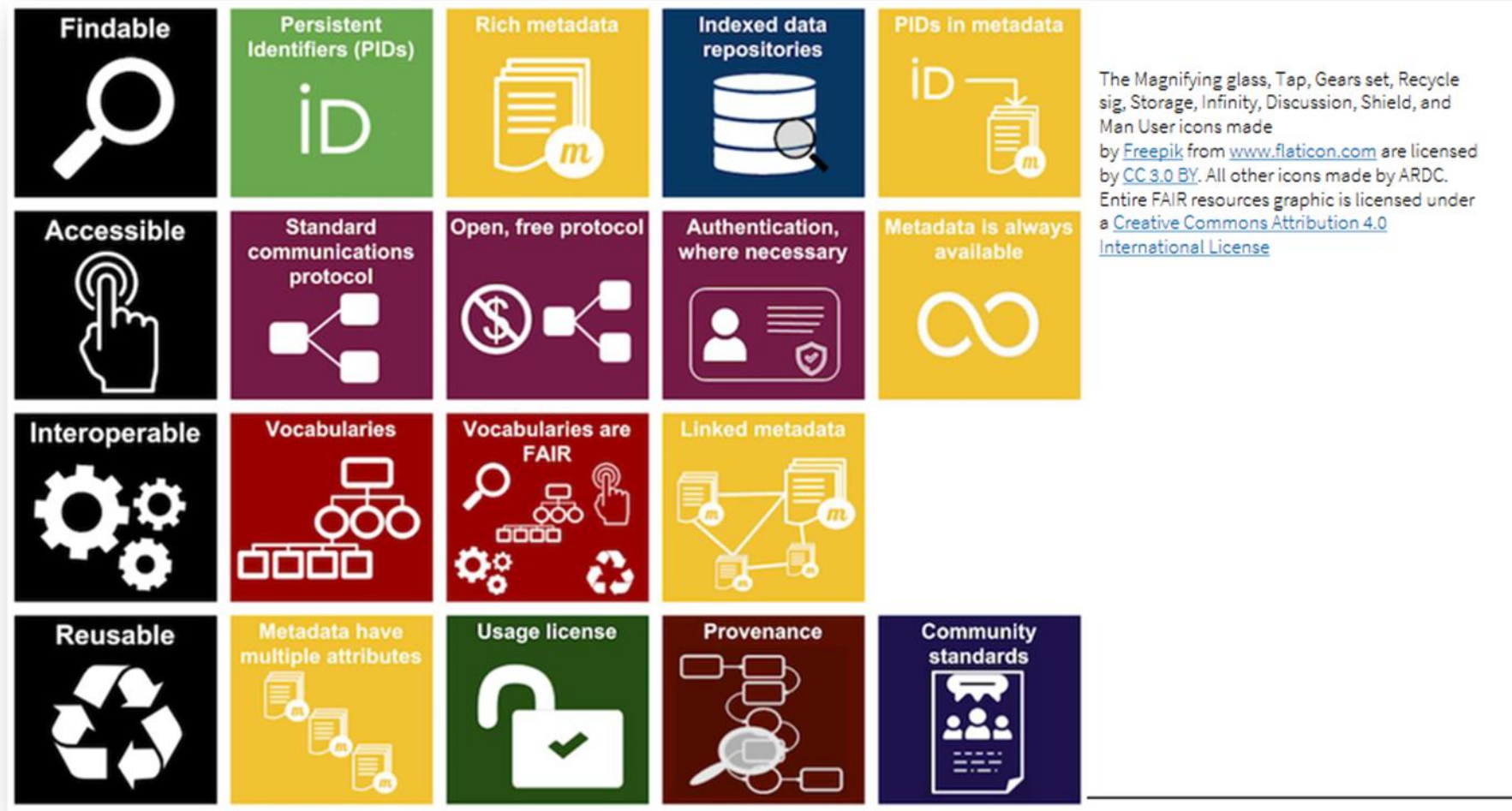
Rechtlich: Urheber- und Nutzungsrechte

Technisch: Dateiformate, Speicherort etc.

LZA-spezifisch: Entstehungsgeschichte des digitalen Objekts,
Checksummen

- Zentrale Frage: Wie hilft KI, wissenschaftliche Inhalte effizienter zu erschließen und nutzbar zu machen?
- Voraussetzung für „AI Readiness“: FAIR
 - **F**INDABLE
 - **A**CCESSIBLE
 - **I**NTEROPERABLE
 - **R**EUSABLE

Die FAIR-Prinzipien





LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

DIGITALE DIENSTE / FORSCHUNGSUNTERSTÜTZENDE DIENSTE



Künstliche Intelligenz

19. Mai 2025



Was ist Künstliche Intelligenz?



„Künstliche Intelligenz (KI), englisch artificial intelligence, daher auch artifizielle Intelligenz (AI), ist ein Teilgebiet der Informatik, das sich mit der Automatisierung intelligenten Verhaltens und dem maschinellen Lernen befasst. Der Begriff ist schwierig zu definieren, da es verschiedene Definitionen von Intelligenz gibt.“

Wikipedia (https://de.wikipedia.org/wiki/K%C3%BCnstliche_Intelligenz)

„Künstliche Intelligenz ist die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren.“

Europäisches Parlament (<https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt>)

„Sammelbezeichnung“ u.a.

- Automatisierung von Arbeitsabläufen
- Computerlinguistische Mustererkennung
- Wissensmodellierung
- Maschinelles Lernen
- Künstliche neuronale Netze (LLMs, Deep Learning)
- Retrieval-Augmented Generation (RAG)

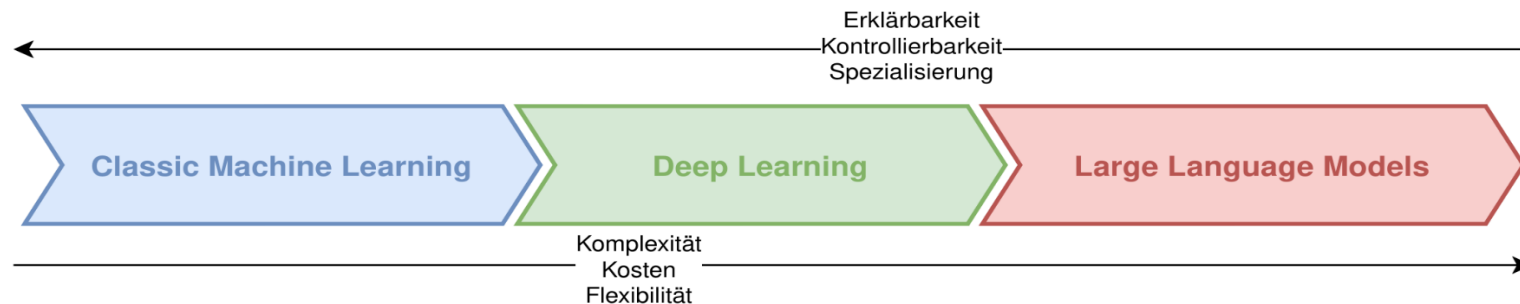


Bild aus: Zinnen, M., & Zöllner, G. (2025, April 30). SODa Forum: Maschinelles Lernen und Provenienzforschung, 10. April 2025. Zenodo. <https://doi.org/10.5281/zenodo.15308989>



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

DIGITALE DIENSTE / FORSCHUNGSUNTERSTÜTZENDE DIENSTE



Metadaten und KI

19. Mai 2025



Zentrale Aufgaben, bei denen KI hilft:

- Automatische formale und inhaltliche Erschließung (z. B. Titel, Abstract, Schlagwörter)
 - Auch Audio-, Bild- und Videoinhalte via Sprach- und Bilderkennung
- Named Entity Recognition (NER) (z. B. Personen, Institutionen, Orte)
- Klassifikation (z. B. DDC)
- Extraktion strukturierter Metadaten aus PDFs
- Sprachmodell-basierte Annotationen (z. B. durch LLMs)

Automatische Erschließung



- Nutzung von Natural Language Processing (NLP) zur Analyse von Titeln, Abstracts, Volltext
- Klassifikationsmodelle zur Fachgebietszuteilung
- Extraktion von Schlagwörtern und Kurztiteln und Zusammenfassungen bzw. Abstracts
- Beispielprojekte:
 - Automatische Erschließung bei der DNB:
https://www.dnb.de/DE/Professionell/ProjekteKooperationen/Projekte/KI/ki_node.html
 - Erschließung Grauer Literatur der Nationalbibliotheken von Norwegen und Finnland:
<https://forum.swib.org/t/automating-metadata-extraction-and-cataloguing-experiences-from-the-national-libraries-of-norway-and-finland/>
 - Automatisierte Erschließung bei der ZBW für das EconBiz-Portal:
<https://www.zbw.eu/de/ueber-uns/wissensorganisation/automatisierung-der-erschliessung-mit-methoden-der-kuenstlichen-intelligenz/>

Extraktion von Entitäten



Extraktion von Entitäten

- Ziel: Erkennung von Entitäten im Text (Personen, Organisationen, Orte...)
- Beispiel:

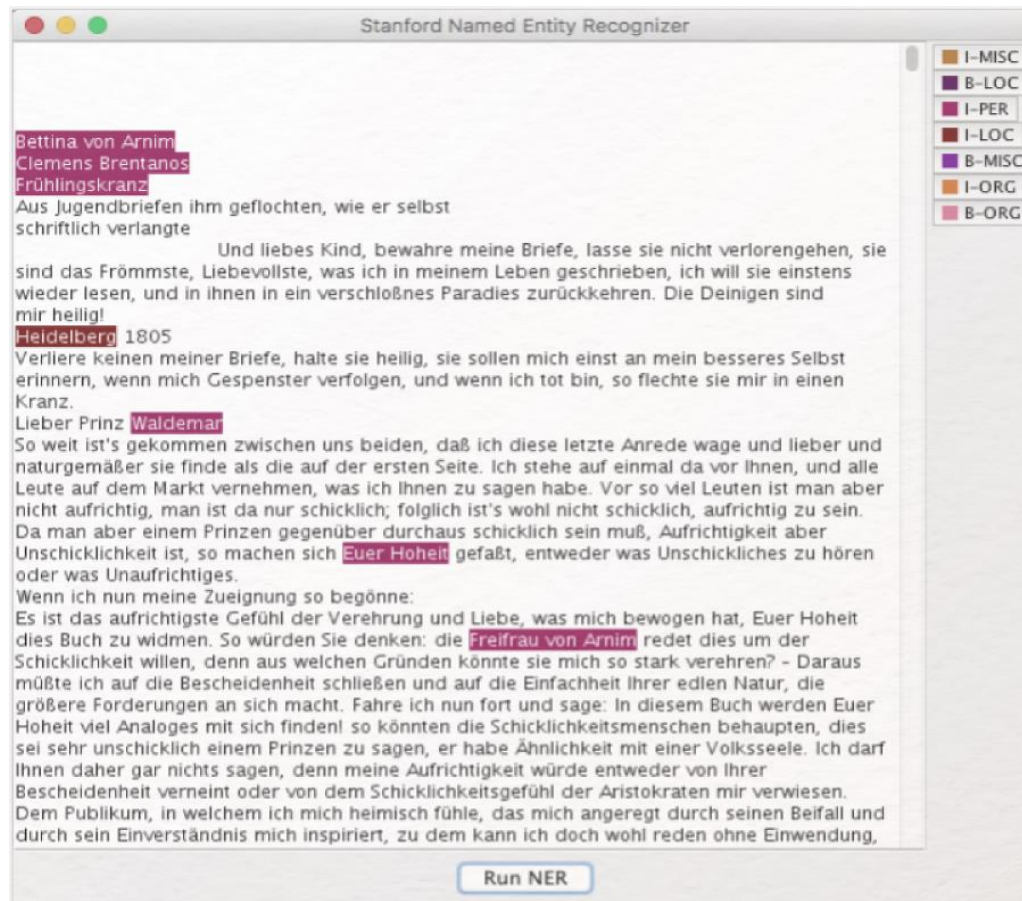


Bild aus: Mareike Schumacher (2018): „Named Entity Recognition (NER)“. In: forTEXT. Literatur digital erforschen. URL: <https://fortext.net/routinen/methoden/named-entity-recognition-ner> [Zugriff: 14. Mai 2025].

AI Readiness

Bereitstellung von (Meta)daten



- Maschinenlesbar und skalierbar
- Nur strukturierte Metadaten ermöglichen präzises maschinelles Verständnis
- Sie ermöglichen (idealerweise) semantische Suche, Topic Modelling und Named Entity Linking
- Offen und untereinander verlinkt (Linked Data)
- Wichtige Formate: XML, JSON LD und RDF

- Linked Open Data ist Linked Data mit offener Lizenz – also frei zugänglich, wiederverwendbar und durch Dritte nutzbar
- LOD ist eine Methode zur Veröffentlichung strukturierter Daten, die über URIs identifizierbar und miteinander verknüpfbar sind – und semantisch interpretierbar durch Maschinen
- LOD ist Brücke zwischen isolierten Datensilos und einem vernetzten, semantischen Web
- Die vier Linked-Data-Prinzipien von Tim Berners-Lee:
 - Verwende URIs zur Identifikation von Dingen
 - Verwende HTTP-URIs, damit Menschen und Maschinen sie aufrufen können
 - Liefere nützliche Informationen (idealerweise in RDF) zurück, wenn jemand eine URI aufruft.
 - Verlinke auf andere URIs, um den Zusammenhang zwischen Daten zu zeigen.

RDF (Resource Description Framework)

- RDF ist ein W3C-Standard zur semantischen Repräsentation von Wissen. Es beschreibt Informationen in Form von sogenannten Tripeln:

Subjekt — Prädikat — Objekt

„Artikel“ — hatAutor — „Max Mustermann“

„Max Mustermann“ — hatGNDId — <https://d-nb.info/gnd/1143543866>

- Offener Standard für interoperable, verlinkte Daten
- URI-basiert – jeder Begriff und jede Entität hat eine eindeutige Webadresse
- Grundlage für das „Linked Open Data“-Prinzip
- Gespeichert in „Triple Store“-Datenbanken

JSON-LD (JSON for Linked Data)

- JSON-LD ist ein vereinfachtes Austauschformat für RDF-Daten
- Maschinen und menschenlesbar (~lesefreundlich)
- Von Suchmaschinen und vielen anderen Systemen verwendet
- Direkt in HTML einbettbar oder für API-Abfragen verwendbar

```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "Artikel",
  "author": {
    "@type": "Person",
    "name": "Max Mustermann",
    "identifier": "https://d-nb.info/gnd/1143543866"
  }
}
```

- Ontologien sind formale, maschinenlesbare Darstellungen von Wissen in einem bestimmten Bereich
- Umfassen und definieren Klassen (Entitys), ihre Eigenschaften (Attributes) und Zusammenhänge
- Ermöglichen semantisches Verstehen durch die KI mittels Bedeutung und Kontext
- Diese Beziehungen werden meist in OWL (Web Ontology Language) oder RDFS (RDF Schema) beschrieben
- Auch Ontologien sollten „offen“ sein, damit die damit verknüpften Daten sinnvoll interpretiert werden können (LOUD: Linked Open Usable Data)

Beispiel: Bedeutungsunterschiede — „author“ - „creator“ - „contributor“

```
@prefix : <http://example.org/ontology#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

:Resource a owl:Class ;
  rdfs:label "Resource" ;
  rdfs:comment "A digital or physical resource." .

:Document a owl:Class ;
  rdfs:domain :Resource ;
  rdfs:label "Document" ;
  rdfs:comment "A digital or physical written document." .

:Person a owl:Class .

:creator a owl:ObjectProperty ;
  rdfs:domain :Resource ;
  rdfs:range :Person ;
  rdfs:label "creator" ;
  rdfs:comment "General agent responsible for creating a resource." .

:author a owl:ObjectProperty ;
  rdfs:subPropertyOf :creator ;
  rdfs:domain :Document ;
  rdfs:range :Person ;
  rdfs:label "author" ;
  rdfs:comment "A person who wrote the intellectual content of the document." .

:contributor a owl:ObjectProperty ;
  rdfs:domain :Resource ;
  rdfs:range :Person ;
  rdfs:label "contributor" ;
  rdfs:comment "A person who contributed to the creation of a resource, but is not the main creator." .
```

- CIDOC CRM
 - CIDOC Conceptual Reference Model
 - 1999 vom Comité international pour la documentation (Internationales Komitee zur Dokumentation) entwickelt
 - Standard-Ontologie im Bereich des kulturellen Erbes
 - <https://cidoc-crm.org/>
- schema.org
 - 2011 von Suchmaschinen-Betreibern Google, Microsoft und Yahoo etabliert
 - „Standardformat“ des Internets
 - <https://schema.org/>

Fazit

- Metadaten sind das Rückgrat datengetriebener KI. Sie ermöglichen Kontext, Kontrolle, Effizienz und Transparenz – und damit letztlich Zuverlässigkeit und Verantwortlichkeit von KI-Systemen.
- Künstliche Intelligenz wird zunehmend zur zentralen Schnittstelle für Informationszugang – wer heute in gute Metadaten investiert, sorgt dafür, dass Inhalte morgen auffindbar, nutzbar und interpretierbar bleiben.
- LLMs wie ChatGPT sind keine Suchmaschinen: Sie generieren Sprache basierend auf Wahrscheinlichkeiten – nicht auf verifiziertem Wissen. Strukturelle und semantische Klarheit im Datenbestand bleibt deshalb entscheidend.

Ausblick

- DFG-Ausschreibung „Datenkorpora für Künstliche Intelligenz“
 - Ziel: „Die Förderung soll den Auf- und Ausbau qualitativ hochwertiger, umfangreicher Datenkorpora ermöglichen, um eine breite und wissenschaftlich fundierte Datenbasis für die (Weiter-)Entwicklung und Anwendung von Methoden der Künstlichen Intelligenz für die Wissenschaft zu schaffen.“ (<https://www.dfg.de/de/aktuelles/neuigkeiten-themen/info-wissenschaft/2025/ifw-25-28>)
- Strukturierte Metadaten können helfen, dass KIs nach Antworten suchen können
 - Beispielprojekt: Wikifunctions (<https://www.wikifunctions.org/>)
 - KI kann in natürlicher Sprache formulierte Fragen in Funktionen Abfragen für Wikidata übersetzen (<https://forum.swib.org/t/keynote-how-knowledge-representation-is-changing-in-a-world-of-large-language-models/>)





LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

DIGITALE DIENSTE / FORSCHUNGSUNTERSTÜTZENDE DIENSTE



Andreas Frech, UB LMU München
andreas.frech@ub.uni-muenchen.de

