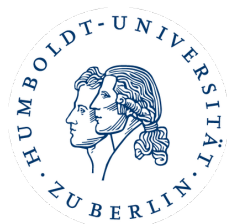
# LAUDATIO: Modellbasierte Entwicklung eines fachbezogenen und institutionellen Forschungsdatenrepositoriums

Carolin Odebrecht, Rolf Guescini, Thomas Krause
Humboldt-Universität zu Berlin

8. DINI/nestor-Workshop "Forschungsdatenrepositorien"
27.-28.11.2017



## Unsere Zielstellung
### Nachhaltigkeit durch Wiederverwendung

- Langfristige Speicherung, Erreichbarkeit und umfassende Dokumentation von historischen Korpora
- Modellbasierte Entwicklung von Softwarekomponenten mit angepasster Anwendungsoberfläche
- In institutioneller und interdisziplinärer Zusammenarbeit

$\rightarrow$ Domänenspezifisches Repositorium LAUDATIO

# Gliederung

1. Historische Korpora
2. Metadaten und Metadatenmodell
3. Anwendung
4. Zusammenfassung, Ausblick

# Historische Korpora

## Annotationen, Erschließung und Wiederverwendung

# Historische Korpora

- Annotierte Sammlung digitalisierter natürlichsprachlicher Äußerungen vergangener Sprachstufen (Claridge 2008; McEnery und Hardie 2012)
- Variation hinsichtlich Design, Größe, Annotationsschemata, Annotationskonzepte, Formate und Architektur, motiviert durch den Forschungskontext (vgl. Lüdeling 2011)
- Annotationen als Interpretationen und explizite Zuweisungen von Kategorien zu einzelnen oder mehreren Einheiten in einem Korpus (Kuebler und Zinsmeister 2015; McEnery und Hardie 2012)

→ Repräsentation strukturierten Wissens

# Beispiele

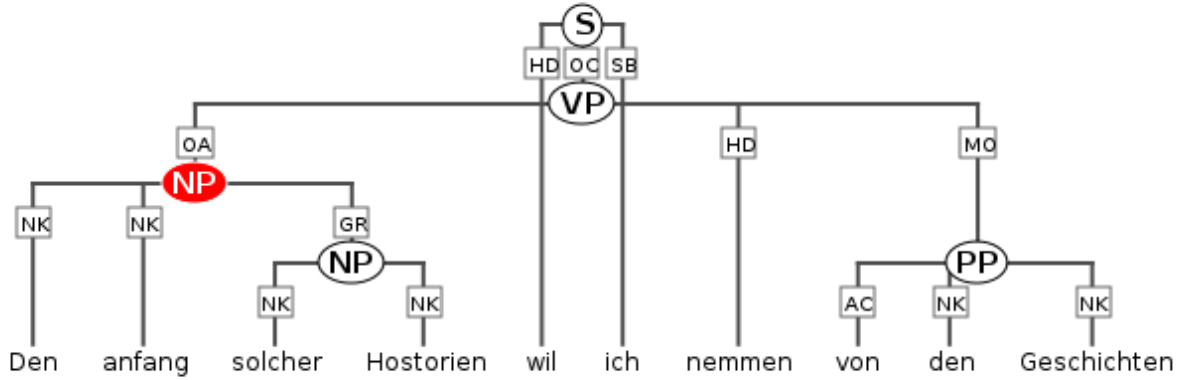



Abbildung 1: Ausschnitt aus Mercurius (Demske 2007).[1]

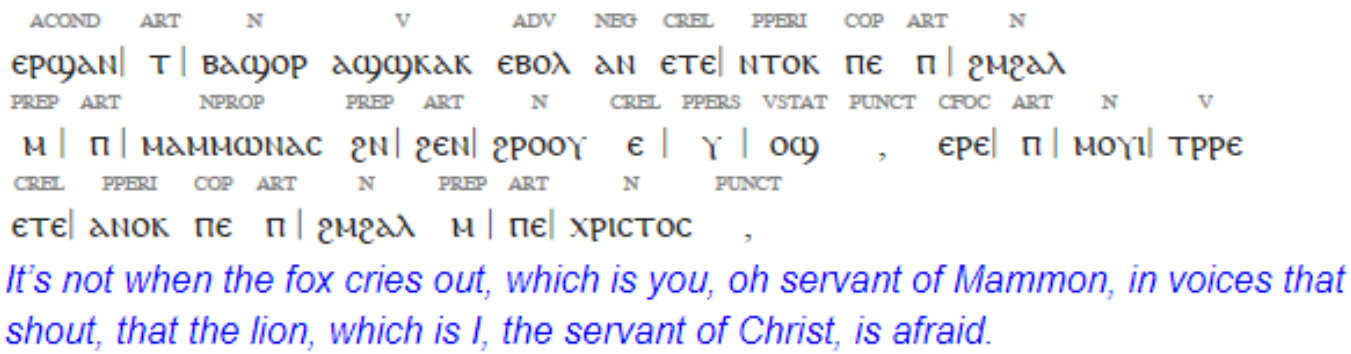



Abbildung 2: Ausschnitt aus Coptic Scriptorium (Schroeder und Zeldes 2016).[2]

# Erschließung von historischen Korpora zum Zweck der Wiederverwendung

→ Sich ein Korpus zu erschließen heißt u. a., dessen Annotationen inklusive der Konzepte, der Relationen und der jeweiligen Realisierungen zu kennen.

- Komplexe Beziehung zwischen Annotationen (inkl. Repräsentationen von Sprache)
- Informationen über u.a. das Design, Größe, Ersteller und Forschungskontext

# Erschließung von historischen Korpora zum Zweck der Wiederverwendung

- Wiederverwendungsszenarien
  - Replikation bestehender Studien
  - Neue, ggf. nicht vom Korpusersteller intendierte Analysen
  - Hinzufügen neuer Annotationen, Änderung bestehender Annotationen im Korpus
  - Erweiterung des Korpus, neues Sampling

→ Historische Korpora als empirische Grundlage für verschiedene Analysen für z. B. Literaturwissenschaften, Geschichtswissenschaften oder Sprachwissenschaften

# Anforderungen für LAUDATIO

- Auffindbarkeit, Erreichbarkeit
- Domänenspezifische, umfassende und strukturierte Dokumentation
- Offen für alle Formate und Annotationsmodelle
- Open Access Langfristige Speicherung/ Archivierung

→ Ziel: Voraussetzungen für die Wiederverwendung von Forschungsdaten schaffen

→ Ziel: Neuentwicklung von nutzerfreundlichen Schnittstellen für die Suche und Auffindbarkeit von historischen Korpora – LAUDATIO 2

# Metamodell für Korpusmetadaten

## Modell, Metadaten und Metadatenmodell

# Metadaten und Modell

- Modell
  - Zweckgebundene Abstraktion und Reduktion der Wirklichkeit
  - Definition des Modells durch ein Metamodell

  (vgl. z. B. Kempa und Mann 2005; Rumpe 2011; Zipser 2009)
- Metadaten
  - Strukturierte Informationen zur Beschreibung, Erklärung, Verwaltung und Identifikation einer Ressource
  - Klassifizierbar nach enthaltenen Informationen, Struktur, Bezugsobjekten und Zweck

  (vgl. z. B. Haynes 2004; Hider 2012; NISO 2004)

# Metamodell für Korpusmetadaten

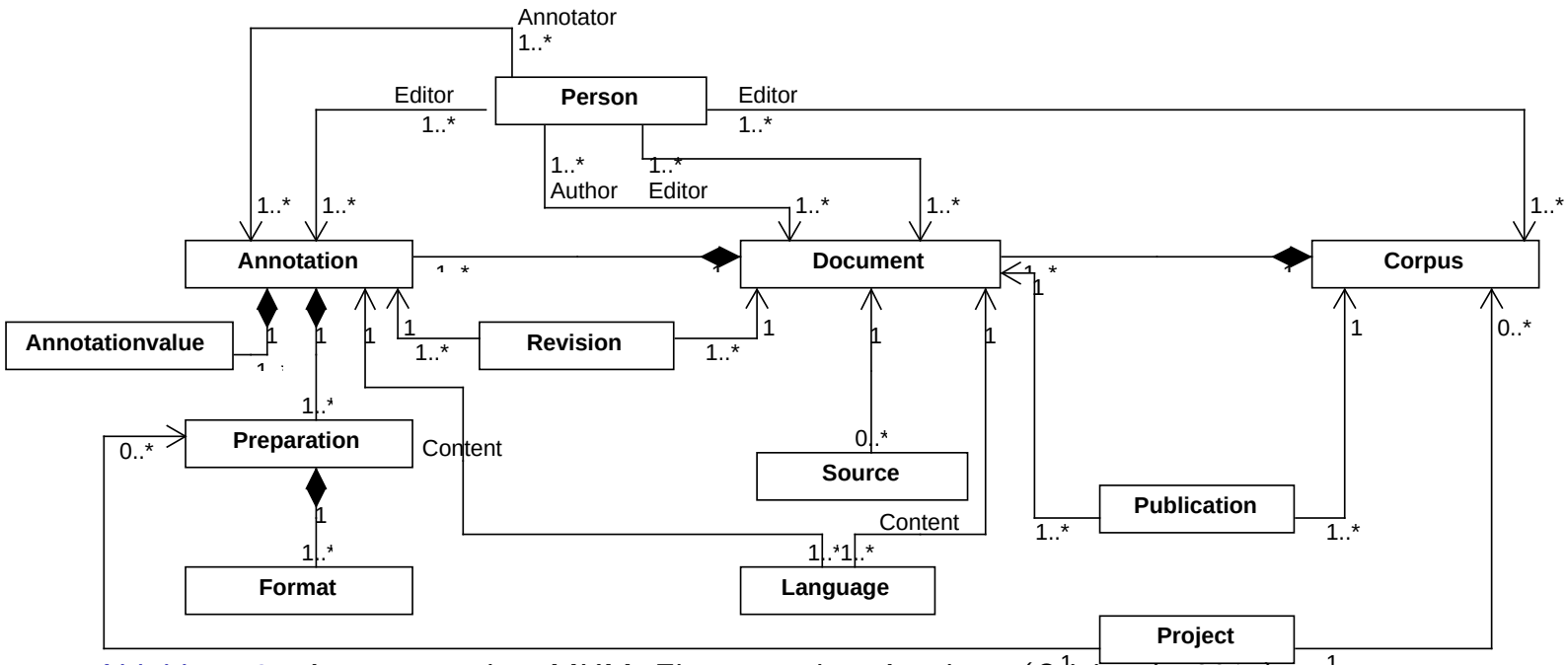



Abbildung 3: Auszug aus dem MKM. Elemente ohne Attribute (Odebrecht 2017).

# Entwicklungsmodell für Software

- Erfüllen der Anforderungen der Domäne
- Entwicklung von nachhaltiger Software
- Kein Widerspruch: Softwarearchitektur mit generischen Komponenten und domänenspezifischen Komponenten
  - Domäne wird über ein Modell (MKM) in die Software integriert
  - Domäne ist austauschbar und konfigurierbar

$\rightarrow$ So Reaktion auf Anforderungen neuer Domänen möglich ohne eine komplette Neuimplementierung

# E n t w i c k l u n g s m o d e l l

## Funktionen, Softwarekomponenten, Softwarearchitektur

# Softwareentwicklung

- Bewusste Trennung von Aufgaben der Erstellung, Speicherung und Suche von Daten
  - Daten und Format unabhängige Speicherung und Repräsentation
  - Daten und Format unabhängiges System
  - Datenmodell informiert das Anwendungsmodell

→ Wiederverwendung von Software durch generische Softwarekomponenten und modellbasierte, domänenspezifische Komponenten

# Softwarearchitektur

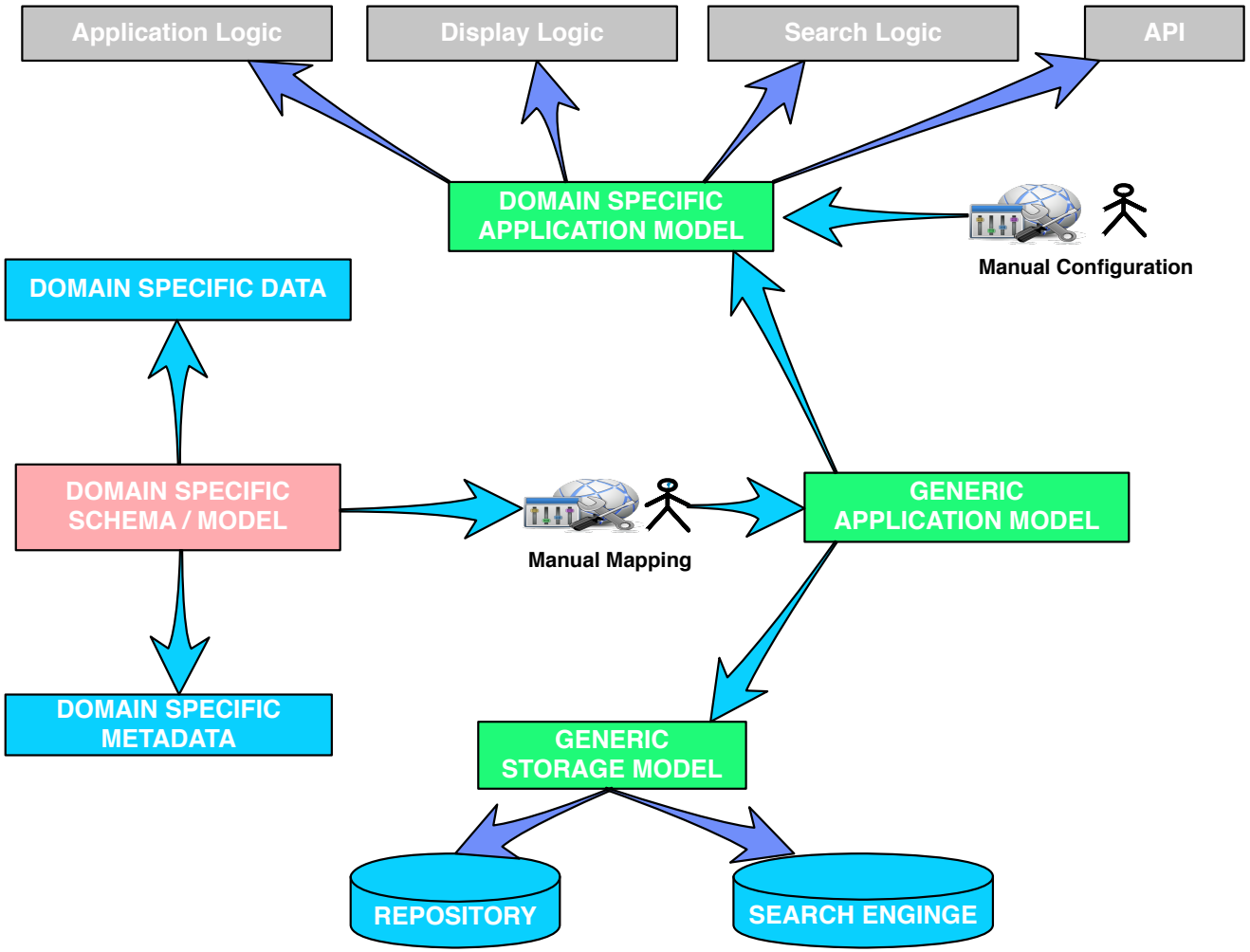



Abbildung 4: Application Model

# Zusammenfassung, Ausblick

# Zusammenfassung, Ausblick

Prinzipien der (neuen) Entwicklung mit Fokus auf das Anwendungsziel:

- Balance zwischen domänenspezifischen Anwendungen und Generalisierbarkeit
- Integration von automatisierten und manuellen Prozessen in der Softwarearchitektur
- Auf Basis eines gemeinsames Modells (MKM als Drehpunkt in der Implementierung)
- In enger Zusammenarbeit mit den Erstellern der Forschungsdaten und mit Kenntnis des wissenschaftlichen Nutzungskontextes
- User Experience als Zusammenarbeitsmodell

# Zusammenfassung, Ausblick

Prinzipien der (neuen) Entwicklung mit Fokus auf Nachhaltigkeit:

- Unabhängigkeit von Formaten, Daten und Forschungskontexten
- Wiederverwendung von Forschungsdaten und Software

Nachhaltigkeit über Projekte und Forschungskontexte hinweg:

- Notwendige technische Expertise zur Entwicklung und Betrieb eines nachhaltigen Repositoriums durch den Computer- und Medienservice der HU als zentrale universitäre Einrichtung

Vielen Dank für Ihre Aufmerksamkeit!

# Referenzen I


Claridge, Claudia (2008). „Historical Corpora“. In: *Corpus Linguistics*. Hrsg. von Anke Lüdeling und Merja Kytö. Bd. 1. Berlin: De Gruyter, S. 242–259.

Demske, Ulrike (2007). „Das Mercurius-Projekt: eine Baumbank für das Frühneuhochdeutsche“. In: *Sprachkorpora*. Hrsg. von Gisela Zifonun und Werner Kallmeyer. Jahrbuch des Instituts für deutsche Sprache 2006. Berlin: De Gruyter, S. 91–104.

Haynes, David (2004). *Metadata for information management and retrieval*. London: facet publishing.

Hider, Philip, Hrsg. (2012). *Information Resource Description: Creating and Managing Metadata*. London: facet publishing.

Kempa, Martin und Zolt?n Ad?m Mann (2005). „Model Driven Architecture“. In: *Informatik-Spektrum* 28.4, S. 298–302.

Kuebler, Sandra und Heike Zinsmeister (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury Academic.

Lüdeling, Anke (2011). „Corpora in Linguistics: Sampling and Annotation“. In: *Going Digital*. Hrsg. von Karl Grandin. Bd. 147. Nobel Symposium. New York: Science History Publications, S. 220–243.

McEnery, Tony und Andrew Hardie (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambride [u.a.]: Cambridge University Press.

NISO (2004). *Understanding Metadatada*. Hrsg. von National Information Standards Organization. Bethesda. URL: http://www.niso.org/publications/press/UnderstandingMetadata.pdf (besucht am 13. 02. 2015).

Odebrecht, Carolin (2017). „MKM - ein Metamodell für Korpusmetadaten: Dokumentation und Wiederverwendung historischer Korpora“. Diss. Berlin: Humboldt-Universität zu Berlin.

# Referenzen II



Rumpe, Bernhard (2011). *Modellierung mit UML*. Berlin, Heidelberg: Springer Berlin Heidelberg.


Schroeder, Carolin T. und Amir Zeldes (2016). „Raiders of the Lost Corpus“. In: *Digital Humanities Quarterly* 10.2. URL: http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html.


Zipser, Florian (2009). *Entwicklung eines Konverterframeworks für linguistisch annotierte Daten auf Basis eines gemeinsamen (Meta-)modells*. Berlin. URL: hal-00606102 (besucht am 25. 09. 2016).

# Referenzen – Korpora


- Coptic SCRIPTORIUM, MONB_XH_204_216, urn:cts:copticLit:shenoute.fox.monbxh_204_216, 2.2.0, 2017-11-20. http://data.copticscriptorium.org/
- Demske, Ulrike; Mercurius (Version 1.1), Universität Potsdam. http://www.uni-potsdam.de/guvdds/projekte/abgproj.html. http://hdl.handle.net/11022/0000-0000-467D-6
- Lüdeling, Anke; Odebrecht, Carolin; Zeldes, Amir; RIDGES-Herbology (Version 6.0), Humboldt-Universität zu Berlin. http://korpling.org/ridges/. http://hdl.handle.net/11022/0000-0005-8620-F