NFDI und Spezialbibliotheken im Gespräch

eine Umfrage des NFDI Konsortiums Text+ zu Katalogdaten von Bibliotheken

José Calvo Tello (SUB Göttingen)
Nanette Rißler-Pipka (Max Weber Stiftung)

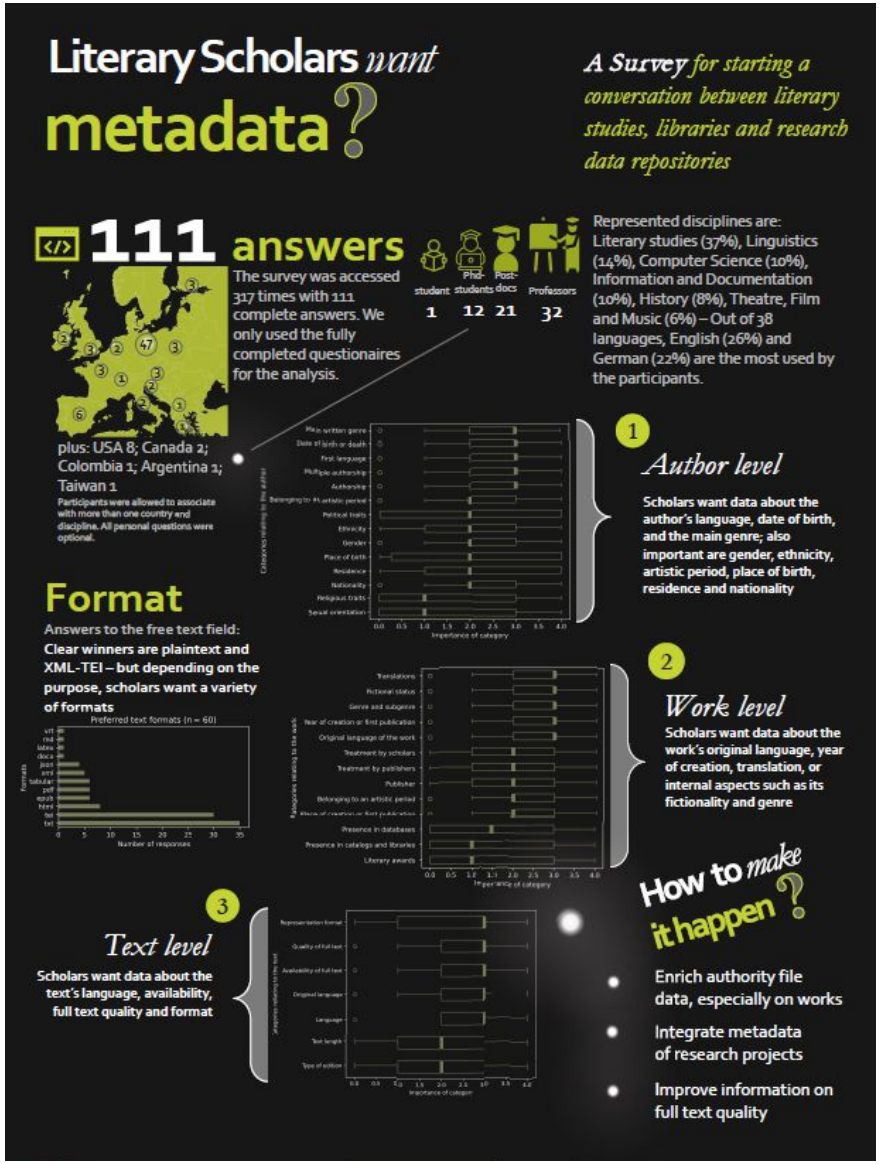
weitere Autor*innen der Umfrage: Patrick Helling (Universität Köln; DFG-Schwerpunktprogramm), Sally Chambers (DARIAH-EU, Ghent University, CLS-Infra), Philippe Genêt (DNB, Text+)

Text+

Computational LiteraryStudies

CLSINFRA

30.10.2024



Agenda

- Kontext und Design der Umfrage
- Ergebnisse: Autor, Werk, Text
- Erste Umsetzungen im TextGrid Repository

https://zenodo.org/records/11202132

Text+

Computational Literary Studies

CLSINFRA

30.10.2024

Kontext: Motivation und Beteiligte

- **Infrastruktur** und **digitale Literaturwissenschaft**
- Zusammenarbeit in **Text+** (NFDI), **CLS Infra** (EU Projekt), **Schwerpunktprogramm CLS** (DFG)
- **Gemeinsames Anliegen**: Finden und Zusammenstellen literarischer (digitaler) Korpora kann durch Metadaten in Katalogen erheblich vereinfacht werden - **Welche Metadaten braucht die Forschung?**

Text+

Computational LiteraryStudies

CLSINFRA

Kontext: Text+

Collections
Lexical Resources
Editions
Infrastructure/
Operations

» NFDI Konsortium für text- und sprachbasierte Forschung (Förderzeitraum 2021-2026)
» Collections: Volltext, Text als Bild, Metadaten
» Infrastructure/Operations: Technische Lösungen

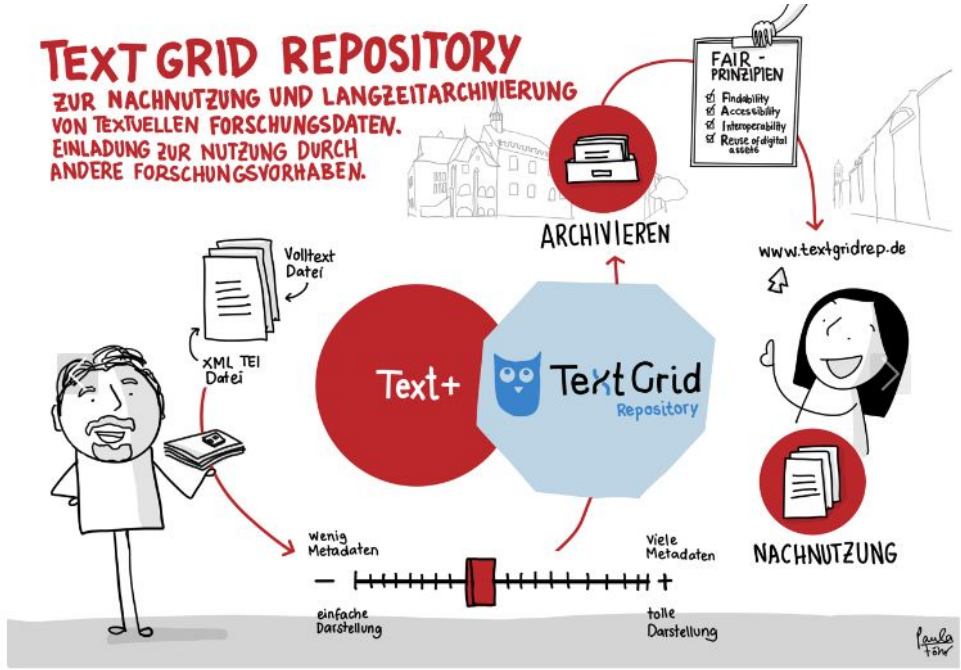


Illustration: Paula Föhr

https://text-plus.org/

30.10.2024

Design der Umfrage

- Zielgruppe: Literaturwiss., die mit **digitalen Korpora** arbeiten
- Motto: **Wünsch dir was** - Umsetzung spielt erstmal keine Rolle
- Auswahl der abgefragten Metadaten in **Vorbereitung mit Wiss.**
- Fragen auf die Bereiche **Autor**, **Werk**, **Text** angewendet
- nicht eingegrenzt auf digitale Bibliotheken, sondern alle **Kataloge**

Kritische Diskussionen

» Warum fragen wir nach Metadaten wie: sexuelle Orientierung der Autor-Person, Religion oder politische Richtung?

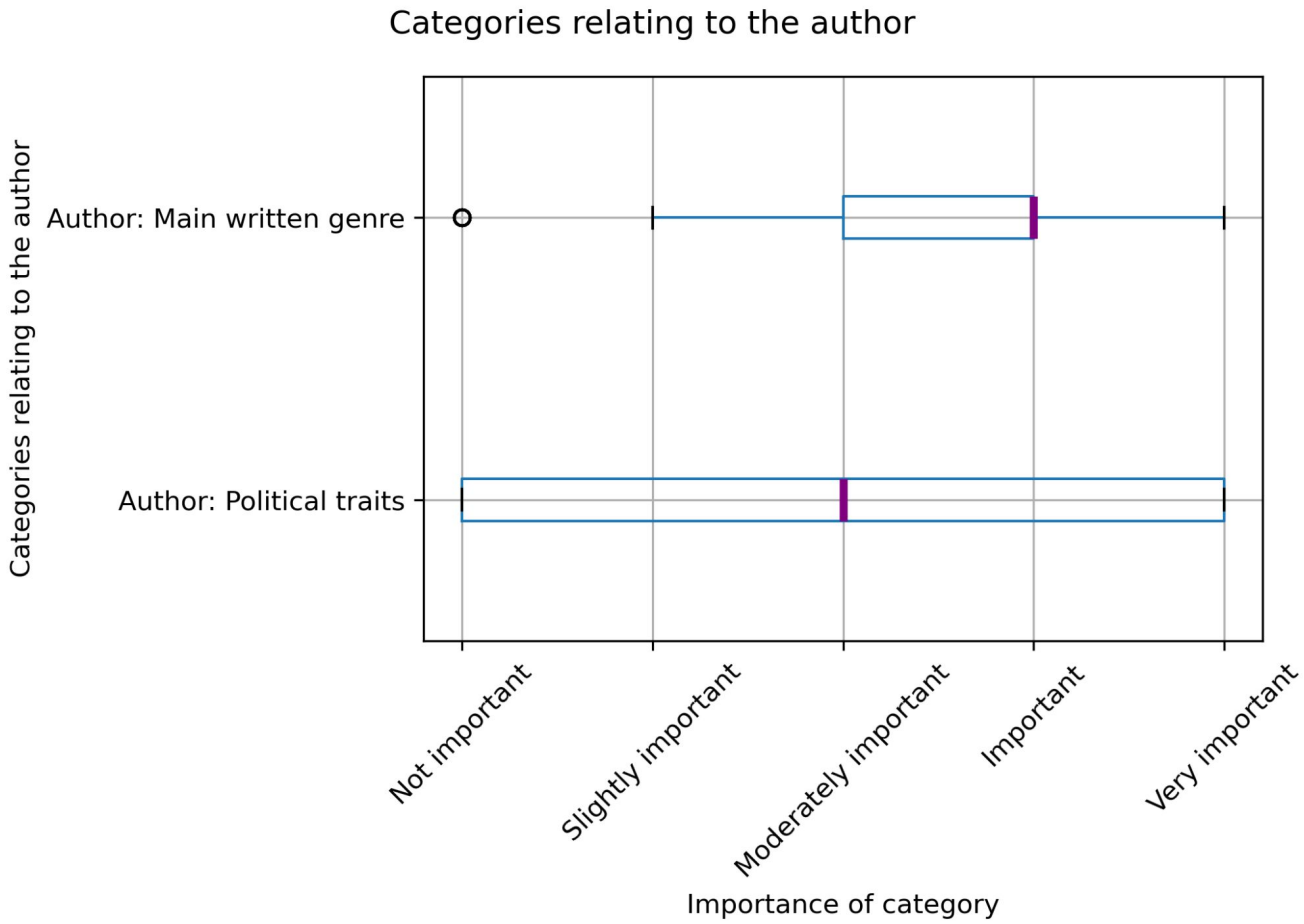
» Warum bitten wir die Teilnehmenden der Umfrage, um persönliche Daten (Ausbildung, Beruf, Geschlecht)?

Information zu den Befragten

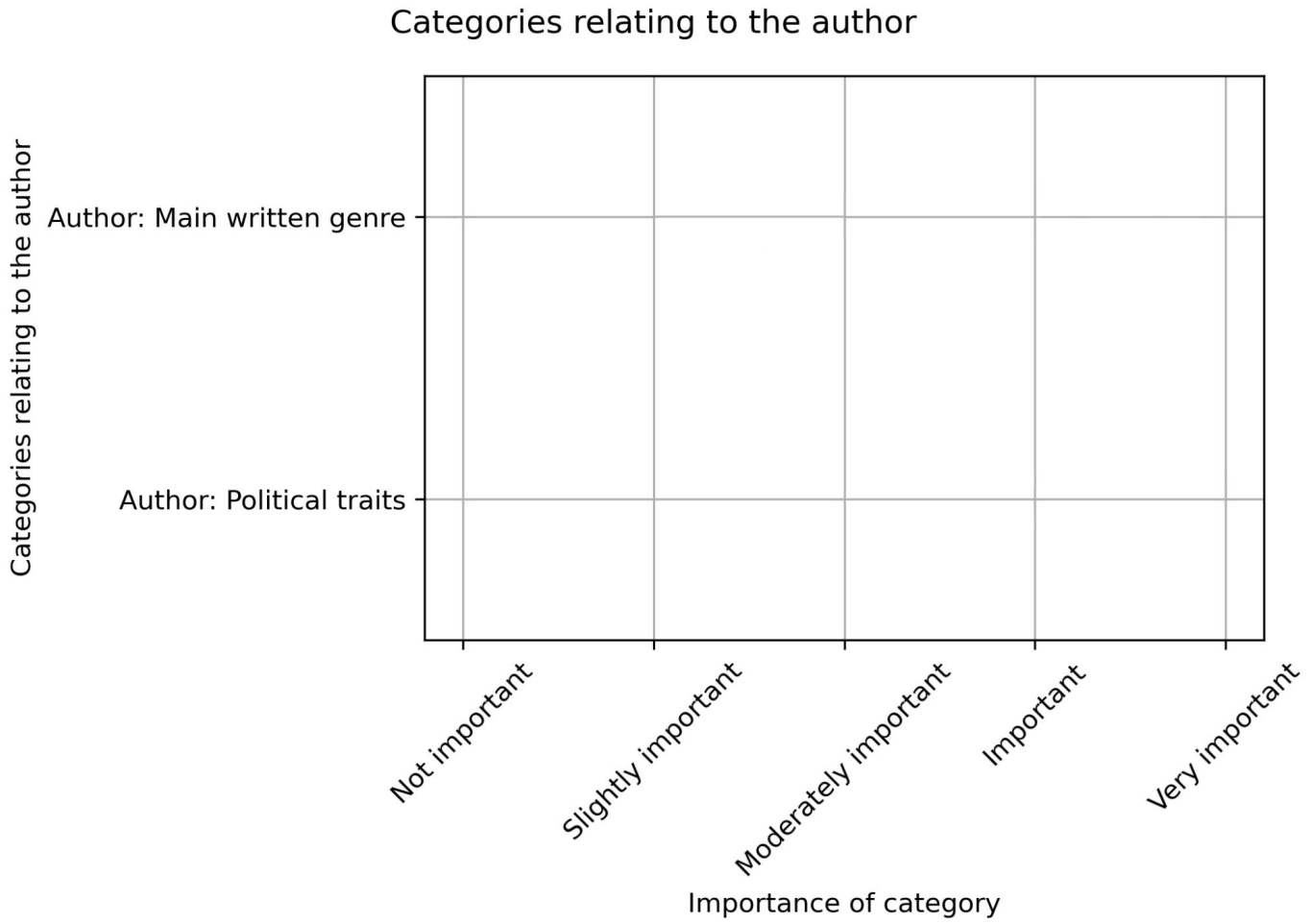
- 317 begonnene Fragebögen -> **111** vollständig ausgefüllt
- Alle persönlichen Fragen waren optional:
- aus **19 Ländern**, davon aber **52 % aus Deutschland**
- **wenig Young Researchers** oder Studierende (nur 12 Doktoranden), Altersdurchschnitt 40+, Gender balanced
- **viele Fächer** neben Literaturwissenschaft vertreten

Ergebnisse der Umfrage

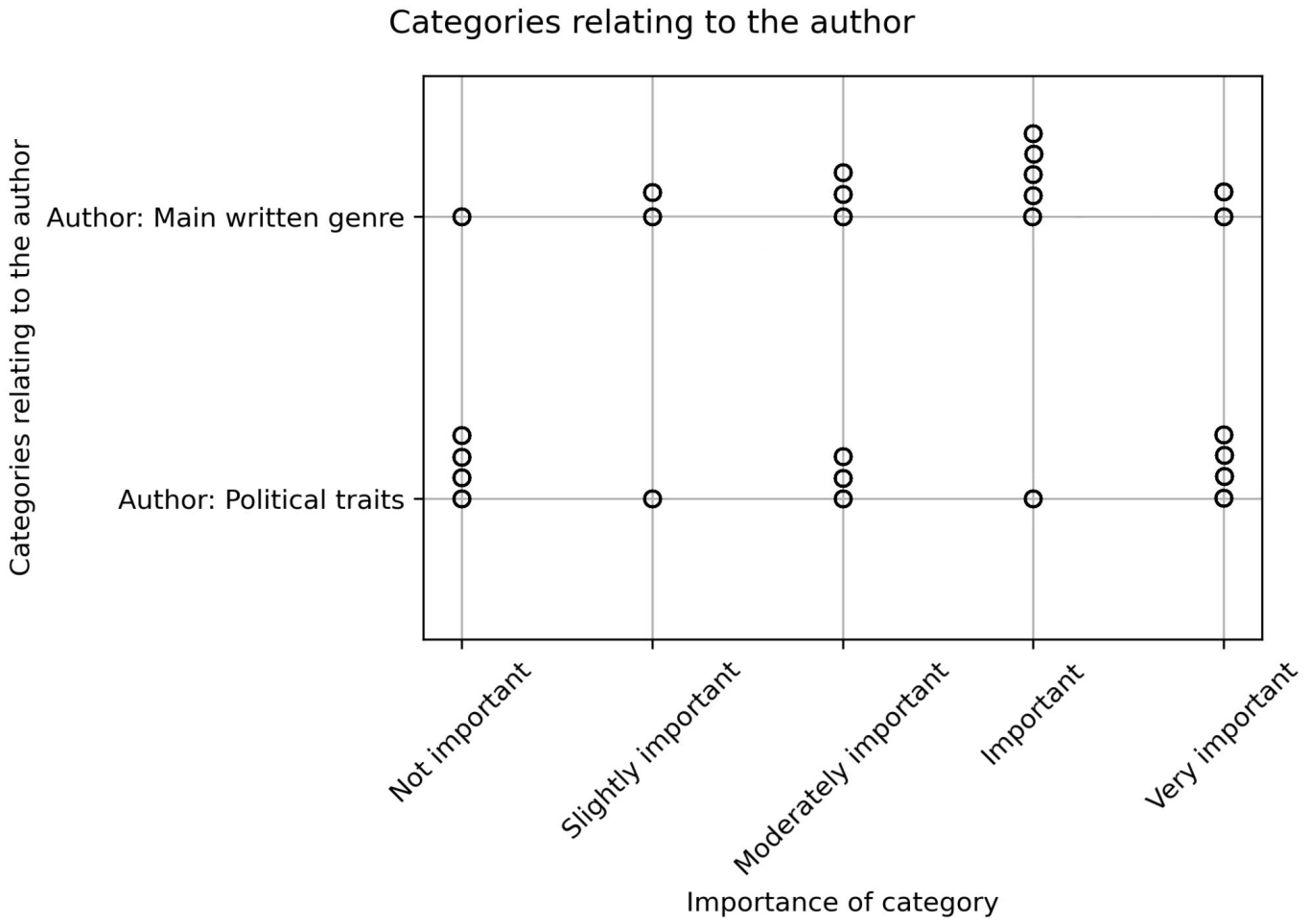
Boxplots


Categories relating to the author
Categories relating to the author
Author: Main written genre
Author: Political traits
Not important
Slightly important
Moderately important
Important
Very important
Importance of category

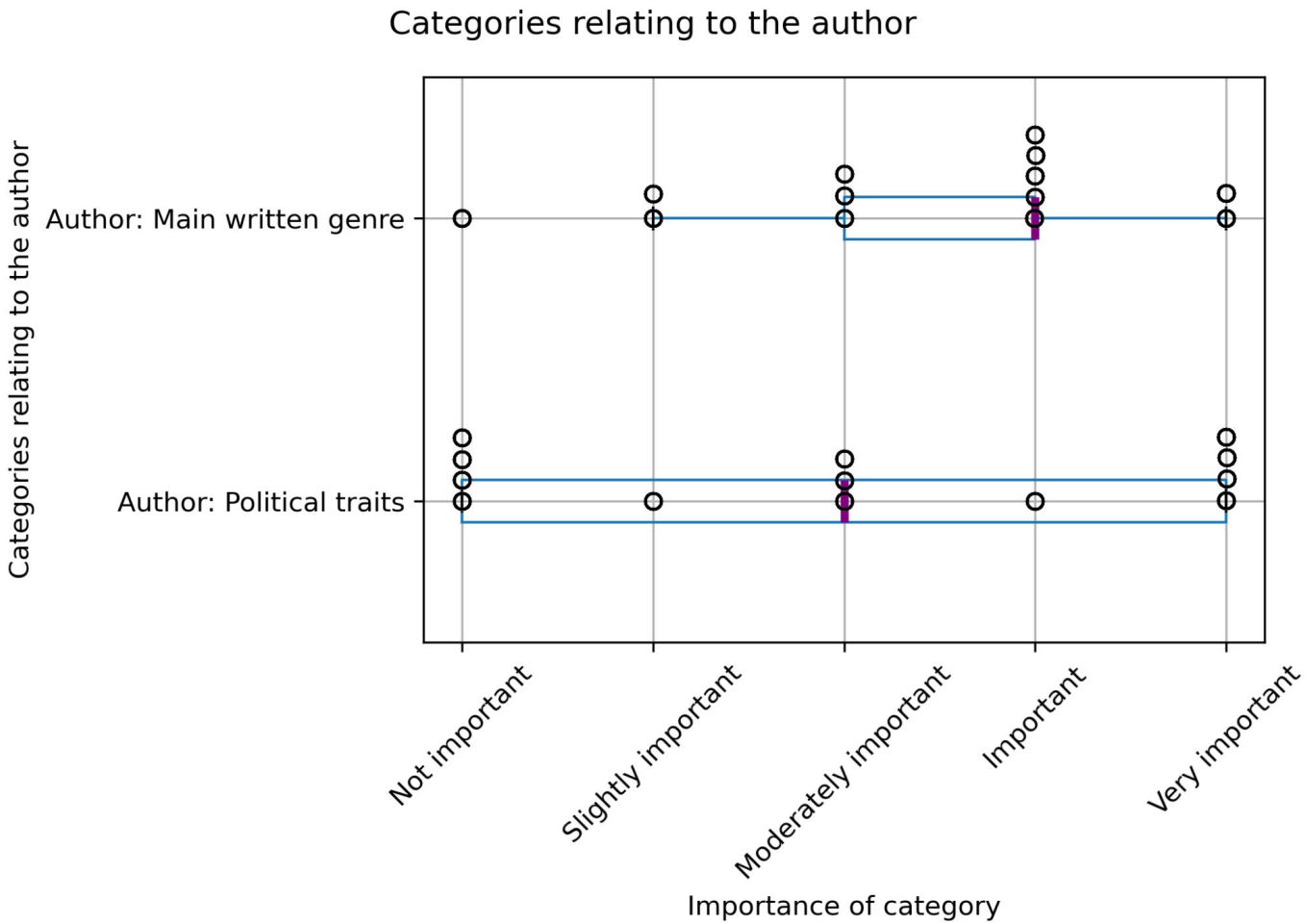
Boxplots


Categories relating to the author
Categories relating to the author
Author: Main written genre
Author: Political traits
Not important
Slightly important
Moderately important
Important
Very important
Importance of category

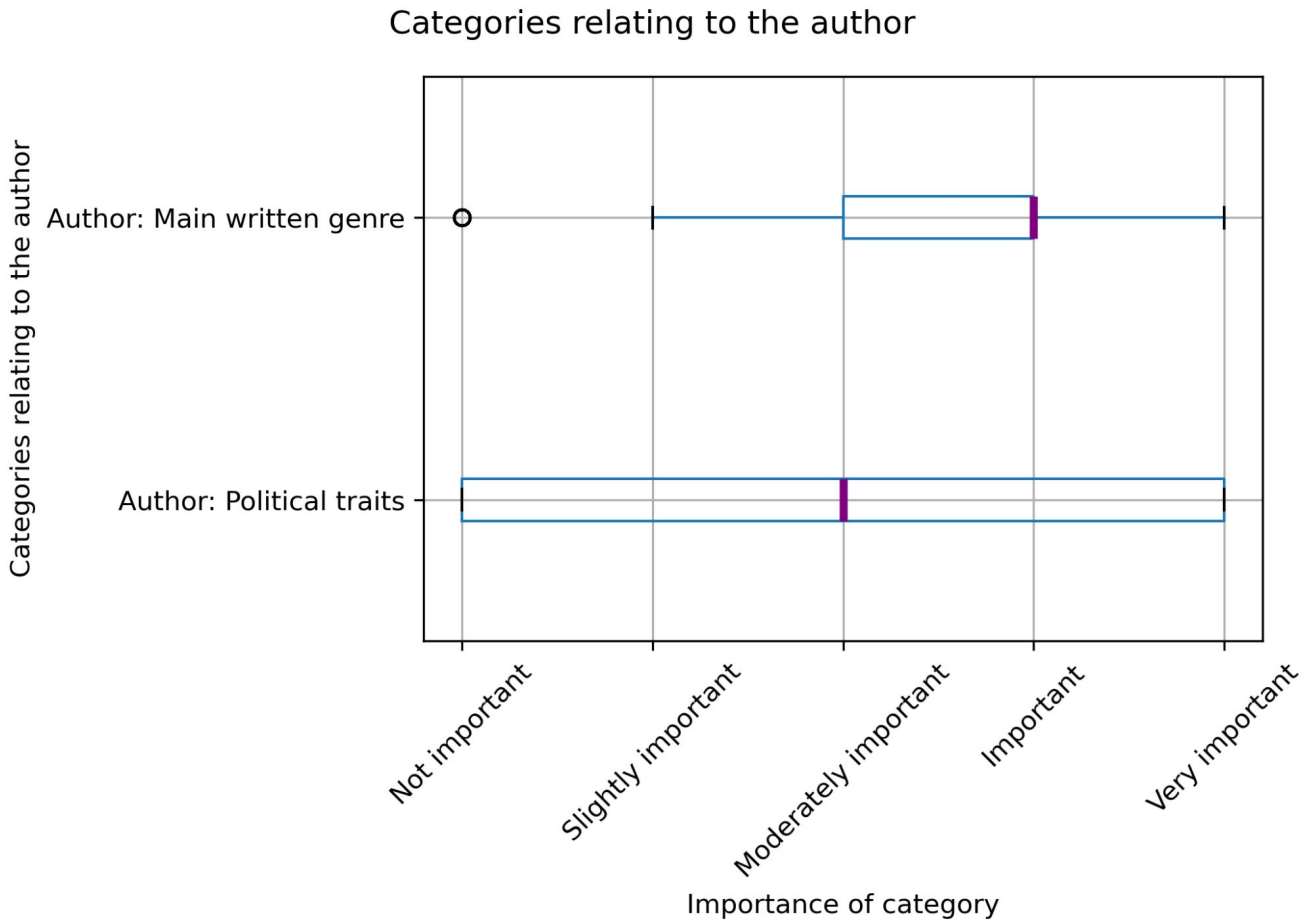
Boxplots


Categories relating to the author
Categories relating to the author
Author: Main written genre
Author: Political traits
Not important
Slightly important
Moderately important
Important
Very important
Importance of category

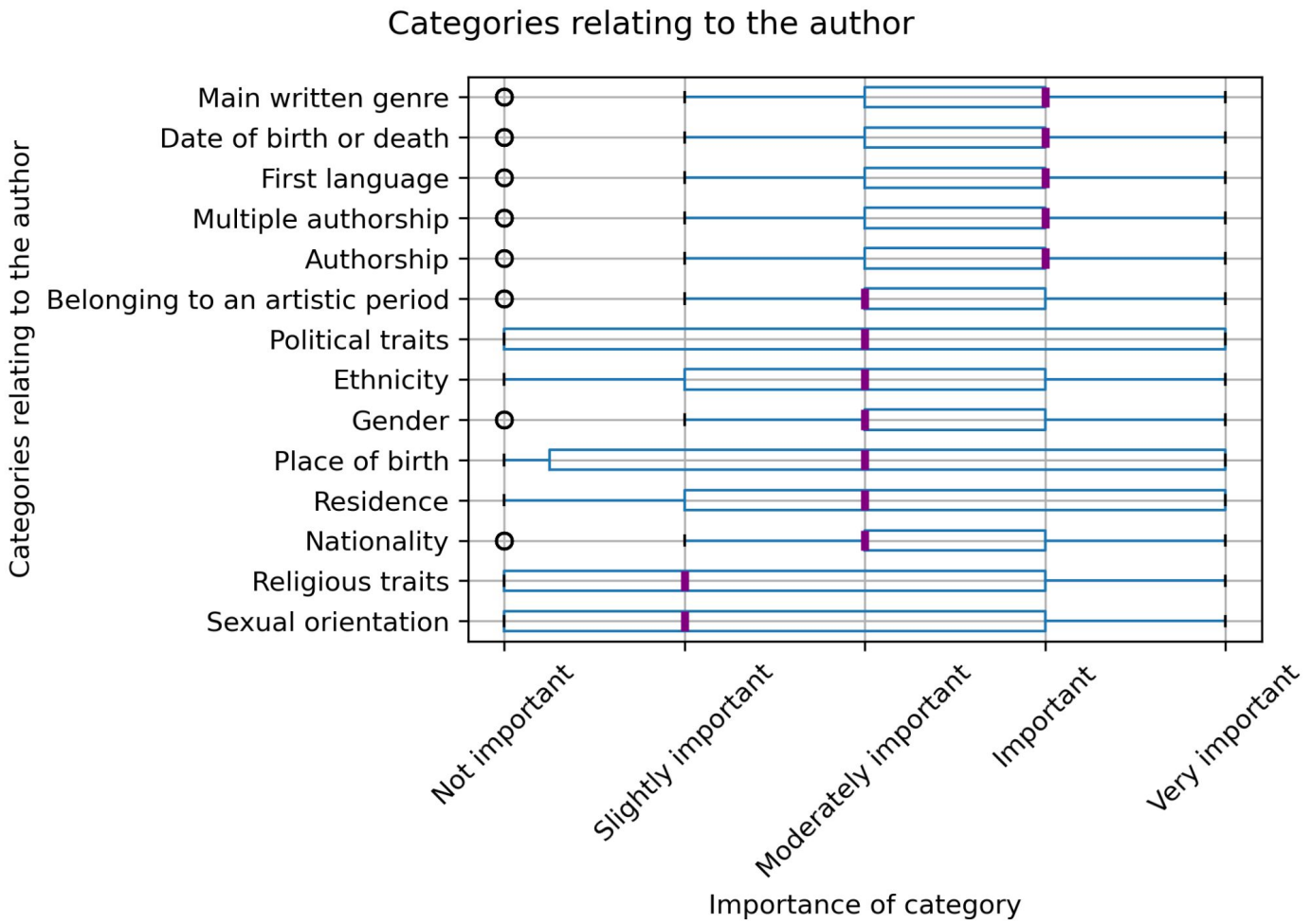
Boxplots


Categories relating to the author
Categories relating to the author
Author: Main written genre
Author: Political traits
Not important
Slightly important
Moderately important
Important
Very important
Importance of category

Boxplots


Categories relating to the author
Categories relating to the author
Author: Main written genre
Author: Political traits
Not important
Slightly important
Moderately important
Important
Very important
Importance of category

Ergebnisse auf Autor-Ebene



Text+

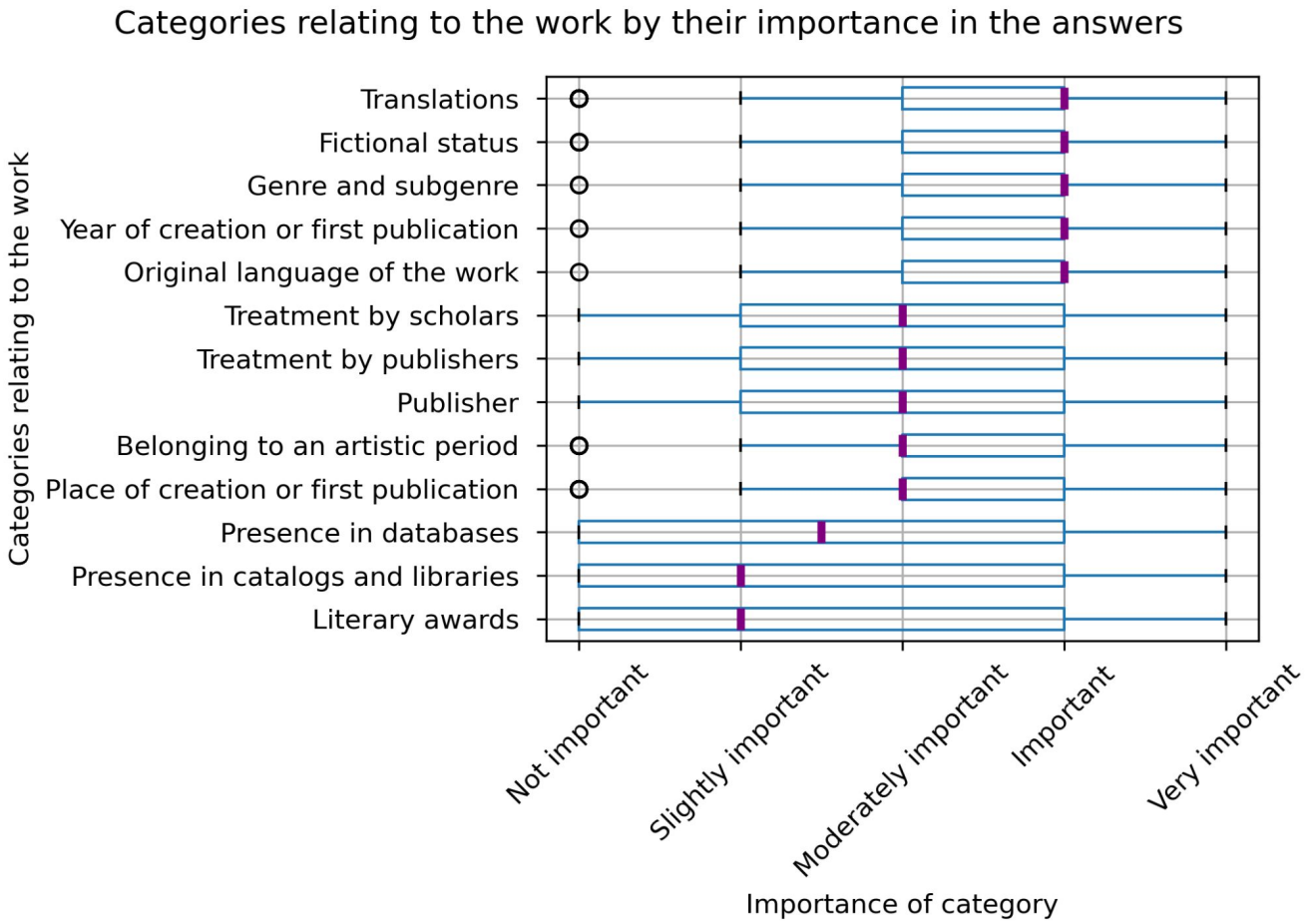
Computational LiterarySTUDIES

CLSINFRA

Ergebnisse auf Autor-Ebene

- Weitere Kategorien:
  - Kulturelle, intellektuelle, bildungsbezogene, berufliche und soziale Gruppen und Zugehörigkeiten
  - Sprachen, in denen der Autor geschrieben hat

Ergebnisse auf Werk-Ebene



Text+

Computational LiteraryStudies

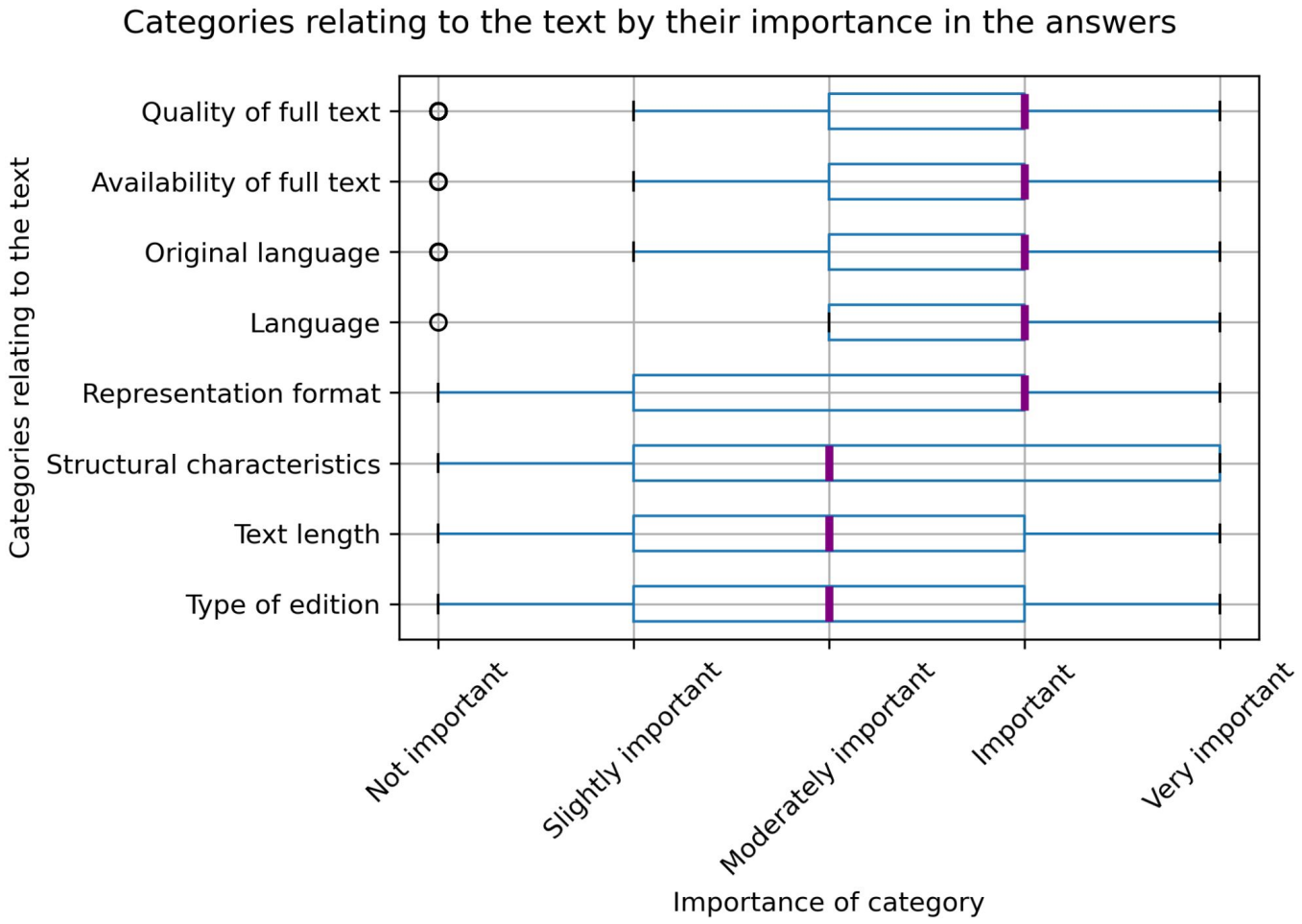
CLSINFRA

30.10.2024

Ergebnisse auf Werk-Ebene

- Weitere Informationen zu Übersetzungen:
  - Übersetzer*in
  - Von welcher Sprache zu welcher Sprache
- Das Werk wurde anonym oder unter einem Pseudonym veröffentlicht
- Verwandte Werke

Ergebnisse auf Text-Ebene



Text+

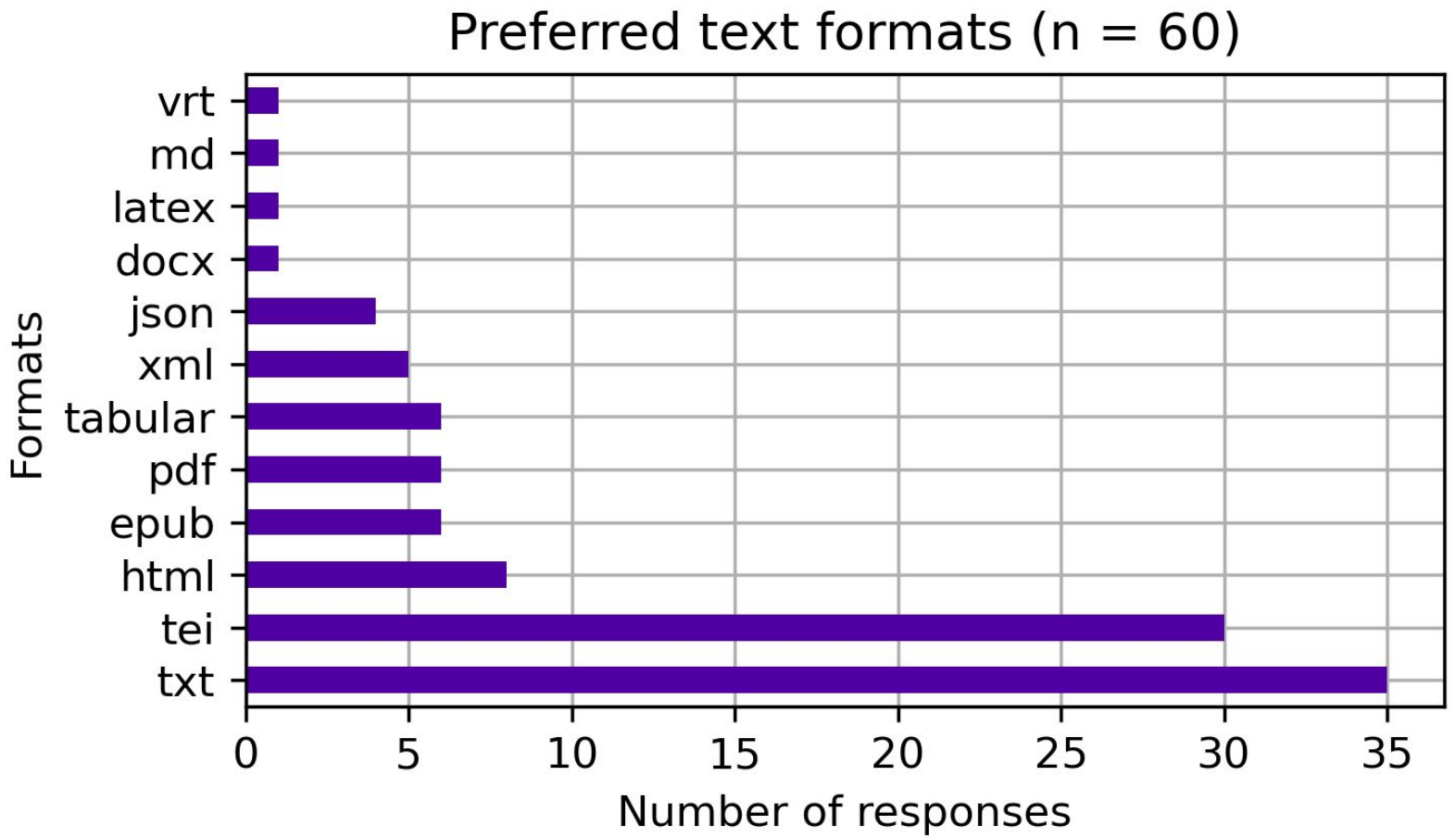
Computational LiteraryStudies

CLSINFRA

Ergebnisse auf Text-Ebene

- Paratexte, Layout, Illustrationen
- Copyright-Status
- Transkription, Rechtschreibung
- Angaben zur Veröffentlichung (Fortsetzungsroman, Sammelband, Miscellany...)

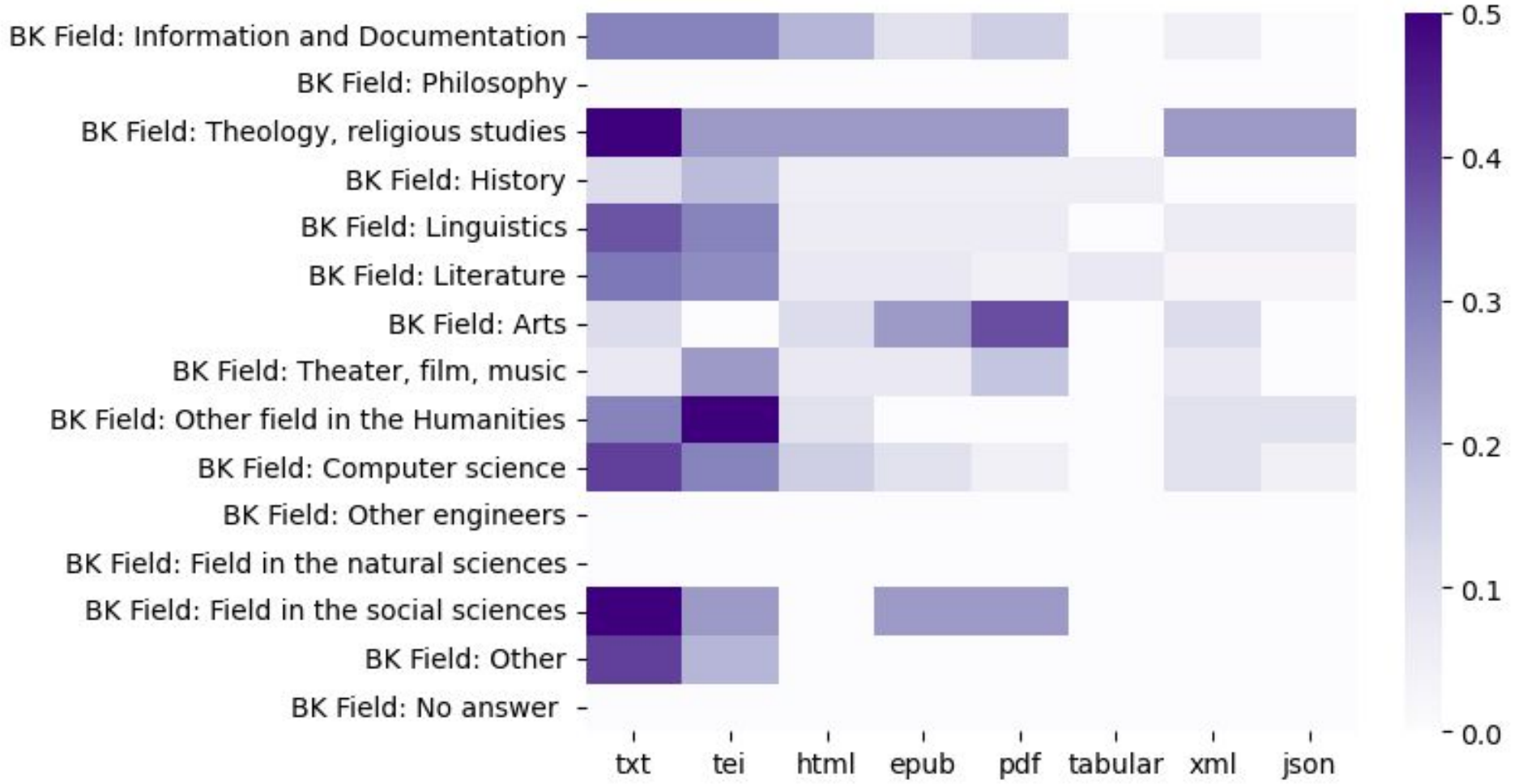
Ergebnisse auf Text-Ebene



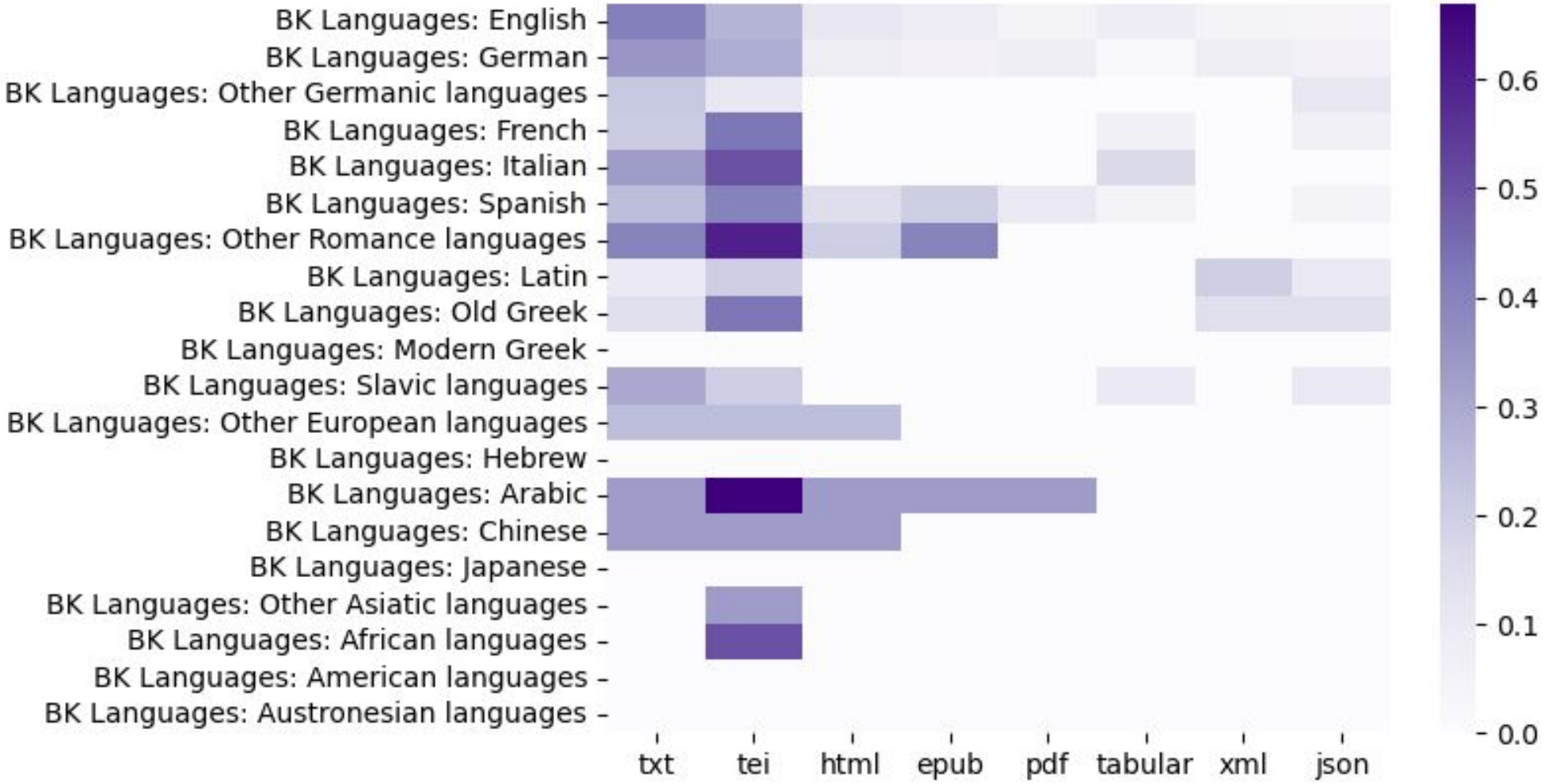
Ergebnisse auf Text-Ebene

- “tei ideal, plain text if tei not available”
- “I analyze texts, so anything I need to OCR is bad”
- “The main criterion is easy "processability": TEI is great (although perhaps too complex), plain text is easy to process (but with some information lost), PDF is often a pain (hard to process with reasonable accuracy, but this depends on the source).”

Ergebnisse auf Text-Ebene



Ergebnisse auf Text-Ebene

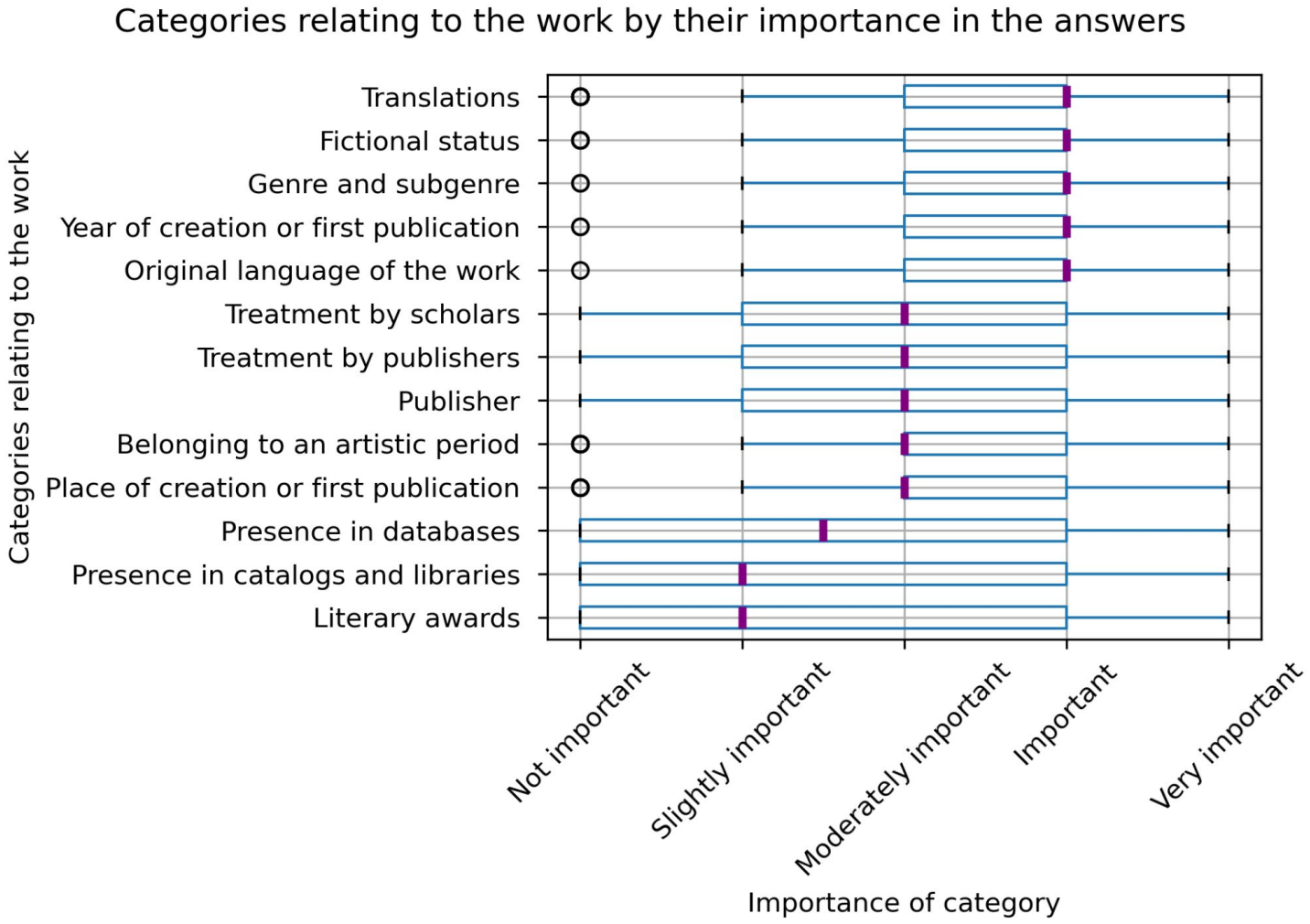


Ergebnisse auf Text-Ebene

» “Anything I can get my hands on”

» “Like, whatever, man. I've analyzed Klingon and Quenya as well as English and German.”

Änderungen im TextGrid Repository



Text+

Computational Literary Studies

CLSINFRA

Änderungen im TextGrid Repository

» Gattung wird zu einer neuen Priorität

Änderungen im TextGrid Repository

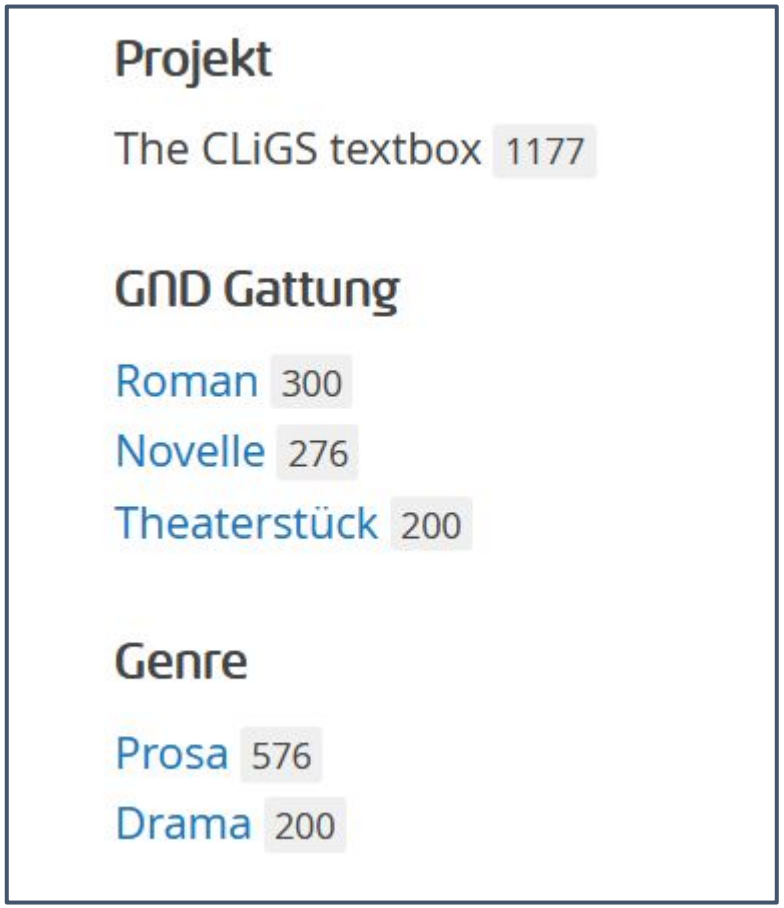
- Repositorium für Forschungsdaten (Korpora, Editionen) in TEI
- Core Trust Seal zertifiziert
- Neue Korpora (CoNSSA, ELTeC, textbox, French Novel 18th century, Kaukasische Folklore...)
- Neuer *fluffiger* Workflow
- Portalconfig, Landing Page
- Projekt-spezifische Optionen
- Python Library
- Zugang über Oxygen
- TEI → HTML → Plaintext
- Bibliothekarische Klassifikationssystem (Basisklassifikation)

Bisherige Gattungen im TextGrid Repository

» Nur 6 Gattungen zur Auswahl: Vers, Prosa, Drama, Fachtext, Nachschlagewerk, Anderes

Neue Gattungen im TextGrid Repository

- 1.326 Gattungen in der GND
- *Subject* Element in den Werk-Ebene in den TextGrid Metadaten Modell
- Mit ID der GND Entität
- In Absprache mit der Text+ GND-Agentur



Text+

Computational LiterarY Studies

CLSINFRA

Drei neue Projekte

- textbox
  - Roman
  - Novelle
  - Theaterstück
- Collection of Eighteenth-Century French Novels 1751-1800
  - Roman
- Kaukasische Folklore
  - Volkserzählung

Weitere neue Projekte

- Mehr Projekte sind geplant
- “Meine Gattung ist nicht vorhanden”
- “Sprich mit der GND!”

Zusammenfassung und weitere Schritte

- **Autor**: Hauptgattung, Geburt, Sprachen, Geschlecht; Epoche, politische Zugehörigkeit, Orte
- **Werke**: Übersetzung, fiktiv, Gattung, 1. Veröffentlichungsjahr und -sprache
- **Text**: Format (txt und TEI), Volltext, Qualität, Sprache

Zusammenfassung und weitere Schritte

- Welche dieser Daten sollen in welche bibliothekarischen Ressourcen aufgenommen werden?
  - Kataloge, Forschungsdatenrepositorien, Normdaten…
- Nicht alles muss in bibliothekarischen Ressourcen vorhanden sein
- Enthalten bibliothekarische Ressourcen diese Metadaten?
- Wenn ja, wie gut ist die Abdeckung?
- Wenn nicht, was muss geändert werden, damit die Forschenden diese Metadaten finden können?

Zusammenfassung und weitere Schritte

- Prioritäten in bibliothekarischen Ressourcen setzen
- Qualität der (Meta)Daten sollte Priorität von Bibliotheken sein
- Welche Informationen sind besonders wichtig? Z.B. **Gattung**
- Welche Arten von Daten und Publikationen sind besonders wichtig?
- Ressourcen dafür einsetzen bzw. umlenken
- Beispiel:
  - Zeit für die Sacherschließung von gedruckten Titeln in den Fachreferaten
  - → andere Daten wie Forschungsdaten oder **Normdaten** (Personen, Werke, Schlagwörter, Klassifikationssysteme) anreichern und bearbeiten
- Gerne mit Vorschlägen von Algorithmen

Zusammenfassung und weitere Schritte

» Bibliotheken als Ort menschlicher Expertise, unterstützt von Technologie, um den Bedürfnissen der Benutzer*innen gerecht zu werden.

Text+

Collections
Lexical Resources
Editions
Infrastructure/
Operations

Danke für Ihre Aufmerksamkeit

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
Project number 460033370
Part of
nfdi Nationale Forschungsdaten Infrastruktur

This presentation was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.