Bedarf und Anforderungen an Ressourcen für Text und Data Mining

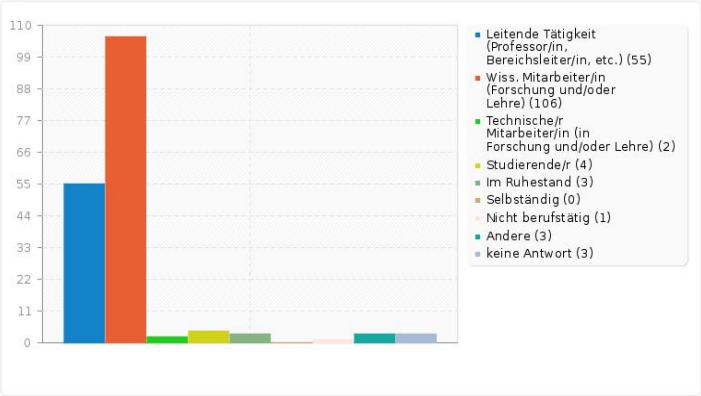
Zusammenfassung der Ergebnisse einer Umfrage aus dem Zeitraum April bis Mai 2015

Durchgeführt von der Schwerpunktinitiative
„Digitale Information“
der Allianz der deutschen Wissenschaftsorganisationen

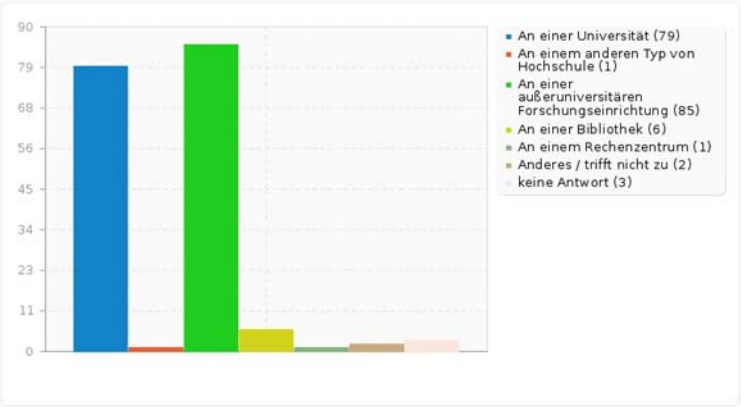
Arbeitsgruppe „Text and Data Mining“ (Katerbow, Mittermaier, Sens, Schöch)
http://www.allianzinitiative.de/handlungsfelder/querschnittsthemen/nutzungsrechte/ad-hoc-arbeitsgruppe.html

Teil A: Befragte Personen (N = 177)

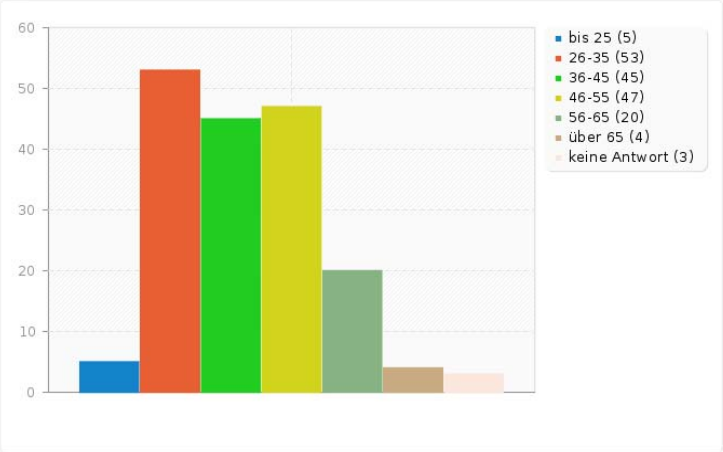
1. Beruflicher Status



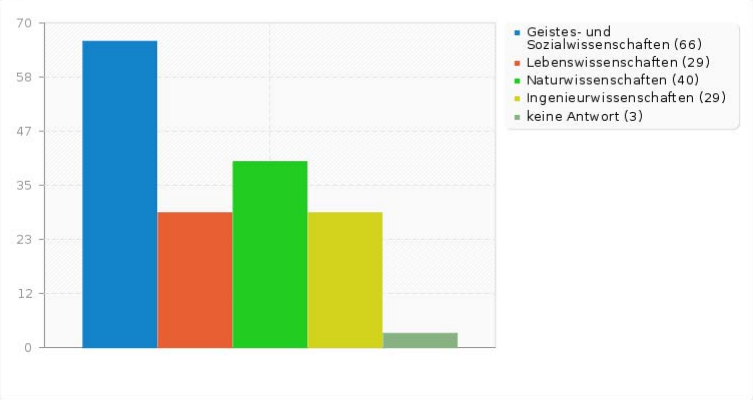
2. Zugehörigkeit



3. Alter

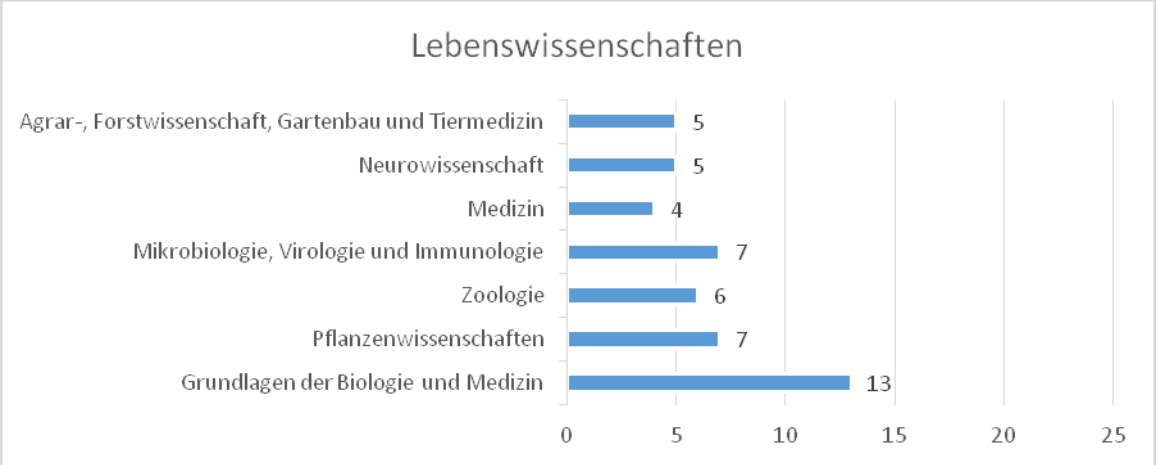


4. Wissenschaftsbereiche und Fächer

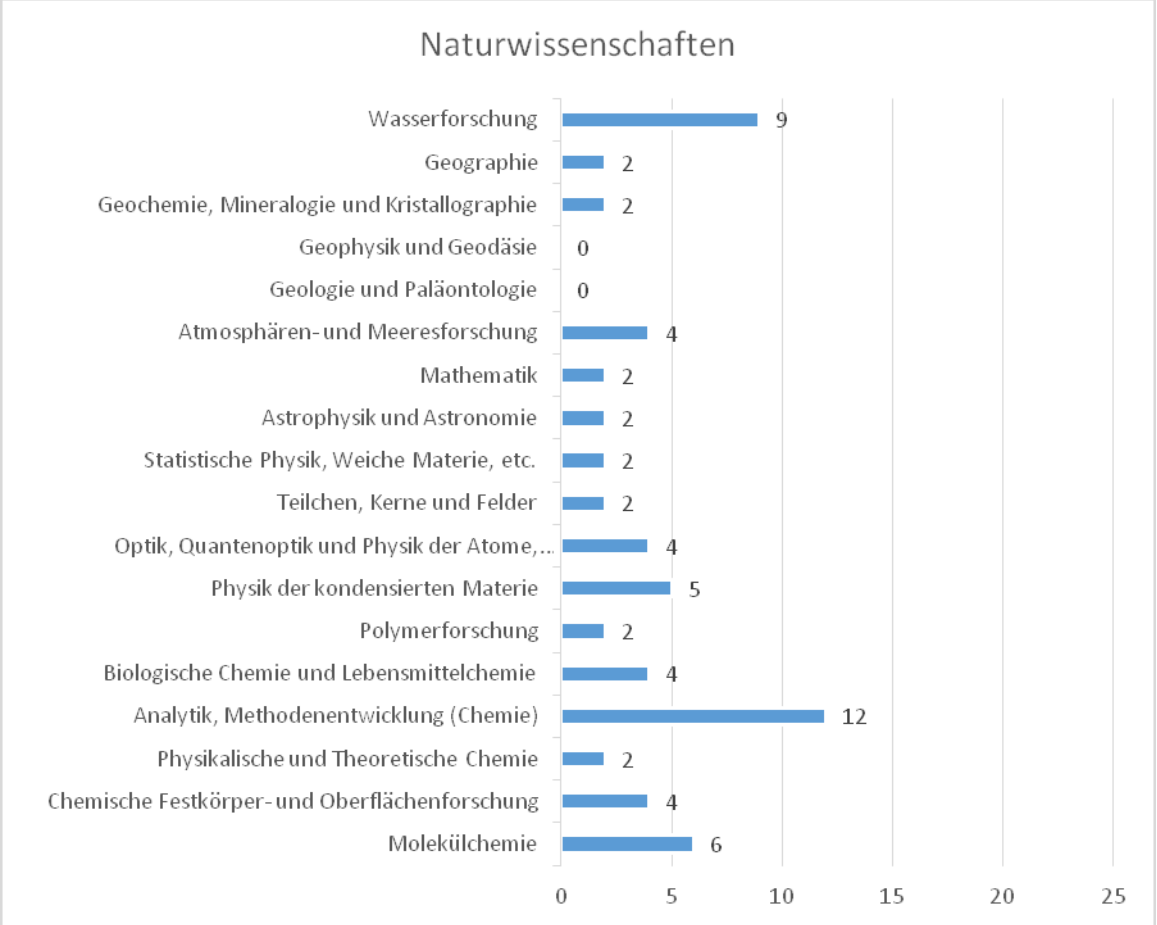


4.a. Geistes- und Sozialwissenschaften (N = 66), aber Mehrfachnennungen

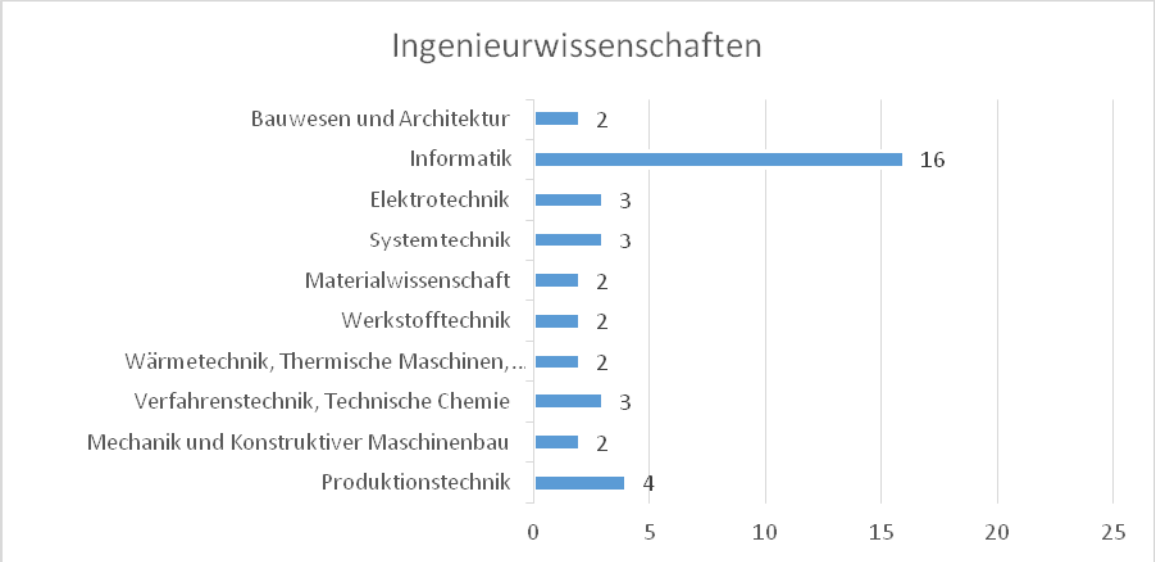
4.b. Lebenswissenschaften (N = 29), aber Mehrfachnennungen



4.c. Naturwissenschaften (N = 40), aber Mehrfachnennungen



4.d. Ingenieurwissenschaften (N = 29), aber Mehrfachnennungen



5. Forschungsformen, mit denen die Befragten arbeiten

Skala: 1 = nicht zutreffend, 2, 3, 4, 5 = sehr stark zutreffend

	1	2	3	4	5	k.A.
Experimentierend	50	19	22	24	48	14
Beobachtend	19	25	27	52	37	17
Hermeneutisch-interpretierend	45	25	15	29	36	27
Begrifflich-theoretisch	42	22	22	34	37	20
Gestaltend	62	27	21	22	11	34
simulierend	45	26	33	29	19	25

Andere Forschungsformen, die genannt wurden:

- partizipatorische
- statistische
- modellierende
- quantitative
- analytisch,
- explorierend
- qualitativ und quantitativ empirisch
- modellierend und ontologisch
- vergleichende
- entwickelnd
- administrative
- strategische

Teil B: Fragestellungen, die mit TDM bearbeitet werden

6. Welche Fragestellungen bearbeiten Sie mit TDM? (Freitextfeld)

Diese Frage haben 80% der Befragten beantwortet, allerdings lautet die Antwort mehrfach auch “keine”. Das Spektrum der genannten Fragestellungen ist sehr breit, eine Zusammenfassung problematisch. Im Folgenden sind eine Reihe von wiederkehrenden Antworten genannt, die eine Fragestellung oder ein Forschungsthema benennen:

Inhaltliche, linguistische und stilistische Analysen von Texten

- Fragen im Rahmen einer korpusgestützten Romangeschichte
- Text/Bild-Anteile in Zeitschriften
- Automatische, unüberwachte Taxonomiebildung auf Basis von Textkorpora und automatische Verschlagwortung.
- Transnationale politische und kulturelle Kommunikation / Berichterstattung im Länder- und im Zeitvergleich; Inhaltsanalysen für verschiedene politikwissenschaftliche Fragestellungen; Extraktion von Fakten aus biomedizinischen Texten
- Begriffsgeschichtliche Fragestellungen; Wie werden spezifische Begriffe in der Bildungsforschung (auch historisch genutzt)?
- Bedeutung von grammatischen Strukturen in bestimmten Zeiträumen, Bedeutungswandel über Epochen hinweg; Extraktion von Grammatiken
- Automatische Erkennung von Urheberschaft; Textähnlichkeitsberechung (Clustering, Similarity, Projection)
- Extraktion von wesentlichen Konzepten als Aufbereitung für intelligente (semantische) Recherchesysteme.
- korpuslinguistisch-diskursanalytische und soziolinguistische
- Named Entity Recognition und Normalisierung zur Verknüpfung von Literaturwissen mit anderen Datenquellen
- Modellierung von Themen / semantischen Strukturen in historischen Textdatenbanken.
- Überprüfung von theoretisch-linguistischen Hypothesen
- Wodurch zeichnet sich die Fachwissenschaftssprache der Ingenieurwiss. aus?

Publikationswesen, Wissenschaftsgeschichte und -Soziologie, Kulturgeschichte, Politikwiss.

- Publikationsverhalten von Wissenschaftlern; Finden von Datenzitationen in wissenschaftlichen Publikationen; Netzwerkstrukturen (Geo/Personendaten) in Zeitschriften
- Bau von Recommendersystemen und Retrievalsystemen auf großen Textkorpora
- Wie entwickeln sich Teildisziplinen der Erziehungswissenschaft auf Basis der von Wissenschaftlern in ihren Publikationen genutzten Begriffe?
- Evaluierung von abgeschlossenen Projekten und Nutzung der Daten im Sinne von Langzeitstudien
- Analyse von Technologie Trends mit Data Mining in digitale Bibliotheken. Technologie- und Ökonomisierungsprozesse im Sprachgebrauch der politischen Öffentlichkeit; 1) Ermitteln Stand der Technik, 2) Ableiten globaler Forschungstrends, 3) Bewertung eigene Forschungsarbeiten; Erkennung emergenter (technologischer) Entwicklungen
- Erstellung von Metaanalysen, Artverbreitungsmodellen, großskalige Analyse von
- Eventdatensammlung aus News Items / Nachrichtenwebsites
- Kulturhistorische Fragestellungen auf Basis naturwissenschaftlicher Analytik.

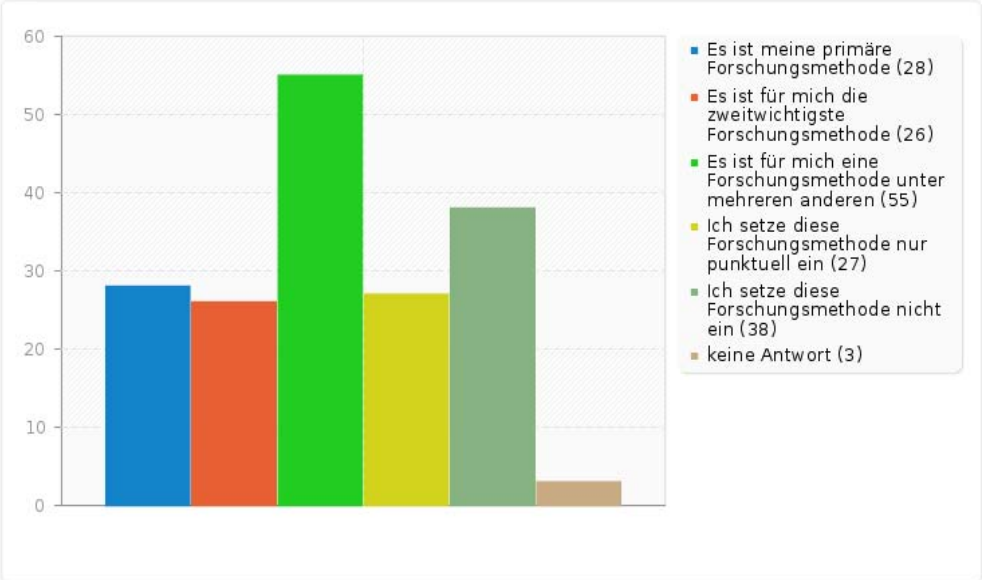
- Stabilität politischer Systeme; Konfliktwahrscheinlichkeit; Motive für Politikveränderungen

Naturwissenschaftliche Fragestellungen

- Suche nach Übereinstimmungen bei Sequenzen in Datenbanken
- Biologische Wirkung von Chemikalien
- Globale Auswirkungen von Klimawandel
- Recherche und Modellierungen von Zusammenhängen zwischen Genen-Krankheiten-Wirkstoffen
- Verständnis bio-geo-chemischer Zusammenhänge
- Informationsextraktion für Zwecke der Modellierung von Krankheiten
- Bioinformatische und systembiologische
- Validierung durch große experimentelle Datensätze
- Neue Hypothesen zur Ätiologie von Krankheiten
- Statistische Analyse von Umweltdaten
- Informationsextraktion zur Erstelleung biomed.Netzwerke und Wissensresourcen
- Struktur und Entwicklung von Sternhaufen
- Suche nach Kooperationspartnern
- Nukleare Daten von Actiniden
- Kernspektroskopische Daten
- Neutronenflußmessungen
- Tumorimmunolgie
- Forschung auf dem Gebiet Komplexe Systeme und Netzwerke, nichtlineare Dynamik und Kontrolle komplexer Netzwerke
- Wie wichtig ist welcher Faktor für die Wirtschaftlichkeit von Fabriken?
- Physiologie, Umweltforschung, Genomanalytik

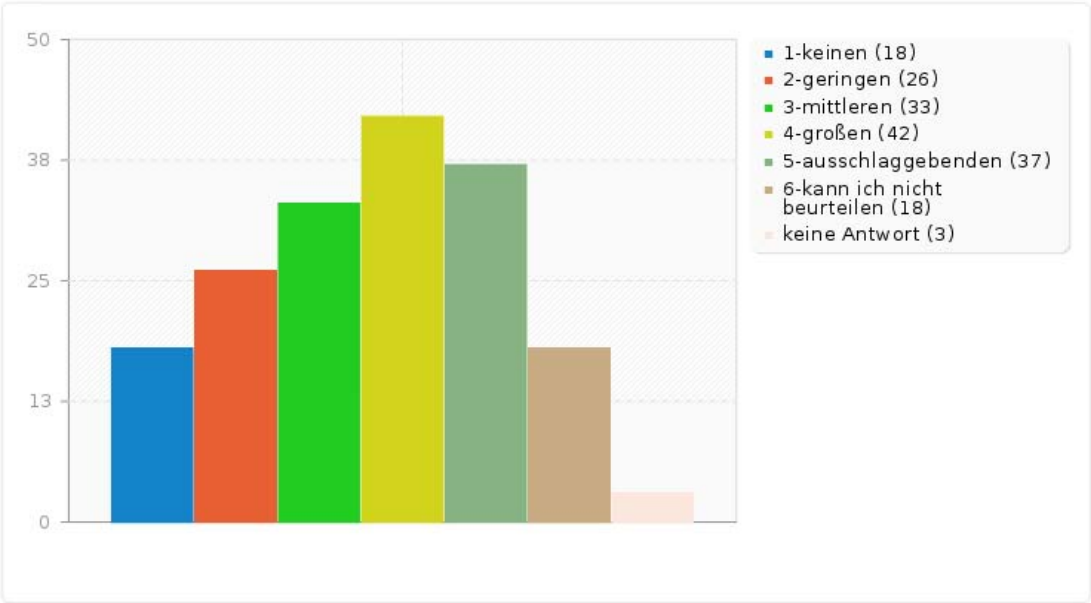
7. Welchen Stellenwert hat TDM für Ihre persönliche Forschung?

Für eine deutliche Mehrheit (55 Antworten) ist TDM “eine Forschungsmethode unter anderen”, wird also ergänzend eingesetzt.



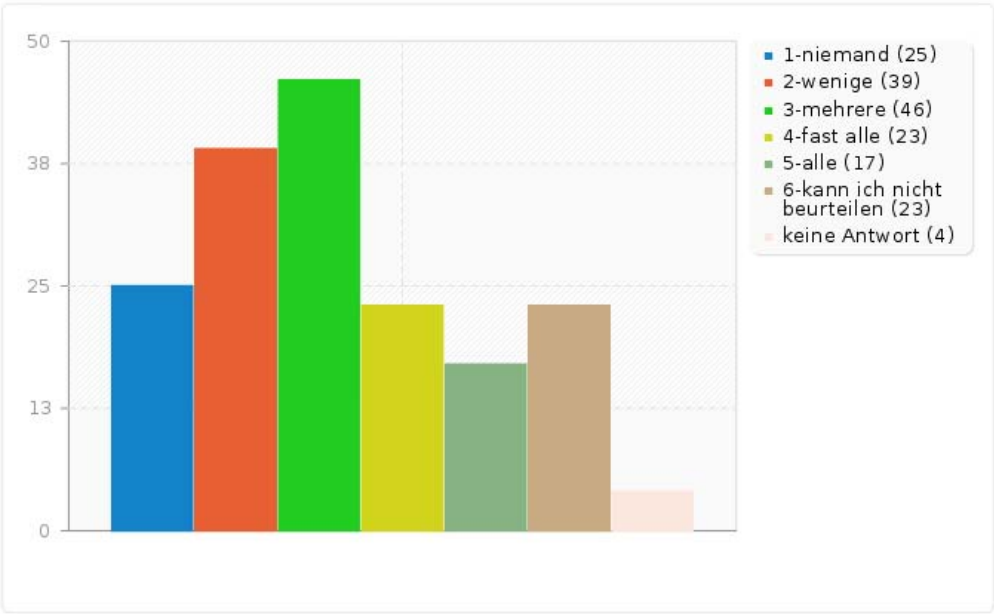
8. Welchen Nutzen hat TDM in Ihrer Forschungspraxis?

Für eine Mehrheit der Befragten hat TDM einen großen (42), sogar ausschlaggebenden (37) oder zumindest mittleren (33) Nutzen.



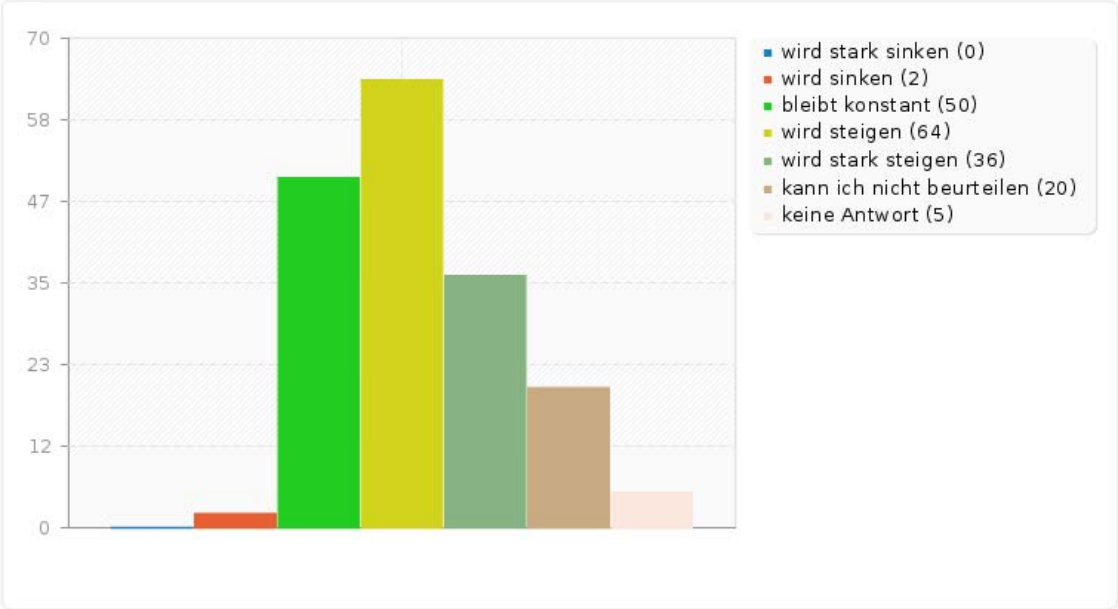
9. Welcher Anteil der KollegInnen in Ihrer Arbeitsgruppe (falls nicht zutreffend: an Ihrem Institut) praktiziert TDM?

Für die Mehrheit der Befragten gilt, dass in der eigenen Arbeitsgruppe mehrere (46) oder wenige (39) KollegInnen TDM praktizieren, deutlich seltener fast alle (23) oder alle (17), seltener aber auch niemand (25).



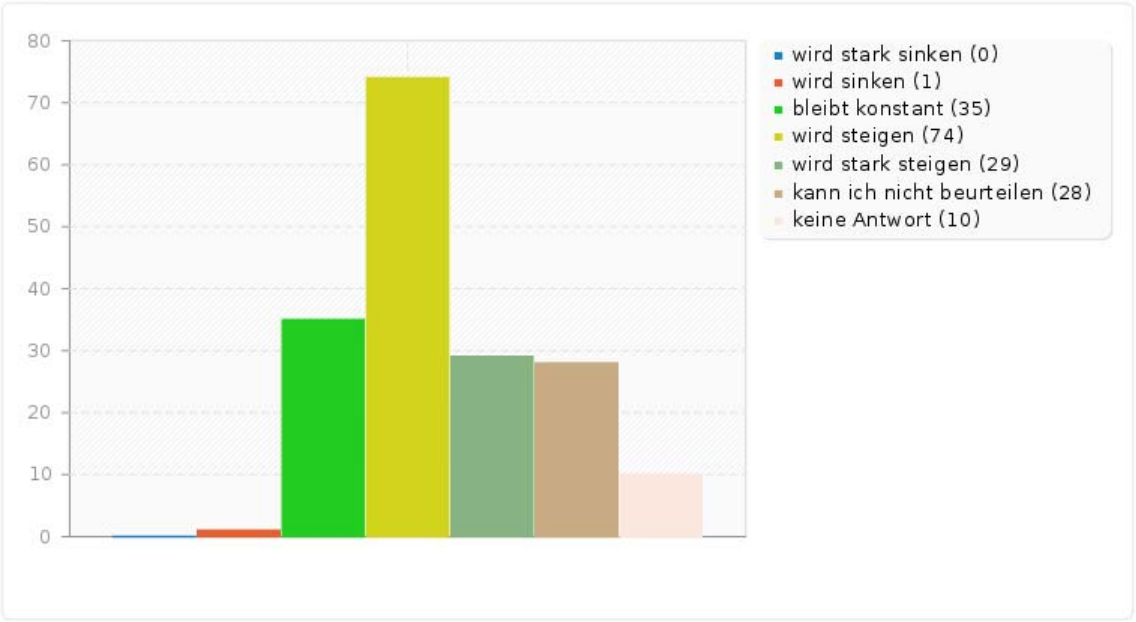
10. Wie wird sich dieser Anteil in den nächsten drei Jahren vorraussichtlich verändern? [für Sie persönlich]

Eine sehr sehr deutliche Mehrheit der Befragten (74) drückt die Erwartung aus, dass in Zukunft der Anteil der TDM an der eigenen Tätigkeit größer werden wird.



11. Wie wird sich dieser Anteil in den nächsten drei Jahren vorraussichtlich verändern? [für Ihre Arbeitsgruppe oder Ihr Institut]

Eine deutliche Mehrheit der Befragten drückt die Erwartung aus, dass in Zukunft der Anteil der TDM betreibenden KollegInnen größer werden wird (64), viele meinen aber auch, dass der Anteil konstant bleiben wird (50).



Teil C

12. Bitte nennen Sie die drei für Sie wichtigsten Quellen von Inhalten (Anbieter, Plattform, Webseiten, etc.) für TDM?

Antworten von 68% der Befragten, die Übrigen haben keine Antwort gegeben. Breites Spektrum.

Arten von Ressourcen mit Beispielen

- Sehr viele Datenbanken für wissenschaftliche Zeitungsartikel, teils Metadaten und Abstracts oder OpenAccess-Angebote, großenteils aber lizenzpflichtige Volltexte: arXiv, DBLP, ssoar.info, ACM DL, pubmed, Factiva, Mediaperspektiven, LexisNexis, SciFinder, Coremine, Genios, OCMiner, Econstor, SAO/NASA ADS.
- Sehr viele naturwissenschaftliche Faktendatenbanken: Reaxys (Chemie), NIST (Chemie), ENA/EMBL (Nucleotide), VizieR (Astronomie), BrainMap, LAILAPS (Biologie), NCBI (DNA, RNA, Proteine), ChemSpider (Chemie), UniProt (Proteine), METLIN (Biologie), IHOP (Proteine), RCSB (Proteine), KMASKER (Genomsequenzen), Chembldb (Molekküldatenbank).
- Einige Datenbanken mit literarischen Texten: TextGrid, Gutenberg, Deutsches Textarchiv, Internet Archive, zeno.org, Ctext (chinesische Texte).
- Einige Textressourcen und Wörterbücher etc. verschiedener Art: Logos Bible Software, Accordance Bible Software, Biblioteca Teubneriana Online, Thesaurus Linguae Graecae, Responsa Project: Global Jewish Database, Wibilex Bibellexikon, etc.
- Mehrere Messinfrastrukturen mit Datenangeboten: Fluxnet (Meteorologie), GAVP (Astrophysik), SOLIS (Sonnenbeobachtung), CIFOR (CO2-Messungsdatenbank)
- Einzeln genannte, andere Datenanbieter / Datentypen: Destatis (Statistisches Bundesamt), Arachne (Archäologische Objektdatenbank), da-ra.de / GESIS (sozialwiss. Forschunsdaten).

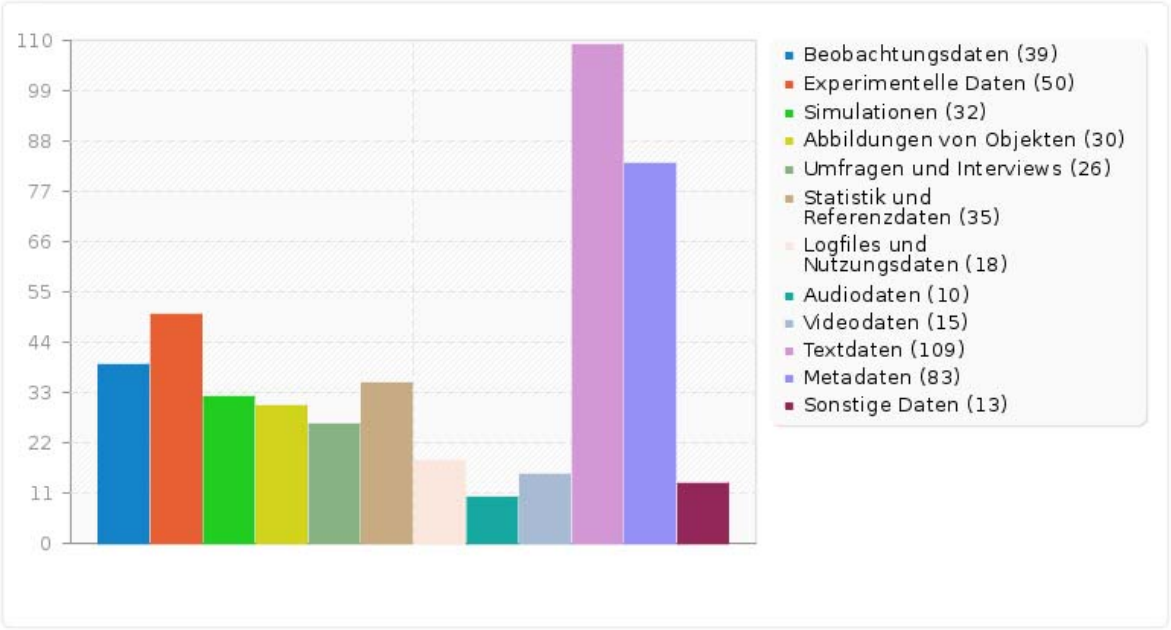
Nach Kategorien

26 mal wurde eine Quelle genannt, 57 mal zwei Quellen und 18mal drei Quellen. Die in den Antwort oft (aber nicht immer) exakt bezeichneten Quellen (z.B. Web of Science; GoogleScholar) lassen sich wie folgt kategorisieren:

Kategorie	Anzahl Nennungen
Datenbank	56
Webportal	33
Suchmaschine	24
verschiedene Textarten	15
Repositorien	14
Verlage	13
Computerprogramm	11
Internet	10
Bibliothek	9
Zeitung	4
sonstiges/unbekannt/unklar	20

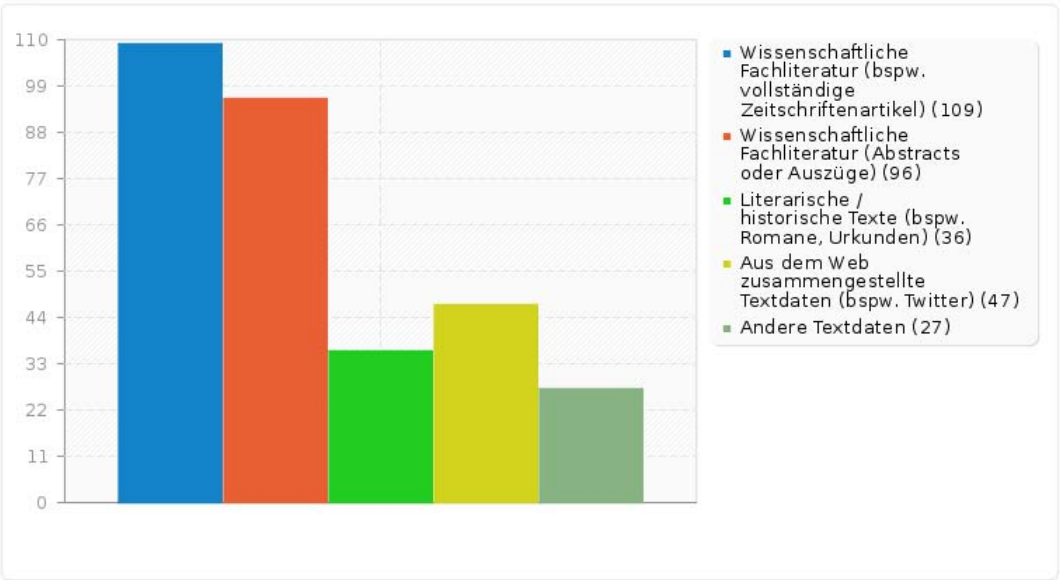
13. Welche Arten von Daten nutzen Sie für TDM?

Deutlich am stärksten genutzt werden Textdaten und Metadaten, gefolgt von Experimentellen Daten und Beobachtungsdaten sowie Statistik- und Referenzdaten. Die hier gar nicht angebotenen Bilddaten werden unter sonstige Daten einmal genannt.



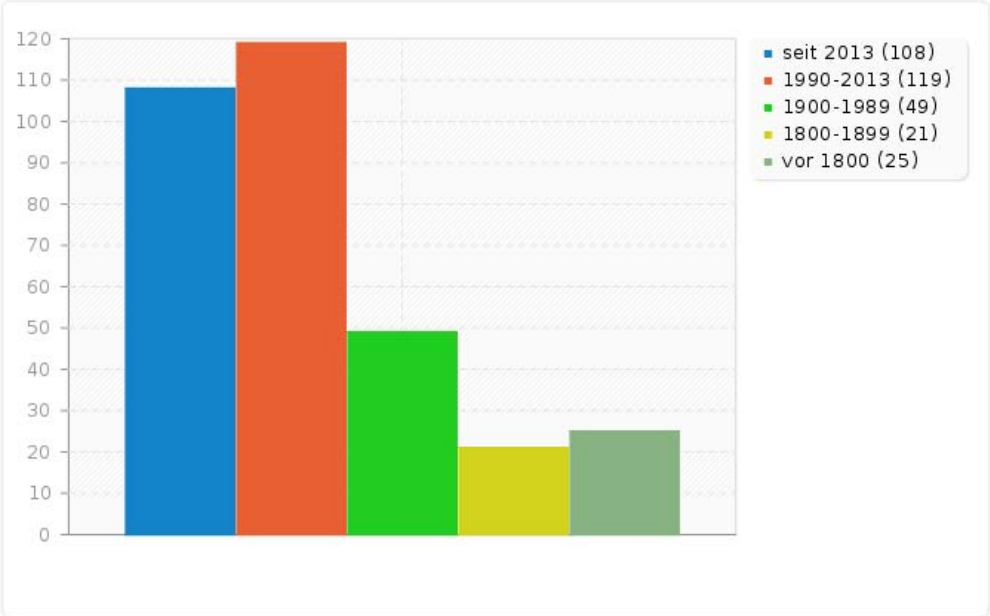
14. Welche Arten von Texten nutzen Sie für TDM?

Fokus auf die viel verwendeten Textdaten: Die Analyse von wissenschaftlicher Fachliteratur überwiegt die Analyse literarischer Text oder historischer Dokumente sowie die von Webdaten deutlich. Bei der Auswertung von wiss. Fachliteratur werden vollständige Texte etwas häufiger analysiert als nur die Abstracts oder Auszüge. Beides zeigt die Bedeutung von in der Regel lizenzpflichtigen Inhalten für das TDM. Dies gilt umso mehr, wenn man die zeitliche Verteilung der Daten mit berücksichtigt (siehe Frage 3.3).



15. In welcher Zeit sind die Daten, die Sie analysieren, ursprünglich entstanden?

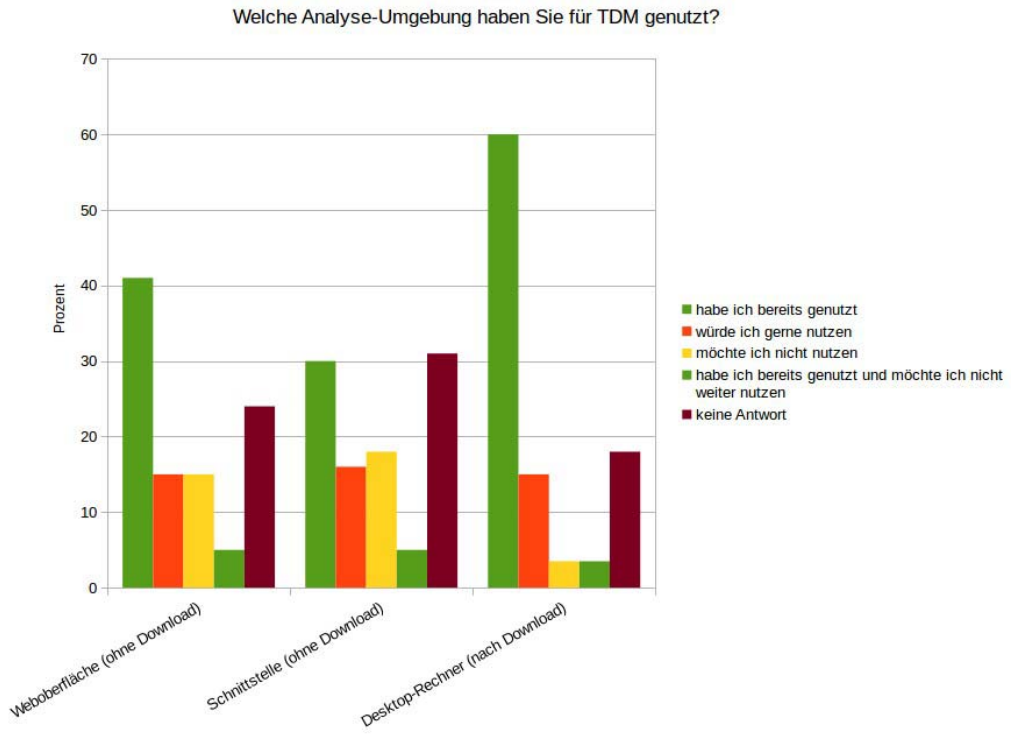
Die ganz überwiegende Mehrheit der Daten, die von den Befragten analysiert werden sind sehr aktuell (seit 2013), recht aktuell (1990-2013) oder (etwas seltener) noch aus einem Zeitrahmen, in dem bei urheberrechtsrelevanten Materialien der Urheberrechtschutz überwiegend noch greift (nach 1900). Nur ein relativ geringer Anteil der Befragten analysiert Daten, die in aller Regel gemeinfrei sein sollten, sofern nicht Datenbankschutzrechte vorliegen (vor 1900; insgesamt 46 Antworten). In Kombination mit dem großen Interesse an Volltexten ergibt sich hier ein großer Lizensierungsbedarf.



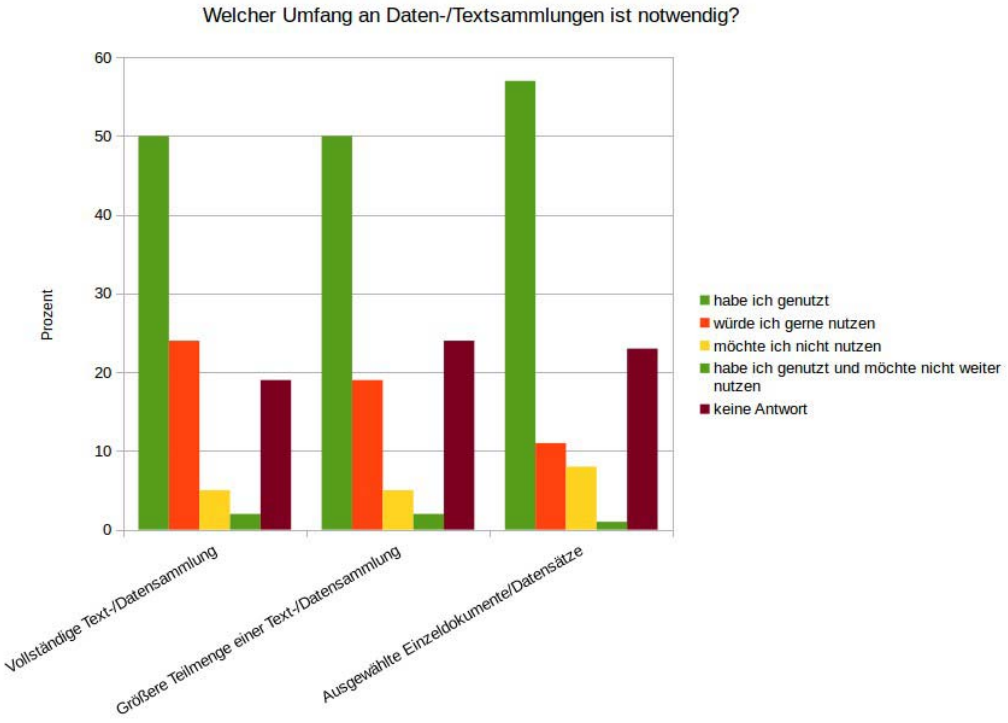
Teil D

16. Welche Analyse-Umgebung wurde bereits genutzt?

Weboberfläche ohne Download, Schnittstelle ohne Download, oder Analyse auf dem Desktop nach Download. Letztere Variante wird am stärksten bevorzugt, aber die beiden anderen werden auch häufig genutzt.

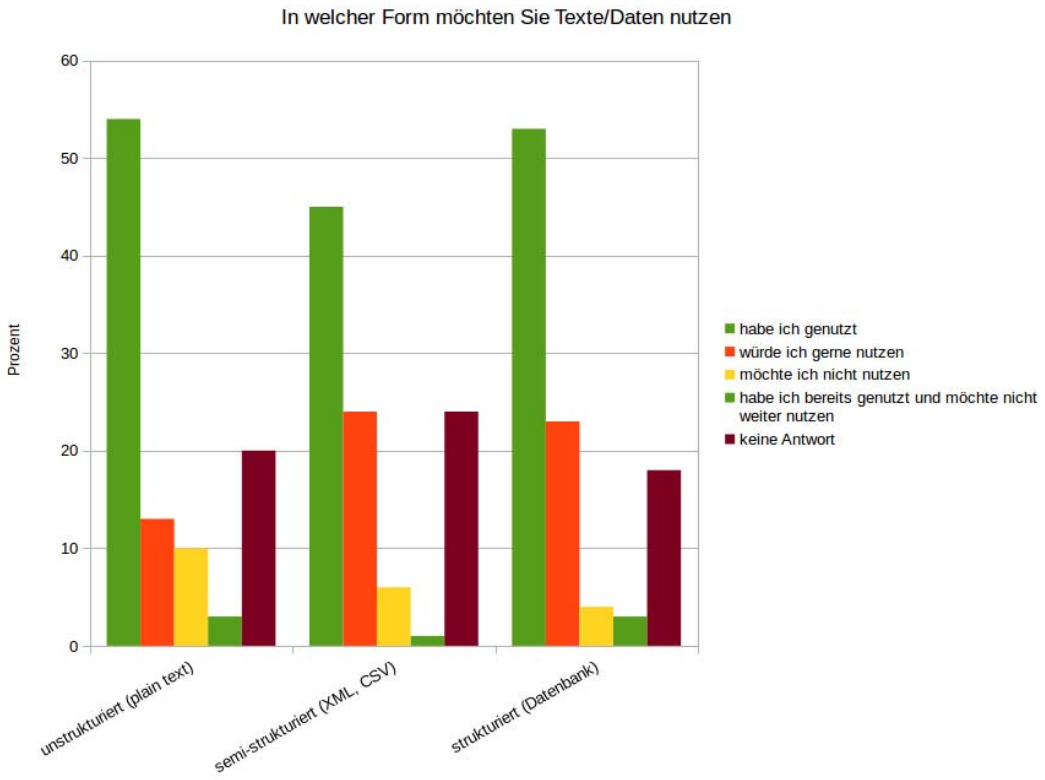


17. Welcher Umfang an Daten-/Textsammlungen entspricht Ihren Bedürfnissen?



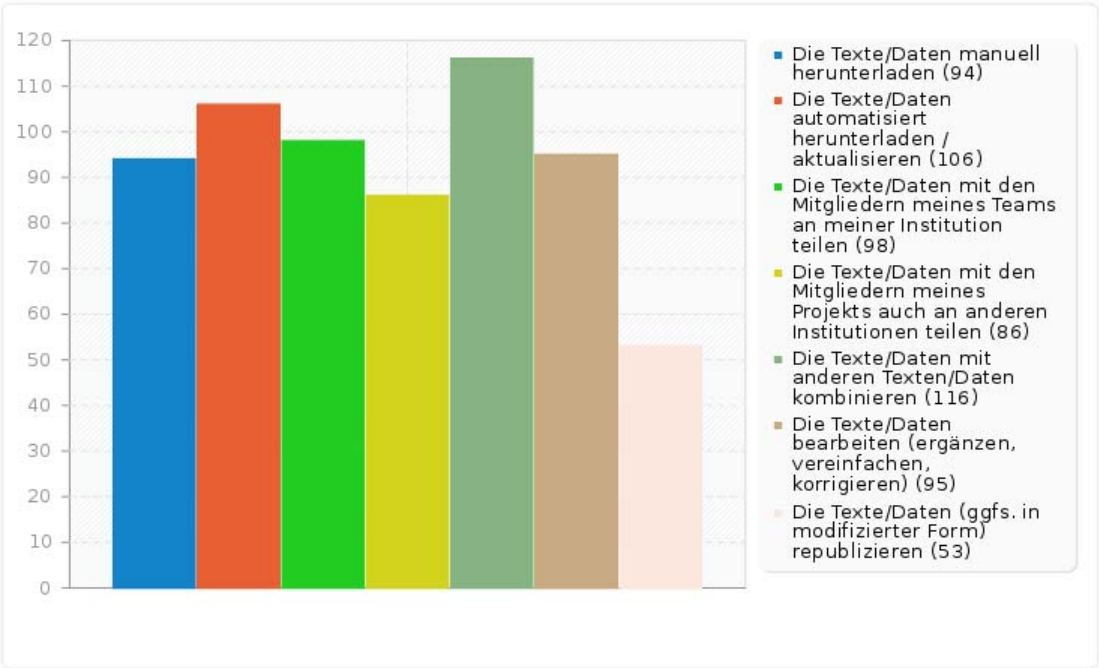
18. In welcher Form möchten Sie Texte und/oder Daten nutzen?

Keine sehr großen Unterschiede.



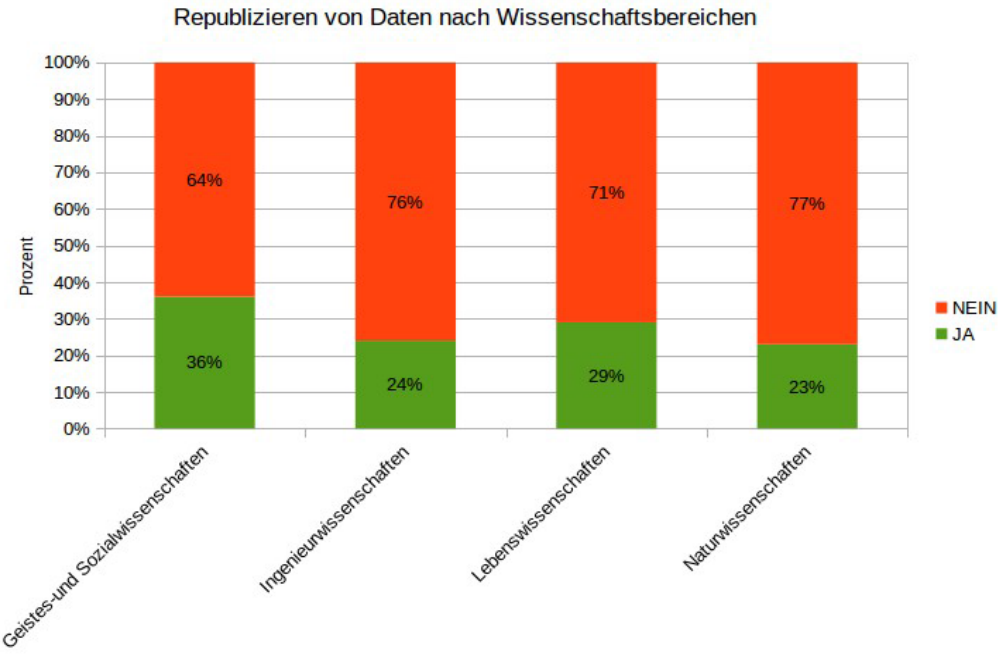
19. Welche der folgenden Nutzungsszenarien sind für Sie wichtig?

Alle genannten Szenarien sind sehr wichtig, insbesondere aber die Daten/Texte mit anderen kombinieren zu können und die Texte/Daten automatisiert herunterladen zu können. Auch alle anderen Szenarien werden von mindestens 80 Befragten als wichtig eingestuft. Am niedrigsten, aber mit 86 Antworten immer noch sehr hoch, liegt noch die Möglichkeit, die Texte/Daten auch über die eigene Institution hinweg teilen zu dürfen. Deutlich weniger nachgefragt ist die Möglichkeit, verwendete Daten zu republizieren.



20. Welche der folgenden Nutzungsszenarien sind für Sie wichtig? Die Texte republizieren (nach Wissenschaftsbereichen)

Es gibt zwischen den Wissenschaftsbereichen mittelstark ausgeprägte Unterschiede im Wunsch, verwendete Daten zu republizieren.



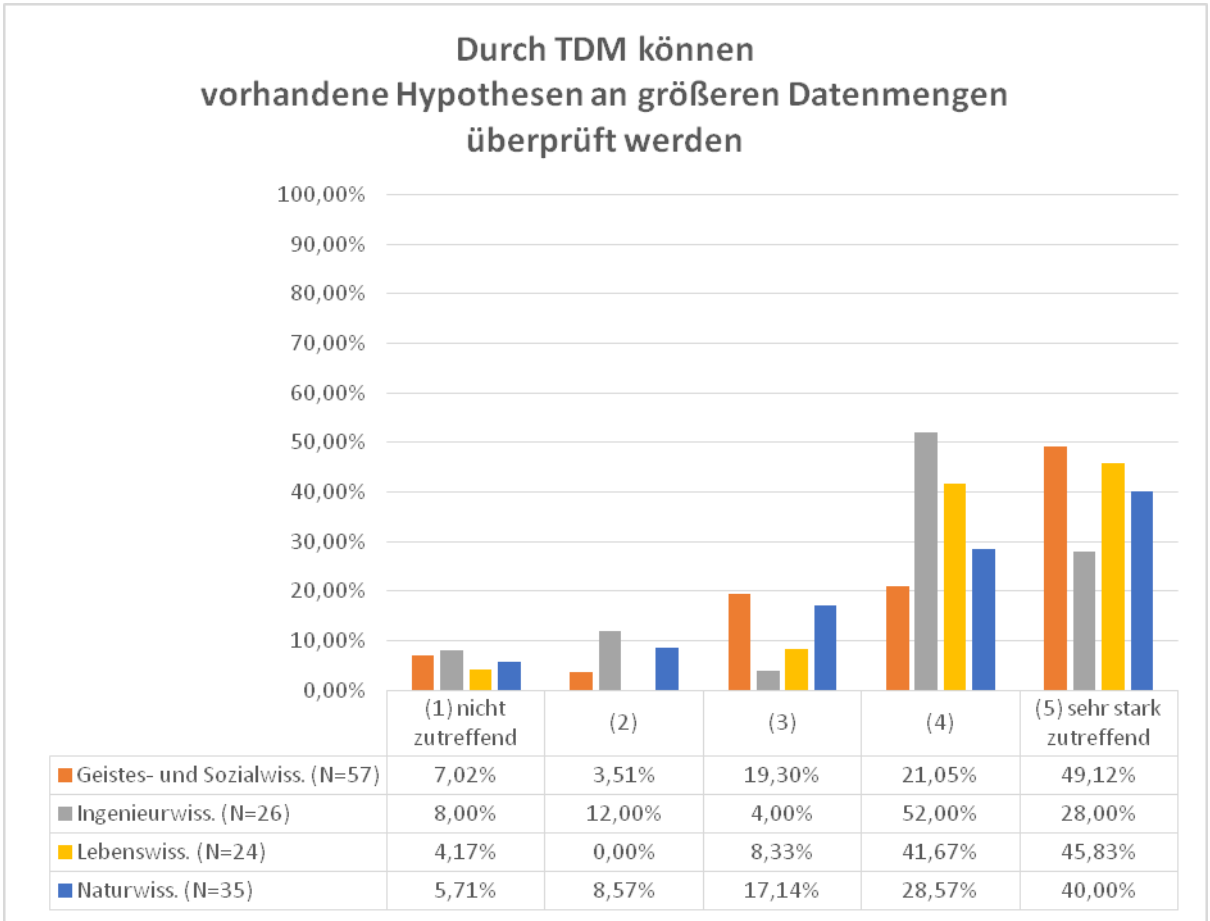
Teil E: Aktuelle Vorteile und Hindernisse von TDM

21. Vorteil 1: Durch TDM können vorhandene Hypothesen an größeren Datenmengen überprüft werden

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	57	26	24	35
(1) nicht zutreffend	11	4	2	1	2
(2)	8	2	3	0	3
(3)	23	11	1	2	6
(4)	46	12	13	10	10
(5) sehr stark zutreffend	63	28	7	11	14
keine Antwort	26	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



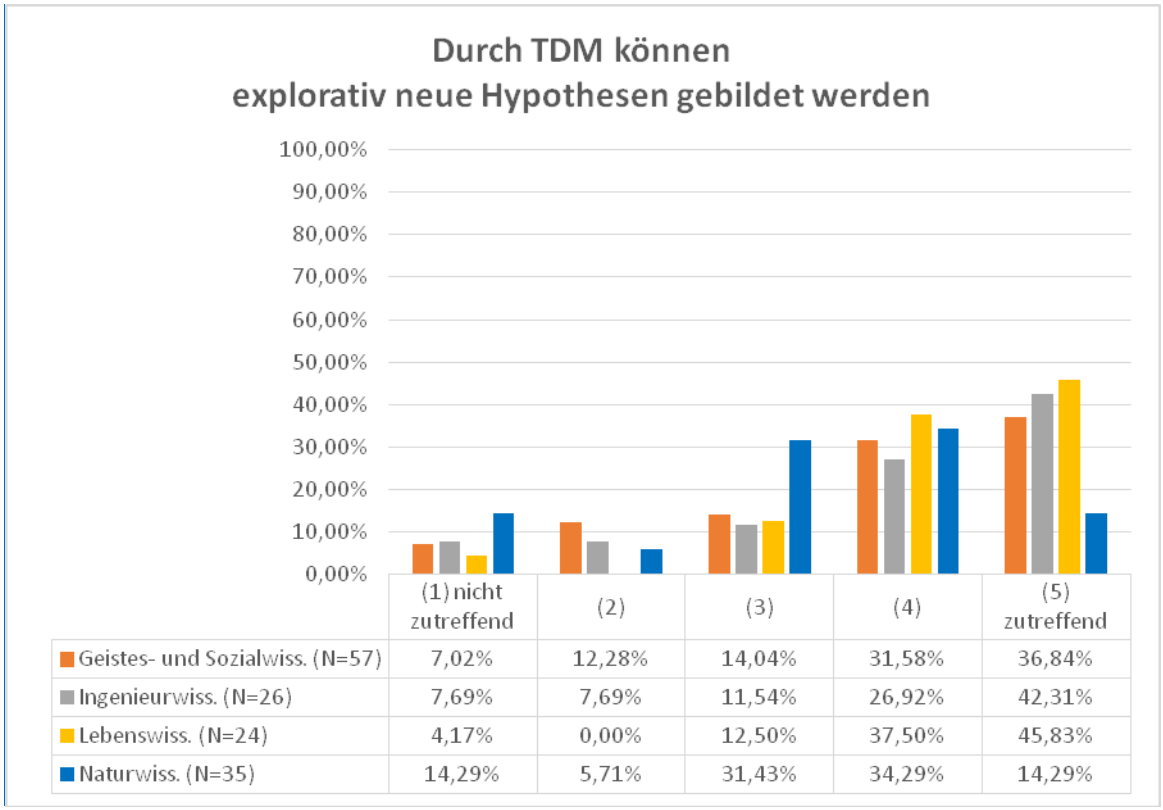
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=57)	7,02%	3,51%	19,30%	21,05%	49,12%
Ingenieurwiss. (N=26)	8,00%	12,00%	4,00%	52,00%	28,00%
Lebenswiss. (N=24)	4,17%	0,00%	8,33%	41,67%	45,83%
Naturwiss. (N=35)	5,71%	8,57%	17,14%	28,57%	40,00%

22. Vorteil 2: Durch TDM können explorativ neue Hypothesen gebildet werden

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	57	26	24	35
(1) nicht zutreffend	13	4	2	1	5
(2)	12	7	2	0	2
(3)	28	8	3	3	11
(4)	50	18	7	9	12
(5) sehr stark zutreffend	48	21	11	11	5
keine Antwort	26	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



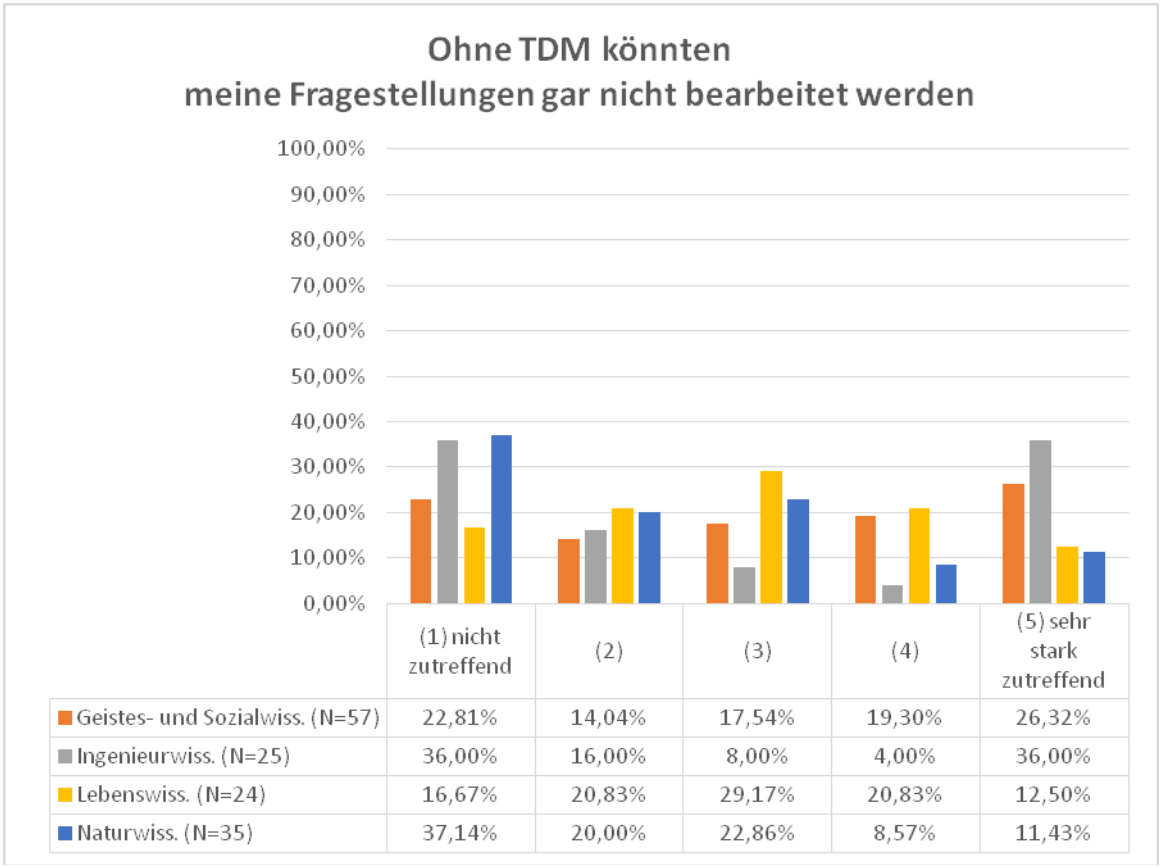
	(1) nicht zutreffend	(2)	(3)	(4)	(5) zutreffend
Geistes- und Sozialwiss. (N=57)	7,02%	12,28%	14,04%	31,58%	36,84%
Ingenieurwiss. (N=26)	7,69%	7,69%	11,54%	26,92%	42,31%
Lebenswiss. (N=24)	4,17%	0,00%	12,50%	37,50%	45,83%
Naturwiss. (N=35)	14,29%	5,71%	31,43%	34,29%	14,29%

23. Vorteil 3: Ohne TDM könnten meine Fragestellungen gar nicht bearbeitet werden

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	57	25	24	35
(1) nicht zutreffend	41	13	9	4	13
(2)	26	8	4	5	7
(3)	31	10	2	7	8
(4)	21	11	1	5	3
(5) sehr stark zutreffend	32	15	9	3	4
keine Antwort	26	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



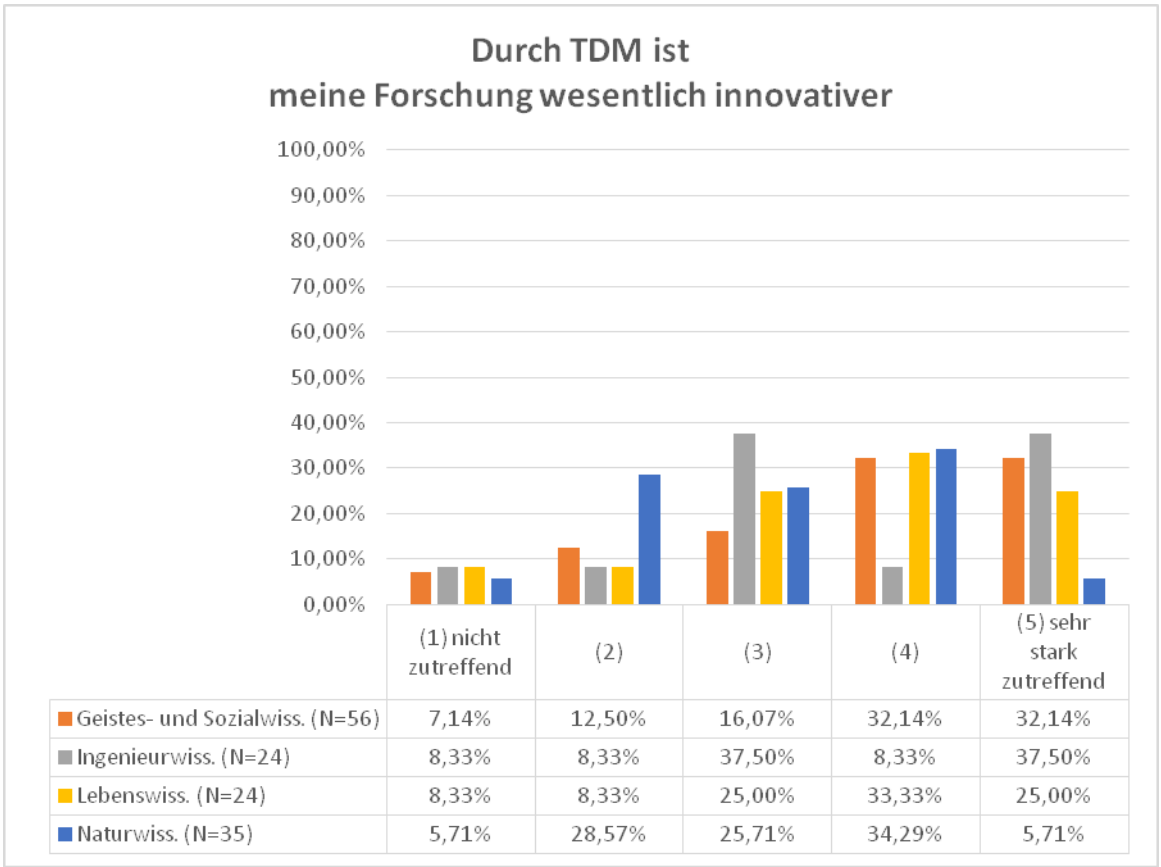
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=57)	22,81%	14,04%	17,54%	19,30%	26,32%
Ingenieurwiss. (N=25)	36,00%	16,00%	8,00%	4,00%	36,00%
Lebenswiss. (N=24)	16,67%	20,83%	29,17%	20,83%	12,50%
Naturwiss. (N=35)	37,14%	20,00%	22,86%	8,57%	11,43%

24. Vorteil 4: Durch TDM ist meine Forschung wesentlich innovativer

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	56	24	24	35
(1) nicht zutreffend	10	4	2	2	2
(2)	23	7	2	2	10
(3)	36	9	9	6	9
(4)	43	18	2	8	12
(5) sehr stark zutreffend	36	18	9	6	2
keine Antwort	29	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=56)	7,14%	12,50%	16,07%	32,14%	32,14%
Ingenieurwiss. (N=24)	8,33%	8,33%	37,50%	8,33%	37,50%
Lebenswiss. (N=24)	8,33%	8,33%	25,00%	33,33%	25,00%
Naturwiss. (N=35)	5,71%	28,57%	25,71%	34,29%	5,71%

25. Weitere Vorteile von TDM

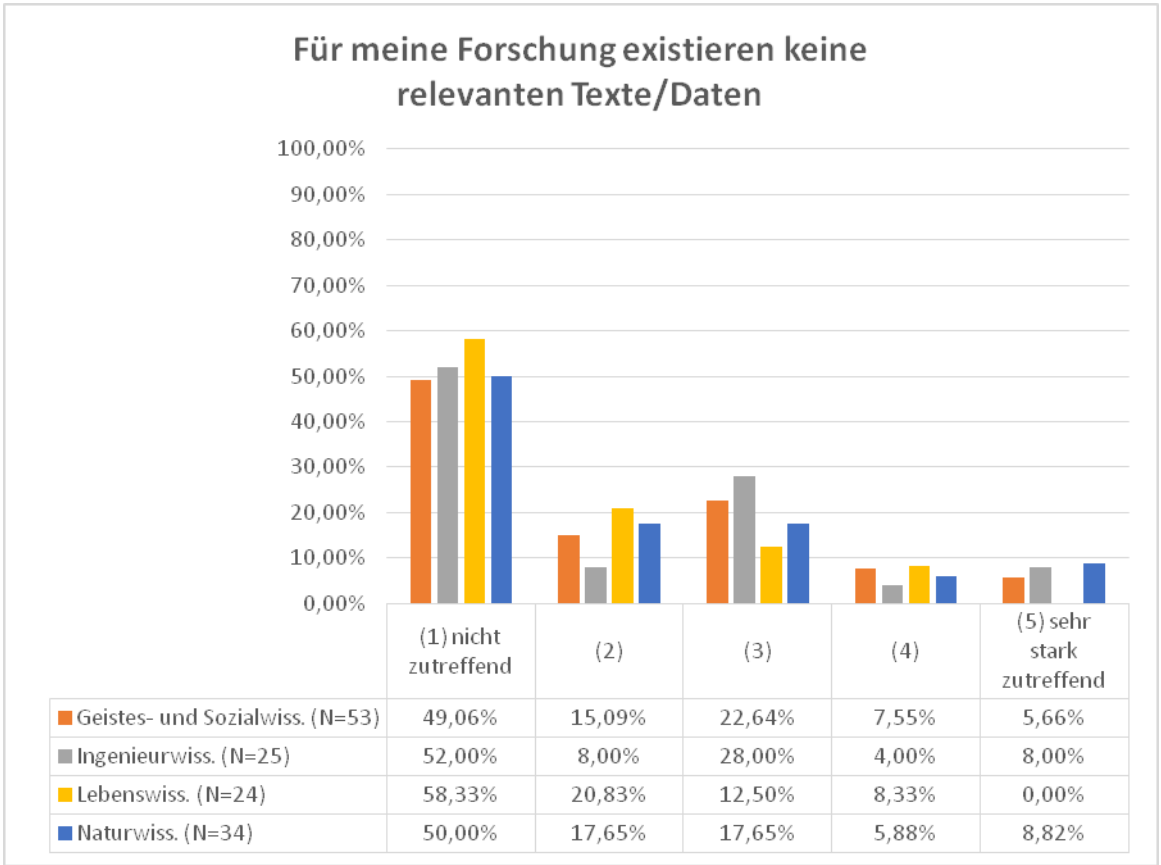
- höhere Replizierbarkeit der Forschung
- Nachnutzung der Daten für andere Forschungsfragen
- Tätigkeit im Ausland, daher Bibliotheksbeschaffung andere Prioritäten bzw. nicht in einem Leben machbar
- Metadatenanalyse - also die Zusammenschau mehrerer Studien zu einem Thema und die gemeinsame Analyse aller dieser Ergebnisse
- Ich erforsche und entwickle neue TDM Lösungen.
- Serendipity-Effekte werden bei der Auswertung möglich
- Produktivitätssteigerung
- Die Datenflut kontrollieren
- Die Originale digitaler Texte bzw. Textabbildungen (Rara) kann ich sonst nur in Bibliotheksräumen nutzen.
- Analyse von Marktrends
- Ohne Literaturdatenbanken wäre produktive Forschung nicht möglich.
- schnelles und einfaches Auffinden von Daten/Texten/Inhalten
- Kostenersparnis bei der Existenz-überprüfung von Objektdetails ohne eine Reise machen zu müssen
- TDM erleichtert mir die Suche nach spezifischen Inhalten in großen Textmengen. Es erlaubt darüber hinaus quantitative Aussagen z.B. über die Häufigkeit von Begriffsverwendungen u.ä.
- Die Debatten der gegenwärtigen Frontstellung zwischen innovativem und konservativem Umgang mit neuen Technologien sowie zwischen den mit Geisteswissenschaften und Naturwissenschaften verknüpften epistemologischen Grundannahmen liefert mir soziologisches und wissenschaftspolitisches Anschauungsmaterial über meine in der vormodernen Wissenschaftsgeschichte angesiedelte Forschungsfrage.
- schnellere, einfachere Auswertung/Bearbeitung von Texten/Daten
- Bei Ausfall von Datenquellen - sei es nun willkürlich oder ungewollt - stehen die Informationen trotzdem über die eigene Datenbank zur Verfügung
- Durch TDM kann ich Forschungstrends erkennen und formulieren

26. Hindernis 1: Für meine Forschung existieren keine relevanten Texte/Daten

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	53	25	24	34
(1) nicht zutreffend	76	26	13	14	17
(2)	22	8	2	5	6
(3)	30	12	7	3	6
(4)	9	4	1	2	2
(5) sehr stark zutreffend	9	3	2	0	3
keine Antwort	31	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



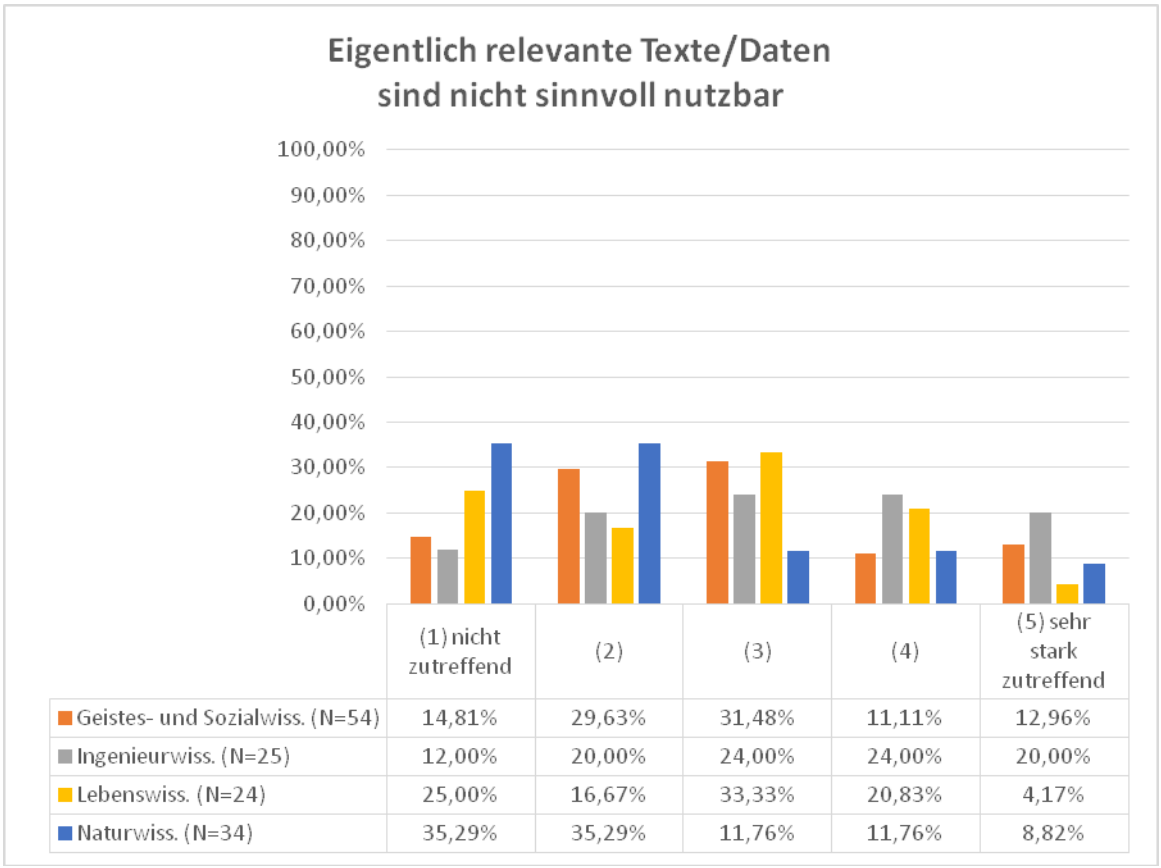
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=53)	49,06%	15,09%	22,64%	7,55%	5,66%
Ingenieurwiss. (N=25)	52,00%	8,00%	28,00%	4,00%	8,00%
Lebenswiss. (N=24)	58,33%	20,83%	12,50%	8,33%	0,00%
Naturwiss. (N=34)	50,00%	17,65%	17,65%	5,88%	8,82%

27. Hindernis 2: Eigentlich relevante Texte/Daten sind nicht sinnvoll nutzbar

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	54	25	24	35
(1) nicht zutreffend	32	8	3	6	12
(2)	39	16	5	4	12
(3)	37	17	6	8	4
(4)	23	6	6	5	4
(5) sehr stark zutreffend	17	7	5	1	3
keine Antwort	29	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



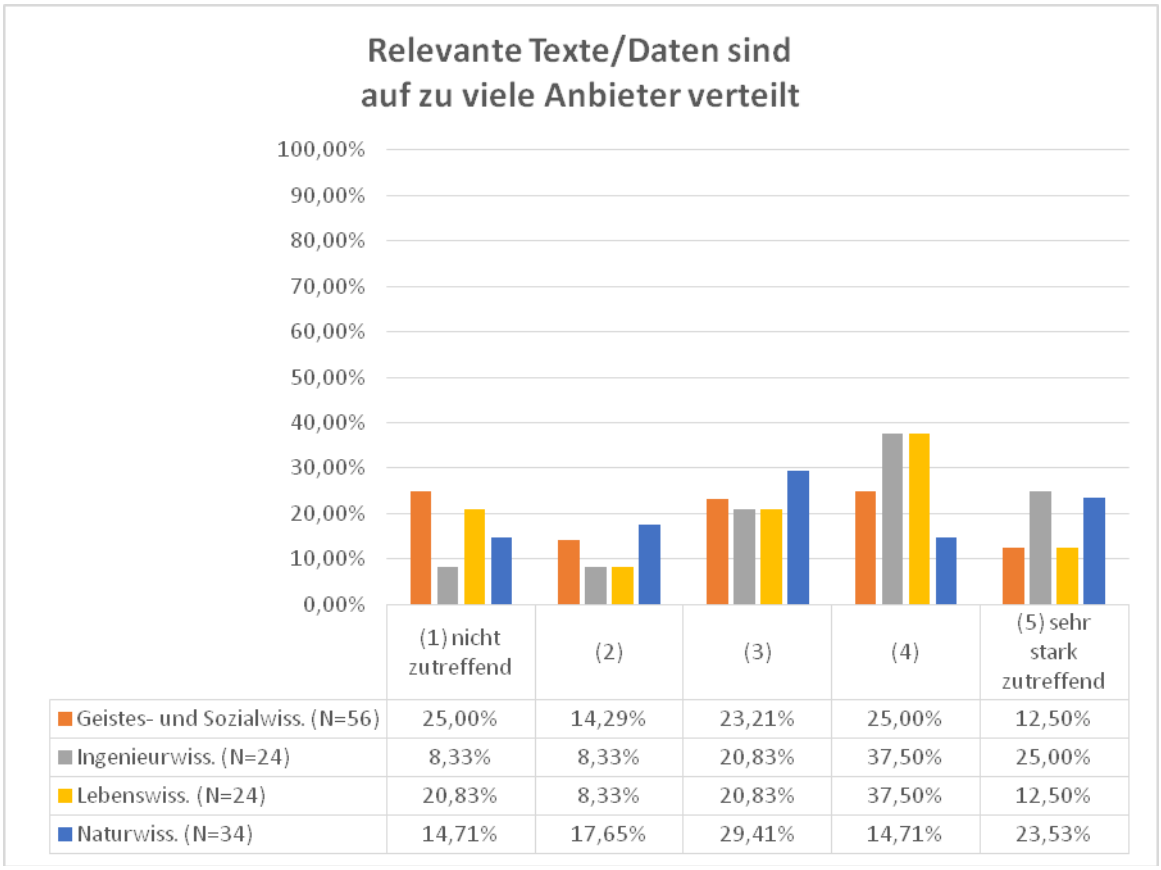
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=54)	14,81%	29,63%	31,48%	11,11%	12,96%
Ingenieurwiss. (N=25)	12,00%	20,00%	24,00%	24,00%	20,00%
Lebenswiss. (N=24)	25,00%	16,67%	33,33%	20,83%	4,17%
Naturwiss. (N=34)	35,29%	35,29%	11,76%	11,76%	8,82%

28. Hindernis 3: Relevante Texte/Daten sind auf zu viele Anbieter verteilt

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	56	24	24	34
(1) nicht zutreffend	27	14	2	5	5
(2)	19	8	2	2	6
(3)	34	13	5	5	10
(4)	41	14	9	9	5
(5) sehr stark zutreffend	27	7	6	3	8
keine Antwort	29	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



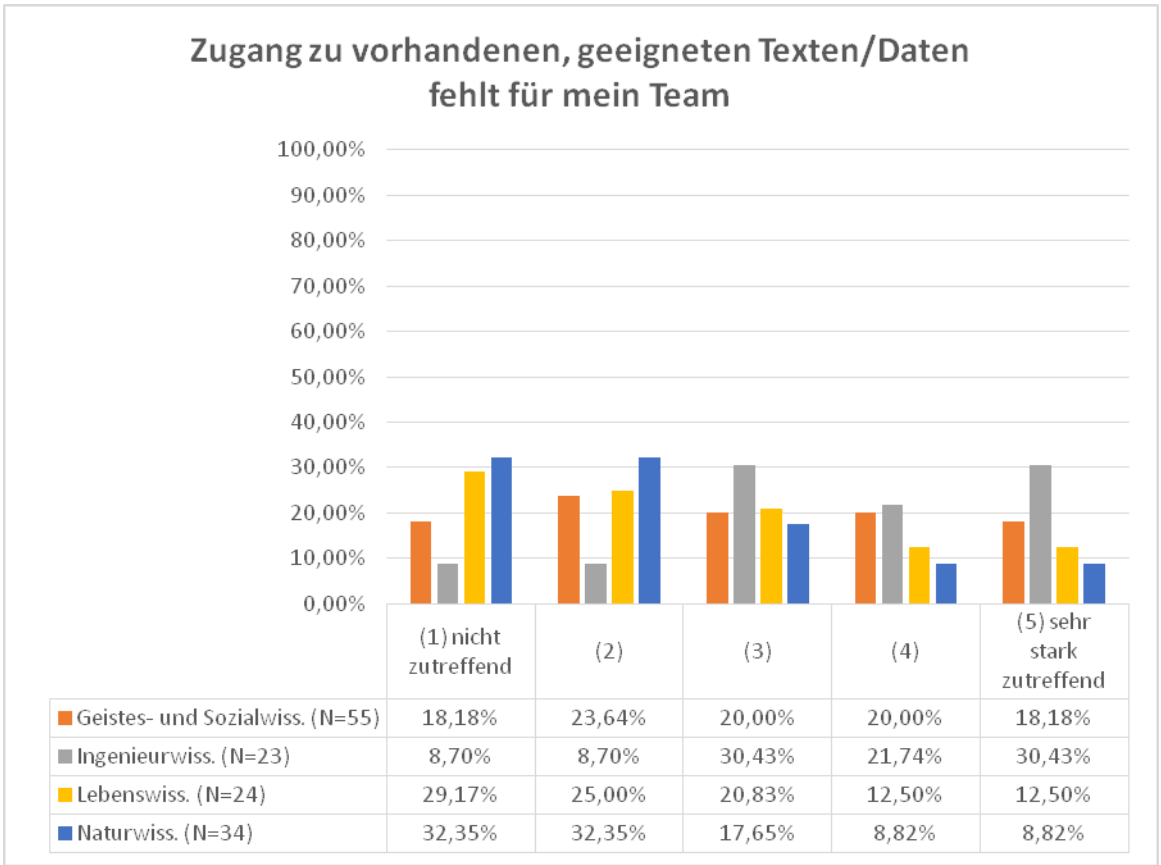
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=56)	25,00%	14,29%	23,21%	25,00%	12,50%
Ingenieurwiss. (N=24)	8,33%	8,33%	20,83%	37,50%	25,00%
Lebenswiss. (N=24)	20,83%	8,33%	20,83%	37,50%	12,50%
Naturwiss. (N=34)	14,71%	17,65%	29,41%	14,71%	23,53%

29. Hindernis 4: Zugang zu vorhandenen, geeigneten Texten/Daten fehlt für mein Team

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	55	23	24	34
(1) nicht zutreffend	33	10	2	7	11
(2)	34	13	2	6	11
(3)	29	11	7	5	6
(4)	24	11	5	3	3
(5) sehr stark zutreffend	26	10	7	3	3
keine Antwort	31	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



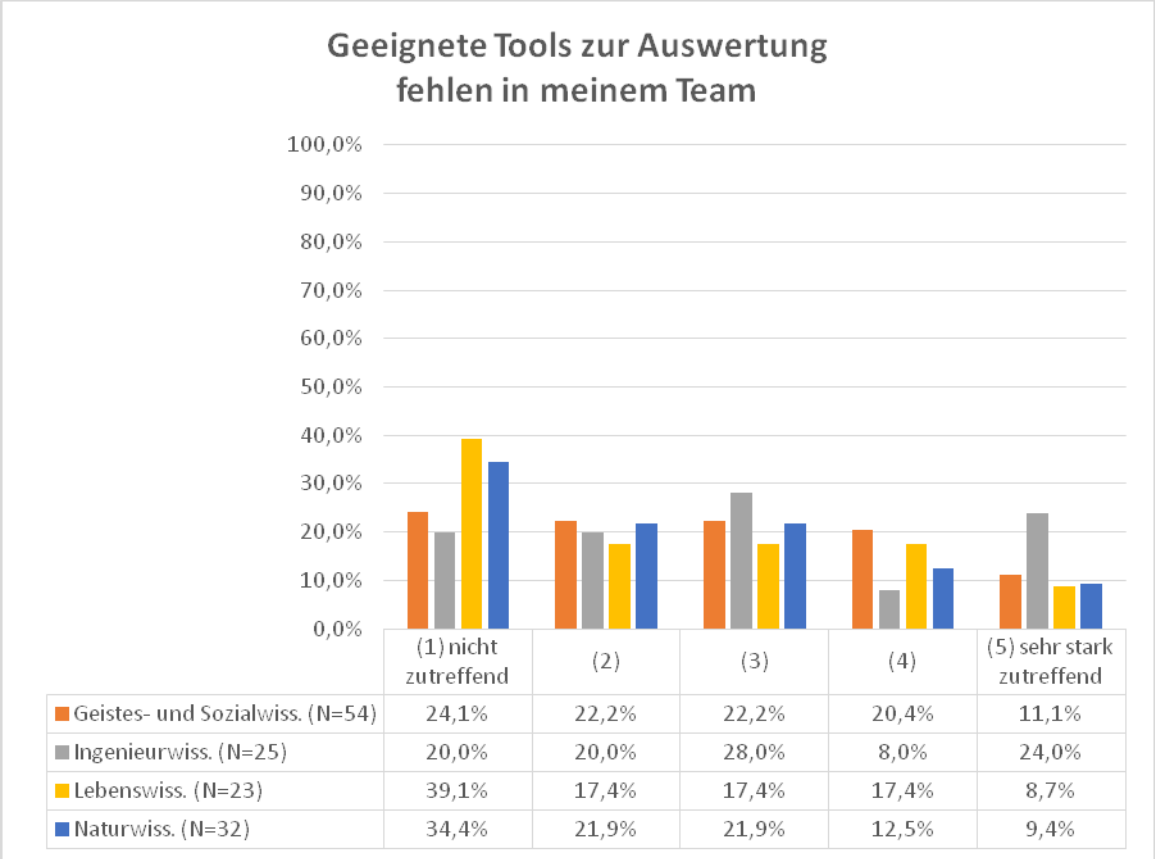
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=55)	18,18%	23,64%	20,00%	20,00%	18,18%
Ingenieurwiss. (N=23)	8,70%	8,70%	30,43%	21,74%	30,43%
Lebenswiss. (N=24)	29,17%	25,00%	20,83%	12,50%	12,50%
Naturwiss. (N=34)	32,35%	32,35%	17,65%	8,82%	8,82%

30. Hindernis 5: Geeignete Tools zur Auswertung fehlen in meinem Team

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	54	25	23	32
(1) nicht zutreffend	38	13	5	9	11
(2)	30	12	5	4	7
(3)	32	12	7	4	7
(4)	23	11	2	4	4
(5) sehr stark zutreffend	21	6	6	2	3
keine Antwort	33	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



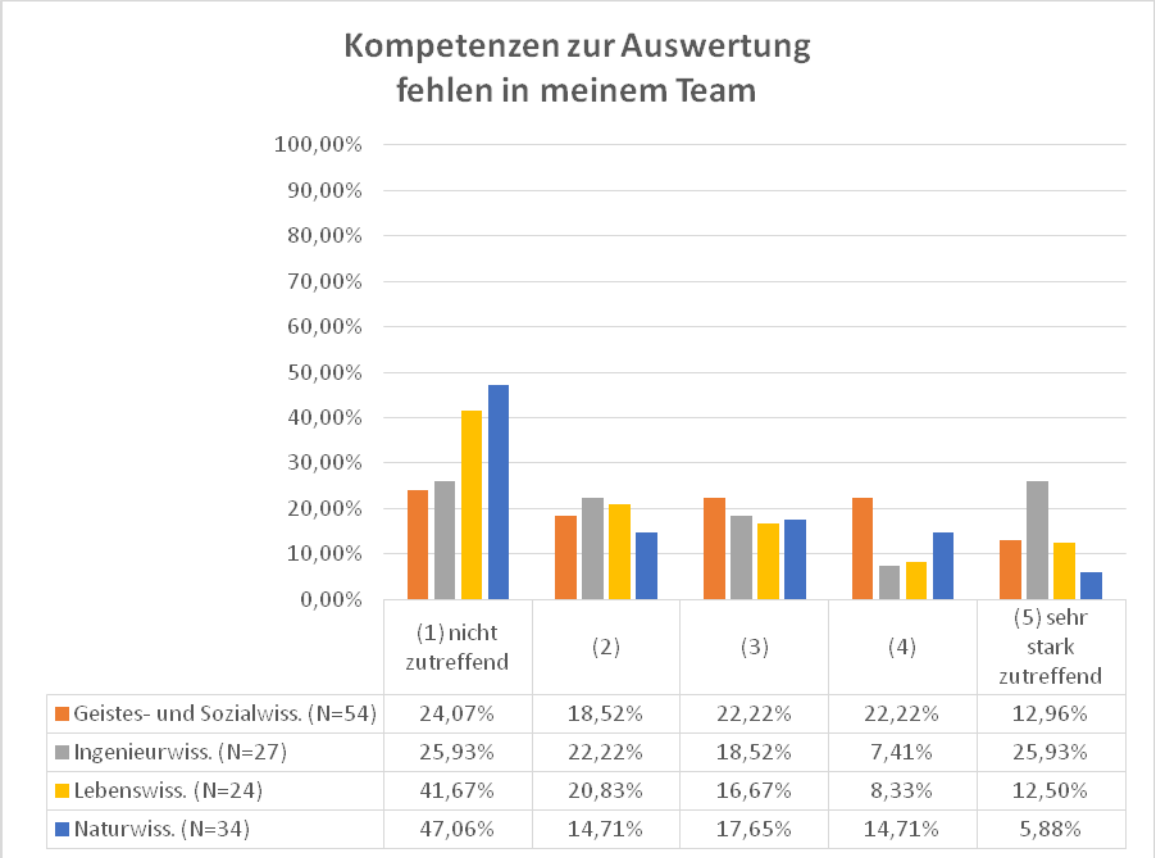
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=54)	24,1%	22,2%	22,2%	20,4%	11,1%
Ingenieurwiss. (N=25)	20,0%	20,0%	28,0%	8,0%	24,0%
Lebenswiss. (N=23)	39,1%	17,4%	17,4%	17,4%	8,7%
Naturwiss. (N=32)	34,4%	21,9%	21,9%	12,5%	9,4%

31. Hindernis 6: Kompetenzen zur Auswertung fehlen in meinem Team

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	54	27	24	34
(1) nicht zutreffend	47	13	7	10	16
(2)	27	10	6	5	5
(3)	28	12	5	4	6
(4)	26	12	2	2	5
(5) sehr stark zutreffend	21	7	7	3	2
keine Antwort	28	?	?	?	?

Relative Zahlen nach Wissenschaftsbereichen



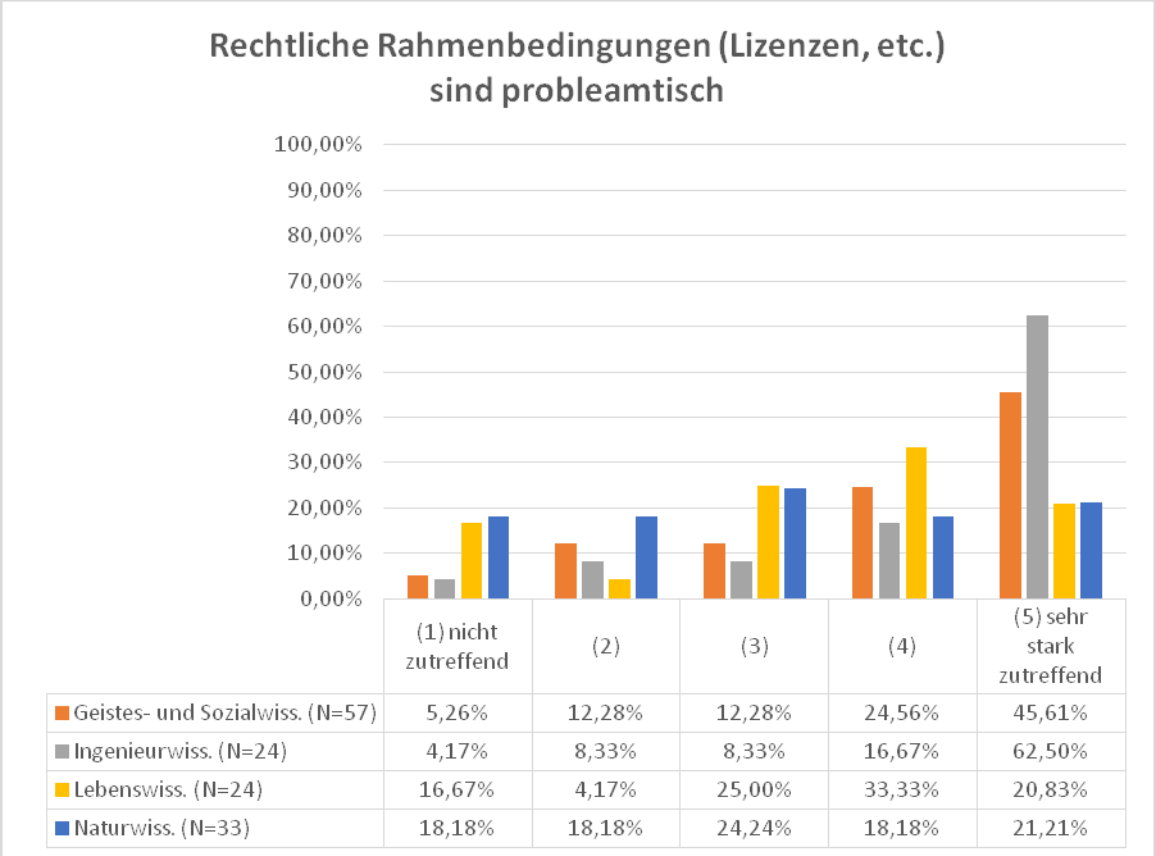
	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=54)	24,07%	18,52%	22,22%	22,22%	12,96%
Ingenieurwiss. (N=27)	25,93%	22,22%	18,52%	7,41%	25,93%
Lebenswiss. (N=24)	41,67%	20,83%	16,67%	8,33%	12,50%
Naturwiss. (N=34)	47,06%	14,71%	17,65%	14,71%	5,88%

32. Hindernis 7: Rechtliche Rahmenbedingungen (Lizenzen, etc.) sind problematisch

Absolute Zahlen

	Gesamt	Geistes- und Sozialwiss.	Ingenieurwiss.	Lebenswiss.	Naturwiss.
N	177	57	24	24	33
(1) nicht zutreffend	14	3	1	4	6
(2)	17	7	2	1	6
(3)	25	7	2	6	8
(4)	36	14	4	8	6
(5) sehr stark zutreffend	56	26	15	5	7
keine Antwort	29				

Relative Zahlen nach Wissenschaftsbereichen



	(1) nicht zutreffend	(2)	(3)	(4)	(5) sehr stark zutreffend
Geistes- und Sozialwiss. (N=57)	5,26%	12,28%	12,28%	24,56%	45,61%
Ingenieurwiss. (N=24)	4,17%	8,33%	8,33%	16,67%	62,50%
Lebenswiss. (N=24)	16,67%	4,17%	25,00%	33,33%	20,83%
Naturwiss. (N=33)	18,18%	18,18%	24,24%	18,18%	21,21%

33. Weitere Hindernisse für TDM

- Finden zusätzlicher Informationen ist mit ein Schlüsselproblem
- Knowledge Gaps in Kollaborationen, insbesondere im Verhältnis zwischen Data Scientists und Domain Experts.
- User-centered Design in Data Science ist Teil meiner Forschungstätigkeit.
- PDF Format von Volltext-Artikeln hindert Textextraktion
- Fehlende Abbildung der Untergliederung in Computer-lesbaren Format (wie z.B. in PMC gegeben)
- Nicht nur die Tools zur Auswertung, auch zum Management und zur Strukturierung fehlen.
- Entsprechende Instrumente und Methoden sind noch zu wenig für den praktischen Entsatz entwickelt
- Kompetenzen zur optimalen Speicherung großer Text und Bild-kombinierter Datenmengen fehlen in meinem Team
- In meinem Fachbereich steckt die Digitalisierung relevanter Daten (vorwiegend Texte) noch in den Kinderschuhen.
- Neue Technologien werden vor allem im "kritischen" Flügel der Geistes- und Sozialwissenschaften von den entsprechenden Entscheidungsträgern systematisch ausgebremst. Klassischer Generationenkonflikt und Paradigmenwechsel.
- Chemie: Qualität publizierter exp. Daten sehr inhomogen, teilweise vollkommener "Schwachsinn" - offensichtlich beim 1. Hinschauen !
- Wildwuchs an Journalen - jeder, dem ein Paper abgelehnt wird, scheint ein neues Journal zu gründen (welches nach kurzer Zeit (=einige Jahre) wieder verschwindet) -> Mehrfachpublikation derselben Sachverhalte, Zersplitterung der Literatur!
- Das PDF-File Format ist der Fluch der Chemie, weil automatisierte Auswertung schwierig bis nahezu unmöglich!

Teil F

Möchten Sie andere Anmerkungen machen? Es steht Ihnen frei, hier weitere Sachinformationen einzubringen oder auch die Umfrage selbst zu kommentieren.

- Besonders problematisch ist die fehlende legale Möglichkeit, Datensätze mit ForscherInnen an anderen Universitäten zu teilen und die Tatsache, dass wir unsere Datensätze nicht legal bei der GESIS o.ä. archivieren lassen können.
- Sozialwissenschaftlich relevante Textquellen sind nicht digital &legal archivierbar, solange das Urheberrecht (nach deutscher Lesart) keine Freiheit der Wissenschaft kennt sondern nur die ökonomische Verwertung.
- OPEN ACCESS!!!
- Internationale rechtliche Rahmenbedingungen sind extrem kompliziert, ebenso wie die Verlinkung von institutionellen und internationalen rechtl. Bedingungen
- vorsturkturierte Datenpools werden immer wichtiger. Bücher werden verschwinden. Datenpools und ihrer Auswertung gehört die Zukunft.
- Es gibt schon tolle Tools (Ontochem, Reaxys etc.), aber die Nutzung ist zu teuer für ein einzelnes Institut, da oft nur eine Abteilung auf einem Gebiet arbeitet und dann die Mindestsummen der Anbieter zu hoch sind.
- Es wäre für den sprachwissenschaftlichen Bereich bzw. für die Analyse großer Fachtexte-Korpora (mit dem Ziel, das wissenschaftliche Schreiben an Universitäten zu unterstützen)von großem Nutzen, wenn für diese Zwecke der Zugang zu entsprechend großen Textdaten (z.B. E-Dissertationen) ermöglicht würde (Stichwort Urheberrecht).
- Größtes Hindernis sind die üblichen Copyrights der wiss. Verlage.

- Tolle Initiative! Würde mich über ein großes, alle historischen Sprachstufen abdeckendes Korpus sehr freuen!
- Ich arbeite mit genomischen Sequenzierdaten und habe alle Fragen basierend darauf beantwortet. Soweit ich das beurteilen kann ist diese Community trotz der großen Datenmengen an Transparenz, Zugänglichkeit, Reproduktion und Nutzung durch Dritte stark interessiert. Die Daten müssen veröffentlicht werden und werden von den großen bioinformatischen Zentren in USA, Japan und Europa verfügbar gemacht. Die Rohdaten sind damit oft einfacher zugänglich als die dazugehörigen Publikationen.
- Normalisierung auf semantische Konzepte ist sehr wichtig, sonst ist Aggregation fast unmöglich
- Download-Raten werden derzeit schon zum Bottleneck da die Datenmengen sehr groß sind. OpenDap oder ähnliche Dienste sind viel zu selten, so dass ich meist alle Daten downloaden muss. Viel geschieht über Webinterfaces, was bei größeren Datenmengen in vielen Dateien sehr, sehr unhandlich wird.
- Die Verarbeitung von Volltexten ist in unserem Bereich wichtig und daher ist Open-Access ein entscheidender Vorteil. OA sollte viel stärker gefördert werden. Es wäre auch hilfreich, personenbezogenes Datamining (z.B. intelligente Expertensuche) zu entwickeln.
- Ich habe sehr große Textsammlungen die ich aufbereitet habe und würde sie gerne der Wissenschaft zur Verfügung stelle scheitere aber an Urheberrecht. Was kann man hier machen?
- Es scheint ein Teufelskreis: Natürlich möchte man als Wissenschaftler optimalen Zugang zu Daten haben und dadurch nicht durch den eigenen Geldbeutel limitiert werden. Andererseits stecken viele nicht-institutsinterne Geisteswissenschaftler selbstfinanzierte Forschungsarbeit in unvergütet publizierte Aufsätze. Lizenzen könnten daher eventuell auch helfen, unbezahlte Forschung für Tagungsbeitrage zu honorieren, da sich gerade aus dem Studium kommende Nachwuchswissenschaftler kaum trauen, dafür ein Honorar vom Herausgeber zu verlangen, der dies mit dem Verlag absprechen müsste.
- In meinem Fach scheint mir TDM noch sehr selten zu sein, da wenig Material zur Verfügung steht. In der Fachgeschichte/Wissenschaftsgeschichte erweitert sich durch die Möglichkeit verschiedene Datenbanken mit hist. Daten/Texten zu nutzen der Forschungshorizont deutlich.
- Forschung zur Optimierung von Informationsinfrastrukturen ist vom Fragebogen nicht leicht zu erfassen.
- Die momentan häufig vorzufindenden Organisationsstrukturen bieten Personen mit Doppelqualifikationen in STM und Geisteswissenschaften kaum Möglichkeiten wirklich passend Anschluss zu finden, da die Stellen(ausschreibungen) nach wie vor nach dem interdisziplinären Schema: Geisteswissenschaftler denkt, Informatiker programmiert (und, Seniortät bzw. Rang regiert,) konzipiert sind.
- Ich bin in Kürze promovierter Sinologe und Computerlinguist und würde es wirklich auf die Entwicklungsmöglichkeiten in dieser Fächerkombination ankommen lassen, ob meine Zukunft weiterhin an einer deutschen Universität stattfinden wird.
- Ehrlich gesagt, habe ich von dem Begriff Data-Mining zwar schon gehört aber ich war der Meinung, es handelt sich nur um ein Schlagwort, nicht um eine eigenständige Forschungsmethode. Für die Umfrage habe ich es für mich als Informationssuche aus im Internet verfügbaren Texten definiert.
- Chemische Strukturen gehören *NICHT* als Bild ins PDF, sondern zB. als MOLfile, Spektren als Abbildung --> sinnlos, Original(mess)daten hinterlegen, was bei XRAY funktioniert, sollte auch bei NMR/MS/IR möglich sein !
- Text-Mining ist kostenfrei im Wesentlichen für biomedizinische Absracts verfügbar (Coremine medical etc. analysieren die Abstracts aus Pubmed!); Wichtig wäre es, derartige Verfahren auch auf Umweltwissenschaften (Ökologie, Systematik, etc.)

anwenden zu können, deren Publikationen nur sehr unvollständig in der Pubmed-datenbank verfügbar sind. Dafür ist vermutlich eine der Pubmed-Datenbank vergleichbare Abstractsammlung aus dem Umweltwissenschaften nötig.

- Seitens der eigenen Abteilung ist TDM bislang nicht genutzt worden. Wenngleich eine künftige Nutzung keineswegs auszuschließen ist, gibt es dafür aktuell keine konkreten Planungen, so dass viele der vorhergehenden Fragen unbeantwortet bleiben müssen.