



**SLUB**

Wir führen Wissen.

# Texte als Forschungsdaten

**Status Quo und Potentiale in der Kooperation zwischen  
Forschung/Lehre und Informationsinfrastrukturen am Beispiel der  
Erschließung und Analyse historischer Texte an der TU Dresden / SLUB**

*Beitrag zur Tagung „Forschungsdaten in der Geschichtswissenschaft“  
(Paderborn 7.-8.6. 2018)*

**7. Juni 2018**

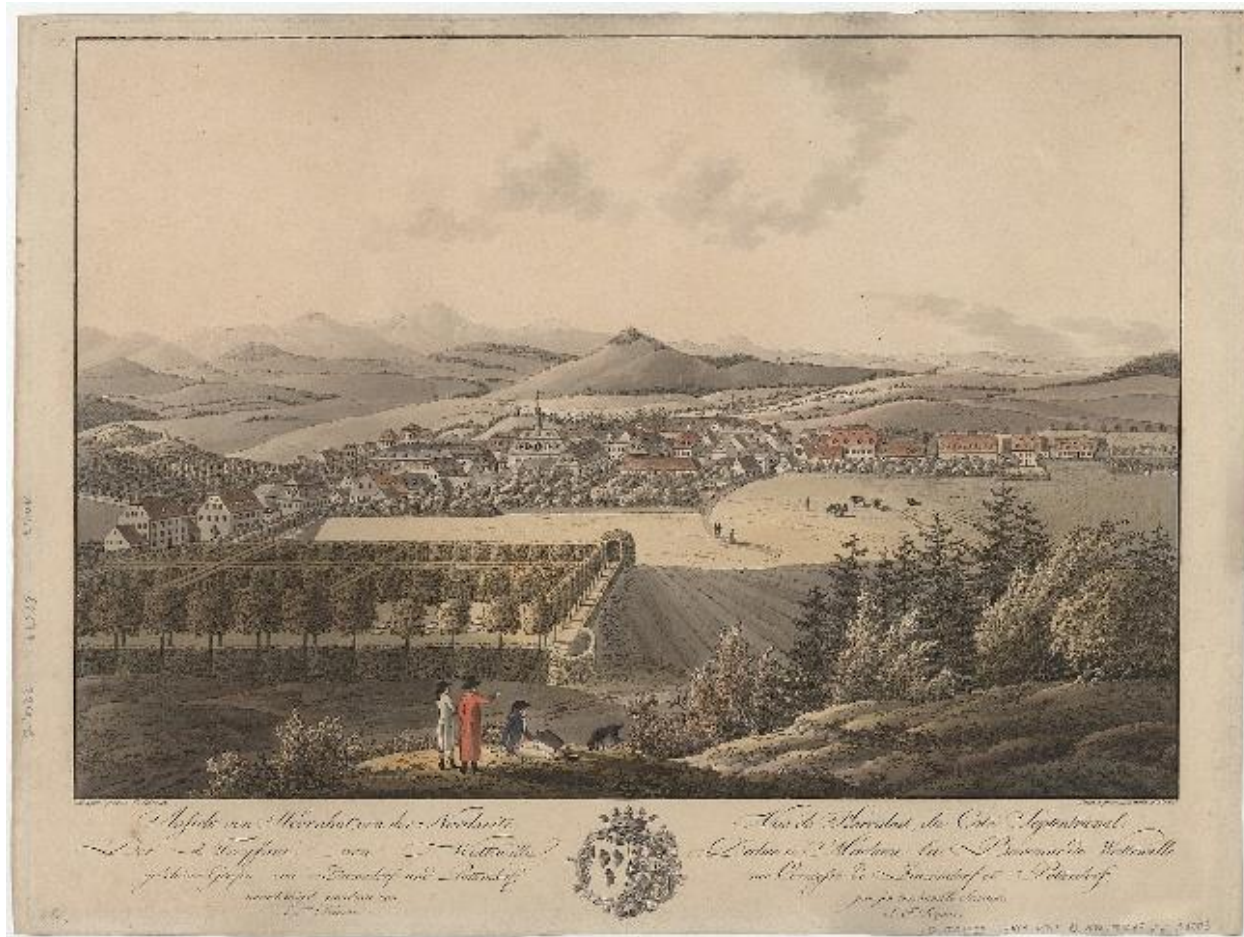
Matti Stöhr, SLUB Dresden

# Anlass



Wachau-Seifersdorf, Dorfkirche (1604/1605, Restaurierung 1892; C. G. Schramm). Chorraum mit Altar. © SLUB / Deutsche Fotothek / Lüttig, Matthias - <http://www.deutschefotothek.de/documents/obj/90095381>

# Anlass



Heinrich Friedrich Laurin nach Ludwig Friedrich Schmuz: Ansicht von Herrnhut von der Nordseite. Dresden: kolorierte Radierung, 1801. TS Mp.3.5. CC-BY-SA 4.0, SLUB / Deutsche Fotothek -

<http://www.deutschefotothek.de/documents/obj/70400329>

# Anlass

Auch Opfer gehören zur Religion der Indianer. Ihre Absicht ist, Gott, und die übrigen guten Geister zu versöhnen. Sie sind von den ältesten Zeiten unter ihnen gewöhnlich, und ihnen so wichtig und heilig, daß sie glauben, sie würden sich selbst und ihrer ganzen Freundschaft unfehlbar allerley Krankheiten, Unglück und selbst den Tod und Untergang zuziehen, wenn sie die Opfer unterließen, oder sie nur nachlässig, und nicht zur rechten Zeit verrichteten. Eigentliche Opferpriester und Tempel haben sie nicht. Bey großen Opfern, woran viele Theil haben, vertreten die ältesten Männer die Stelle der Priester; bey kleinern thut<sup>s</sup> derjenige, der das Opfer gibt. Zu großen Opfern wird ein großes und geräumiges Wohnhaus zubereitet.

Grundlage des Seminars: Georg Heinrich Loskiel (1789):  
*Geschichte der Mission der Evangelischen Brüder unter den Indianern in Nordamerika*. Barby.

<http://digital.slub-dresden.de/werkansicht/dlf/205632/>

# These

Forschungsdaten in der Geschichtswissenschaft werden (bestmöglich) aus der Kooperation von Forschungs- und Infrastruktureinrichtung(en) generiert. Digital nachnutzbare Forschungsdaten sind ein Konglomerat aus Daten *für* die Forschung und Daten *aus* der Forschung.

# Agenda

1. Begriffliche Positionierung
2. Prozess des Datenmanagements mit Fokus Texte
3. Akteure und Rollen
4. Lösungsansätze – (Mögliche) Standards / Verfahren / Tools
5. Fazit und Diskussion



# Begriffliche Positionierung

## Texte als Forschungsdaten

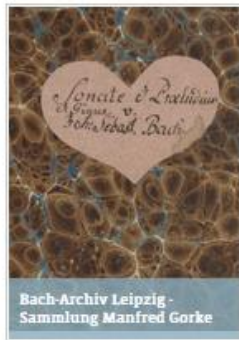
*„Forschungsdaten sind durch eine spezifische Methode und Schritte der Operationalisierung systematisch gewonnene, strukturierte Informationen, die (computergestützt) ausgewertet und verarbeitet werden können. Sie bilden die Grundlage des Forschungsprozesses. Daten können sowohl quantitativen wie qualitativen Charakter tragen.“*

Vgl. <http://www.geschichte.uni-halle.de/struktur/hist-data/datenmanagement/>

# Begriffliche Positionierung

## Texte als Forschungsdaten

### Digitale Sammlungen



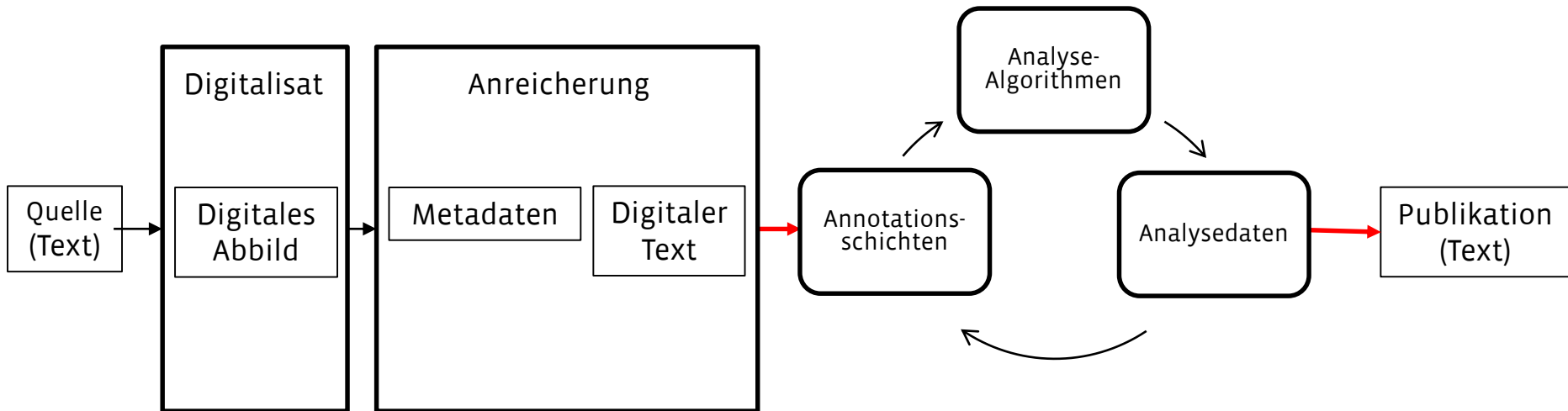
Vgl. <http://digital.slub-dresden.de/kollektionen>





# Prozess des Datenmanagements

## Texte in der „Forschungsdatenkette“



# Akteure und Rollen

## Infrastrukturprojekte / Institutionen / Forschende



[Home](#) [Auffinden](#) [Auswerten](#) [Aufbereiten](#) [Facharbeitsgruppen](#) [Blog](#) [Über...](#) [Hilfe](#) [Aktuelles](#)

### WILLKOMMEN

**Auffinden**

**Auswerten**

**Aufbereiten und Aufbewahren**

Vgl. <https://www.clarin-d.net/de/>

# Akteure und Rollen

## Infrastrukturprojekte / Institutionen / Forschende

Anmelden (DTAQ) DWDS dlexDB CLARIN-D

**DTA** Werke im Deutschen Textarchiv Texte ▼ Projekt ▼ Dokumentation ▼ Impressum

in den Titeldaten  im Korpus  in der Dokumentation [Hilfe](#)

Beispielfragen: `ehelichen with $p=VVINF` `ehelich with $p=ADJA` `"@gefunden #2 @werden" #random`

### Deutsches Textarchiv

#### GRUNDLAGE FÜR EIN REFERENZKORPUS DER NEUHOCHDEUTSCHEN SPRACHE

Das Deutsche Textarchiv stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 bereit. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern erstellten und ausführlich kommentierten, umfangreichen Bibliographie. In Ergänzung wurden einschlägige Literaturgeschichten und (Fach-)Bibliographien ausgewertet. Aus der Gesamtliste der auf diesem Wege ermittelten Titel wurde von der DTA-Projektgruppe ein hinsichtlich der repräsentierten Textsorten und Disziplinen ausgewogenes Korpus zusammengestellt (weitere Informationen zur Textauswahl).

Um den historischen Sprachstand möglichst genau abzubilden, werden als Vorlage für die Digitalisierung in der Regel die Erstausgaben der Werke zugrunde gelegt. Das elektronische Volltextkorpus des DTA ist über das Internet frei zugänglich und dank seiner Aufbereitung durch (computer-)linguistische Methoden schreibweisentolerant über den gesamten jeweils verfügbaren Bestand durchsuchbar. Sämtliche Texte stehen zum Download zur Verfügung.

#### Das DTA in Zahlen

- 3 353 Werke
- 638 511 digitalisierte Seiten
- 157 054 630 fortlaufende Wortformen
- 1 095 130 031 Zeichen (Unicode)
- 1 057 weitere Werke in DTAQ

#### Das DTA am 7. Juni 2018



Nachweis: Richard Dighton: George "Beau" Brummell, Aquarell (1805) [Quelle: Wikipedia]

Am 7. Juni 1778 wird George Bryan Brummell in London geboren. Unter seinen Zeitgenossen wurde er als der *Beau* oder *Dandy* bezeichnet. Erste Belege für das Wort *Dandy* sind seit 1780 vorhanden (siehe Oxford English Dictionary). Wie Brummell wahrgenommen wird, illustriert Fürst Hermann Ludwig Heinrich von Pückler-Muskau in ...

Vgl. <http://www.deutschestextarchiv.de/>

# Akteure und Rollen

Infrastrukturprojekte / Institutionen / Forschende



Vgl. <https://de.dariah.eu/>



# Akteure und Rollen

## Infrastrukturprojekte / Institutionen / Forschende

**TextGrid**  
Virtuelle Forschungsumgebung  
für die Geisteswissenschaften

Suche English

Registrierung Download Community Support Über TextGrid

### Digital edieren – forschen – archivieren

#### Laboratory

Open-Source-Werkzeuge und -Services unterstützen GeisteswissenschaftlerInnen im gesamten Forschungsprozess – insbesondere beim Erstellen digitaler Editionen.

Download

#### Repository

Im Langzeitarchiv für Forschungsdaten können vielfältige digitale Materialien – u. a. XML/TEI-kodierte Texte, Bilder und Datenbanken – sicher gespeichert, publiziert und durchsucht werden.

Die "Digitale Bibliothek" bei TextGrid

Besuchen

#### Community

Online-Hilfen, Text- und Videotutorials, Mailinglisten, Bug Tracker und Source Code ermöglichen einen schnellen Einstieg in die Arbeit mit TextGrid.

Besuchen

Vgl. <https://textgrid.de/>

# Akteure und Rollen

## Infrastrukturprojekte / Institutionen / Forschende

### Digitale Sammlungen



Vgl. <http://digital.slub-dresden.de/kollektionen/>

# Akteure und Rollen

## Infrastrukturprojekte / Institutionen / Forschende

SLUB Dresden > Service > Open-Science-Service > Open Data/Forschungsdaten

### Open Data/Forschungsdaten

#### Unser Forschungsdaten-Service

Die SLUB bietet ihren Nutzern in Zusammenarbeit mit dem [Zentrum für Informationsdienste und Hochleistungsrechnen \(ZIH\)](#) und weiteren Partnern an der TU Dresden über die gemeinsame [Kontaktstelle Forschungsdaten](#) eine umfangreiche Beratung zum Management von [Forschungsdaten](#) aus einer Hand an. Die TU Dresden regelt in ihren [Richtlinien zur Sicherung guter wissenschaftlicher Praxis](#) (speziell §5) die "*Sicherung und Aufbewahrung von Primärdaten*". Darauf aufbauend informieren wir zu den Anforderungen der TU Dresden und der Forschungsförderer, beraten zu geeigneten Daten- und Metadatenformaten und Repositorien und bieten Informationen zur Publikation von Forschungsdaten sowie Unterstützung bei der Erstellung von Datenmanagement-Plänen für Projektanträge. Vereinbaren Sie auch gerne einen persönlichen [Beratungstermin](#) in unserer Wissensbar.

Informationen zu allen Aspekten des Themas Forschungsdaten und das gemeinsame Service-Angebot finden Sie auf der Webseite der [Kontaktstelle Forschungsdaten](#)



#### OPEN-SCIENCE-SERVICE

- Open Access
- Open Evaluation/Bibliometrie
- Open Data/Forschungsdaten
- Open Educational Resources
- Citizen Science
- Veranstaltungen

#### ANSPRECHPARTNER

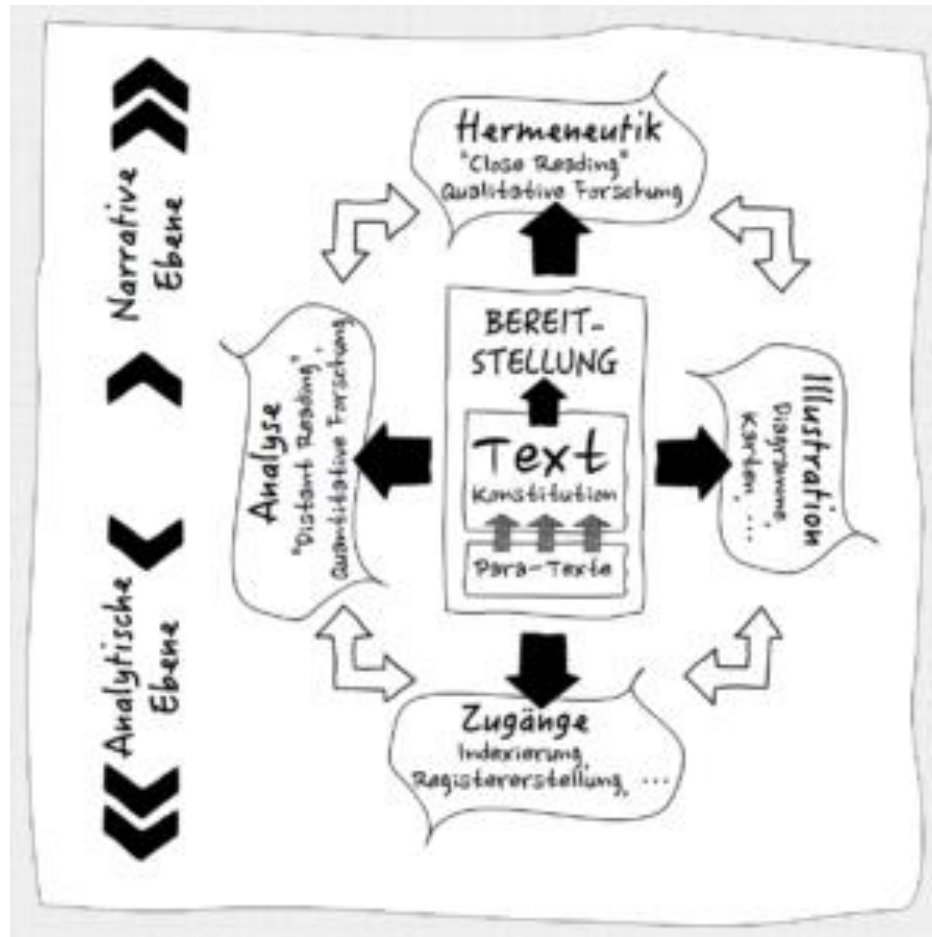
Dr. Jan Polowinski,  
Manuela Queitsch,  
Matti Stöhr  
[Kontaktstelle Forschungsdaten](#)  
Tel.: +49 351 4677-378  
E-Mail: [forschungsdaten@slub-dresden.de](mailto:forschungsdaten@slub-dresden.de)  
Persönlichen [Beratungstermin](#) buchen



Vgl. <https://www.slub-dresden.de/service/open-science-service/open-dataforschungsdaten/>

# Akteure und Rollen

Infrastrukturprojekte / Institutionen / Forschende



Schema Editorik – Quelle: [https://f-origin.hypotheses.org/wp-content/blogs.dir/1535/files/2018/05/editorik\\_schema\\_c.png](https://f-origin.hypotheses.org/wp-content/blogs.dir/1535/files/2018/05/editorik_schema_c.png)

# (Mögliche) Standards / Verfahren / Tools

## Empfehlungen für Forschungsdaten, Tools und Metadaten in der DARIAH-DE Infrastruktur

Erstellt von Johanna Puhl, zuletzt geändert von Hanna Meiners am Feb 22, 2017



### Inhalt

- 1 Grundsätzliches
- 2 Dateiformate für Langzeitarchivierung UND Nachnutzung
  - 2.1 Kriterien für die Langzeitarchivierbarkeit
  - 2.2 Kriterien für die Nutzbarkeit
  - 2.3 Empfohlene Dateiformate
- 3 Metadatenstandards
  - 3.1 Kriterien für die Eignung von Metadatenstandards
  - 3.2 Administrative, deskriptive Metadatenstandards
  - 3.3 Fachwissenschaftliche Metadatenstandards (Content)
- 4 Tools und Verfahren für die digitalen Geisteswissenschaften
- 5 Empfohlene Lizenzen
  - 5.1 Lizenzen für Content
  - 5.2 Lizenzen für Code
  - 5.3 Lizenzen für Dokumentation

Vgl. <https://wiki.de.dariah.eu/pages/viewpage.action?pageId=38080370>



# (Mögliche) Standards / Verfahren / Tools

## TEI / DTA-Basisformat



The screenshot shows the TEI P5 Guidelines website. At the top, there is a blue header with the TEI logo and the text "< Text Encoding Initiative >". Below the header is a search bar with a dropdown menu set to "P5 Guidelines — Deutsch" and a "Search" button. The main content area is titled "P5: Richtlinien für die Auszeichnung und den Austausch elektronischer Texte" and includes the version "Version 3.3.0. Last updated on 31st January 2018, revision f4d8439". There are language links for English, Deutsch, Español, Italiano, Français, 日本語, 한국어, and 中文. Below these are icons for PDF, XML, and Amazon. The page is organized into three columns: "Titelei", "Textkörper", and "TEI sourcecode".

**TEI** < Text Encoding Initiative >

P5 Guidelines — Deutsch Search

**P5: Richtlinien für die Auszeichnung und den Austausch elektronischer Texte**  
Version 3.3.0. Last updated on 31st January 2018, revision f4d8439

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [한국어] [中文]

Titelei	Textkörper	TEI sourcecode
<a href="#">Titel</a>	⊕ 1 <a href="#">The TEI Infrastructure</a>	• <a href="#">Getting and Using the TEI Sources.</a>
i. <a href="#">Releases of the TEI Guidelines</a>	⊕ 2 <a href="#">The TEI Header</a>	• <a href="#">TEI GitHub Repository</a>
ii. <a href="#">Dedication</a>	⊕ 3 <a href="#">Elements Available in All TEI Documents</a>	• <a href="#">Bug Reports, Feature Requests, etc.</a>
iii. <a href="#">Preface and Acknowledgments</a>	⊕ 4 <a href="#">Default Text Structure</a>	
⊕ iv. <a href="#">About These Guidelines</a>	⊕ 5 <a href="#">Characters, Glyphs, and Writing Modes</a>	
⊕ v. <a href="#">A Gentle Introduction to XML</a>	⊕ 6 <a href="#">Verse</a>	
⊕ vi. <a href="#">Languages and Character Sets</a>	⊕ 7 <a href="#">Performance Texts</a>	
<b>Anhang</b>	⊕ 8 <a href="#">Transcriptions of Speech</a>	
	⊕ 9 <a href="#">Dictionaries</a>	

Vgl. <http://www.tei-c.org/release/doc/tei-p5-doc/de/html/>

# (Mögliche) Standards / Verfahren / Tools

## TEI / DTA-Basisformat



Vgl. <http://www.deutschestextarchiv.de/doku/basisformat/>

# (Technische) Herausforderungen

- Verbesserte Qualität der (Retro-)Digitalisierung – insbes. OCR und Fraktur
- Entwicklung, Identifikation, (Nach-)Nutzbarkeit, Verbreitung geeigneter Annotations- und Analysewerkzeuge – Open Source!?
- Automatisierung in der Vergabe und Auszeichnung von (Meta-)Datenelementen
- Standards und Schnittstellen
- Langzeitarchivierung
- Publikationsformate – Datenpublikation (Renommee!)
- ...

# Fazit und Diskussion

## Fazit

- Services / Verfahren / Tools für den standardisierten, systematischen Umgang mit textuellen Forschungsdaten prinzipiell vorhanden
- Kenntnis, Verbreitung und Nutzung von existierenden Tools intensivieren
- Begleitung des Forschungsprozesses -> Zusammenarbeit mit Germanistik-Lehrstuhl der TU Dresden und SLUB in Berücksichtigung existierender Datenstandards und Tools exemplarisch für Kooperation zwischen Forschung/Lehre und Bibliothek im Umgang mit Texten als Forschungsdaten

# Fazit und Diskussion

## These

Forschungsdaten in der Geschichtswissenschaft werden (bestmöglich) aus der Kooperation von Forschungs- und Infrastruktureinrichtung(en) generiert. Digital nachnutzbare Forschungsdaten sind ein Konglomerat aus Daten *für* die Forschung und Daten *aus* der Forschung.



# Fazit und Diskussion

## Diskussion

- Welche (neuen) Anforderungen stellen Sie an die Verfügbarkeit, Auffindbarkeit und Charakteristik textueller Forschungsdaten?
- Welche Werkzeuge bzw. Services benötigen oder vermissen Sie zum (kollaborativen) Management von Textdaten?
- An welchen Stellen der „Forschungsdatenkette“ wünschen Sie sich mehr Unterstützung von Informationsinfrastruktureinrichtungen, insbesondere von Bibliotheken?
- Welche (weiteren) Automatisierungen sind in Format und Prozessierung textueller Forschungsdaten vorstellbar?



**SLUB**

Wir führen Wissen.

**Vielen Dank für Ihre Aufmerksamkeit!!**

**Matti Stöhr**

**SLUB Dresden**

**Abteilung Benutzung und Information**

**Referat Open Science – Forschungsdaten**

**[matti.stoehr@slub-dresden.de](mailto:matti.stoehr@slub-dresden.de)**

**Tel.: 0351 / 4677 437**

**<https://www.slub-dresden.de/open-science/open-dataforschungsdaten/>**