TIB | Technische Informationsbibliothek

TDM und Herausforderungen für Bibliotheken

Elke Brehm

vdb-Fortbildung für Fachreferentinnen und Fachreferenten der Naturwissenschaften

28.09.2015

Bedarf und Anforderungen an Ressourcen für Text und Data Mining

April - Mai 2015: **Umfrage der Arbeitsgruppe „Text und Data Mining“** der Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen

September 2015: **Workshop mit Wissenschaftlern**, DFG zwecks Auswertung und Vertiefung von offenen Fragen

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Ziel der Umfrage

- Grund: Verlage versuchen Regelungen zu Text und Data Mining in Lizenzverträge aufzunehmen
- Frage: Welche inhaltlichen und technischen Voraussetzungen müssen seitens der Wissenschaft erfüllt sein?
  (Feststellung des tatsächlichen Bedarfs: Was wäre zu 100% optimal?)
- Ziel: Treffen von Festlegungen im Sinne der Wissenschaft zur Durchführung von TDM, um nicht auf Dauer ungeregelte Zustände zu haben
- Es geht hier nicht um Mittel und Möglichkeiten zur Anpassung der Rahmenbedingungen (rechtliche Aspekte, Lizenzen, etc.)

Schwerpunkt Initiative „Digitale Information“, Arbeitsgruppe „Text and Data Mining“: Präsentation Katerbow, Mittermaier, Schöch, Sens

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Definition

"Text und Data Mining" meint hier:

- die Nutzung von Software
- zur Analyse
- digital vorliegender Inhalte
(Publikationen oder Forschungsdaten, u.a. in Form von Zahlen, Texten, Bildern, Metadaten, etc.)
- mit dem Ziel, wissenschaftliche Fragestellungen zu bearbeiten

Schwerpunkt Initiative „Digitale Information“, Arbeitsgruppe „Text and Data Mining“: Präsentation Katerbow, Mittermaier, Schöch, Sens

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Befragte Personen

(Teil A)

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Befragte Personen

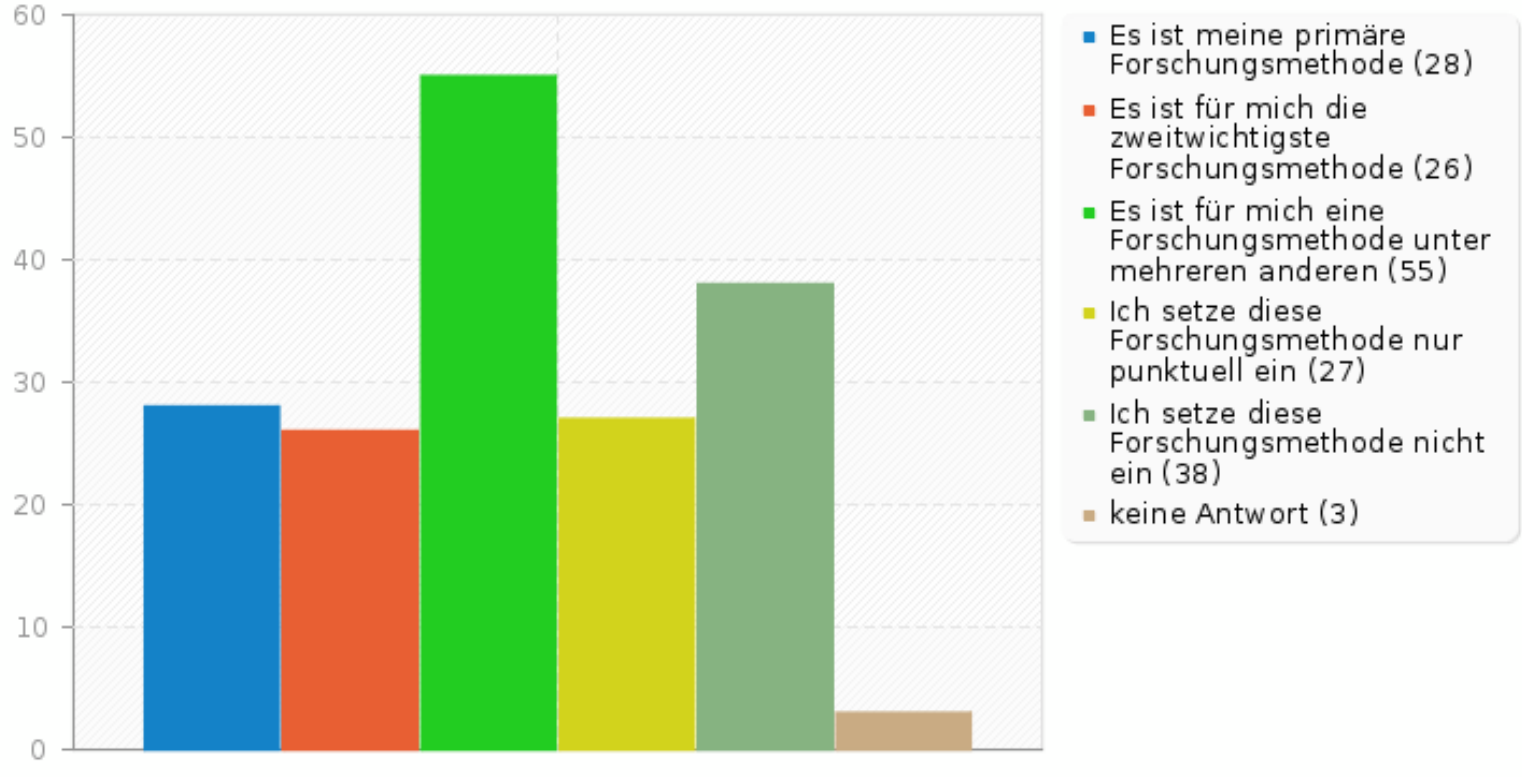
- N = 177
- Überwiegend wiss. MA (106) und leitende Personen (55)
- Universitäten (79) und außeruniversitäre Einrichtungen (85)
- 4 Wissenschaftsbereiche:
  - Geistes- und Sozialwissenschaften
  - Naturwissenschaften
  - Ingenieurwissenschaften
  - Lebenswissenschaften

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Stellenwert von TDM

(Teil B)

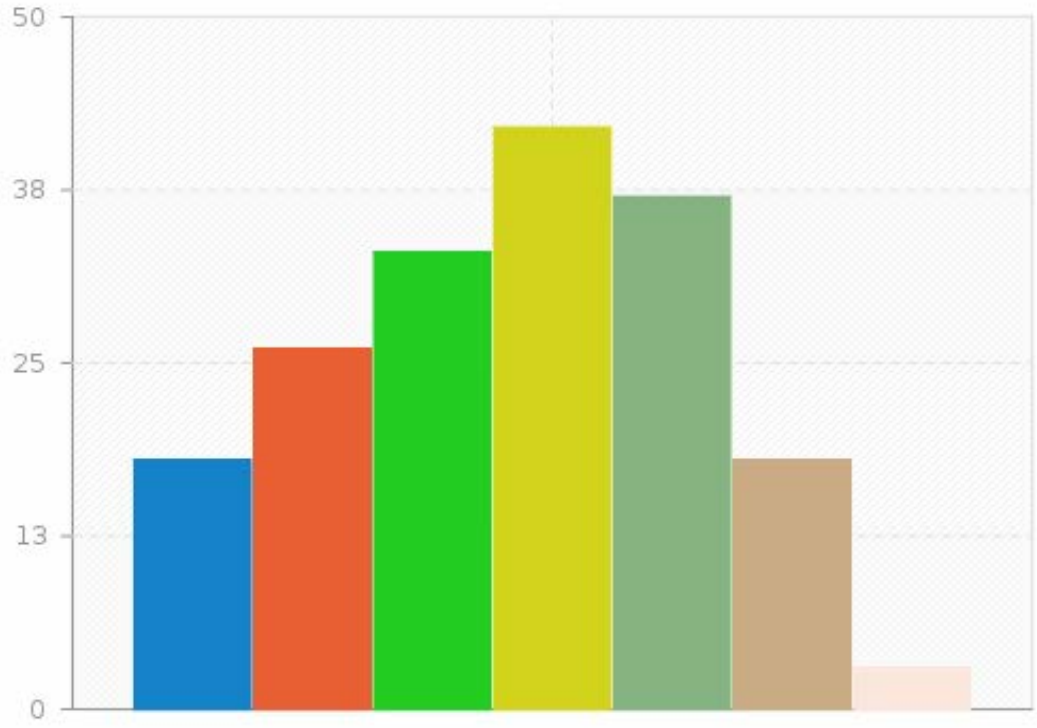
TDM als Forschungsmethode



Schwerpunkt Initiative „Digitale Information“, Arbeitsgruppe „Text and Data Mining“: Präsentation Katerbow, Mittermaier, Schöch, Sens

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Nutzen von TDM



- 1-keinen (18)
- 2-geringen (26)
- 3-mittleren (33)
- 4-großen (42)
- 5-ausschlaggebenden (37)
- 6-kann ich nicht beurteilen (18)
- keine Antwort (3)

Schwerpunkt Initiative „Digitale Information“, Arbeitsgruppe „Text and Data Mining“: Präsentation Katerbow, Mittermaier, Schöch, Sens

TIB Technische Informationsbibliothek

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Methoden in den Natur- und Ingenieurwissenschaften

Genutzte Methoden:

- Klassische Recherche
- Bibliometrische Methoden und Altmetrics
- Beobachtung
- Laborarbeiten
- Computational Science
- Modellierung und Simulationen

Text und Data Mining zur Unterstützung bei der Validierung/Verifikation der Ergebnisse eingesetzt, relevanter ist Data Mining

Big/Smart Data benötigt

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Daten für TDM

(Teil C)

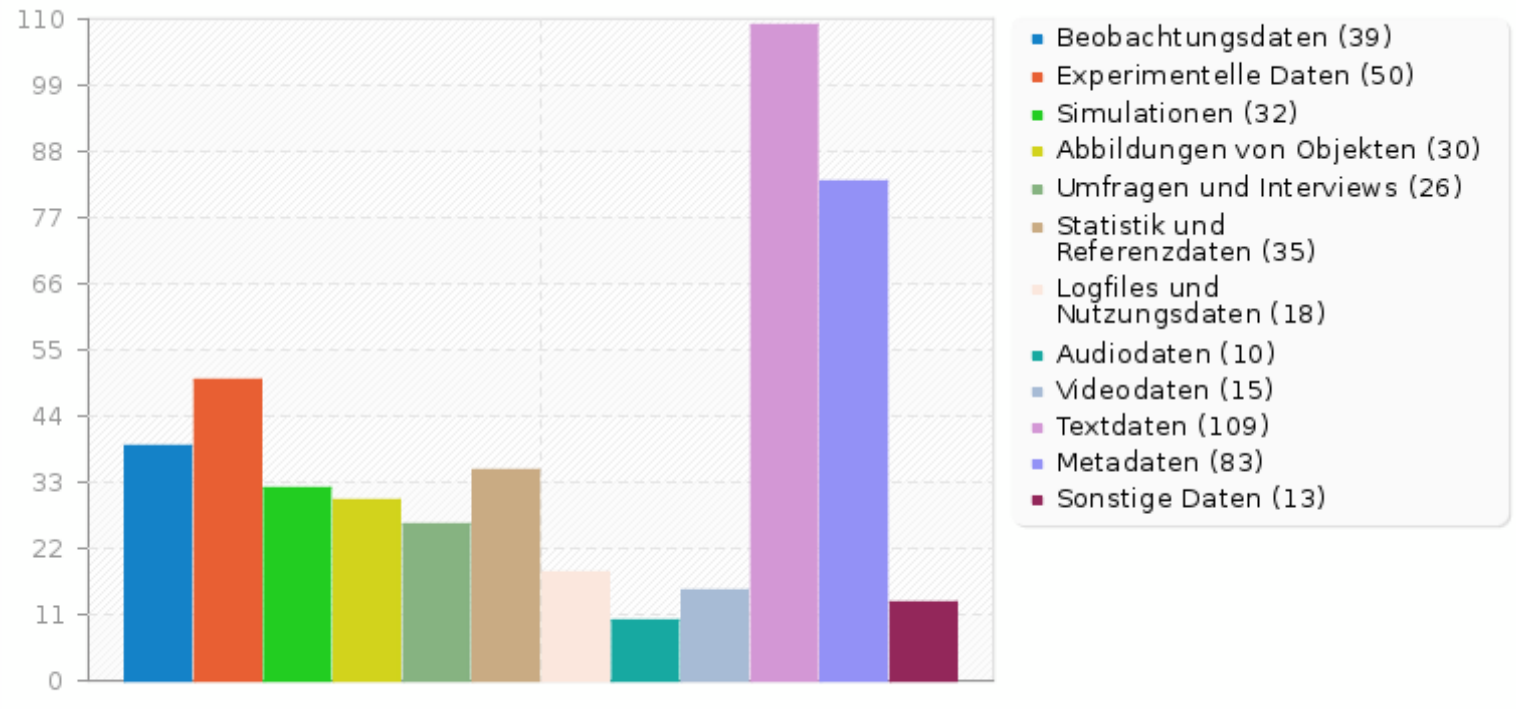
CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Datenquellen

- Sehr viele Datenbanken für wissenschaftliche Zeitschriftenartikel
- Sehr viele naturwissenschaftliche Faktendatenbanken
- Einige Datenbanken mit literarischen Texten
- Einige Textressourcen und Wörterbücher
- Mehrere Messinfrastrukturen mit Datenangeboten

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Datentypen



Schwerpunkt Initiative „Digitale Information“, Arbeitsgruppe „Text and Data Mining“: Präsentation Katerbow, Mittermaier, Schöch, Sens

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Datentypen in den Natur- und Ingenieurwissenschaften

Daten:

- strukturierte Daten dominieren, teilweise auch historische Daten, wenn aus Texten strukturierte Daten rekonstruiert werden sollen.
- wenig Texte: in der Regel nur, wenn Daten aus Texten extrahiert werden müssen
- Zugang zu Texten kein Problem (meist digital vorliegend)
- Datenmanagement wichtig, insbesondere bei heterogenen Datenquellen
- „Weltdatenzentrum“ für Informatik gewünscht

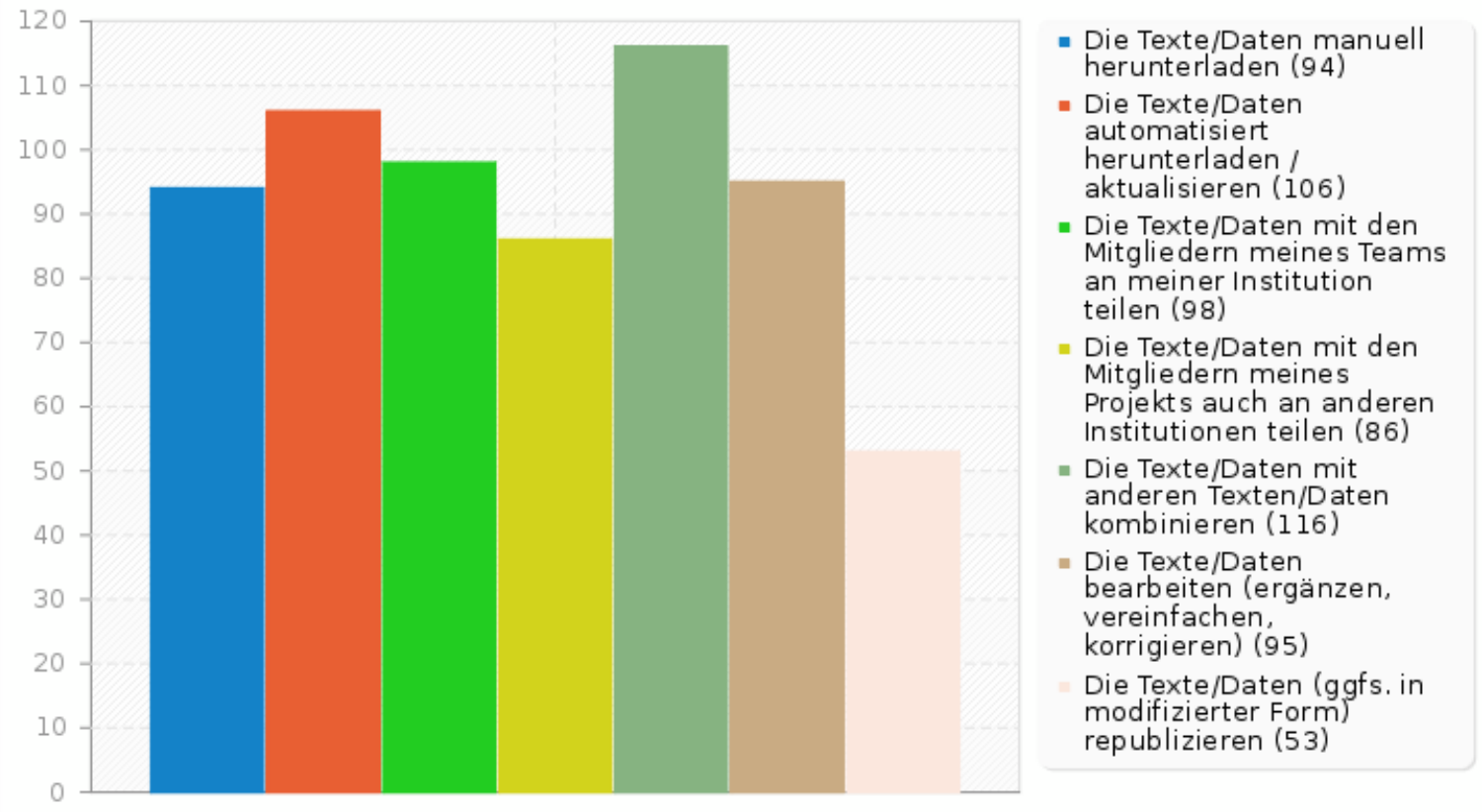
CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Datenanalyse mit TDM

(Teil D)

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Datenanalyse mit TDM: Nutzungsszenarien



Schwerpunkt Initiative „Digitale Information“, Arbeitsgruppe „Text and Data Mining“: Präsentation Katerbow, Mittermaier, Schöch, Sens

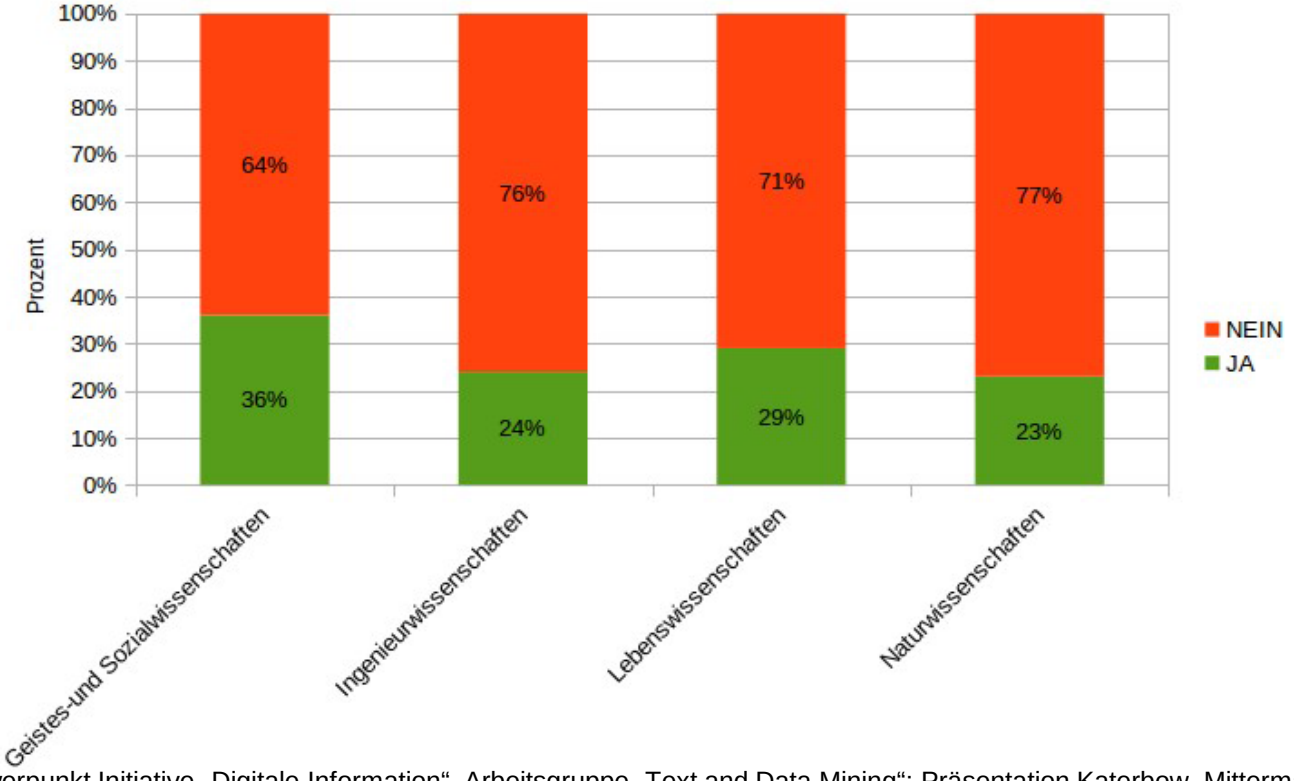
CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

„Vollständigkeit“ von Daten in Natur- und Ingenieurwissenschaften?

„Datenmenge ist nie vollständig. Wichtig ist, ob Datenmenge ausreicht, um vorliegende Frage zu beantworten“

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Republizieren von Daten?



Schwerpunkt Initiative „Digitale Information“, Arbeitsgruppe „Text and Data Mining“: Präsentation Katerbow, Mittermaier, Schöch, Sens

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

„Republikation“ von Daten in Natur- und Ingenieurwissenschaften

- In einzelnen Disziplinen sehr unterschiedlich
  - Meereswissenschaften: alle Daten offen
  - Chemie: keine offenen Daten
  - Ingenieurwissenschaften: wichtige Industriedaten nicht verfügbar

Begriff „Republikation“ unpassend: Es geht um erstmalige Publikation selbst produzierter Daten

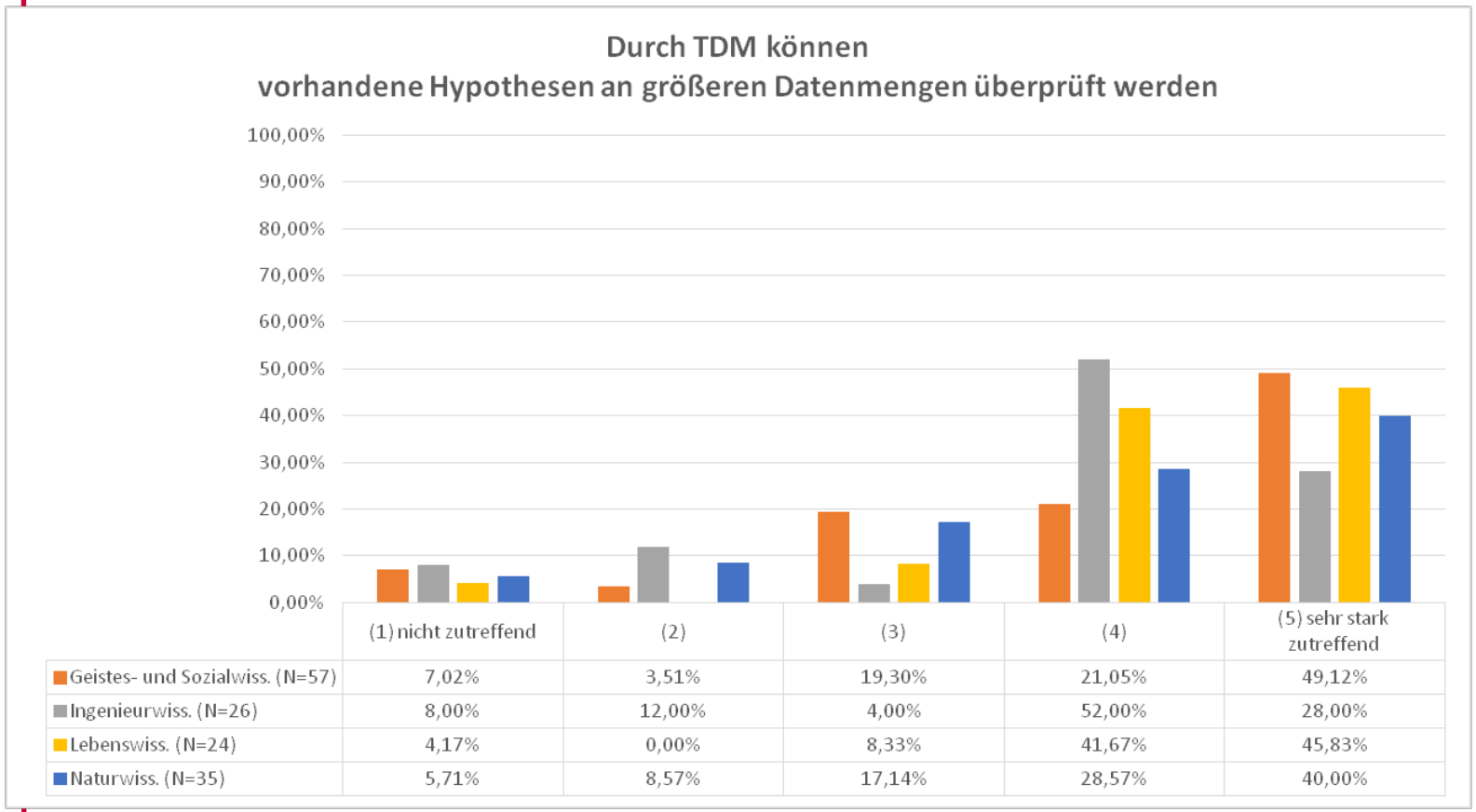
CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Vorteile und Hindernisse von TDM

(Teil E)

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

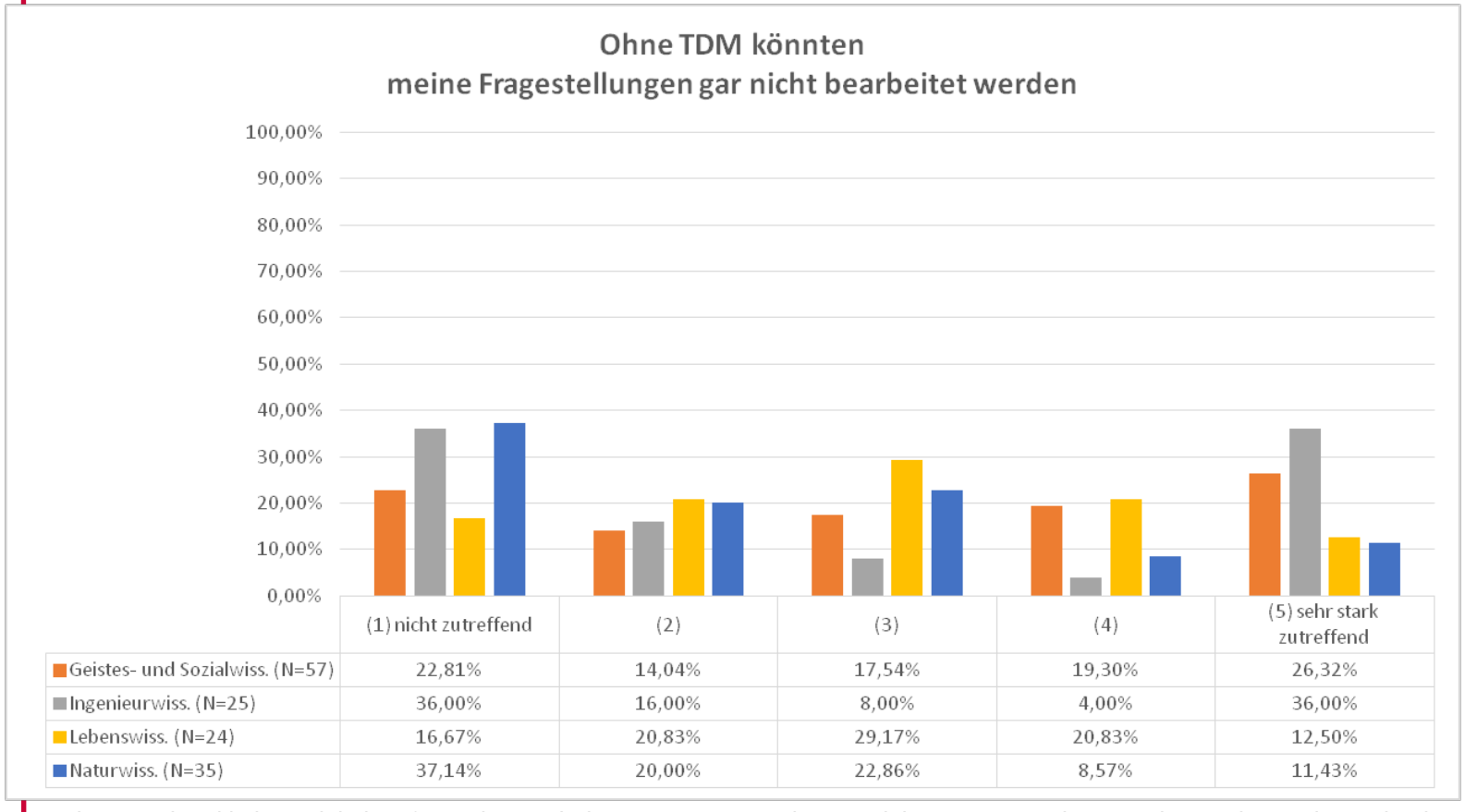
Nutzen von TDM



	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=57)	7,02%	3,51%	19,30%	21,05%	49,12%
Ingenieurwiss. (N=26)	8,00%	12,00%	4,00%	52,00%	28,00%
Lebenswiss. (N=24)	4,17%	0,00%	8,33%	41,67%	45,83%
Naturwiss. (N=35)	5,71%	8,57%	17,14%	28,57%	40,00%

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

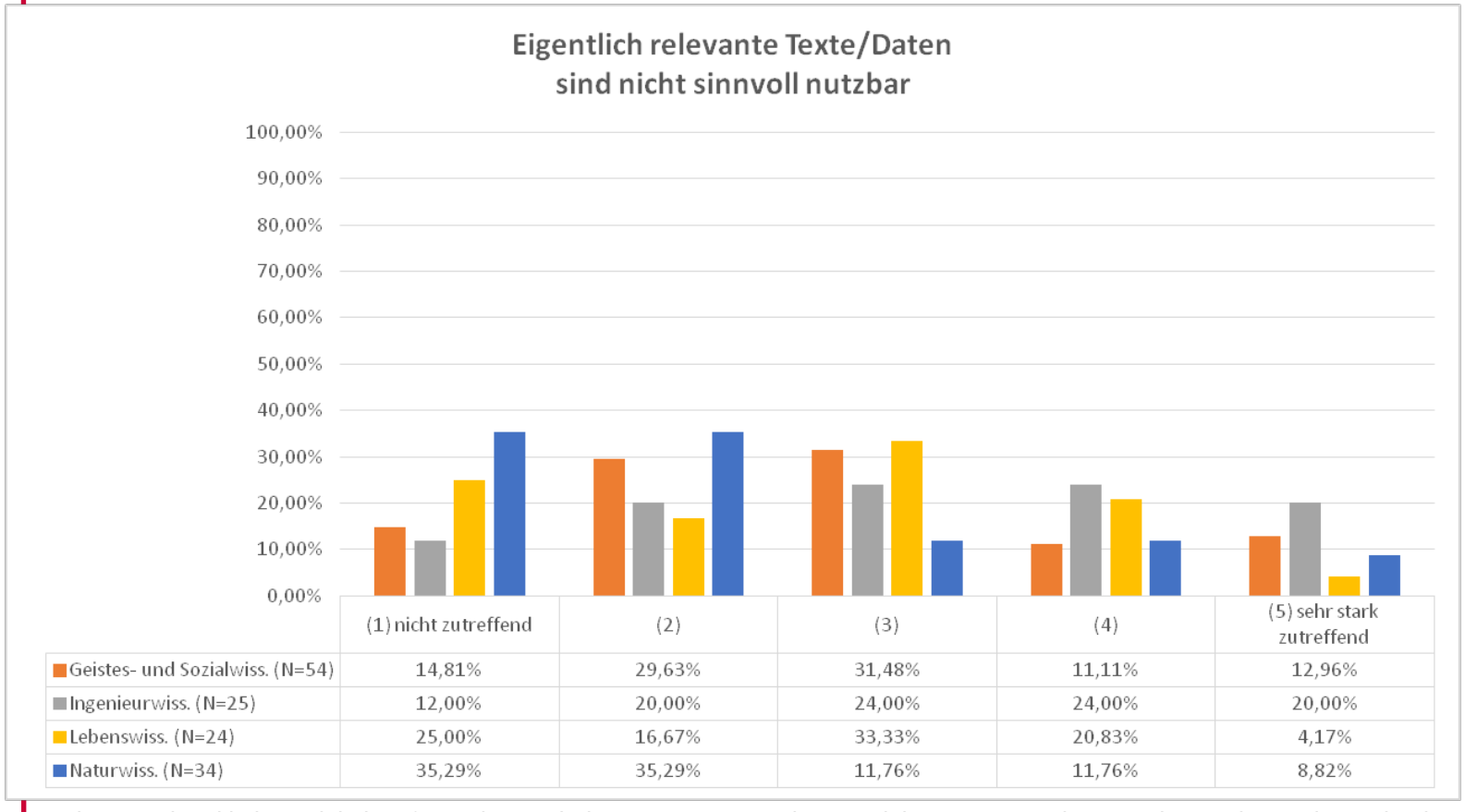
Nutzen von TDM



	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=57)	22,81%	14,04%	17,54%	19,30%	26,32%
Ingenieurwiss. (N=25)	36,00%	16,00%	8,00%	4,00%	36,00%
Lebenswiss. (N=24)	16,67%	20,83%	29,17%	20,83%	12,50%
Naturwiss. (N=35)	37,14%	20,00%	22,86%	8,57%	11,43%

TIB Technische Informationsbibliothek
CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

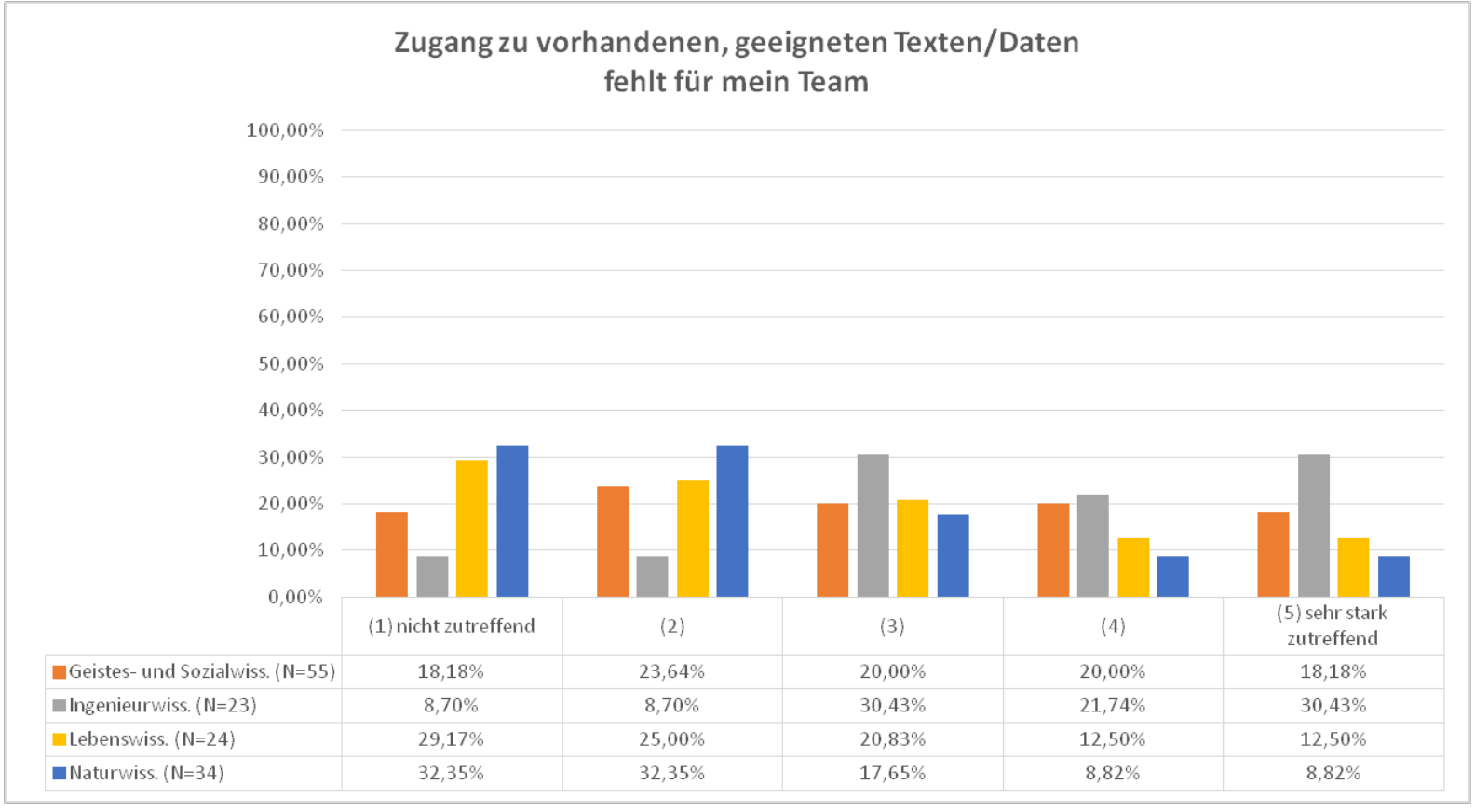
Nutzungshindernisse bei TDM



	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=54)	14,81%	29,63%	31,48%	11,11%	12,96%
Ingenieurwiss. (N=25)	12,00%	20,00%	24,00%	24,00%	20,00%
Lebenswiss. (N=24)	25,00%	16,67%	33,33%	20,83%	4,17%
Naturwiss. (N=34)	35,29%	35,29%	11,76%	11,76%	8,82%

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Nutzungshindernisse bei TDM



	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=55)	18,18%	23,64%	20,00%	20,00%	18,18%
Ingenieurwiss. (N=23)	8,70%	8,70%	30,43%	21,74%	30,43%
Lebenswiss. (N=24)	29,17%	25,00%	20,83%	12,50%	12,50%
Naturwiss. (N=34)	32,35%	32,35%	17,65%	8,82%	8,82%

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Herausforderungen für die Nutzung von Text und Data Mining aus Sicht der Wissenschaftler

- Benötigt: Big/Smart data
- Heterogene Datenquellen (Datenqualität oft nicht ausreichend)
- keine unbearbeiteten Primärdaten
- Lizenzen erlauben kein TDM
- Datenschutz bei amtlichen Datensammlungen
- Datenmanagement: Datenformate und Datendokumentation
- Vollständigkeit nie erreichbar
- Für Ingenieurwissenschaften werden Industriedaten benötigt, die nicht verfügbar sind
- Texte liegen digital und in ausreichender Qualität vor, sind aber aus lizenzrechtlichen Gründen nicht benutzbar

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Mögliche Player aus Sicht der Wissenschaftler

- Forschungsförderorganisationen
- Universitäten (eher Bibliotheken als Zentralverwaltung)
- Lizenzverhandler (Bibliotheken, Allianzinitiative Digitale Information, Konsortien, …)
- Fachgesellschaften

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Stand Text und Data Mining für **Texte**

- Player sind neben Wissenschaftlern, Bibliotheken, Forschungsförderern auch Verlage
- TDM für Texte muss lizenziert werden.
- Großbritannien hat Recht zu TDM in UrhG aufgenommen, aber nur nicht-kommerziell
- Stand: TDM-Ausnahme in UrhG oder auf europäischer Ebene?
- Verlage bieten zT TDM zu nichtkomm. Zwecken an:
  Elsevier und CrossRef
  Stellungnahme von LIBER zur TDM-Vertrag von Elsevier
- Open Access!

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

TDM für **Forschungsdaten**: ToDo´s

- Player können neben Wissenschaftlern, Forschungsförderern auch Bibliotheken sein, noch nicht Verlage
- Forschungsdatenmanagement: Datenqualität, Datenformate, Lizenzen, Metadaten
- Mehr Offenheit: Rohdaten, offene Formate (keine Bilddateien), offene Lizenzen
- Policies – gute wissenschaftliche Praxis

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de

Herausforderungen für Bibliotheken

- **Infrastrukturen** für Publikation von Daten und Texten (Repositories)
- **Verfügbarkeit von Texten** für TDM (geeignete Lizenzen und Formate, Metadaten, Maschinenlesbarkeit)
- **Publikationsberatung** für Forschungsdaten und Texte: Metadaten, offene Lizenzen und Datenformate, Maschinenlesbarkeit (!), Primärdaten

CC by 4.0 International https://creativecommons.org/licenses/by/4.0/deed.de