

Resourcenbeitrag: HathiTrust Digital Library

Marie Flüh  ¹

1. Universität Hamburg

forTEXT

Thema:	Bibliografie	DOI:	10.48694/fortext.3792
Jahrgang:	1	Ausgabe:	11
Erscheinungsdatum:	30-11-2024	Erstveröffentlichung:	2019-04-01 auf fortext.net
Lizenz:			open & access

Allgemeiner Hinweis: Rot dargestellte *Begriffe* werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Kurzbeschreibung

Die HathiTrust Digital Library ist eine Online-Volltextdatenbank, die Primär- wie Sekundärliteratur bedeutender US-amerikanischer Forschungs- und Universitätsbibliotheken und Forschungsinstitutionen zur Verfügung stellt.

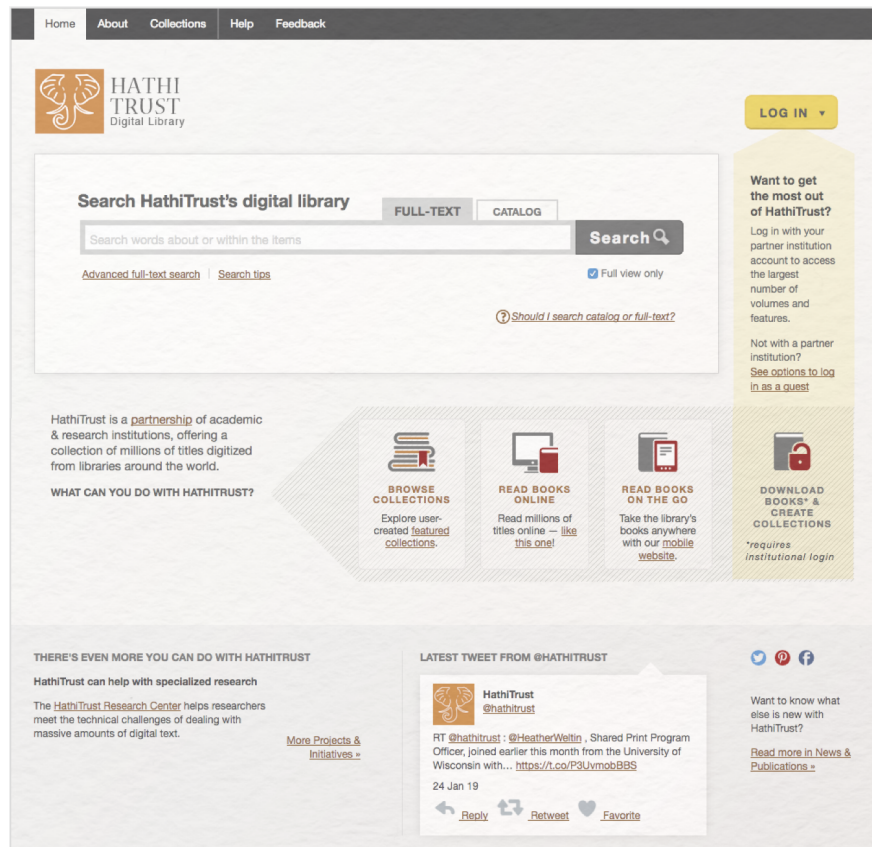


Abb. 1: Benutzeroberfläche der HathiTrust Digital Library

Steckbrief

- <https://www.hathitrust.org/2>
- circa 17.000.000 Bände bzw. circa 8.000.000 Bücher
- Schwerpunkte: Sprach- und Literaturwissenschaft, Geschichtswissenschaft, Theologie, Philosophie
- etwa die Hälfte der Werke in englischer Sprache, zweitgrößte Sprachgruppe Deutsch, gefolgt von Französisch, Spanisch, Chinesisch, Russisch und vielen weiteren Sprachen
- Berichtszeitraum: 1500 bis Gegenwart, Schwerpunkt bei den gemeinfreien Werken (vgl. [Open Access](#)) zwischen 1800 und 1923

- **Metadaten:** Autor, Sprache, Veröffentlichungsort und -jahr, Verlag, Themenbereiche, Seitenzahl, Buchmaß, ISBN (OCLC, LCCN), Urheberrechtsvermerk, URL (vgl. **URI**)
- Ziel: Erhalt des kulturellen Gedächtnisses durch Organisation, Digitalisierung, Langzeitarchivierung und Bereitstellung v. a. von Büchern und Zeitschriften (ursprünglich als Back-up von Google Books angelegt), Koordination einer gemeinsamen Speicherstrategie der Bibliotheken
- Zusammensetzung: Digitalisate aus der Zusammenarbeit der über 140 HathiTrust-Mitglieder mit Google, dem Internet Archive und Microsoft und eigene Digitalisate aller beteiligten Bibliotheken
- Möglichkeit der Zusammenstellung eigener Textsammlungen und deren Verwaltung (Bearbeiten, Speichern, Teilen, Herunterladen der Metadaten als TSV- oder JSON-Datei)
- Zugriffsmöglichkeiten: HathiTrust-Mitglieder: Zugriff auf alle urheberrechtlich geschützten Werke, Herunterladen vollständiger Bücher; andere: Zugriff auf alle gemeinfreien Werke, Herunterladen einzelner Seiten im **PDF**-Format

2. Anwendungsbeispiel

Sie möchten die Zeitschrift *Die Fackel* von Karl Kraus nach Äußerungen über Heinrich Heine untersuchen, um Rhetorik und Darstellungsweise des Herausgebers und Satirikers zu analysieren. Um einen Einstieg in Ihre Recherche zu finden, möchten Sie sich außerdem einen Überblick über die relevante Forschungsliteratur zu Karl Kraus verschaffen.

3. Diskussion

3.1 Kann ich die HathiTrust Digital Library für wissenschaftliche Arbeiten nutzen?

Ja, der Verlag HathiTrust verpflichtet sich dazu, die hohe Qualität digitalisierter Texte und Metadaten zu sichern. Das Repositorium wurde vom *Center for Research Libraries* (CRL) als ein vertrauenswürdiges und die *TRAC-Kriterien* („Trustworthy Repositories Audit & Certification“; Kriterienkatalog für vertrauenswürdige Langzeitarchive) erfüllendes digitales Langzeitarchiv ausgezeichnet. Verbindliche Richtlinien (<https://www.hathitrust.org/ingest7>), an die sich Partnerorganisationen bei der Beisteuerung ihrer Digitalisate halten müssen, zielen ebenfalls auf den Erhalt einer hohen Qualität ab. Im Rahmen einer umfassenden Digitalisierungsstrategie werden fortlaufend Qualitätskontrollen durchgeführt, in denen die zufriedenstellende Berücksichtigung der ratifizierten Standards zur Bewahrung digitalisierter Objekte überprüft wird. Die Qualität der Beiträge wird gemessen und dokumentiert, um darauf aufbauend eine fortschreitende Qualitätsverbesserung zu initiieren. Bei der großen Bandbreite an Inhalten, die von zahlreichen Forschungseinrichtungen beigesteuert werden, ist eine gleichwertige Qualität sämtlicher gescannter Dokumente nicht immer gegeben. Mangelnde Bildqualität oder **OCR**-Probleme werden jedoch bearbeitet, nachdem sie von Nutzer*innen gemeldet wurden.

Das Verbundprojekt HathiTrust unterstützt im Rahmen einer vereinbarten Speicherstrategie die Digitalisierung und Aufnahme digitaler Bücher, Journals und buchähnlicher Materialien wie Manuskripte. Auch der Import von Audio- und Bilddateien ist ausdrücklich erwünscht.

Die HathiTrust Digital Library ermöglicht Ihnen eine äußerst dynamische Texterschließung, die die Arbeit mit umfangreichen Primär- und Sekundärquellen bereichert. Die multifunktionale Menüleiste beinhaltet neben gängigen Funktionen wie dem Heran- und Herauszoomen auch unterschiedliche Seitenansichten (Buch, Thumbnail, Einzelseite, Full Screen). Innerhalb der Referenz können Sie über die Menüleiste direkt an die für Sie relevante Textpassage springen.

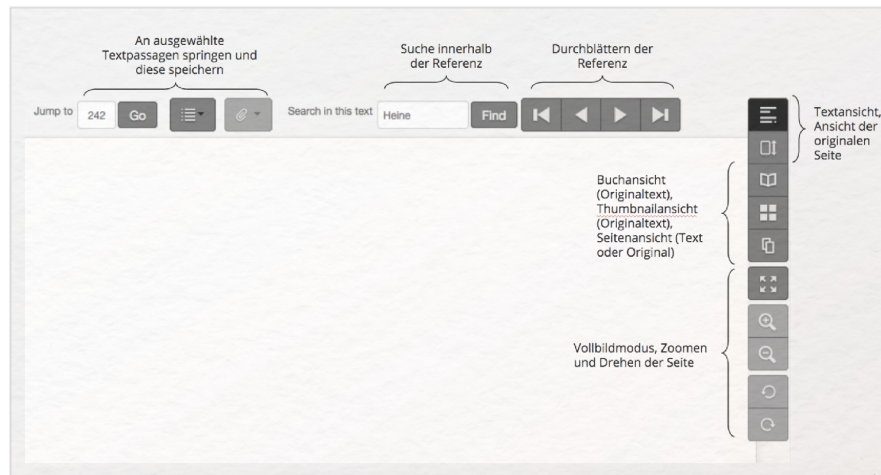


Abb. 2: Interface der Menüleiste der HathiTrust Digital Library

Sie können darüber hinaus beispielsweise die originalen Transkripte der *Fackel* ansehen und zwischen unterschiedlichen Textdarstellungen wählen: Über den „Plain-Text“-Button (vgl. [Reintext-Version](#)) können Sie zwischen der reinen Textansicht und dem originalen Seitenlayout hin und her springen. In den einzelnen Digitalisaten sind Volltextsuchen über das „Search in this text“-Eingabefeld möglich. Ergebnisse werden im Transkript angezeigt.



Abb. 3: Per Suchanfrage lassen sich Textdokumente nach Begriffen durchsuchen. Links: Die Ansicht der „Plain Text“-Variante mit der Hervorhebung des Suchbegriffs; Rechts: Via „Scroll“-Button können Sie die entsprechende Seite im Original ansehen

3.2 Wie benutzerfreundlich ist die Arbeit mit der HathiTrust Digital Library?

Die Website der HathiTrust ist übersichtlich gestaltet und für Erstnutzer*innen intuitiv bedienbar. Der Fokus richtet sich nach dem Aufrufen der Startseite unmittelbar auf das Wesentliche des digitalen Langzeitarchivs: Das Suchfeld und die unterschiedlichen Recherchemöglichkeiten (Volltextsuche oder Katalogsuche (vgl. [Query](#))). Ebenfalls selbsterklärend ist die Sondierung der Ergebnisse von Suchanfragen und die Erkundung der frei

zugänglichen Referenzen. Ein englischsprachiges digitales Handbuch, in dem unter anderem häufig gestellte Fragen zur Benutzung der Online-Bibliothek sowie Hinweise für eine erfolgreiche Suche oder Copyright-Vermerke erläutert werden, steht Ihnen zur Verfügung (<https://www.hathitrust.org/help13>) und erleichtert erste Recherche-Schritte. Darüber hinaus ist der HathiTrust Verlag für Anfragen aller Art per E-Mail, Telefon, Kontaktformular oder Post erreichbar. Das primäre Engagement der HathiTrust gilt den US-amerikanischen Universitäts- und Forschungsbibliotheken, deren Status über einen Check der IP-Adresse überprüft wird. Auf urheberrechtlich geschützte Werke können ausschließlich Mitglieder der Partnerbibliotheken zugreifen. Das Herunterladen ganzer Bücher ist ebenfalls den Mitgliedern vorbehalten. Der Zugang zu Werken, die bis 1923 in den USA veröffentlicht wurden, ist nur mit einer US-amerikanischen IP-Adresse möglich. Sämtliche vor 1872 außerhalb der USA veröffentlichten Werke werden als gemeinfreie Werke eingestuft und sind deshalb weltweit abrufbar. Freie Nutzer*innen können zwar in der gesamten Bibliothek recherchieren, müssen hierbei jedoch die eingeschränkten Zugriffsmöglichkeiten auf den Gesamtbestand der HathiTrust Digital Library berücksichtigen, was die Recherche erheblich einschränken kann. Über den „Find in a Library“-Button können externe Nutzer*innen allerdings herausfinden, in welcher Bibliothek oder welcher Datenbank Referenzen zur Verfügung stehen.

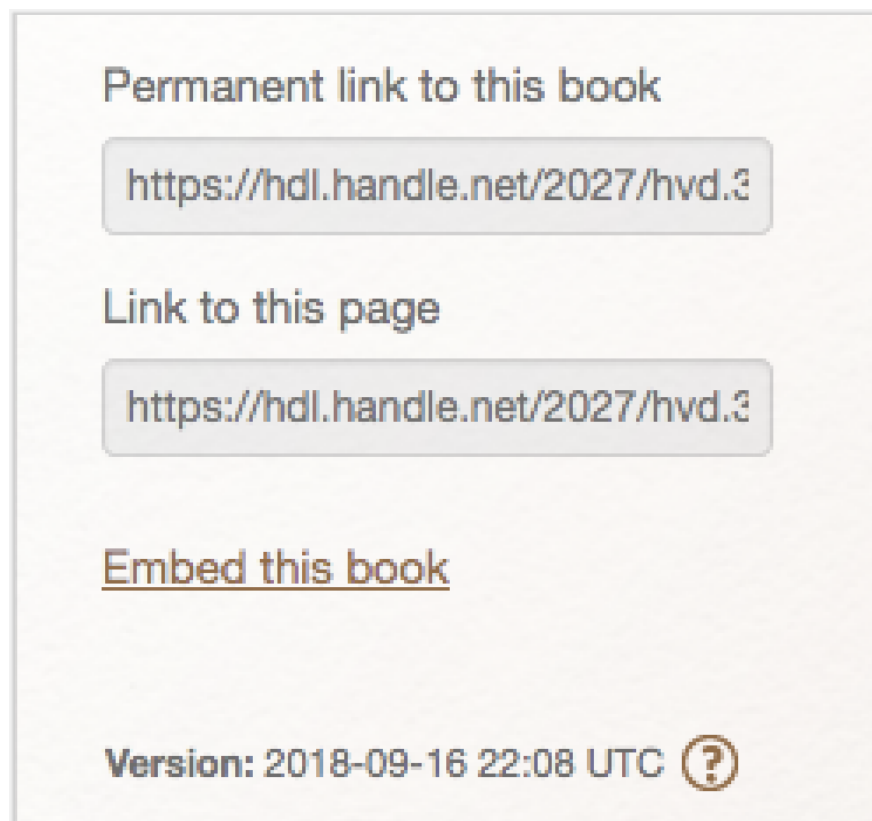


Abb. 4: Durch den permanenten Link können Sie zeitunabhängig auf die entsprechende Referenz zurückgreifen

Einzelne Seiten lizenzierter Werke können Sie als PDF herunterladen oder als permanenten Link speichern, der Sie zur der entsprechenden Seite zurückführt.

Externe Nutzer*innen können außerdem einen kostenlosen Gast-Account erstellen (Zugang entweder über existierenden Account bei Google, Facebook, Twitter, Yahoo, LinkedIn, Microsoft oder nach der Erstellung eines Accounts bei der University of Michigan), in dem Sie eigene Literatursammlungen erstellen, verwalten und mit anderen Nutzer*innen teilen können. Auch hier gilt allerdings die Beschränkung auf als gemeinfrei eingestufte Werke, die vor 1872 außerhalb der USA erschienen sind. Aktuelle relevante Forschungsliteratur wird zwar angezeigt und kann seitenweise gelesen werden, für einen Volltextzugang müssen Sie jedoch den Umweg über den „Find in a Library“-Button gehen.

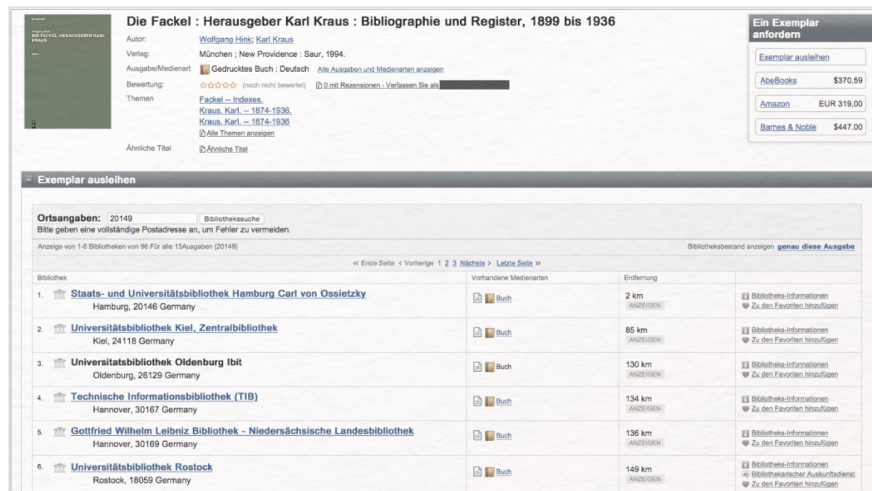


Abb. 5: Via „Find in a library“-Button werden Ihnen die Standorte einer Referenz aufgelistet

4. Wie funktioniert die Textsuche in der HathiTrust Digital Library?

Innerhalb des Gesamtkatalogs können Sie über die Freitexteingabe eine Volltextsuche durchführen. Die Katalogsuche können Sie nach Titel, Autor, Thema, ISBN/ISSN, Herausgeber oder Serientitel spezifizieren. Ergebnisse der Suchanfrage werden als scrollbare Literaturliste dargestellt. Als hilfreiche Ergänzung der Recherche erweist sich die „Refine Results“-Spalte, in der unter Filtern wie beispielsweise Sprache, Thema oder Datum der Veröffentlichung spezifizierte Rechercheergebnisse abgerufen werden können. Über den „Catalog Record“-Button gelangen Sie zu der Detailansicht der Referenz, in der Ihnen auch die Metadaten zur Verfügung gestellt werden. Diese basieren auf den Metadaten der unterschiedlichen Bibliotheken und sind deshalb nicht immer einheitlich. Grundlegende Metadaten wie Titel, Referenzart und Datum werden in jedem Fall zur Verfügung gestellt und lassen sich direkt in Ihr Literaturverwaltungsprogramm importieren. Hier wird die weitere Suche nach Forschungsliteratur erleichtert, indem in der „Similar Items“-Spalte weitere Publikationen des Autors angezeigt werden.

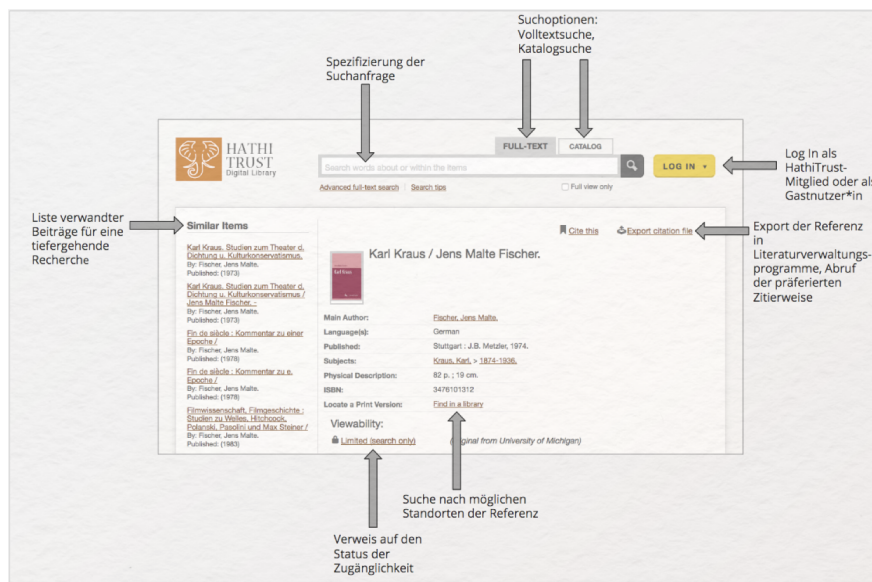


Abb. 6: Detailansicht der Recherche in der HathiTrust Digital Library

Externe und weiterführende Links

- HathiTrust: <https://web.archive.org/web/20241106120018/https://www.hathitrust.org/> (Letzter Zugriff: 06.11.2024)
- HathiTrust Handbuch: <https://www.hathitrust.org/help> (Letzter Zugriff: 06.11.2024)
- HathiTrust Help Center: <https://web.archive.org/web/20241106120052/https://hathitrust.atlassian.net/servicedesk/customer/portals> (Letzter Zugriff: 06.11.2024)

- Richtlinien für Digitalisate: <https://web.archive.org/web/20241106120104/https://www.hathitrust.org/member-libraries/contribute-content/> (Letzter Zugriff: 06.11.2024)

Bibliographie

- Christenson, Heather. 2011. HathiTrust. A Research Library at Web Scale. *Library Resources and Technical Services* 55, Nr. 2: 93–102.
- Cobine, Ryan, J. Stephen Downie, Inna Kouper, Robert McDonald, Beth Pale, Beth Sandore Namachchivaya, Yiming Sun und John Unsworth. 2013. HathiTrust Research Center: Computational Access for Digital Humanities and Beyond. In: *JCDL 13 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 395–396.
- Harvey, Ross. 2012. *Current Topics in Library and Information Practice*. Berlin/Boston: de Gruyter.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch **Machine-Learning-Verfahren** durchgeführt wird. Ein klassisches Beispiel ist das automatisierte **PoS-Tagging** (Part-of-Speech-Tagging), welches oftmals als Grundlage (**Preprocessing**) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv`, sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von **Webbrowsern** dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche **Metainformationen** enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen **Preprocessing**-Schritten in der Textverarbeitung. Dabei werden alle Wörter (**Token**) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z. B. **HTML**, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch **Annotationen** durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch **Annotationen** bzw. **Markup** sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigename, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

- Open Access** Open Access bezeichnet den freien Zugang zu wissenschaftlicher Literatur und anderen Materialien im Internet.
- PDF** PDF steht für *Portable Document Format*. Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein **OCR**ter Scan oder ein am Computer erstellter Text.
- POS** PoS steht für *Part of Speech*, oder „Wortart“ auf Deutsch. Das PoS- **Tagging** beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger **Preprocessing**-Schritt, beispielsweise für die Analyse von **Named Entities**.
- Preprocessing** Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden **lemmatisiert**.
- Query** *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen *Queries* zusammen die *Query Language* eines Tools.
- Reintext-Version** Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und **CSV**.
- Type/Token** Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.
- URI** *Uniform Resource Identifier* (URI) ist ein Identifikator zur eindeutigen Erkennung von Online-Ressourcen wie Webseiten. Im „Raum“ des Internets können so alle Inhalte eindeutig identifiziert werden, unabhängig davon, ob es sich dabei beispielsweise um eine Seite mit Text oder Video handelt. Die am häufigsten verwendete Form eines URI ist die Webseitenadresse, die URL.