Die Rolle von Textsammlungen und Wörterbüchern für eine NFDI

PD Dr. Alexander Geyken,
Berlin-Brandenburgische
Akademie der Wissenschaften

GEFÖRDERT VOM

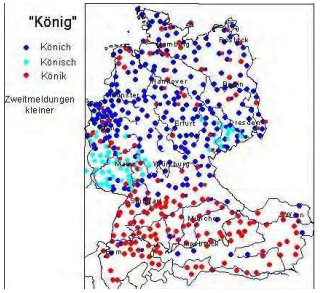
Bundesministerium
für Bildung
und Forschung

Gliederung

1. Charakterisierung
2. Funktion und Stellenwert von Textsammlungen und Wörterbüchern
3. Grade der Digitalisierung
4. Plattformen
5. Wer und was muss in ein NFDI-Konsortium?

1a. Charakterisierung Wörterbücher

- Kategorisierung nach:
  - Gegenstandsbereich: *Sprache*, *Dialekte, …*
  - Typologie: *einsprachig*, *mehrsprachig*, *Sachwörterbücher,* …
  - Informationspositionen: *Rechtschreibung*, *Bedeutung*, *Etymologie*, …
- Wörterbücher können multimodal sein
  - *Gebärdensprache*
  - *Aussprachewörterbücher*



1b. Charakterisierung Textsammlungen

Begriff Textsammlungen: keine einheitliche Definition

i) „breite“ Bedeutung:

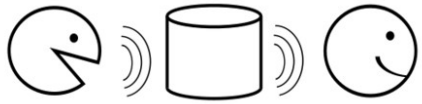
- die von den Gedächtnisinstitutionen bereitgestellten Volltextdigitalisate
- z. B. BSB, SPK, DDB

ii) „enge“ Bedeutung:

- für bestimmte Forschungsfragen aufbereitet
- Verlässliche (zitierbare) Strukturierung und Volltextqualität
- z. B. *DeReKo*, *Deutsches Textarchiv*, (*Textgrid-Rep*)…

1b. Charakterisierung Textsammlungen

- Kategorisierung nach Dimensionen:
  - *Referenzkorpora vs. Spezialkorpora, historische Korpora vs. synchrone Korpora, retrodigitalisiert vs. digital* erstellt
- Textsammlungen können multimodal sein
  - *Korpora zur Gebärdensprache*
  - *Transkribierte gespr. Sprache*



besondere Bedeutung besitzen Textsammlungen für Wörterbücher: Häufig sind Belege in Wörterbüchern rückgebunden an Texte.

2. Funktion und Stellenwert

a. Wörterbücher
   - zentrales Instrument für die Erschließung und Interpretation von Texten,
   - aber auch eigene Darstellungsform für historische Semantik

b. Textsammlungen
   - Grundlage für wissenschaftliche Untersuchungen (FAIR-Prinzipien)

a+b sind disziplinenübergreifend: u.a. Sprach- und Literatur-, Kultur-, Geschichts- und Politikwissenschaften

2. Stellenwert Wörterbücher

- Wörterbücher: verlässliche Informationen für
  - die Wissenschaft, das Sprachlernen, die interessierte Öffentlichkeit
- Potenziell sehr hohe Reichweiten
  - Kommerzielle Plattformen unter den 100 populärsten *Webseiten in DE: *duden*, *linguee, dict.cc, pons, leo*
  - *dwds*, *wortschatz* (Uni Leipzig), *woerterbuchnetz (Trier)*, *owid (IDS)*

(Quellen: sistrix bzw. alexa)*

3. Grade der Digitalisierung

- Wörterbuchtext ist ohne tiefere Annotationen nicht „auswertbar“
- Textsammlungen ohne qualitativ hochwertige Aufbereitung und Standardisierung nicht interoperabel und nachnutzbar (FAIR-Prinzipien)

Einschub: Digitalisierungsklassen bei Textsammlungen

a. Digitalisate mit standardisierten Metadaten
b. Volltexte unstrukturiert bzw. nicht standardisierte Strukturierung (z. B. aus Roh-OCR)
c. Volltexte mit standardisierter Strukturierung (entweder per OCR/OLR oder double keying)
   – anerkannte, insbesondere in DFG-Richtlinien empfohlene Best-Practice Formate: Alto, DTA-Basisformat, TEI simple

Quelle: Stand der Kulturgutdigitalisierung in Deutschland: Eine Analyse und Handlungsvorschläge des DARIAH-DE Stakeholdergremiums “Wissenschaftliche Sammlungen” (Klaffki, Schmunk, Stäcker 2018)

Brücke schließen zwischen a. und c.

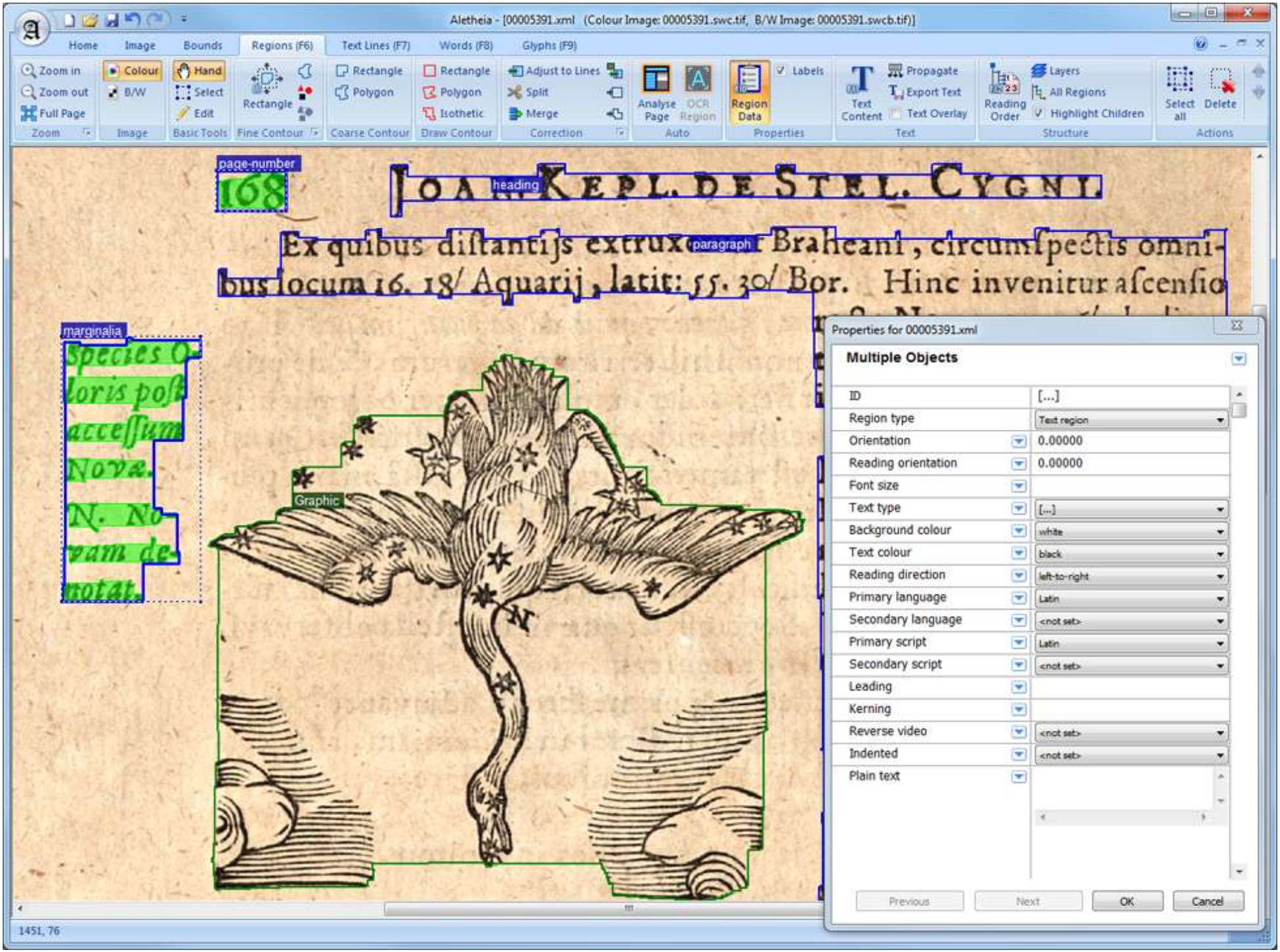
a. Digitalisate mit standardisierten Metadaten

→

c. Volltexte mit standardisierter Strukturierung

durch die von der DFG geförderte „Initiative zur Verbesserung von Verfahren der OCR: **OCR-D**“: Koordinierungsprojekt + 8 Modulprojekte

URL: *www.ocr-d.de*


Aletheia - [00005391.xml (Colour Image: 00005391.swc.tif, B/W Image: 00005391.swcb.tif)]
Home
Image
Bounds
Regions (F6)
Text Lines (F7)
Words (F8)
Glyphs (F9)
Zoom in
Zoom out
Full Page
Zoom
Colour
B/W
Image
Hand
Select
Edit
Basic Tools
Rectangle
Fine Contour
Rectangle
Polygon
Coarse Contour
Rectangle
Polygon
Isothetic
Draw Contour
Adjust to Lines
Split
Merge
Correction
Analyse Page
OCR Region
Auto
Region Data
Labels
Properties
Text Content
Propagate
Export Text
Text Overlay
Text
Reading Order
Layers
All Regions
Highlight Children
Structure
Select all
Delete
Actions
page-number
168
heading
paragraph
marginalia
Graphic
Properties for 00005391.xml
Multiple Objects
ID
[...]
Region type
Text region
Orientation
0.00000
Reading orientation
0.00000
Font size
Text type
[...]
Background colour
white
Text colour
black
Reading direction
left-to-right
Primary language
Latin
Secondary language
<not set>
Primary script
Latin
Secondary script
<not set>
Leading
Kerning
Reverse video
<not set>
Indented
<not set>
Plain text
Previous
Next
OK
Cancel
1451, 76

4. Plattformen

- Wörterbuch- und Textsammlungsprojekte benötigen heutzutage Arbeits- und Präsentationsplattformen
- Benötigt wird: übergreifende (verteilte) Infrastruktur, in denen die Informationen einheitlich zugreifbar sind

5. Wer und was muss in ein NFDI-Konsortium?

- Retrodigitalisierte Wörterbücher, Enzyklopädien und Textsammlungen
- Ergebnisse laufender Projekte
  - Universitäten: *BAS*, *Wortschatz Leipzig*, *Wörterbuchnetz-Trier*, *GermaNet-Tübingen*, *Francfort Latin Lexicon…*)
  - Akademien: *Altägyptisch* bis *Thesaurus Linguae Latinae*, von *Althochdt. Wörterbuch* bis *DWDS); von altägyptischen Kursivschriften bis DTA und LTA*
  - Leibniz-Institute (*IDS: OWID, DEREKO*)
  - …

5. Wer und was muss in ein NFDI-Konsortium?

- Wörterbücher und Textsammlungen sind fachübergreifend und damit „integrierend“ für geisteswissenschaftliche NFDI-K.
- Wissenschaftliche Akteure sind universitär und außeruniversitär (Akademien, Leibniz)
- Internationale Ebene: elexis, DARIAH-EU, CLARIN-EU
  - Ziel: Umsetzung FAIR-Prinzipien für Wörterbücher/Textsammlungen auf europäischer Ebene
  - Ein nationales NFDI-K. kann Einfluss auf diese Prozesse nehmen und Schnittstellen schaffen

Vielen Dank für Ihre Aufmerksamkeit.

Die Rolle von Textsammlungen und Wörterbüchern für eine NFDI

PD Dr. Alexander Geyken,
Berlin-Brandenburgische
Akademie der Wissenschaften

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung