1 *RIDGES Herbology 9.0* — inhaltliche Modellierung und technische Repräsentation

1.1 Inhalt und Modellierung

Das *RIDGES Herbology-Korpus* — Register in Diachronic German Science — zielt auf die Untersuchung der „Entstehung und Entwicklung der deutschen Wissenschaftssprache seit Ende des 15. Jahrhunderts bis ins 20. Jahrhundert“ und wird am Lehrstuhl für Korpuslinguistik und Morphologie der Humboldt-Universität zu Berlin entwickelt. Auf Grundlage von Texten/Textausschnitten sollen diese analysiert werden, um Entwicklungen und Tendenzen der Sprachverwendung identifizieren und beschreiben zu können (vgl. Odebrecht et al., 2020: S. 7). Um diesem Anspruch gerecht werden zu können, sollte ein über die Jahrhunderte durchgängig vorhandenes Themengebiet gefunden werden: Entschieden wurde sich unter diesem Gesichtspunkt für Kräutertexte. Um nun die Untersuchung durchführen zu können, wird im Rahmen des Projektes ein diachrones Mehrebenenkorpus aufgebaut, welches zum einen ausführliche Metadaten zu allen Texten enthält und zum anderen tiefenannotierte linguistische Informationen zu jedem Token liefert (vgl. Odebrecht et al., 2017: S. 697-701). Im besten Sinne des Humboldtschen Universitätsideal der Einheit von Forschung und Lehre (vgl. Paletschek, 2002: S. 184) werden unter Mitwirkung von Studierenden beständig weitere Texte in das Korpus aufgenommen.
Im Folgenden wird mit der Version 9.0 gearbeitet, die insgesamt 305.056 Token beinhaltet, welche sich auf 73 Texte/Textausschnitte aus dem Zeitraum 1482-1914 verteilen. Dies bedeutet im Durchschnitt 4.180 Token pro Text, wobei die untere Grenze bei 270 und die obere bei 10.282 liegt. Tokenisiert wurde automatisiert via Treetagger 3.2 auf Basis der diplomatischen Transkription der Faksimiles (vgl. Odebrecht et al., 2020: S. 7). Das Metadatenschema — und damit die Modellierung des Gegenstandes — basiert auf den einzelnen Texten/Textausschnitten, zu denen je 25 Einträge (vgl. Tabelle 1.1) erhoben werden, die ich im Rahmen dieser Arbeit in vier Kategorien einteile:

1. bibliographische Angaben (14)

   bibl, author, title, publisher, place, date, orig_place, orig_date, editor, edition_first, issue, translator, trans_from, repository

2. textlinguistische Angaben (6)

   register, lingualism, lang_type, lang_area, text_type, lyric_type

3. inhaltliche Angaben (4)

   maintopic, topic, wormwood, herb_sorting

4. korpusinterne Angaben (1)

   version

Bezeichnung	Beispiel
title	Alchymistische Practic Das ist Von künstlicher Zubereytung der vornembsten Chymischen Medicinen
author	Andreas Libavius
translator	NA
trans_from	NA
date	1603
place	Frankfurt am Main
publisher	Johann Saurn
bibl	Libavius, Andreas (1603) Alchymistische Practic Das ist Von künstlicher Zubereytung der vornembsten Chymischen Medicinen. Frankfurt am Main. Johann Saurn. 5-26.
version	1.0
editor	Petrus Kopff
edition_first	NA
issue	NA
maintopic	science
register	herbology
topic	AlBM
lingualism	multiling
orig_date	NA
orig_place	NA
repository	http://reader.digitale-sammlungen.de/de/fs1/object/display/bsb10872546_00009.html
lang_type	enhg
lang_area	md
text_type	prose
lyric_type	NA
wormwood	no
herb_sorting	no

Tabelle 1: Metadatenschema und Beispiel, basierend auf den TEI-Metadaten und der Datei 'AlchymistischePractic_1603_Libavius.xlsx' (vgl. Lüdeling et al., 2022).

Da aber nun in einigen Fällen mehrere Texte/Textausschnitte aus ein und demselben Werk stammen, tauchen identische bzw. nahezu identische Angaben wiederholt auf. Wenn nun die Metadaten im Hinblick auf die einzelnen Werke bereinigt werden, wird deutlich, dass von der Anzahl der Texte nicht direkt auf die Anzahl der unikalen Werke geschlossen werden kann: Die 73 Texte entstammen 42 Werken (vgl. **??**).

Modelliert wird hier also ein Teil der Welt (vgl. Flanders & Jannidis, 2019: S. 181), in dem Texte/Textausschnitte als einzelne Objekte existieren, denen klar definierte Eigenschaften zugewiesen werden. Es wurden im Korpus-Design Entscheidungen über die Modellierung des Gegenstands getroffen, die bestimmen, auf welchen Ebenen dieser wahrgenommen wird und welche Eigenschaften modelliert werden. Die Modellierung des Gegenstandes im Rahmen des Korpus ist also gar nicht in der Lage, ihn in seiner Gesamtheit zu erfassen. Dies ergibt sich schon aus den Grundeigenschaften eines Modells:

> „[A] model is a representation of something by someone for some purpose at a specific point in time. It is a representation that concentrates on some aspects—features and their relations—and disregards others. The selection of these aspects is not random but functional: it serves a specific function for an individual or a group. And a model is usually only useful and only makes sense in the context of these functions and for the time that they are needed."
> (Flanders & Jannidis, 2019: S. 28)

Die Anforderungen, denen die Modellierung gerecht werden muss, folgen direkt aus der Fragestellung des Projektes: „[W]issenschaftliche Texte [sollen] auf allen sprachlichen Ebenen (Syntax, Wortbildung, Lexik, Phraseologie, Textstruktur etc.) analysiert werden, um Entwicklungen und Tendenzen identifizieren und beschreiben zu können" (Odebrecht et al., 2020: S. 7). Dementsprechend ist es nicht verwunderlich, dass hier der Fokus auf dem Abbilden vorhandener bibliographischer Metadaten — wie bspw. bereitgestellt vom Münchner Digitalisierungszentrum (vgl. https://www.digitale-sammlungen.de/de/details/bsb10872546; zuletzt abgerufen am 20.11.2022) — und dem Ergänzen linguistischer Informationen liegt. Dieselbe Beobachtung lässt sich auch im Hinblick auf die Annotationsebenen der Textinhalte feststellen:

dipl	mir	das	judicium	Athenienſium	von	Hippocrate	belieben	laſſen	/	welchen	ſie
clean	mir	das	judicium	Atheniensium	von	Hippocrate	belieben	lassen	/	welchen	sie
norm	mir	das	Judicium	Atheniensium	von	Hippocrates	belieben	lassen	/	welchen	sie
abbr	no	no	no	no	no	no	no	no	no	no	no
lang	deu	deu	lat	lat	deu	deu	deu	deu	deu	deu	deu
lb	lb					lb					
lemma	ich	die	Judicium	Atheniensium	von	Hippocrates	belieben	lassen	/	welche	sie
pb	pb										
pb_n	3										
persname						Hippokrates_von_Kos					
pos	PPER	ART	NN	NN	APPR	NN	VVINF	VVINF	$(	PWS	PPER
quote	no	no	no	no	no	no	no	no	no	no	no
script	blackletter	blackletter	roman	roman	blackletter	blackletter	blackletter	blackletter	blackletter	blackletter	blackletter
title			Beschluss_der_Athener								

Abbildung 1: Visualisierung einer Abfrage in der ANNIS-Architektur ('persname'; https://korpling.german.hu-berlin.de/annis/#_q=cGVyc25hbWU&ql=aql&_c=UklER0VTX0hlcmJvbG9neV9WZXJzaW9uOS4w&cl=5&cr=&s=0&l=10&_seg=ZGlwbA).

In diesem Beispiel beinhalten zwei der 14 Ebenen (*persname* und *title*) Informationen, die über den konkreten Text hinausgehen und ihn mit weiteren Entitäten verknüpfen; die anderen zwölf bezeichnen die Transkription (*dipl*), die Normalisierungen (*clean, norm*) und strukturelle sowie linguistische Eigenschaften (vgl. Odebrecht et al., 2020: S. 48-150).

Vorstellbar ist aber auch eine andere bzw. erweiterte Modellierung: Aus einer literaturwissenschaftlichen Perspektive wäre beispielsweise eine weitere Einordnung der erkannten Personennamen in 'historische' und 'literarische' denkbar. Ebenso würden Metadaten über Buchwidmungen weitere Ebenen für eine kritische Analyse öffnen — gerade für Bücher der (frühen) Neuzeit sind Widmungen ein fast schon strukturales Merkmal (vgl. Horch, 2014: S. 69-74). Auch würde es sich für eine werk- statt textbasierte weiterführende Analyse anbieten, wenn Metadaten zu den einzelnen Werken vorhanden wären.

Diese rein exemplarischen Überlegungen zeigen, wie wichtig die Interoperabilität und Nachnutzbarkeit der eigenen Modellierung und ihrer technischen Repräsentation für weitere (interdisziplinäre) Forschungsfragen ist. Denn die Möglichkeit, unterschiedliche Modelle und Datenbanken verknüpfen zu können, öffnet die Forschungsdaten für ganz neue Erkenntnisprozesse.

Doch bevor dazu eine mögliche Lösung vorgestellt werden kann, ist es nötig, die technischen Repräsentationen des Korpus RIDGES Herbology 9.0 zu betrachten.

1.2 Repräsentation und Abfrage: ANNIS

Ein Weg, um Abfragen an das Korpus stellen zu können, ist *ANNIS*: „an open source, cross platform (Linux, Mac, Windows), web browser-based search and visualization architecture for complex multi-layer linguistic corpora with diverse types of annotation“ (corpus-tools.org; zuletzt abgerufen am 20.11.2022). Der Lehrstuhl für Korpuslinguistik und Morphologie stellt über einen eigenen Webserver (https://korpling.german.hu-berlin.de/annis/) die ANNIS-Architektur sowie Zugang zu verschiedenen Korpora — darunter auch RIDGES Herbology — frei zur Verfügung.

Grundsätzlich sind die meisten linguistischen Korpora mehr als nur einfache Zeichenketten, da sie viel mehr strukturelle Informationen umsetzen müssen — spätestens dann, wenn linguistische Konzepte wie 'Sätze' dargestellt werden sollen (vgl. Krause, 2019: S. 9):

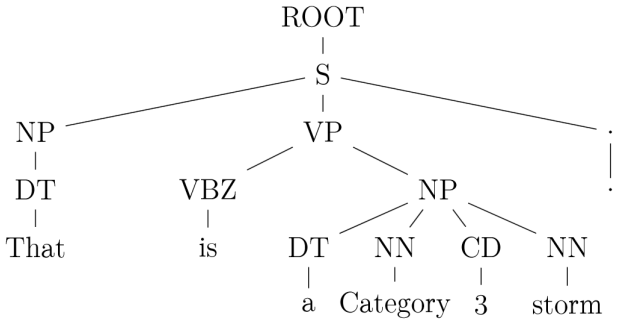

Abbildung 2: Eine exemplarische hierarchische Annotation im 'Penn Treebank Bracket'-Format (Krause, 2019: S. 11, Figure 2.2.).

Hierarchische Annotationen — sog. 'Baumdiagramme' — zeigen sehr gut, weshalb reine Zeichenketten nicht ausreichen, wenn unterschiedliche Korpora, ihre entsprechenden Annotationen und schlussendlich sprachliche Phänomene repräsentiert und abgefragt werden können sollen. Allein schon die Vielzahl linguistischer Theorien und Modellierungen von 'Sprache' im weitesten Sinne lässt eine möglichst generalisierende technische Lösung sinnvoll erscheinen. Und genau dafür bietet sich eine graphbasierte Lösung an, da dort einzelne Knoten über Kanten miteinander in Beziehung gesetzt werden können, ohne dass ein bereits bestehendes Framework die Anwendung einschränkt:

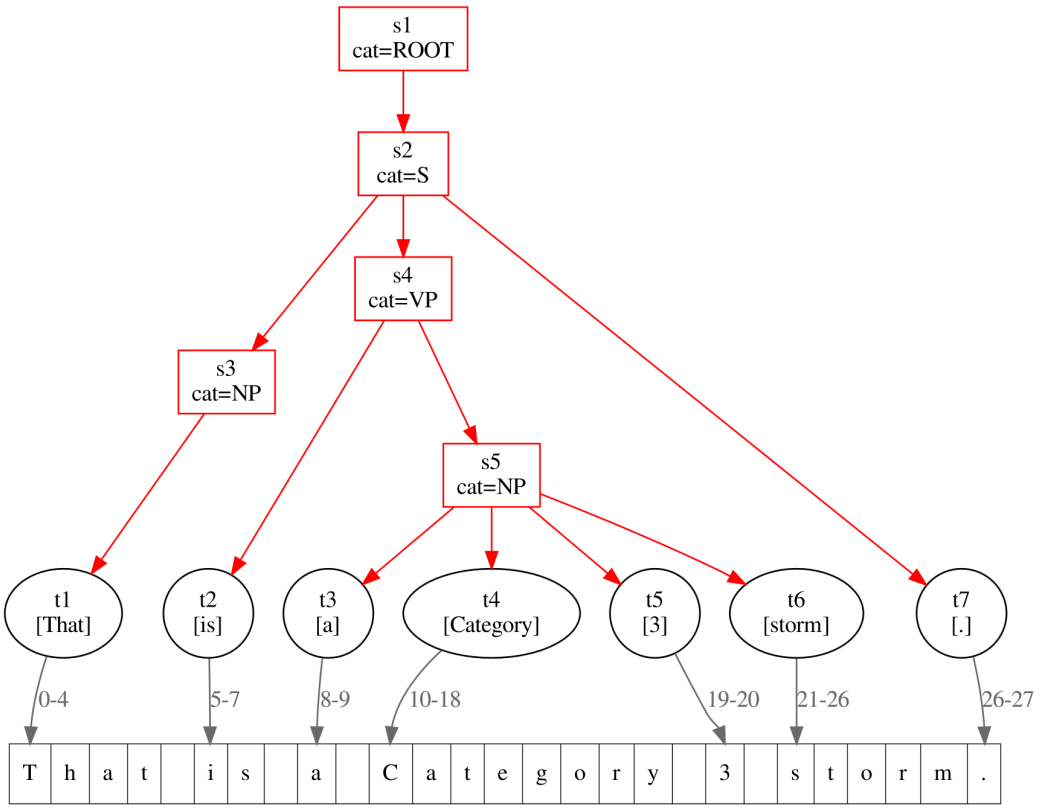

Abbildung 3: Eine exemplarische Repräsentation des Beispiels aus Abb. 1 als Graph: Die kastenförmigen Knoten stellen strukturelle Knoten dar, die runden Knoten einzelne Token; die roten Kanten bilden Dominanzverhältnisse ab (Krause, 2019: S. 25, Figure 3.3.).

Graphbasierte Datenbanksysteme fallen unter die Kategorie der NoSQL-Datenbanken, die über ihre Bezeichnung klar von 'traditionellen' relationalen SQL-Datenbanken abgegrenzt werden. Letztere bestehen aus zweidimensionalen Tabellen, in denen Zellen über Angabe der Zeile und Spalte eindeutig definiert sind (vgl. Gaspar & Coric, 2017: S. 6).

Für graphbasierte NoSQL-Datenbanken gilt ganz besonders: „[They] excel at dealing with highly interconnected data. They focus on relationships rather than data.“ (Gaspar & Coric, 2017: S. 104). Ihre kleinste strukturelle Einheit — Knoten - Kante - Knoten — wird auch als 'triple stores' bezeichnet, für deren Beschreibung das 'Ressource Description Framework' (RDF) als Standard fungiert (vgl. Gaspar & Coric, 2017: S. 104). Für die Linguistik existiert seit 2012 der ISO-Standard 'Linguistic Annotation Framework' (LAF), der ebenfalls auf Graphen basiert und damit als Grundlage zum Harmonisieren bestehender

'language ressources' — darunter auch Korpora — fungieren soll (vgl. Krause, 2019: S. 10).

Auf dieser Grundlage basieren auch ANNIS und die Abfragesprache AQL (ANNIS Query Language) auf einer graphbasierten Repräsentation der Korpusdaten. Für die beiden Versionen vor der aktuellen Version 4 bestand eine technische Besonderheit: Das Datenbankmanagementsystem PostgreSQL wurde genutzt, um intern die Korpusdaten zu speichern und Abfragen auszuführen. Die vom Nutzer in AQL formulierte Abfrage wurde in eine SQL-Abfrage übersetzt, auf der PostgreSQL-Datenbank (*relANNIS*) ausgeführt und die Ergebnisse wiederum graphbasiert abgebildet. Dafür war es nötig, die Graphstruktur in der relationalen Datenbank abzubilden (vgl. Krause, 2019: S. 32f.):

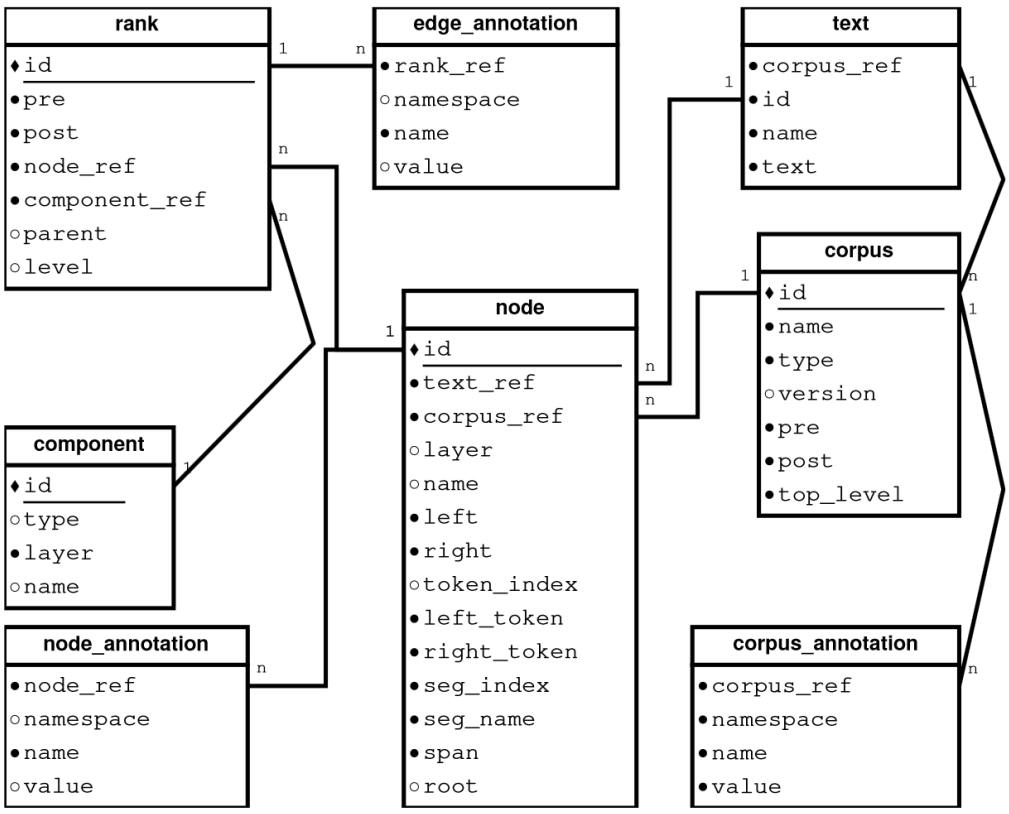

Abbildung 4: Diagramm des PostgrelSQL-Schemas (*relANNIS*), normalisiert (Krause, 2019: S. 34, Figure 3.9.).

Deutlich wird hier, wie kompliziert eine graphbasierten Struktur in einer relationalen Datenbank ist; die eigentliche Simplizität und 'Eleganz' der Tripel gehen verloren. Daher wurde die Version 4 (*graphANNIS*) komplett graphbasiert modelliert (vgl. Krause, 2019: S. 37f.).

1.3 Anforderungen an ein interoperables Datenbanksystem zur Nachnutzung und Anreicherung

Ausgehend von den bisherigen Überlegungen habe ich nun einen Anforderungskatalog an ein Modell und seine technische Repräsentation entwickelt. Ziel ist es, einer größtmöglichen Interoperabilität und Nachnutzbarkeit den Weg zu ebnen:

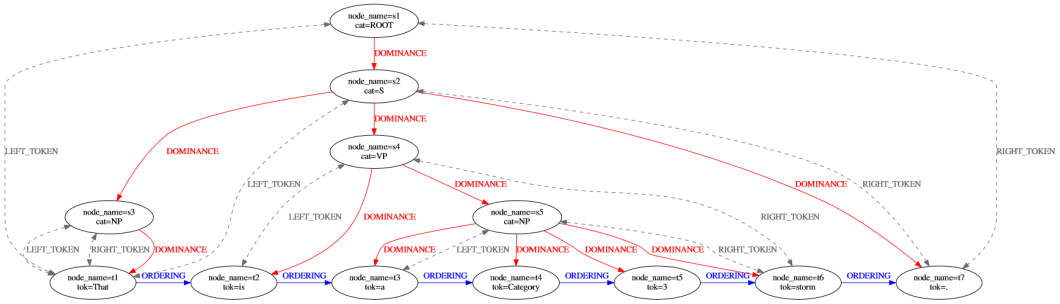

Abbildung 5: Das Beispiel aus Abb. 1 wie es in ANNIS 4 (*graphANNIS*) technisch repräsentiert wird (Krause, 2019: S. 41, Figure 4.5.).

1. Abbilden der ursprünglichen Modellierung und Inhalte

   Das neue Modell muss in der Lage sein, die ursprüngliche Modellierung abzubilden, damit sämtliche Inhalte, die nachgenutzt werden sollen, auch nachgenutzt werden können. Es kann zwar sein, dass nicht alle Aspekte des ursprünglichen Modells für die Nachnutzung relevant sind; trotzdem muss es auf der technischen Ebene möglich sein.

   Gleichzeitig bedeutet diese Anforderung aber auch eine klare Dokumentation der ursprünglichen Modellierung in Verbindung mit der neuen, die die Unterschiede und Gemeinsamkeiten deutlich macht. Wenn sich entschieden wurde, nur einen Teil der Modellierung zu übernehmen, ist auch diese Entscheidung zu dokumentieren.

2. Anpassen und Erweitern der Modellierung

   Weiterhin müssen das neue Datenbanksystem und sein Datenmodell Möglichkeiten zum Anpassen und Erweitern der ursprünglichen Modellierung bieten, denn ohne die Möglichkeit würden die Potenziale der Abbildung von Inhalten in einem neuen System nicht ausgenutzt. Gerade die Unterschiede in der Modellierung zwischen verschiedenen Wissensdomänen bietet einen Mehrwert der interdisziplinären Arbeit. Es erleichtert die Arbeit ungemein, wenn nicht in den Modellen einer jeden Domäne die jeweils anderen mitgedacht werden müssen.

   Dazu gehört aber auch, dass es eindeutige Verknüpfungen zwischen den Inhalten gibt, die in der ursprünglichen sowie in der neuen Modellierung vorkommen. Das neue Datenmodell und seine technische Umsetzung sollen zwar unabhängig von der ursprünglichen Umsetzung sein, aber trotzdem die Referenz auf Daten des Quelldatensatzes erlauben.

3. Umsetzen der FAIR-Prinzipien

   Die eindeutige Referenz auf Daten des Quelldatensatzes ist auch für die Umsetzung der FAIR-Prinzipien im Rahmen der neuen Lösung wichtig: So wird über die Referenz eine Interoperabilität auf Ebene der Daten hergestellt. Dasselbe sollte

auch auf Ebene des Datenmodells geschehen: Es ist also nötig, Objekte, Klassen und Aussagen auf ihre semantisch äquivalenten Gegenstücke in anderen Datenmodellen/Vokabularien abzubilden.

Zusätzlich muss sich auch die neue technische Umsetzung an den Prinzipien Auffindbarkeit, Zugänglichkeit und Nachnutzbarkeit messen.

4. Ermöglichen von additivem und (potenziell) kollaborativem Arbeiten

   Zu guter Letzt sollte es aus meiner Sicht technisch erlaubt sein, stückweise zu arbeiten: Nicht jedes Forschungsvorhaben kann es sich arbeitsökonomisch leisten, erst mit viel Aufwand ein umfassendes Datenmodell zu entwickeln, bevor mit konkreten Daten gearbeitet werden kann. Außerdem stellt sich oft erst im Arbeiten heraus, wie das eigene Datenmodell am besten gestaltet sein sollte. Ebenso ist nicht immer der Umfang der konkreten Gegenstände von Anfang an festgelegt, neue und andersartige können im Forschungsprozess dazukommen und eine Änderung des Modells nötig machen.

   Für additives Arbeiten ist eine Versionierung nicht nur des Gesamtinhalts in größeren Intervallen, sondern auch eine Versionierung jeder einzelnen Änderung wichtig. So kann zu jeder Zeit nachvollzogen werden, wie der Ausgangspunkt war und welche Änderung vorgenommen wurde. Noch wichtiger wird dies, wenn an einem Forschungsvorhaben mehr als nur eine Person arbeitet. Um die Versionierung der Änderung jeder Person festhalten zu können, ist zudem eine Benutzer- und Rechteverwaltung vonnöten.

   Die Verwaltung von Lese- und Schreibberechtigungen würde zudem eine Auffindbarkeit, Zugänglichkeit und Nachnutzbarkeit im laufenden Betrieb gestatten: Unfertige Einträge könnten bspw. nur einem internen Redaktionsteam zugänglich sein, während gleichzeitig fertige Einträge öffentlich zugänglich sind.

Literatur

corpus-tools.org. ???? ANNIS (corpus-tools.org). https://corpus-tools.org/annis/.

Flanders, Julia & Fotis Jannidis (eds.). 2019. *The Shape of Data in the Digital Humanities: Modeling texts and text-based resources* Digital research in the arts and humanities. London and New York: Routledge. doi:10.4324/9781315552941.

Gaspar, Drazena & Ivica Coric. 2017. *Bridging Relational and NoSQL Databases.* Hershey, Pennsylvania (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA): IGI Global. Backup Publisher: IGI Global.

Horch, Andre. 2014. *Buchwidmungen der Frühen Neuzeit als Quellen der Stadt-, Sozial- und Druckgeschichte. Kritische Analyse der Dedikationen in volkssprachlichen Mainzer Drucken des 16. Jahrhunderts. Unter Verwendung statistischer, netzwerkanalytischer und textinterpretatorischer Methoden* (Mainzer Studien zur Neueren Geschichte Band 32). Frankfurt am Main: Peter Lang Edition.

Krause, Thomas. 2019. ANNIS: A graph-based query system for deeply annotated text corpora doi:10.18452/19659. https://edoc.hu-berlin.de/handle/18452/20436. Publisher: Humboldt-Universität zu Berlin.

Lüdeling, Anke, Carolin Odebrecht, Thomas Krause, Gohar Schnelle & Catharina Fischer. 2022. RIDGES Herbology. doi:10.34644/LAUDATIO-DEV-PYSSCNMB7CARCQ9CNKFY. Type: dataset. https://www.laudatio-repository.org/browse/corpus/PySSCnMB7CArCQ9CNKFY/corpora.

Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke Lüdeling & Thomas Krause. 2017. RIDGES Herbology. Designing a diachronic multi-layer corpus. *Language Resources and Evaluation* 51(3). 695–725. doi:10.1007/s10579-016-9374-3. http://link.springer.com/10.1007/s10579-016-9374-3.

Odebrecht, Carolin, Gohar Schnelle, Catharina Fischer & Laura Perlitz. 2020. Dokumentation und Annotationsrichtlinien für das Korpus Ridges Herbology Version 9.0 (ANNIS-und PAULA-Format) auf Grundlage des Metadatenframeworks nach LAUDATIO. Stand 31.03.2020. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/pubs/ridgesv9-2020-03.pdf.

Paletschek, Sylvia. 2002. Die Erfindung der Humboldtschen Universität. Die Konstruktion der deutschen Universitätsidee in der ersten Hälfte des 20. Jahrhunderts. *Historische Anthropologie* 10(2). 183–205. doi:10.7788/ha.2002.10.2.183. https://www.vr-elibrary.de/doi/10.7788/ha.2002.10.2.183.