Datensilos anzapfen!

Ein Vorschlag und Versuch zur interdisziplinären Nachnutzung und Anreicherung korpuslinguistischer Forschungsdaten auf Basis von Python und Wikibase

Eigentlich:

Interdisziplinäre Nachnutzung und Anreicherung korpuslinguistischer Forschungsdaten.

Wikibase als Fundament zur Abbildung, Anreicherung und Auswertung ausgewählter Daten des Korpus RIDGES Herbology 9.0

Datensets, Scripts und Einleitung auf Github:

https://github.com/Schoeneh/ridges-herb-wikibase

Forschungsdaten

„Forschungsdaten sind eine wesentliche Grundlage für das wissenschaftliche Arbeiten. Die Vielfalt solcher Daten entspricht der Vielfalt unterschiedlicher wissenschaftlicher Disziplinen, Erkenntnisinteressen und Forschungsverfahren.“

(Senat der Deutschen Forschungsgemeinschaft, 2015: S. 1)

„Unter digitalen Forschungsdaten verstehen wir dabei alle digital vorliegenden Daten, die während des Forschungsprozesses entstehen oder ihr Ergebnis sind. […]“

(Kindling & Schirmbacher, 2013: S. 130)

Boris Queckbörner (2019): Forschungsdaten und Forschungsdatenmanagement in der Geschichtswissenschaft. Gegenwärtige Praxis und Perspektiven am Beispiel ausgewählter Sonderforschungsbereiche; Berlin: Humboldt-Universität zu Berlin, Masterarbeit. https://edoc.hu-berlin.de/handle/18452/21227.

Herausforderung I:

‚**Datensilos**‘ — Modellierung der Forschungsdaten

explizit; v.a. aber **implizit**

jede Forschungsfrage und -methode erzeugt eine eigene Modellierung des jeweiligen Gegenstands
(vgl. Thalheim & Nissen, 2015: S.615-617)

FAIR-Prinzipien:

- Findability, Accessibility
- Interoperability, Reusability

(vgl. Harrower et al., 2020: S. 3)

Herausforderung II:

„[H]ow is it possible to adopt a quasi-universal conceptualization, an ontology, in order to ensure the interoperability of the data produced by historians?“
(Beretta, 2021: S. 280)

<u>Verknüpfte – ‚*föderierte*‘ – Datenbanksysteme</u>

nicht eine Ontologie für die gesamte Welt

je Projekt/Frage eigene Ontologie, aber als ‚*Föderation*‘

Konkret I:

RIDGES Herbology 9.0 – **R**egister **i**n **D**iachronic **Ge**rman **S**cience

(https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt)

Lüdeling, Anke, Odebrecht, Carolin, Krause, Thomas, Schnelle, Gohar, and Fischer, Catharina. 2022. 'RIDGES Herbology'. Humboldt-Universität zu Berlin. https://doi.org/10.34644/LAUDATIO-DEV-PYSSCNMB7CARCQ9CNKFY

- Kräutertexte
  - 305.056 Token
  - 73 Texte/Textausschnitte aus dem Zeitraum 1482-1914

  - Transkribiert via OCR4all
  - Tokenisiert via Treetagger 3.2
  - weitere Annotationen (linguistische, strukturelle, inhaltliche) durch Studierende

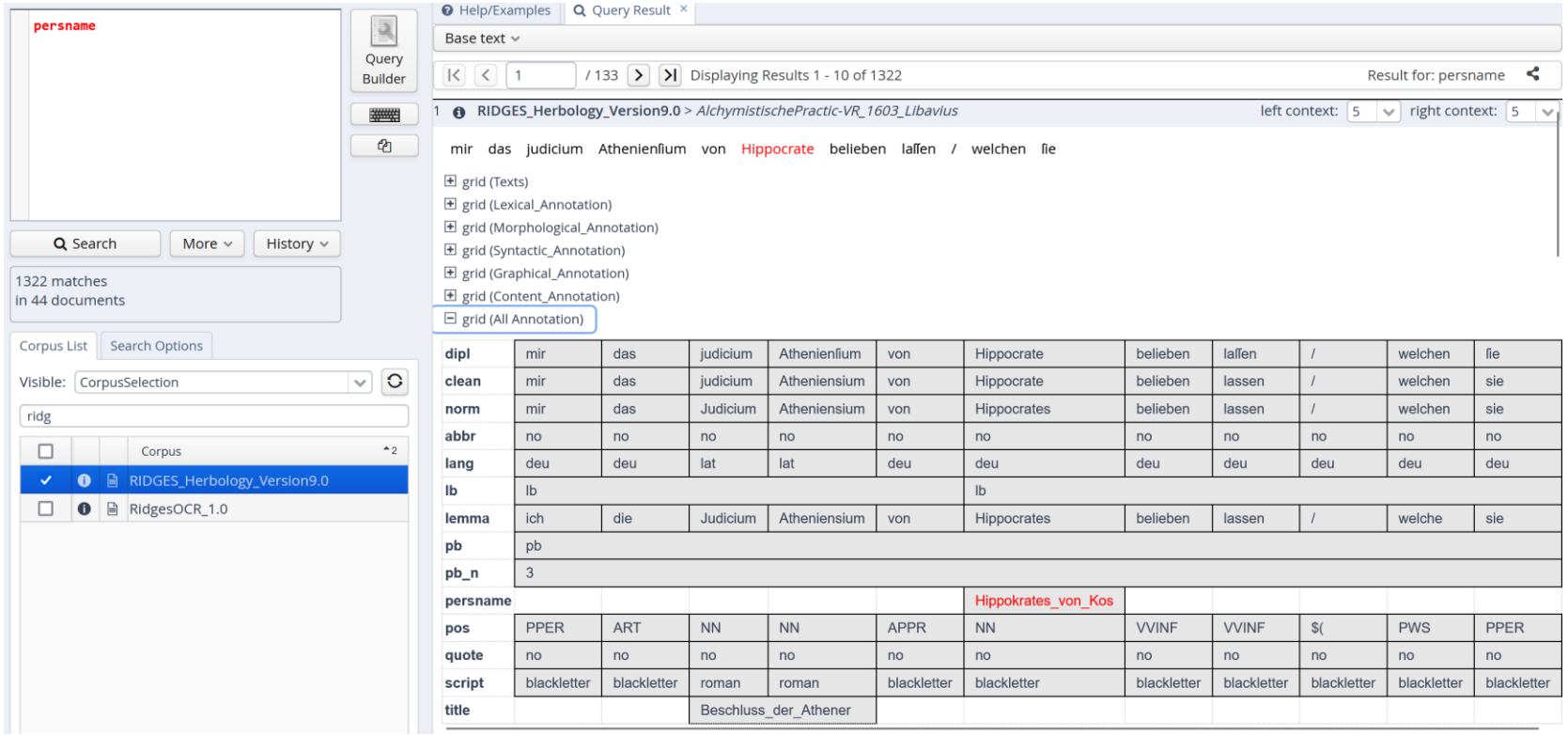
Konkret II:

- verschiedene Perspektiven (bspw.):
  - **Linguistik**
    - morphologische, syntaktische und weiteren Eigenschaften einzelner **Texteinheiten**
  - **Literaturwissenschaft**
    - **Paratext** sowie interne und externe **Intertextualität** des Textes und seiner Teile
  - **Geschichtswissenschaften**
    - Verknüpfung der im Text und in den Metadaten erwähnten Personen, Orte und Ereignisse mit ihren entspr. historischen **Entitäten**

Konkret III:

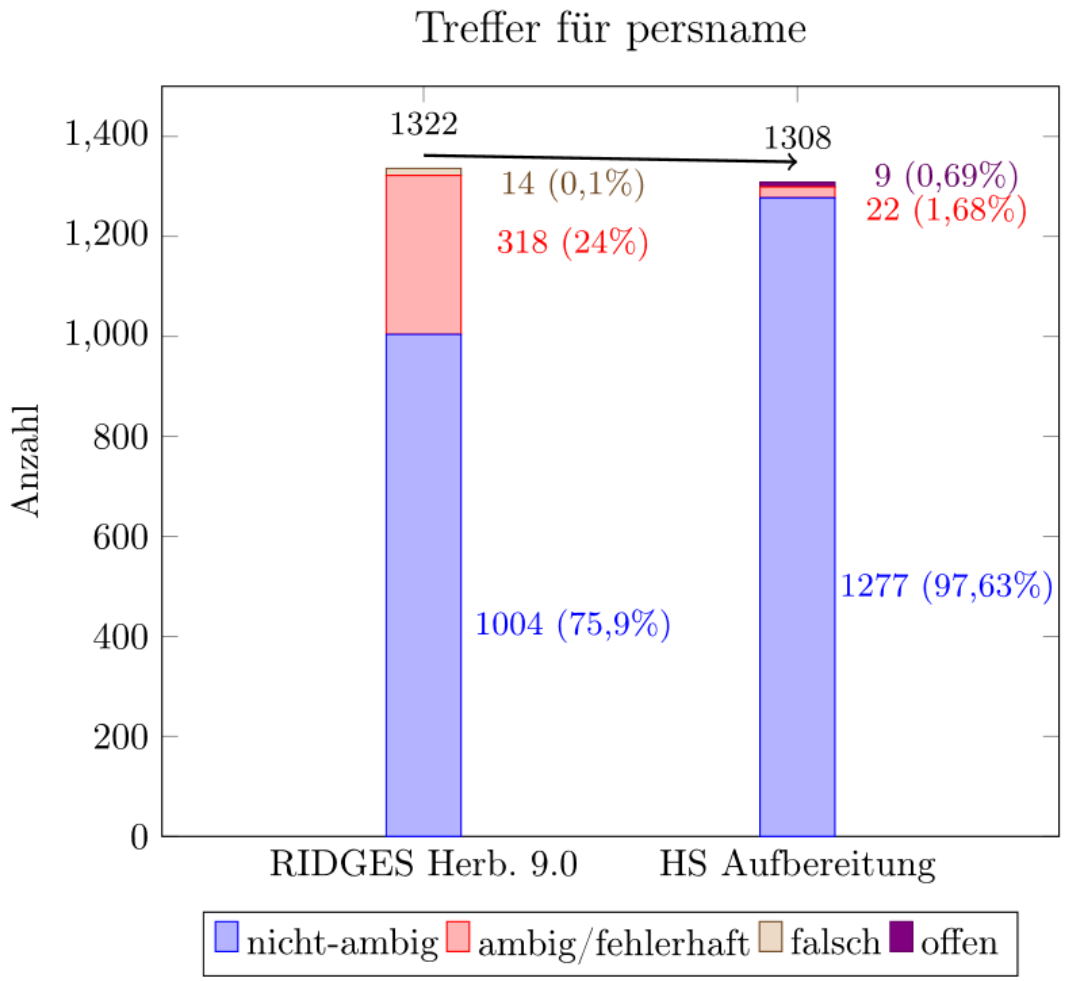
Findability & Accessibility: LAUDATIO-Repository

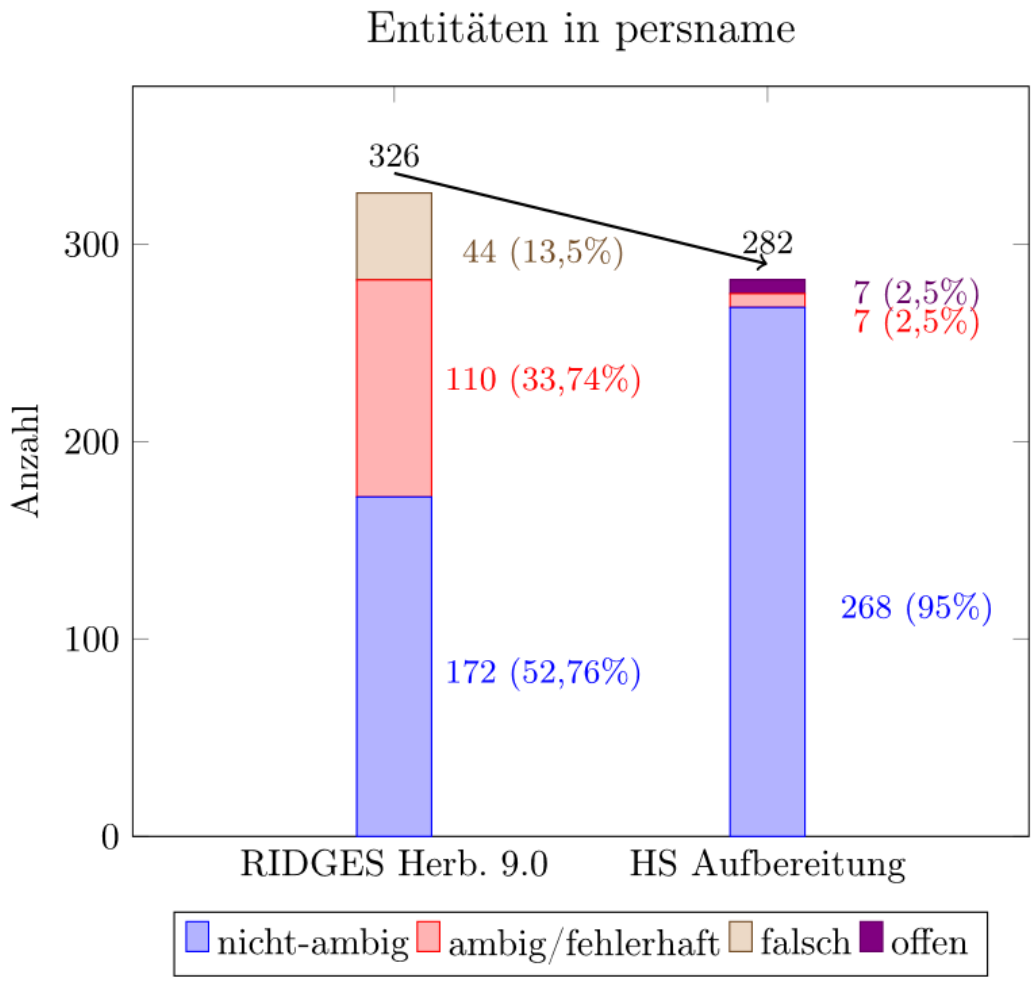
Interoperability & Reusability: ANNIS - ANNotation of Information Structure



Thomas Krause (2019): ANNIS: A graph-based query system for deeply annotated text corpora; Berlin: Humboldt-Universität zu Berlin, Dissertation. https://edoc.hu-berlin.de/handle/18452/20436.

Konkret IV – Disambiguieren:





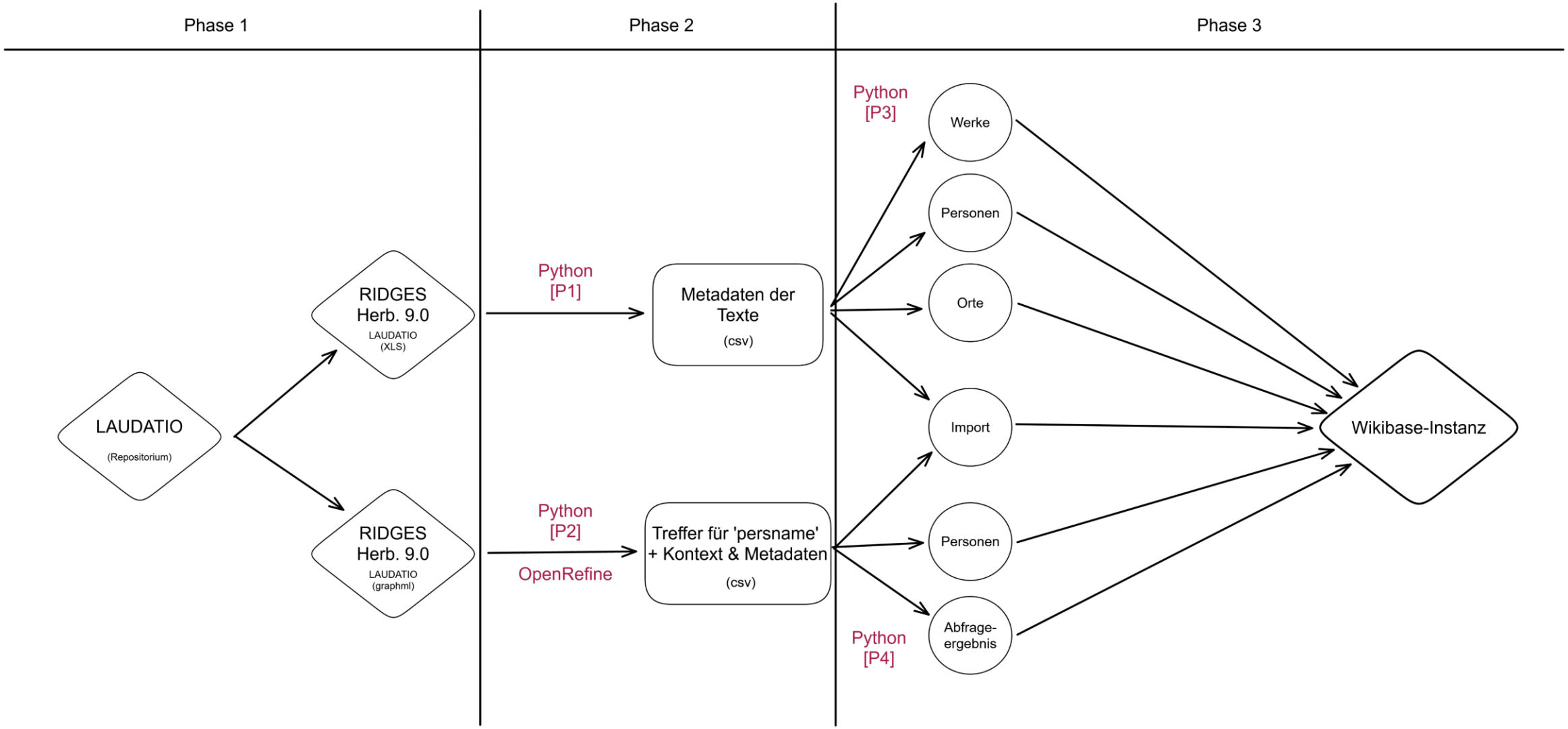
Ausblick

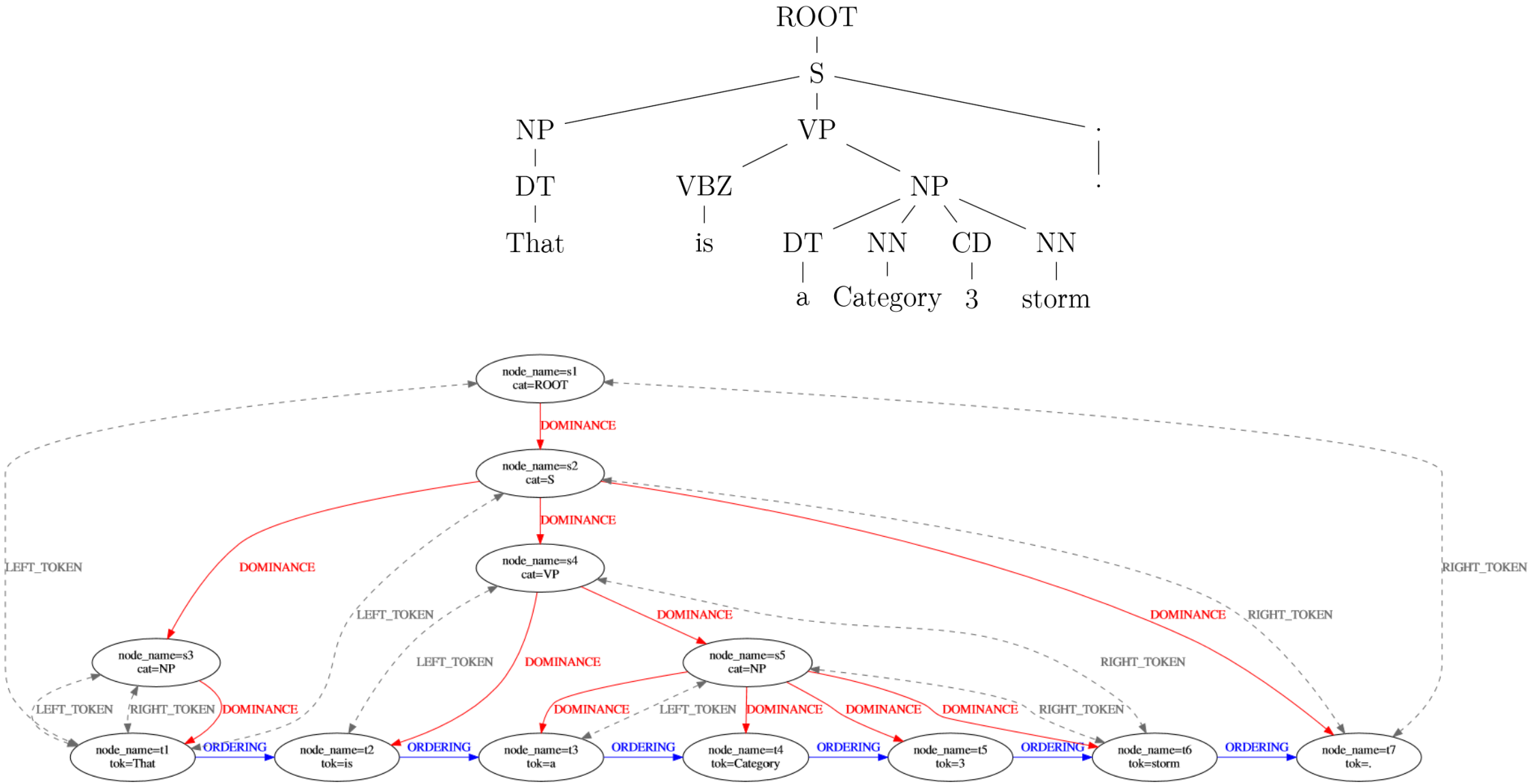
- Anforderungen an eine mögliche Lösung:
  - Abbilden der ursprünglichen Modellierung und Inhalte
  - Anpassen und Erweitern der Modellierung
  - Umsetzen der FAIR-Prinzipien
  - Ermöglichen von additivem und (potenziell) kollaborativem Arbeiten

- Wikibase?
  - https://ridges-herb.wikibase.cloud/wiki/Main_Page

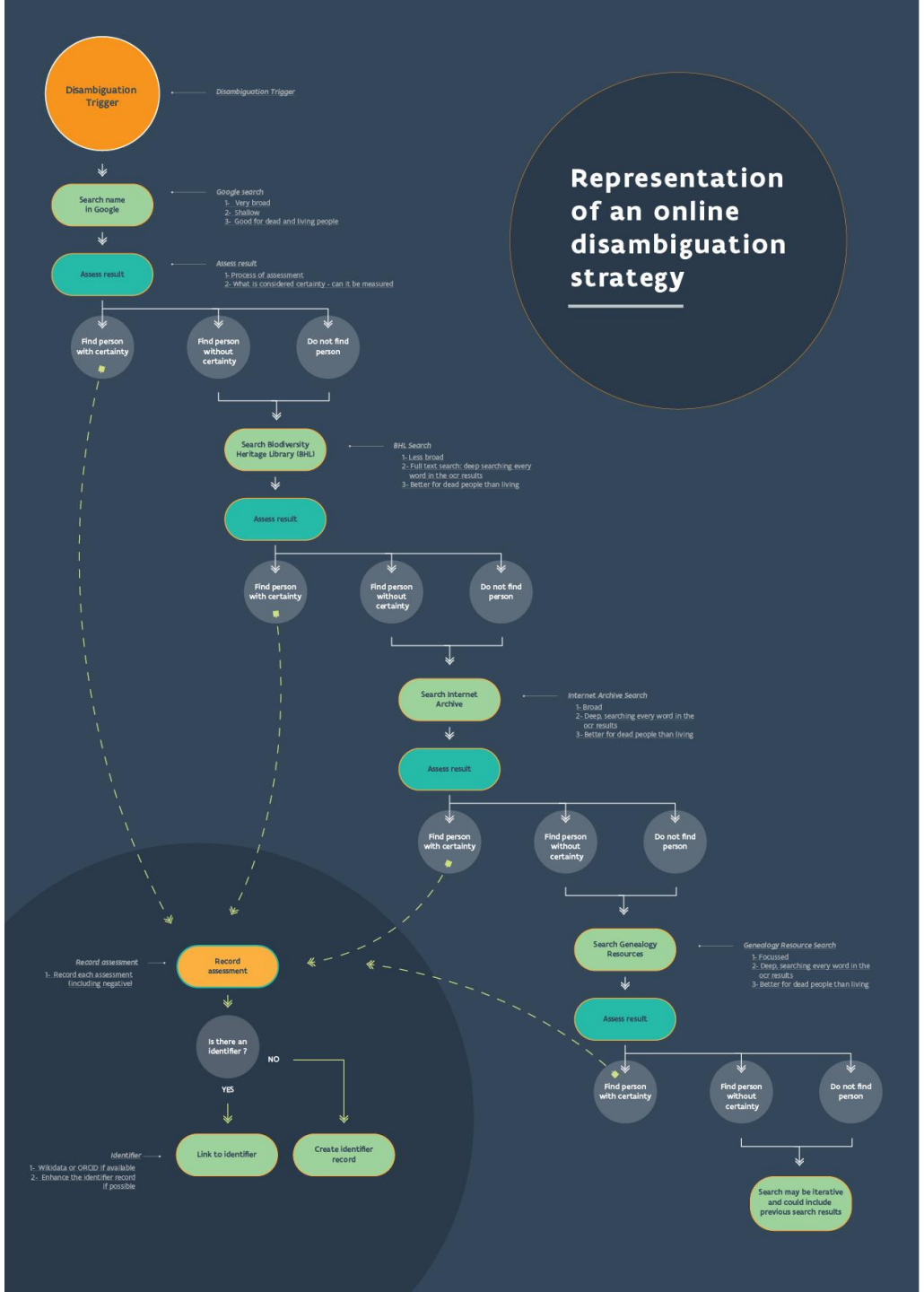
Zitierte Literatur

- Beretta, Francesco. 2021. ‘A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)’. *Semantic Web* 12 (2): 279–94. https://doi.org/10.3233/SW-200416.
- Harrower, Natalie, Maciej Maryl, Timea Biro, and Beat Immenhauser. 2020. ‘Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities’. Edited by ALLEA - All European Academies, Berlin. Digital Repository of Ireland. https://doi.org/10.7486/DRI.TQ582C863.
- Kindling, Maxi, and Peter Schirmbacher. 2013. ‘Die Digitale Forschungswelt Als Gegenstand Der Forschung / Research on Digital Research / Recherche Dans La Domaine de La Recherche Numérique’. *Information - Wissenschaft & Praxis* 64 (2–3): 127–36. https://doi.org/10.1515/iwp-2013-0017.
- Krause, Thomas. 2019. ‘ANNIS: A graph-based query system for deeply annotated text corpora’. Humboldt-Universität zu Berlin. https://edoc.hu-berlin.de/handle/18452/20436.
- Lüdeling, Anke, Odebrecht, Carolin, Krause, Thomas, Schnelle, Gohar, and Fischer, Catharina. 2022. ‘RIDGES Herbology’. Humboldt-Universität zu Berlin. https://doi.org/10.34644/LAUDATIO-DEV-PYSSCNMB7CARCQ9CNKFY.
- Queckbörner, Boris. 2019. ‘Forschungsdaten Und Forschungsdatenmanagement in Der Geschichtswissenschaft: Gegenwärtige Praxis Und Perspektiven Am Beispiel Ausgewählter Sonderforschungsbereiche’. Masterarbeit, Berlin: Humboldt-Universität zu Berlin. https://doi.org/10.18452/20460.
- Thalheim, Bernhard, and Ivor Nissen. 2015. ‘The Notion of a Model’. In *Wissenschaft & Kunst Der Modellierung. Kieler Zugang Zur Definition, Nutzung Und Zukunft*, edited by Bernhard Thalheim and Ivor Nissen, 1. Auflage, 615–18. Philosophische Analyse = Philosophical Analysis, Band 64. Boston: De Gruyter.
- Senat der Deutschen Forschungsgemeinschaft. 2015. ‘Leitlinien Zum Umgang Mit Forschungsdaten’. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/leitlinien_forschungsdaten.pdf.

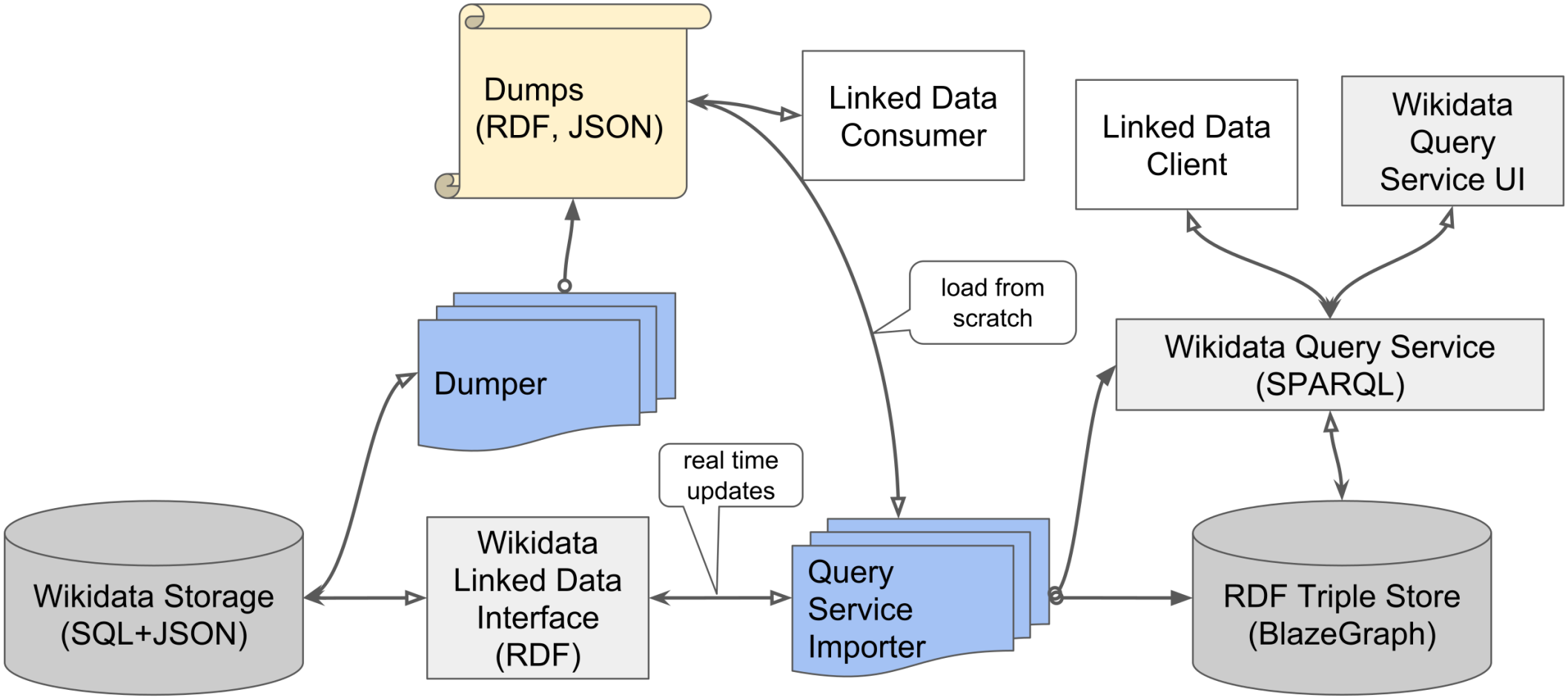

Phase 1
Phase 2
Phase 3
LAUDATIO
(Repositorium)
RIDGES
Herb. 9.0
LAUDATIO
(XLS)
RIDGES
Herb. 9.0
LAUDATIO
(graphml)
Python
[P1]
Python
[P2]
OpenRefine
Metadaten der
Texte
(csv)
Treffer für 'persname'
+ Kontext & Metadaten
(csv)
Python
[P3]
Python
[P4]
Werke
Personen
Orte
Import
Personen
Abfrage-
ergebnis
Wikibase-Instanz



aus Krause 2019



Quentin Groom, Christian Bräuchler, Robert Cubey, Mathias Dillen, Pieter Huybrechts, Nicole Kearney, Niels Klazenga, et al. (2022): 'The Disambiguation of People Names in Biological Collections'. *Biodiversity Data Journal* 10. https://doi.org/10.3897/BDJ.10.e86089.



https://commons.wikimedia.org/wiki/File:Wikidata_Architecture_Overview_-_Query_Service.svg