Patrick D. Brookshire und Jonas Richter

Im Netz mittelhochdeutscher Wörter
Das ‚Mittelhochdeutsche Wörterbuch' und seine Quellen

Kontakt

Patrick D. Brookshire,
Mittelhochdeutsches Wörterbuch/
Digitale Akademie,
Akademie der Wissenschaften und der Literatur Mainz,
Geschwister-Scholl-Str. 2,
55131 Mainz,
patrick.brookshire@adwmainz.de
https://orcid.org/0000-0002-5843-7577

Dr. Jonas Richter,
Mittelhochdeutsches Wörterbuch,
Niedersächsische Akademie der Wissenschaften zu Göttingen,
Geiststraße 10,
37073 Göttingen,
jonas.richter@adwgoe.de
https://orcid.org/0000-0002-2924-6026

Abstract This paper illustrates how existing interdependencies between various Middle High German dictionaries have, over time, lead to a dense network with 'Mittelhochdeutsches Wörterbuch' ('Middle High German Dictionary') at its core. This is due to the fact that it offers multiple interfaces that can be used to connect Middle High German words between different projects. One such interface is the lemma list, which has been around since the early days of the project, and another more recent one is the Federated Content Search of Text+ (a consortium of the National Research Data Infrastructure in Germany). In addition, we show how the digitally available primary source corpus can be reused by presenting a tool for cooccurrence analyses which also considers metadata, as well as a pilot study on Named Entity Recognition. With all of this, we not only highlight existing cooperation and collaboration efforts but also want to foster future interdisciplinary research to let this network grow even more.

Keywords Corpus Linguistics; Digital Medieval Studies; Federated Content Search; Middle High German; Named Entity Recognition

1 Hintergrund

Im Sommer 1867, nur ein Jahr nach der Vollendung des ‚Mittelhochdeutschen Wörterbuchs' von Benecke-Müller-Zarncke (BMZ), bat der Verleger Salomon Hirzel Matthias Lexer, ein Handwörterbuch auszuarbeiten, das zugleich als Supplement und als alphabetischer Index zum BMZ dienen sollte.[1] Für die Ausarbeitung benötigte Lexer ein Vielfaches der Zeit, die Hirzel 1869 veranschlagt hatte, und das ‚Handwörterbuch' überstieg auch den geplanten Umfang deutlich. Hinzu kamen 400 Spalten mit Nachträgen, die Lexer seinem ‚Handwörterbuch' als Anhang beifügte.[2] In der Altgermanistik etablierten sich BMZ und Lexer als Zweigespann, mit denen das Netz mittelhochdeutscher Wörter zu erkunden war – wenn auch nicht ohne Klagen darüber, dass der Zugriff über die zwei Alphabete im Lexer (Hauptteil und Nachträge) sowie im Nebeneinander der zwei Wörterbücher umständlich sei.[3] Um die Verstrickungsgefahr in diesem Netz zu verringern, wurden 1997 bis 2000 DFG-finanziert die beiden Wörterbücher sowie das ‚Findebuch'[4] retrodigitalisiert, miteinander verlinkt und als ‚Mittelhochdeutsche Wörterbücher im Verbund' auf CD-ROM und im Internet veröffentlicht. Dieser Verbund ist mittlerweile im ‚Wörterbuchnetz' aufgegangen.[5]

Trotz dieser jüngeren, elektronischen Entwicklung sollte aber nicht vergessen werden, dass die systematische, werkübergreifende Vernetzung des mittelhochdeutschen Wortschatzes mit Matthias Lexer begann, der in seinem Handwörterbuch konsequent nachwies, wo im BMZ der entsprechende Eintrag zu finden ist. In gewisser Weise kann man das von Hirzel initiierte und von Lexer ausgearbeitete ‚Handwörterbuch' auch als einen kleinen Schritt in Richtung der FAIR-Prinzipien[6] betrachten, da er die Auffindbarkeit (*Findability*) der BMZ-Daten erhöhte und sie für seine eigenen Wörterbucheinträge nachnutzte bzw. darauf aufbaute (*Reusability*).

Die digitalen Möglichkeiten erlauben aber eine noch engere Verzahnung, als sie zu dieser Zeit erreicht werden konnte. In der erwähnten Retrodigitalisierung Ende der 1990er Jahre wurden die Einträge von BMZ, Lexer und ‚Findebuch' mit Hyperlinks verbunden, außerdem die Quellensiglen mit Nellmanns ‚Verzeichnis

1 Lexer 1872, S. v.

2 Vgl. Gärtner 1993.

3 Vgl. Nellmann 1991, S. 255.

4 Vgl. Gärtner, Gerhardt, Jährling u. a. 1992.

5 https://woerterbuchnetz.de/ (Zugriff: 11.03.2025).

6 ‚FAIR' (Wilkinson, Dumontier, Aalbersberg u. a. 2016) steht für ‚Findability', ‚Accessibility', ‚Interoperability' und ‚Reusability'.

der benutzten Werke‘[7] verknüpft.[8] Im neuen ‚Mittelhochdeutschen Wörterbuch‘ (MWB), das nach Vorarbeiten in den 1990ern seit 2000 als Langzeitprojekt im Akademienprogramm geführt wird und 2032 abgeschlossen sein soll, kommen weitere Verlinkungen hinzu.[9]

Einen besonderen Mehrwert stellt nach Gärtner und Plate die Verknüpfung von Belegstellen in den Wörterbuchartikeln mit Volltexten dar:[10] Viele Quellen des MWB sind elektronisch ins System eingepflegt, so dass die verwendeten Belegzitate mit dem Volltext verknüpft bleiben. Bei der Nutzung des Online-Wörterbuchs kann also bei Bedarf eine Referenz angeklickt werden, um den Kontext des Belegs anzusehen. Die Verlinkung ist dabei bidirektional: Neben dem Volltext wird angezeigt, in welchen Artikeln die jeweiligen Textstellen zitiert sind.

Als *born-digital* Wörterbuch mit hybrider Publikationsform (elektronisch und gedruckt) bemüht sich das Projekt, neben der Hauptarbeit, dem Erstellen der Wörterbuchartikel, seine Inhalte auch als nachnutzbare und vernetzungsfähige Daten aufzubereiten. Dazu dienen die Verwendung etablierter Standards der Datenmodellierung und die Bereitstellung verschiedener Schnittstellen (API), die einen automatisierten Zugriff auf die Daten des Projektes (und nicht nur auf die Wörterbuchartikel) ermöglichen. Durch digitale Methoden können interoperable Daten deutlich einfacher über Projektgrenzen hinaus geteilt oder über föderierte Architekturen gemeinsam durchsucht werden, ohne dass die Datensouveränität und die Sichtbarkeit der Einzelprojekte verlorengehen.

Im folgenden Abschnitt wird zunächst auf Aspekte der Datenmodellierung innerhalb des MWBs eingegangen. Anschließend werden im 3. Abschnitt bereits existierende Schnittstellen zu anderen Projekten am Beispiel der Gesamtlemmaliste des Projekts sowie der föderierten Suche von Text+ vorgestellt. Als weitere Nachnutzungsmöglichkeiten von MWB-Daten thematisiert der 4. Abschnitt einerseits ein Tool zur Kookkurenzsuche und andererseits eine Pilotstudie zur automatischen Eigennamenerkennung in mittelhochdeutschen Texten.

2 Datenmodellierung

Die beiden zentralen Ressourcen des MWB (Wörterbuchartikel und Primärquellen im Volltext) sind nach den Richtlinien der *Text Encoding Initiative* (TEI) in XML codiert und damit in hohem Maß nachnutzbar und interoperabel. Die Wortartikel

7 Vgl. Nellmann 1997.

8 Vgl. Fournier 2000.

9 Vgl. Diehl u. Hansen 2016.

10 Vgl. Gärtner u. Plate 2013.

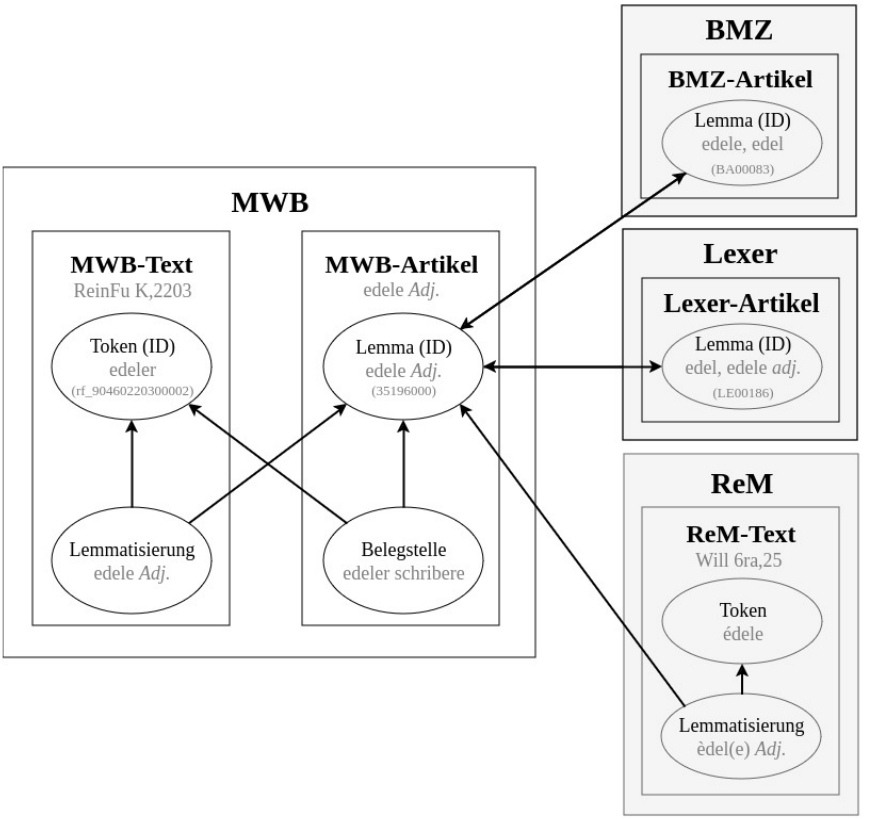

Abb. 1 | Beispielhafte Verlinkung zwischen MWB, BMZ, Lexer und ReM. Graphik: Patrick D. Brookshire und Jonas Richter.

liegen dabei als TEI Lex-0 vor, einem Substandard für lexikalische Ressourcen.[11] Die Primärquellen im MWB-Textarchiv folgen hingegen dem generischeren TEI-P5-Format, wobei streng zwischen dem tokenisierten Volltext mit vor allem typographischen Auszeichnungen und *Stand-off*-Annotationen für Lemmatisierungen unterschieden wird, sodass sich beide Ebenen unabhängig voneinander betrachten lassen. Die Nutzung sogenannter *Feature Structures* ermöglicht eine potentielle (spätere) Anreicherung mit beliebigen weiteren Annotationslayern.[12] Hier können künftige korpuslinguistische Projekte ansetzen, die Flexions- oder Dependenzinformationen eingliedern, oder literaturwissenschaftliche Projekte mit narratologischen Annotationen. Die in der Frühphase des Projekts geschaffene Basis

11 Vgl. Tasovac, Romary, Banski u. a. 2018.

12 Vgl. Griesinger u. Stolz 2019, S. 123 f.

eines lemmatisierten Belegarchivs wird im Rahmen der Wörterbucharbeit laufend ergänzt, indem bedarfsweise weitere Textstellen mit den Lemmata über IDs verknüpft werden (vgl. Abb. 1). Automatische Verfahren wie der TreeTagger[13] mit einer mittelhochdeutschen Parameterdatei[14] ergänzen die manuellen Lemmatisierungen, auch wenn dieses Verfahren Lemmaformen der ‚Mittelhochdeutschen Begriffsdatenbank' (MHDBDB) ohne die im MWB verwendeten grammatischen Angaben erzeugt. Darüber hinaus wurden neuere Tools wie der RNN-Tagger[15] bereits angetestet. Die *TEI Feature Structures* erlauben eine entsprechende Unsicherheitsmarkierung, sodass in den Daten selbst festgehalten wird, dass im MWB zwar jedes Token lemmatisiert ist, aber nicht alle Lemmatisierungen manuell geprüft sind.

3 Schnittstellen

Das Netz mittelhochdeutscher Wörter und Projekte zeichnet sich seit der Jahrtausendwende durch Schnittstellen wie die Mittelhochdeutsche Gesamtlemmaliste aus, mit der IDs zwischen Projekten ausgetauscht werden können. Darüber hinaus ist es über die föderierte Architektur des NFDI-Konsortiums Text+ seit dem Jahr 2023 möglich, mittelhochdeutsche Daten über Projektgrenzen hinaus gemeinsam abzufragen. Mit diesen beiden Schnittstellen befassen sich exemplarisch die folgenden Unterabschnitte.

3.1 Mittelhochdeutsche Gesamtlemmaliste

In der Vorbereitungsphase des MWB (1994–1999) wurde aus den Einträgen der Vorgängerwörterbücher (BMZ, Lexer samt Nachträgen und Findebuch) eine Lemmaliste kompiliert, die als zentrale Referenz von Anfang an eine Art Rückgrat des Projekts gebildet hat.[16] Einander entsprechende Lemmata aus den älteren Wörterbüchern wurden dabei zusammengeführt. Das Resultat wurde für die Zwecke des MWB bearbeitet und stellt daher kein Gesamtverzeichnis aller Einträge des Vorgängerwörterbuchs dar, da diese unter anderem auch Eigennamen und Verweisformen enthielten. In dem resultierenden Lemmagerüst hat jedes Stichwort eine feste ID und ist mit den zugehörigen Einträgen in den Vorgängerwörterbüchern sowie den nach und nach im MWB entstehenden Artikeln verlinkt. Lemmatisierungen im Volltextarchiv sind

13 Vgl. Schmid 1994.

14 Vgl. Echelmeyer, Reiter u. Schulz 2017.

15 Vgl. Schmid 2019.

16 Vgl. Plate u. Recker 2001, S. 177–181.

ebenfalls mit den Stichwörtern in dieser Gesamtlemmaliste verknüpft, die dadurch hervorragend als Rechercheinstrument für Fragestellungen zur mittelhochdeutschen Lexikographie und darüber hinaus taugt. So empfiehlt auch Hinkelmanns, die IDs der MWB-Lemmaliste, auf die auch das ‚Referenzkorpus Mittelhochdeutsch' (ReM) verweist, zur eindeutigen Identifizierung mittelhochdeutscher Lemmaformen zu nutzen.[17] Wie sich daraus ein Netz mittelhochdeutscher Wörter entwickelt hat, illustriert Abbildung 1. Sie zeigt, dass im MWB-Textarchiv allen Wortformen (Tokens) eindeutige IDs zugewiesen sind wie zum Beispiel *edeler* in *Saget, edeler schribere* (ReinFu K, 2203).[18] Diese ID wird dann über die *Stand-off*-Lemmatisierung mit der ID des MWB-Lemmas *edele* verlinkt, das über die Lemmaliste bidirektional mit den Äquivalenten *edele, edel* im BMZ bzw. *edel, edele* im Lexer verknüpft ist. Zusätzlich ist diese Textstelle auch ein Beleg im MWB-Artikel, sodass hier ebenfalls eine projektinterne bidirektionale Verlinkung vorliegt. Diese Vernetzung ist jedoch nicht auf Wörterbuchartikel beschränkt, wie beispielsweise das Token *édele* in *édele uuîntrûbo uóne cýproie* (Will 6ra, 25) aus dem ReM zeigt. Insbesondere Projekte wie das ReM, die ihre Annotationen nicht alphabetisch vornehmen, profitieren dabei von der Tatsache, dass mit der Mittelhochdeutschen Gesamtlemmaliste das MWB auch Verweisziele für noch nicht publizierte Artikel bereithält, um so das Netz der mittelhochdeutschen Ressourcen weiter vergrößern zu können.

3.2 Föderierte Suche

An dem von Hinkelmanns beschriebenen Befund, dass gerade lexikalische Ressourcen sich durch eine Vielzahl unterschiedlicher Datenformate auszeichnen,[19] hat sich in den letzten Jahren nur bedingt etwas geändert. Umso wichtiger sind daher Bestrebungen, standardisierte Schnittstellen zu implementieren, die noch mehr Vernetzungen zwischen Projekten ermöglichen. Das MWB hat sich dabei wie das ‚Wörterbuchnetz' zunächst im Rahmen des Projektes ‚eHumanities-Zentrum für historische Lexikographie' (ZHistLex)[20] und jüngst durch Anbindung an die Föderierte Suche (FCS, *Federated Content Search*) von Text+[21] engagiert. Letztere eröffnet neue Möglichkeiten, da hiermit nicht nur lexikalische Ressourcen wie das MWB und seine Vorgängerwörterbücher, sondern auch die Quellentexte gemeinsam durchsucht werden können. Das kommt der Vision von Griesinger und

17 Vgl. Hinkelmanns 2019, S. 122.

18 Siglen folgen der MWB-Form – vgl. http://mhdwb-online.de/quellenverzeichnis.php (Zugriff: 11.03.2025)

19 Vgl. Hinkelmanns 2019, S. 133.

20 Vgl. ebd., S. 140 f., und Pietsch 2020 sowie https://zhistlex.de/ (Zugriff: 11.03.2025).

21 Vgl. Brookshire u. Recker-Hamm 2023.

Stolz, digitale Editionen und lexikalische Ressourcen unter Beachtung gemeinsamer Standards enger miteinander zu verbinden, schon recht nahe.[22]

Die FCS basiert auf standardisierten Abfragesprachen (OASIS-CQL[23] für einfache Suchen, FCS-QL[24] für Korpussuchen und LexCQL[25] für lexikalische Suchen) sowie standardisierten Ergebnisformaten[26] und Fehlerdiagnostiken. Ein projektspezifischer Endpunkt wie der des MWB-Textarchivs übersetzt dabei eine Suchanfrage gemäß der FCS-Syntax in eine interne (XQuery-)Syntax und transformiert die gefundenen (TEI-)Fragmente in das gemeinsame FCS-Format. Dadurch abstrahiert die gemeinsame Suche sowohl von der Datenhaltung als auch dem jeweiligen technischen Setup der Einzelprojekte. So findet zum Beispiel eine Suche nach ‚Terramer' über den FCS-Aggregator von Text+[27] nicht nur Treffer in mittelhochdeutschen Projekten wie BMZ, MWB und ReM, sondern unter anderem auch im ‚Deutschen Zeitungsportal der Deutschen Digitalen Bibliothek', da die Ausgabe vom 4.10.1941 der ‚Kölnischen Zeitung' einen Auszug aus Wolframs von Eschenbach ‚Willehalm' enthält. Dieser Zufallsfund könnte für an der Rezeptionsgeschichte Interessierte eine wichtige Rolle spielen, zumal die entsprechende Quelle im Aggregator verlinkt ist. Besteht hingegen Interesse an der Person, kann mit derselben Suche der BMZ-Artikel gefunden werden, aus dem hervorgeht, dass er „[S]ohn des Kanabêus" war,[28] während die Treffer im MWB und ReM zu weiteren Belegstellen führen. Einer dieser MWB-Belege stammt aus WhvÖst (7724 f.) und wird im Aggregator bei einer Multi-Layer-Korpussuche im sogenannten AdvancedDataView folgendermaßen angezeigt:

Tabelle 1 | Beispiel für FCS-AdvancedDataView-Layer.[29]

Layer	**...**	**s8**	**s9**	**s10**	**s11**	**s12**	**s13**	**s14**	**s15**	**...**
text	...	der	kůnc	Terramer/	gewůnne	nie	so	manigen	man./	...
lemma	...	der	künic	Terramer	ge-winnen	nie	sô	manigen	man	...
pos	...	DET	NOUN	PROPN	VERB	ADV	ADV	ADJ	NOUN	...

22 Vgl. Griesinger u. Stolz 2019, S. 118, 121.

23 http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/os/part5-cql/searchRetrieve-v1.0-os-part5-cql.html (Zugriff: 11.03.2025).

24 Vgl. van Uytvanck, Olsson, Schonefeld u. a. 2023, S. 39–42.

25 Vgl. Körner, Eckart, Herold u. a. 2023, S. 5 f.

26 Vgl. van Uytvanck, Olsson, Schonefeld u. a. 2023, S. 18–25.

27 https://fcs.text-plus.org/?&query=Terramer&queryType=cql (Zugriff: 11.03.2025).

28 https://woerterbuchnetz.de/?sigle=BMZ&lemid=T00393&hitidlist=2705807 (Zugriff: 11.03.2025).

29 https://fcs.text-plus.org/?&query=%5B%20word%20%3D%20%22Terramer%22%20%5D&queryType=fcs (Zugriff: 11.03.2025).

In Tabelle 1 sieht man, dass die Korpussuche mit Layern arbeitet und sich auf diese Weise nicht nur durch Anbinden weiterer Projekte horizontal erweitern lässt, sondern auch vertikal durch weitere Annotationslayer. So ließen sich potentiell auch Angaben zur im MWB ausgezeichneten (Un-)Sicherheit der Lemmatisierungen integrieren, um die Nachnutzbarkeit zu erhöhen. Des Weiteren bieten sich auch die für das Netz mittelhochdeutscher Lemmata so zentralen IDs künftig als ein weiterer Layer an. Allerdings kann umgekehrt die Einführung eigener projektspezifischer Layer die Interoperabilität reduzieren, weshalb beide bislang noch nicht eingebunden wurden.

4 Nachnutzungsperspektiven

Das MWB kann Anknüpfungspunkt für eine Reihe anderer digitaler mediävistischer Forschungsprojekte sein,[30] wodurch sich wiederum Möglichkeiten für weitere Vernetzungen ergeben. Aus diesem Grund werden in den folgenden Teilabschnitten Nachnutzungsperspektiven der MWB-Daten aufgezeigt, die über die reine Wörterbucharbeit als Hilfestellung zum Verständnis mittelhochdeutscher Texte und die Vernetzung der Daten über IDs hinausgehen. So liegen etwa, wie bereits erwähnt, alle zentralen Ressourcen des MWB als TEI-XML vor, wodurch sie per se nachnutzbar und interoperabel sind.[31] Die automatischen Lemmatisierungen der Primärquellen ermöglichen darüber hinaus statistische Analysen, die einen hohen *Recall* einfordern. Projektintern stehen den Redakteuren bereits Möglichkeiten zur Gruppierung von Belegen nach Textmetadaten zur Verfügung, die gerade bei großen Treffermengen die Sichtung erleichtern, womit sich Abschnitt 4.1 befasst. Hier könnten in Zukunft auch Tasks aus dem Bereich *Natural Language Processing* wie eine Genre-Identifikation oder *Topic Modelling* ansetzen. Mit den manuellen Annotationen lassen sich *Deep Learning*-Modelle umgekehrt durch ihre hohe *Precision* nicht nur zur Lemmatisierung, sondern zum Beispiel auch zur Eigennamenerkennung (nach-)trainieren und/oder evaluieren, wozu im Abschnitt 4.2 eine Pilotstudie präsentiert wird.

4.1 Kookkurrenzanalysen

Ein Tool zur Suche nach Kookkurrenzen, das demnächst auch öffentlich verfügbar sein wird, erlaubt die Suche nach Wortformen oder Lemmata in Kombination mit Bedingungen für den Kontext (zum Beispiel ein bestimmtes Wort oder eine

30 Zur Forschungslandschaft vgl. z. B. Bleier, Fischer, Hiltmann u. a. 2019.

31 Siehe oben unter 2.

Wortart im rechten Kontext). Dabei ist es möglich, alle Lemmatisierungen auszuwerten oder die Suche auf die nicht-automatischen, lexikographisch geprüften Lemmatisierungen zu beschränken.

So zeigt die Umgebungssuche beispielsweise, dass unter 1.504 Treffern zum Adjektiv *reine* häufig Kookkurrenzen mit den Substantiven *maget* (140), *wîp* (111) und *got* (135) vorkommen. Befunde, die die Wörterbücher nahelegen oder die der eigenen Erwartungshaltung entsprechen, werden in solchen Suchergebnissen intersubjektiv überprüfbar. Andererseits können auch weniger erwartete Kookkurrenzen wie *reine man* (150) gefunden werden. Der Großteil der Belege hierfür stammt aus dem ‚Märterbuch' und Rudolfs von Ems ‚Weltchronik', in denen mit dem Ausdruck überwiegend Heilige und biblische Gestalten bezeichnet werden. Bei der Artikelredaktion helfen solche Informationen, um verbreitete Verwendungstypen zu finden. Auch lassen sich gezielt Textstellen suchen, in denen *reine* zum Beispiel in der Nähe von Wortformen steht, die im *Part-of-Speech* (PoS) *Tagging* als Adjektive markiert sind, um zu prüfen, ob es auffallende Verbindungen wie mögliche (Teil-)Synonyme, Antonyme oder Paarformeln gibt.

Das Tool ermöglicht zudem Suchanfragen, die über lexikographische Interessen hinausgehen. So könnte das Textarchiv beispielsweise nach Vorkommen bestimmter Phraseologismen durchsucht oder die Verwendung eines Wortes wie *würfel* (oder semantisch verwandter Wörter) in der Heldenepik unter die Lupe genommen werden, um eine Materialbasis für literaturgeschichtliche oder sprachwissenschaftliche Studien aufzubauen.

Die Suchergebnisse lassen sich auch nach weiteren Kriterien sortieren. Die dafür nötigen Annotationen zur Entstehungsdatierung oder schreibsprachlichen Einordnung der Quellentexte entstanden im Zusammenhang mit dem ZHistLex-Projekt und bedürfen weiterer Ausarbeitung. Ähnliches gilt für die Textsorten-Annotation der rund 1.800 Primärquellen im MWB-Quellenverzeichnis: Diese Daten stammen zu großen Teilen noch aus der Anfangsphase des MWB. Um sie zur Gruppierung oder Filterung der Texte beispielsweise in den Ergebnissen der Umgebungssuche verwenden zu können, ist eine Vereinheitlichung und ontologische Aufbereitung notwendig. Dafür stehen im Rahmen der Wörterbucharbeit jedoch weder Zeit noch Mittel zur Verfügung, künftige Projekte könnten aber diese Anlagen weiter ausbauen. So wäre etwa die Verknüpfung der MWB-Quellen mit der Textreihentypologie von Zeppezauer-Wachauer und Heiles, einem SKOS-Vokabular,[32] wünschenswert. Der Nutzen solcher Datenanreicherung ist offensichtlich, würde er doch nicht nur eine Filterung des Primärquellenverzeichnisses ermöglichen, sondern auch anhand der Quellenzitate im Wörterbuch eine Auswertung des Wortschatzes nach diesen Metadaten erlauben.

32 Vgl. Zeppezauer-Wachauer u. Heiles 2023. SKOS (*Simple Knowledge Organization System*) ist ein vom W3C empfohlener Standard zur Kodierung kontrollierter Vokabulare.

Hier zeigt sich ein generelles Problem von Langzeitprojekten, die in ihrer Laufzeit zwar an einer umfangreichen Aufgabe sehr intensiv arbeiten können, andererseits aber an einen Projektrahmen gebunden sind, der über viele Jahre – oft Jahrzehnte – bestehen bleibt und nicht flexibel auf neue Anforderungen reagiert, die sich in der Fachwelt etablieren. Dies hat zur Folge, dass Forschungsdaten im Rahmen eines vor langer Zeit gestarteten Projekts – wenn überhaupt – nur in der Form eines wenig befriedigenden Kompromisses an konzeptuelle oder technische Gegebenheiten (etwa Datenstandards), die heutzutage als offensichtliches Desiderat gelten mögen, angepasst werden können.

4.2 Eigennamenerkennung

Während Kookkurrenzanalysen ein Beispiel für eine stärker korpuslinguistisch motivierte Nachnutzungsperspektive der MWB-Daten sind, bildet eine automatische Erkennung von Eigennamen (*Named Entity Recognition,* NER) Anknüpfungspunkte für Projekte aus den Bereichen *Digital Humanities* und *Computational Literary Studies*. Ein Beispiel hierfür sind die (Figuren-)Netzwerkanalysen von Braun und Ketschik, die auf der Auswertung von Figurennennungen basieren.[33] „Personendaten" sind mit den Worten von Hinkelmanns „natürlich nicht das Kerngeschäft des MWBs oder der MHDBDB".[34] Nichtsdestotrotz wurden in einigen Texten Personen- und Ortsnamen manuell ausgezeichnet, wenn auch ohne Differenzierung zwischen diesen beiden Entitätentypen. Diese Annotationen lassen sich als Basis für NER-Algorithmen nachnutzen, um in beliebigen unausgezeichneten mittelhochdeutschen Texten entsprechende Wortformen zu finden. Eine Möglichkeit ist dabei ein simples *String Matching* der erfassten Wortformen (Types), eine weitere das Finetuning eines vortrainierten *Deep Learning*-Modells mit den Annotationen. Für eine diesbezügliche Pilotstudie wurden 3.000 zufällige Textstellen aus elf annotierten Texten ausgewählt und diese in Trainings- bzw. Testdaten aufgeteilt.

33 Vgl. Braun u. Ketschik 2019.

34 Hinkelmanns 2019, S. 139.

Tabelle 2 | Anzahl der Textstellen im Trainings- und Testdatenset pro Text (eingeklammert die Anzahl der Textstellen, die mindestens eine Entität enthalten).

Text	Trainings-Textstellen	Test-Textstellen	Summe
Parz	1.032 (724)	263 (185)	1.295 (909)
Wh	625 (473)	150 (117)	775 (590)
UvZLanz	267 (174)	68 (41)	335 (215)
RvEBarl	141 (117)	42 (34)	183 (151)
Er	107 (83)	18 (16)	125 (99)
Wig	104 (97)	20 (17)	124 (114)
VAlex	90 (68)	29 (23)	119 (91)
Vateruns	25 (5)	9 (4)	34 (9)
VEzzo	4 (1)	1 (1)	5 (2)
SuTheol	2 (2)	0 (0)	2 (2)
SüklZw	1 (1)	0 (0)	1 (1)
Summe	2.400 (1746)	600 (438)	3.000 (2184)

Die Textstellen sind hier wie beim Tool zur Kookkurrenzanalyse *Keywords in Context* mit je zehn vorangehenden bzw. nachfolgenden Tokens um eine beliebige großgeschriebene Wortform (vgl. Tab. 3), wodurch sie strukturell Ergebnissen aus der FCS ähneln.[35] Als Keywords wurden dabei bewusst nicht nur Eigennamen gewählt, damit das Sprachmodell nicht einfach nur die Features Großschreibung oder Position in der Textstelle als diskriminierendes Merkmal lernen kann. Letzteres wird zudem durch die Tatsache erschwert, dass im Kontextfenster weitere Eigennamen vorkommen können wie beispielsweise *Pantanor* im linken Kontext in Tabelle 3. Insgesamt umfasst das Datenset 62.916 Tokens, wovon 4.467 (~7,1 Prozent) manuell als Eigenname ausgezeichnet sind.

35 Vgl. Tabelle 1 und 3.

Tabelle 3 | Beispiele für Trainings- bzw. Test-Textstellen (Eigennamen hervorgehoben).

Quelle	linker Kontext					Keyword	rechter Kontext				
Parz 566,1	...	er	in	dienen	wolte.	Da!	wârn	si	doch	unschuldec	...
		O	O	O	O	O	O	O	O	O	
Wh 77,24	...	**Pantanor**	gap	dem	künege	**Salatrê,**	der	gabz	dem	künege	...
		B	O	O	O	**B**	O	O	O	O	

Die Eigennamenannotation der Trainings- und Testdaten folgt dem verbreiteten IOB-Schema, das auf Ramshaw und Marcus zurückgeht,[36] wobei nur zwischen den Labeln *O* (= keine Entität) und *B* (= erstes Token einer Entität) unterschieden wird. Das Label *I* (= Folgetoken einer zusammengesetzten Entität) wird nicht benötigt, da in diesen Texten Mehrwort-Entitäten keine nennenswerte Rolle spielen. Durch den Rückgriff auf dieses Standardformat ist eine spätere Einbindung in verschiedene Tools denkbar.

Mit den Trainingsdaten wurde ein Finetuning zweier deutscher Transformermodelle mit dem Python-Modul Transformers[37] sowie Default-Parametern über drei Epochen durchgeführt.[38] Das erste Modell GHistBERT wurde von Beck und Köllner auf ca. 660.000 Sätzen aus Texten vortrainiert, die zwischen 750 und 1999 entstanden sind.[39] In diesen Daten finden sich nach Angaben der Autorinnen auch 246.000 (~37,5 Prozent) mittelhochdeutsche Sätze aus dem ReM, weshalb die Möglichkeit besteht, dass es einige Textstellen bereits gesehen hat. Das ist bei dem anderen in dieser Pilotstudie gewählten Modell ‚gbase-bert'[40] nicht zu erwarten, weil hier die Trainingsgrundlage 163,4 GB deutschsprachige Internetseiten sind. Um einschätzen zu können, wie gut diese Modelle Entitäten finden können, wurden, wie bei entsprechenden Benchmarks üblich, neben der sogenannten *Majority Baseline*, die überhaupt keine Entitäten erkennt, noch weitere bereits existierende Methoden getestet. Das erste Verfahren ist dabei der TreeTagger von Schmid[41], der eigentlich ein PoS-Tagger ist und als solcher im MWB für automatische Lemmatisierungen genutzt wird.[42] Allerdings entspricht der PoS-Tag PROPN im Prinzip der Eigennamenannotation im MWB, sodass diese Methode einem Mapping auf das IOB-Schema

36 Vgl. Ramshaw u. Marcus 1995.

37 Wolf, Debut, Sanh u. a. 2020.

38 Für entsprechende Daten und Skripte siehe https://github.com/digicademy/mwb-name-recognition (Zugriff: 11.03.2025)

39 Vgl. Beck u. Köllner 2023.

40 Chan, Schweter u. Möller 2020.

41 Vgl. Schmid 1994; ergänzt um die mittelhochdeutsche Parameterdatei von Echelmeyer, Reiter u. Schulz 2017.

42 Siehe oben unter 2.

vergleichbar ist. Als Letztes wurde mit ‚stanza'[43] noch ein eigentliches NER-Modell getestet, das auf Zeitungsartikeln von 1992[44] und damit ebenso wie ‚gbert-base' nicht auf mittelhochdeutschen Texten basiert. Um die Vergleichbarkeit zu gewährleisten, wurden hier ebenfalls nur Personen und Orte berücksichtigt und weitere Typen auf O (= kein Eigenname) gemappt. Die Klassifikationsergebnisse all dieser Verfahren sowie des *String Matchings* bezogen auf die vier üblichen Kennzahlen *Accuracy*, *Precision*, *Recall* und *makro F1-Score* finden sich in Tabelle 4. Die *Accuracy* misst dabei, wie viel Prozent der Testdatensätze insgesamt richtig klassifiziert wurden, während die anderen Maße sich auf die relevanten Modellvorhersagen beziehen. So misst die *Precision*, wie viel Prozent der vom Modell als Eigenname ausgezeichneten Tokens auch wirklich Eigennamen sind, der *Recall*, wie viel Prozent der Eigennamen gefunden wurden, und der F1-Score kombiniert beide Werte.

Tabelle 4 | Klassifikationsergebnisse verschiedener Methoden zur Eigennamenerkennung (beste Ergebnisse fett hervorgehoben; bei Modellen mit * erfolgte ein Finetuning).

Methode	Datengrundlage	Accuracy	Precision	Recall	F1 (makro)
Majority Baseline	-	93,0 %	0,0 %	0,0 %	48,2 %
String Matching	MWB-Wortformenliste	96,6 %	**99,6 %**	52,0 %	83,3 %
TreeTagger	MHDBDB	98,2 %	96,2 %	77,7 %	92,5 %
stanza	CoNLL03	94,1 %	86,8 %	19,3 %	64,3 %
GHisBERT*	MWB-Textarchiv	97,6 %	80,7 %	86,3 %	91,1 %
gbert-base*	MWB-Textarchiv	**99,4 %**	95,6 %	**95,7 %**	**97,7 %**

Erwartungsgemäß sind die Werte aller vier Kennzahlen bei der *Majority Baseline* am niedrigsten. Allerdings verdeutlicht hier die *Accuracy* von rund 93,0 Prozent die erwähnte Seltenheit von Eigennamen in diesem Datenset. Bezogen auf *Accuracy* und F1-Score war das ‚stanza'-NER-Modell für modernes Deutsch am zweitschlechtesten, was primär am niedrigen *Recall* liegt. Grund hierfür dürfte die geringe Anzahl an für mittelhochdeutsche Texte relevanten Eigennamen in heutigen Zeitungsartikeln sein. Einen zwar deutlich besseren, aber ebenfalls im Vergleich niedrigen *Recall* erzielte auch der *String-Matching*-Ansatz, der jedoch wegen seiner mit Abstand höchsten *Precision* von 99,6 Prozent immer noch einen

43 Qi, Zhang, Zhang u. a. 2020, S. 142 f.

44 Tjong Kim Sang u. De Meulder 2003.

hohen Nachnutzwert hat. Ähnliches gilt für den TreeTagger, der zwar etwas weniger präzise war, aber dafür einen deutlich höheren *Recall* einbrachte. Die beiden nachtrainierten Modelle waren in Bezug auf das Verhältnis von *Precision* und *Recall* (und damit den F1-Score) ausgewogener. Interessanterweise war beim diachronen Modell jedoch die *Precision* am zweitniedrigsten, während das mit modernem Deutsch vortrainierte ‚gbert-base' mit einer *Accuracy* von 99,4 Prozent und einem F1 von 97,7 Prozent insgesamt am besten abschnitt. Möglicherweise liegt das daran, dass dieses Modell besser generalisieren kann und daher beim Finetuning der Transfer besser gelang. Somit könnte dieses neue Modell schon jetzt in anderen Projekten zur automatischen Eigennamenerkennung in mittelhochdeutschen Texten genutzt werden. Darüber hinaus unterstreichen die Ergebnisse, dass, selbst wenn im MWB Annotation und Lemmatisierung von Eigennamen nur tentativ erfolgten, diese Anlage nachgenutzt werden kann – sei es innerhalb des Projektes zum *Pretagging* noch unausgezeichneter Eigennamen oder zur Vorbereitung anderer Personen- bzw. Ortsnamen-basierter Analysen.

5 Fazit

Mit der Bitte des Verlegers Salomon Hirzel, dass Matthias Lexer ein Handwörterbuch verfassen möge, das sich eng auf das Wörterbuch von Benecke-Müller-Zarncke bezieht, begann die Vernetzung des mittelhochdeutschen Wortschatzes. Heute ist das Netz gewachsen und weist weit über seinen ursprünglichen Horizont hinaus. Weitverbreitete Datenformate und Standards zu Schnittstellen ermöglichen, Annotationslayer in Einzeltexten oder Textkorpora mit Verweisen auf Einträge in der Lemmaliste des MWB anzureichern und damit nicht nur Wortformen zu lemmatisieren, sondern sie indirekt auch mit den einschlägigen Artikeln in den mittelhochdeutschen Wörterbüchern zu verbinden. Schnittstellen, die zwischen nativen Datenformaten und externen Anfragen ‚übersetzen' können, erlauben der Forschung, mit föderierten Suchen wie der FCS von Text+ gezielte Fragen an eine große Anzahl nicht nur mediävistischer Projekte zu stellen. Nicht immer sind die technischen Grundlagen dafür bei Langzeitprojekten von Anfang an gegeben, wie es beispielsweise bei den IDs der Lemmaliste der Fall ist. In anderen Fällen, wie etwa bei den Schnittstellen, die die Anbindung an die FCS ermöglichen, müssen Drittmittel eingeworben werden. Die Pilotstudie zur Entitätenerkennung lässt aber hoffen, dass auch diejenigen Datenanreicherungen, die nicht mit größter Konsequenz und Einheitlichkeit, sondern eben nur tentativ ‚nebenbei' vorgenommen werden, bereits einen Mehrwert darstellen und als Basis für weiterführende Untersuchungen dienen können. Dazu müssen die Vernetzungsmöglichkeiten prinzipiell offen und erweiterbar angelegt sein.

Die am Beispiel mittelhochdeutscher Daten hier exemplarisch vorgestellte Verbesserung der Zugänglichkeit und Interoperabilität steht dabei für eine generelle Tendenz in der Mediävistik, die in Zukunft zu einer noch stärker interdisziplinären Forschung führen kann. Denn gerade durch die Vernetzung unterschiedlicher Informationen können neue Muster identifiziert und analysiert werden.

Literaturverzeichnis

Beck, Christin u. Marisa Köllner: GHisBERT – Training BERT from Scratch for Lexical Semantic Investigations Across Historical German Language Stages. In: Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change. Singapur 2023, S. 33–45. https://doi.org/10.18653/v1/2023.lchange-1.4 (Zugriff: 11.03.2025).

Bleier, Roman, Franz Fischer, Torsten Hiltmann u. a. (Hgg.): Digitale Mediävistik (Das Mittelalter. Perspektiven mediävistischer Forschung 24). Berlin 2019.

BMZ = Mittelhochdeutsches Wörterbuch, mit Benutzung des Nachlasses von Georg Benecke ausgearb. von Wilhelm Müller und Friedrich Zarncke. Leipzig 1854–66. Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/23, https://www.woerterbuchnetz.de/BMZ (Zugriff: 11.03.2025).

Brookshire, Patrick D. u. Ute Recker-Hamm: Text+Plus, #03: Verlinkungen und standardisierte Suchen: MWB-APIplus. In: Text+ Blog. 2023. https://textplus.hypotheses.org/8245 (Zugriff: 11.03.2025).

Chan, Branden, Stefan Schweter u. Timo Möller: German's Next Language Model. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona 2020, S. 6788–6796. https://doi.org/10.18653/v1/2020.coling-main.598 (Zugriff: 11.03.2025).

Diehl, Gerhard u. Nils Hansen: Zwischen Handschrift und Online-Datenbank. Bemerkungen zur Lexikographie des Mittelhochdeutschen. In: Anja Lobenstein-Reichmann u. Peter O. Müller (Hgg.): Historische Lexikographie zwischen Tradition und Innovation (Studia Linguistica Germanica 129). Berlin 2016, S. 287–305.

Echelmeyer, Nora, Nils Reiter u. Sarah Schulz: Ein PoS-Tagger für „das" Mittelhochdeutsche. In: Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum. Bern 2017, S. 141–147. http://dx.doi.org/10.18419/opus-9023 (Zugriff: 11.03.2025).

Fournier, Johannes: Digitale Dialektik: Chancen und Probleme mittelhochdeutscher Wörterbücher in elektronischer Form. In: Herbert Ernst Wiegand (Hg.): Wörterbücher in der Diskussion IV. Vorträge aus dem Heidelberger Lexikographischen Kolloquium (Lexicographica Series Maior 100). Tübingen 2000, S. 85–108.

Gärtner, Kurt: Das Handexemplar von Matthias Lexers ‚Mittelhochdeutschem Handwörterbuch'. In: Horst Brunner (Hg.): Matthias von Lexer. Beiträge zu seinem Leben und Schaffen (Zeitschrift für Dialektologie und Linguistik, Beiheft 80). Stuttgart 1993, S. 109–131.

Gärtner, Kurt, Christoph Gerhardt, Jürgen Jährling u. a.: Findebuch zum mittelhochdeutschen Wortschatz. Stuttgart 1992.

Gärtner, Kurt u. Ralf Plate: Die Doppelfunktion des digitalen Textarchivs als Wörterbuchbasis und als Komponente der Online-Publikation. Am Beispiel des Mittelhochdeutschen Wörterbuchs. In: Ingelore Hafemann (Hg.): Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011. Berlin 2013, S. 193–204. URN: urn:nbn:de:kobv:b4-opus-24400

Griesinger, Christian u. Michael Stolz: Sprachwissenschaftliche Erschließungsmethoden für digitale Editionen mittelhochdeutscher Texte. In: Bleier, Fischer, Hiltmann u. a. 2019, S. 112–128.

Hinkelmanns, Peter: Mittelhochdeutsche Lexikographie und Semantic Web. Die Anbindung der ‚Mittelhochdeutschen Begriffsdatenbank' an Linked Open Data. In: Bleier, Fischer, Hiltmann u. a. 2019, S. 129–141.

Körner, Erik, Thomas Eckart, Axel Herold u. a.: Federated Content Search for Lexical Resources (LexFCS): Specification. Version 0.1, 2023-05-04. https://doi.org/10.5281/zenodo.7986303 (Zugriff: 11.03.2025).

Lexer, Matthias: Mittelhochdeutsches Handwörterbuch. Zugleich als Supplement und alphabetischer Index zum Mittelhochdeutschen Wörterbuch von Benecke-Müller-Zarncke. Band 1, Stuttgart 1872.

Nellmann, Eberhard: Die mittelhochdeutschen Wörterbücher. Ihre Qualitäten, ihre Grenzen, ihre mögliche Erneuerung. In: Eijiro Iwasaki u. Yoshinori Shichiji (Hgg.): Begegnung mit dem „Fremden": Grenzen – Traditionen – Vergleiche. Akten des VIII. Internationalen Germanisten-Kongresses, Tokyo 1990. Bd. 4. Sektion 4, Kontrastive Syntax; Sektion 5, Kontrastive Semantik, Lexikologie, Lexikographie; Sektion 6, Kontrastive Pragmatik. München 1991, S. 254–263.

Nellmann, Eberhard: Quellenverzeichnis zu den mittelhochdeutschen Wörterbüchern. Ein kommentiertes Register zum ‚Benecke/Müller/Zarncke' und zum ‚Lexer'. Stuttgart, Leipzig 1997.

Pietsch, Andre: Historische Wörterbücher und Wörterbuchsysteme in digitalen Umgebungen: Komponenten, Verknüpfungen, Darstellungsmittel, Nutzungsoptionen. In: ZHistLex-Papiere. 2020. https://zhistlex.de/papiere/pietsch_2020_digitale_woerterbuecher_ZHistLex.pdf (Zugriff: 11.03.2025).

Plate, Ralf u. Ute Recker: EDV für Wörterbuchzwecke und neue lexikographische Arbeitsweisen. Erfahrungen beim Aufbau des elektronischen Text- und Belegarchivs für das mittelhochdeutsche Wörterbuch. In: Stephan Moser, Peter Stahl, Werner Wegstein u. a. (Hgg.): Maschinelle Verarbeitung altdeutscher Texte V: Beiträge zum Fünften Internationalen Symposion, Würzburg 4.–6. März 1997. Berlin, New York 2001, S. 169–184. https://doi.org/10.1515/9783110909494.169 (Zugriff: 11.03.2025).

Qi, Peng, Yuhao Zhang, Yuhui Zhang u. a.: Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, S. 101–108.

https://doi.org/10.18653/v1/2020.acl-demos.14 (Zugriff: 11.03.2025).

Ramshaw, Lance A. u. Mitchell P. Marcus: Text Chunking Using Transformation-Based Learning. In: Proceedings of the Third ACL Workshop on Very Large Corpora. Cambridge 1995, S. 82–94. https://aclanthology.org/W95-0107/ (Zugriff: 11.03.2025).

Schmid, Helmut: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester 1994, S. 44–49. https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (Zugriff: 11.03.2025).

Schmid, Helmut: Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. Brüssel 2019, S. 133–137. https://doi.org/10.1145/3322905.3322915 (Zugriff: 11.03.2025).

Tasovac, Toma, Laurent Romary, Piotr Banski u. a.: TEI Lex-0: A Baseline Encoding for Lexicographic Data. Version 0.9.3. DARIAH Working Group on Lexical Resources. 2018. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html (Zugriff: 11.03.2025).

Tjong Kim Sang, Erik F. u. Fien De Meulder: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 4. Edmonton 2003, S. 142–147. https://doi.org/10.3115/1119176.1119195 (Zugriff: 11.03.2025).

Uytvanck, Dieter van, Leif-Jöran Olsson, Oliver Schonefeld u. a.: CLARIN Federated Content Search (CLARIN-FCS) – Core 2.0. Version 1.0, 2023-04-26. https://office.clarin.eu/v/CE-2017-1046-FCS-Specification-v20230426.pdf (Zugriff: 11.03.2025).

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg u. a.: The FAIR Guiding Principles for Scientific Data Management and Stewardship. In: Scientific Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18 (Zugriff: 11.03.2025).

Wolf, Thomas, Lysandre Debut, Victor Sanh u. a.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, S. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6 (Zugriff: 11.03.2025).

Zeppezauer-Wachauer, Katharina u. Marco Heiles: Eine digitale Textreihentypologie für deutschsprachige Texte des Mittelalters und der Frühen Neuzeit. Showcase eines kontrollierten Vokabulars in SKOS. In: Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte 6 (2023), S. 6–39. https://doi.org/10.26012/mittelalter-30680 (Zugriff: 11.03.2025).