

# Digital Scholarship Services

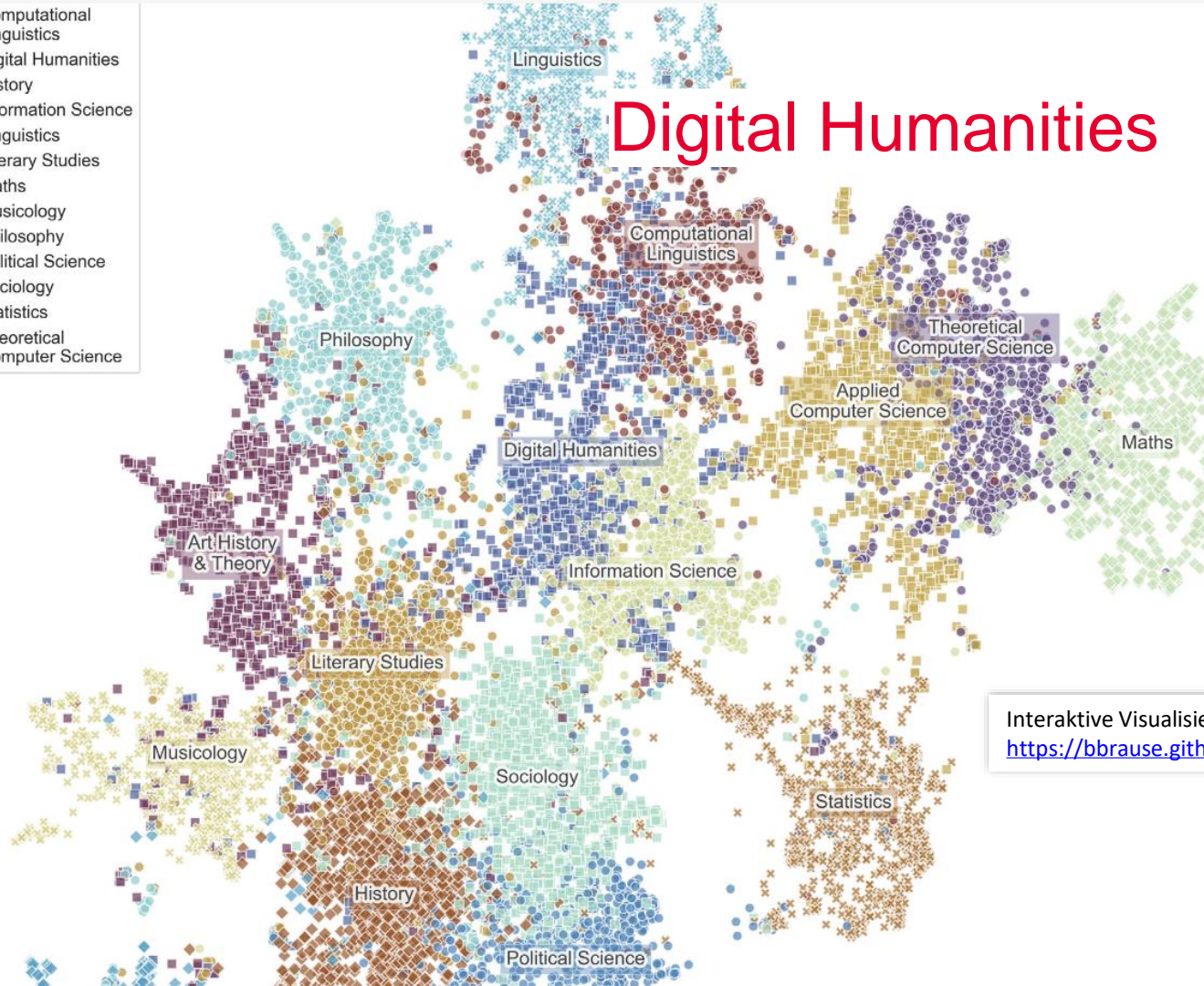
Neue Dienstleistungen von  
wissenschaftlichen Bibliotheken für die  
datenbasierte Forschung

**Kathi Woitas, Universitätsbibliothek Bern**

**01.06.2022 8. Bibliothekskongress Leipzig**

# Digital Humanities

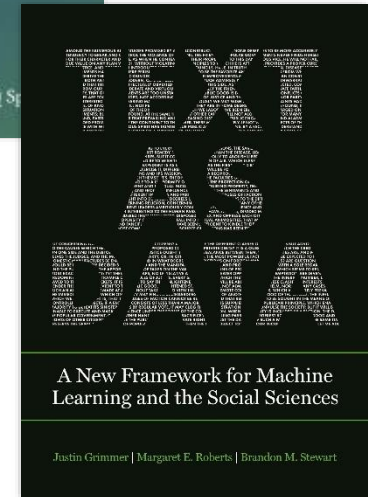
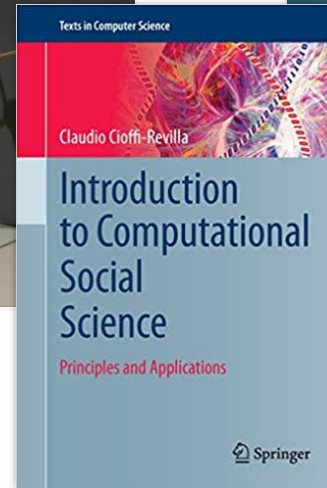
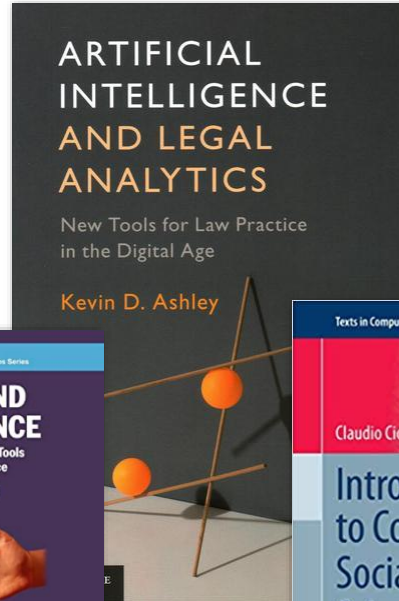
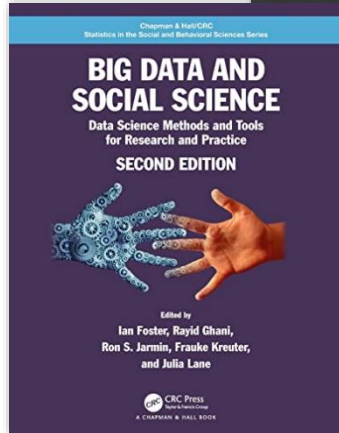
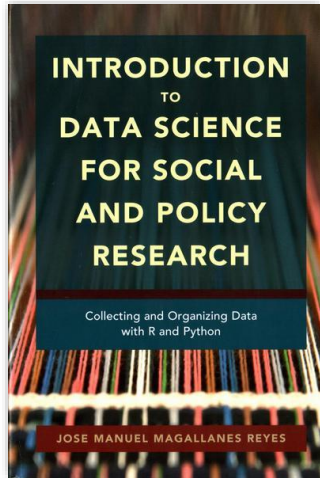
- Computational Linguistics
- Digital Humanities
- ◆ History
- Information Science
- ✱ Linguistics
- Literary Studies
- Maths
- ✱ Musicology
- Philosophy
- Political Science
- Sociology
- ✱ Statistics
- Theoretical Computer Science



Luhmann, J., und M. Burghardt (2021): „Digital Humanities—A discipline in its own right? An analysis of the role and position of Digital Humanities in the academic landscape“. *Journal of the Association for Information Science and Technology*, S. 1-24.  
<https://doi.org/10.1002/asi.24533>

Interaktive Visualisierung:  
[https://bbrause.github.io/dh\\_academic\\_landscape/](https://bbrause.github.io/dh_academic_landscape/)

# Gesellschaftswissenschaften



# Computational Social Science

## Korpora

Text-, Audio-, Videodaten, z.B.

- Medien, z.B. Zeitungen, Websites
- Dokumente aus Exekutive, Legislative, Judikative

Neu im Vergleich zu Inhaltsanalyse:

- Korpusgrösse
- Preprocessing- und Analyse-Methoden

Using Word Embeddings to Analyze how Universities Conceptualize “Diversity” in their Online Institutional Presence

[David Rozado](#) 

Rozado, D. (2019): Using word embeddings to analyze how universities conceptualize “diversity” in their online institutional presence. In: *Society*, 56. Jg., H. 3, S. 256–266. <https://doi.org/10.1007/s12115-019-00362-9>

A new approach to semantic sustainability assessment: text mining via network analysis revealing transition patterns in German municipal climate action plans

[Manuel W. Bickel](#) 

Bickel, M. W. (2017): A new approach to semantic sustainability assessment: text mining via network analysis revealing transition patterns in German municipal climate action plans. In: *Energy, Sustainability and Society*, 7. Jg., H. 1, S. 22. <https://doi.org/10.1186/s13705-017-0125-0>

# Computational Social Science

## Digital Behavioral Data

- Online-Medien: User Generated Content, Nutzungs- und Netzwerkdaten
- Smart Phones, Smart Devices
- Sensordaten, z.B. IoT

Luhmann, M. (2017): Using Big Data to study subjective well-being. In: Current Opinion in Behavioral Sciences, 18. Jg., S. 28–33. <https://doi.org/10.1016/j.cobeha.2017.07.006>

### Using Big Data to measure SWB

Approaches to measure SWB using Big Data can be distinguished in terms of data source, measurement level, and SWB facet (Table 1).

Table 1. Overview of measurement approaches.

Publication	General approach	Data source	SWB facet	Measurement level
Algan <i>et al.</i> [47]	Frequency of specific search terms on Google	Google Trends	Life satisfaction Emotional well-being	Longitudinal trends within one nation (United States)
Carlquist <i>et al.</i> [16]	Closed vocabulary (self-constructed lexicon)	Newspaper articles	Emotional well-being	Longitudinal trends within one nation (Norway)
Collins <i>et al.</i> [12]	Online activity; closed vocabulary (LIWC)	Facebook status updates and likes	Life satisfaction	Individual
Curini <i>et al.</i> [45]	Open vocabulary	Twitter	Emotional well-being	Italian provinces
Doré <i>et al.</i> [46**]	Closed vocabulary (LIWC)	Twitter	Emotional well-being	Local with exact coordinates
Durahim & Coskun [44]	Closed vocabulary (SentiStrength, Turkish version)	Twitter	Emotional well-being	Turkish provinces and national
Hao <i>et al.</i> [24]	Open vocabulary	Weibo posts	Emotional well-being	Individual
Hung <i>et al.</i> [34]	Mobile phone usage	Calling states, app usage	Emotional well-being	Individual
Jones <i>et al.</i> [41]	Closed vocabulary	Twitter	Emotional well-	Regional

# Computational Social Science

## Datenerhebung und -haltung, Datenanalyse

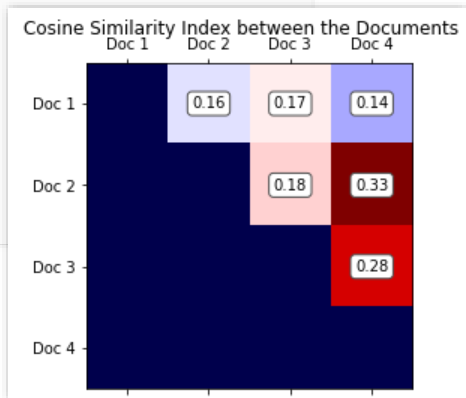
- Webtracking
- Webscraping
- Nutzung von Daten-APIs
- Nutzung von Big-Data-Anwendungen/Infrastrukturen
- z.B. Topic Modeling, Sentiment Analysis, Network Analysis, Named Entity Recognition
- Natural Language Processing (NLP)

# Computational Linguistics

```
# Collecting the unigrams and processing them into `documents`

limit = 500 # Change number of documents being analyzed. Set to `None` to do all documents.
#Limit = None
n = 0
documents = []
document_ids = []

for document in tdm_client.dataset_reader(dataset_file):
    processed_document = []
    document_id = document['id']
    if use_filtered_list is True:
        # Skip documents not in our filtered_id_list
        if document_id not in filtered_id_list:
            continue
    document_ids.append(document_id)
    unigrams = document.get("unigramCount", [])
    for gram, count in unigrams.items():
        clean_gram = process_token(gram)
        if clean_gram is None:
            continue
        processed_document.append(clean_gram)
    if len(processed_document) > 0:
        documents.append(processed_document)
    n += 1
    if (limit is not None) and (n >= limit):
        break
print('Unigrams collected and processed.')
```



- Basiswissenschaft des Natural Language Processing (NLP)
- Analyse von Sprach- und Textcharakteristika
- Datafizierung natürlicher Sprache → Hilfswissenschaft für jede DS auf Textdaten – „Text and Data Mining (TDM)“

# Digital Scholarship?

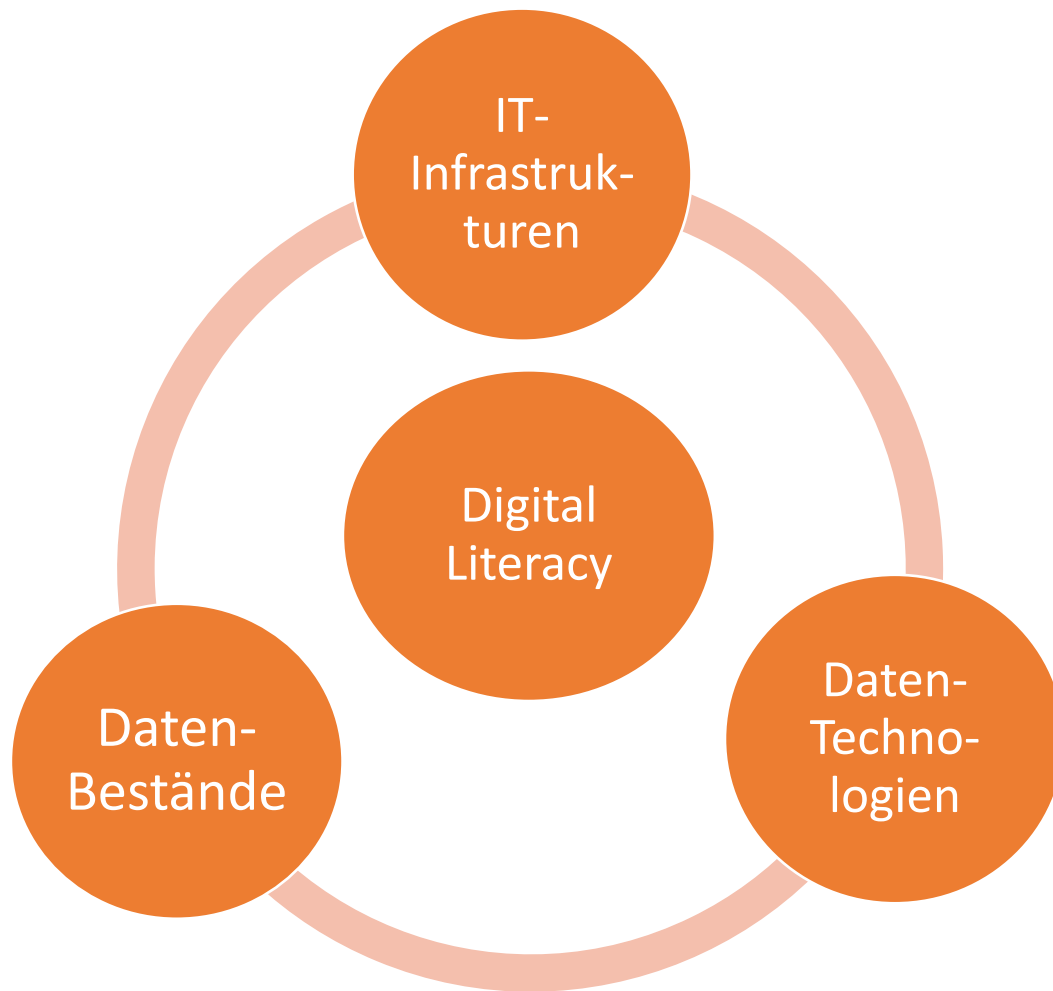
*A broad term that [...] relates to the **variety** of ways in which “**digital evidence**” and **computational methods** are **incorporated into academic research**. [...]*

*[I]t is used to evoke scholarly behavior and practice in which **digital objects, data, or workflows (i.e., analysis techniques)** are a **primary component** of the research project.*



# Digital Scholarship Services?

*This increasing emphasis on **data- and computationally intensive research methods** creates opportunities for **libraries** to contribute to the **education, tools, infrastructure, and communities** that sustain and expand these practices. Given the complexity of big-data analysis and the specialized skills it requires, **educational and consulting services** are essential across the **disciplinary spectrum**.*



# Datenbestände

	Handschriften (Anzahl Laufmeter)		davon: Karten und Pläne	
	F39a	F40		
	5987'218	19'022	733'645	6'444
	14	14	14	
	14	12	14	
	100%	86%	100%	100%
	92'826	4'078	9'456	132'110
	527	21	26'144	6'387
	2'246	101	6'148	945'222
	5'303'359	4'883	5'591	401'589
	33	1	721	0
	0	0	130	84'022
	2'743	490	6'286	132'387
	17'439	718	7'849	17'8
	0	0	99	
	340'989	4'792	342'206	3'6
	0	0	2'841	
	0	4'138	325'517	

Strukturierte  
Daten

```

"author": "Michael",
"description": null,
"publisher": "Paul Haupt Bern",
"contributor": null, "date":
"1949", "type": ["Text", "Journal", "Article"], "source": ["Berner
Zeitschrift f\u00fcr Geschichte
und Heimatkunde", "280461-x",
"0005-9420", "11", "1949", null,
"11", "5"], "language": null,
"relation": null, "coverage":
"rights": null, "form":
"application"
    
```

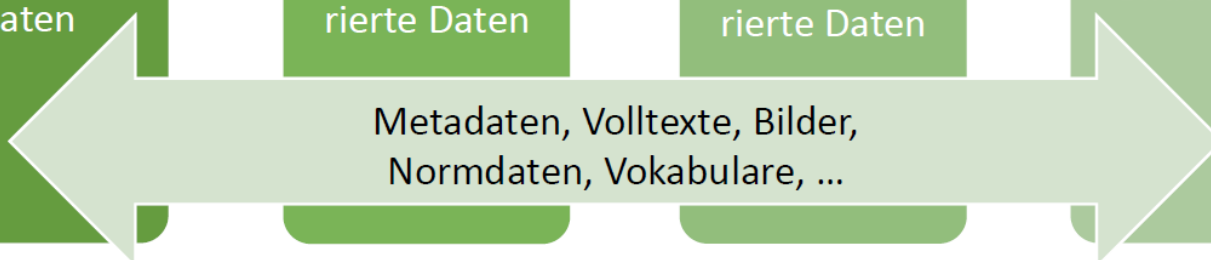
Semistrukturierte  
Daten

... sondern auch \ndiese  
... unsere Mauern gef\u00fcr  
... 1415. Im Verlauf des Konstanz  
... \u00e4ngt K\u00f6nig Sigmund \ndie Re  
... fcher Herzog Friedrich von \u00d6sterr  
... J erl\u00e4\u00dft den \u00e4u\u00dferen  
... \u00fcrsten und \u00f6fken. Dieser veranla  
... \u00dft die Besetzung habsburgi-\nscher Gebie  
... durch die Eidgenossen und die Eroberung des  
... Aargaus durch \ndie Berner. Das Kloster K  
... \u00f6nigsfelden, errichtet an der Stelle, wo  
... \u00f6nig \nAlbrecht einem Meuchelmord zum Opfe  
... gefallen war, wird damit bernisch. \n\n\n\nNac  
... \u00e4tze, mit denen Albrechts Tochter, \nAgn  
... \nUngarn, das Kloster so k\u00f6niglich  
... \nhenkt hatte, nach Bern: der \nmit Jaspj  
... Kristall und Edelstein geschm\u00f6  
... nische Hausaltar \n\n\naus dem Bes  
... Andreas von Ungarn und d  
... \npendien, deren

Unstrukturierte  
Daten



Bilddateien



Metadaten, Volltexte, Bilder,  
Normdaten, Vokabulare, ...

# Datentechnologien

## Data Science



- statistische Grundlagen
- Datenanalyse

### Machine Learning:

- überwachtes Lernen
- unüberwachte Lernen
- Tuning, Evaluation



### Deep Learning

- Neuronale-Netz-Architekturen

# Natural Language Processing (NLP)

## Large-scale language models

A robot wrote this entire article. Are you scared yet, human?

*GPT-3*

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

„A Robot Wrote This Entire Article. Are You Scared yet, Human?“ *The Guardian*, 8. September 2020.

<https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>.

### z.B. GPT-3

- Zusammenfassungen
- Übersetzungen
- Dialog-Generierung
- Semantische Suche
- Vervollständigung von Programm-Code
- ...?

# IT-Infrastrukturen

## Big Data + Cloud Computing

### Big Data

volume, velocity, variety, veracity

Bsp. Batch Data:

Bilder aus Massendigitalisierung

Bsp. Stream Data:

- Social Media Feeds
- Sensordaten

### Cloud Computing

- Cloud-Service-Modelle: IaaS, PaaS, SaaS
- **verteilte** Anwendungen für Datenhaltung, Datenintegration, Datenverarbeitung, -abfrage
- hoch skalierbar, ausfallsicher, performant

ALWAYS ALREADY COMPUTATIONAL - COLLECTIONS AS DATA



[HOME](#) [TEAM](#) [PARTNERS](#) [EVENTS](#) [RESOURCES](#) [UPDATES](#) 🌙 [PART TO WHOLE](#) 🌞

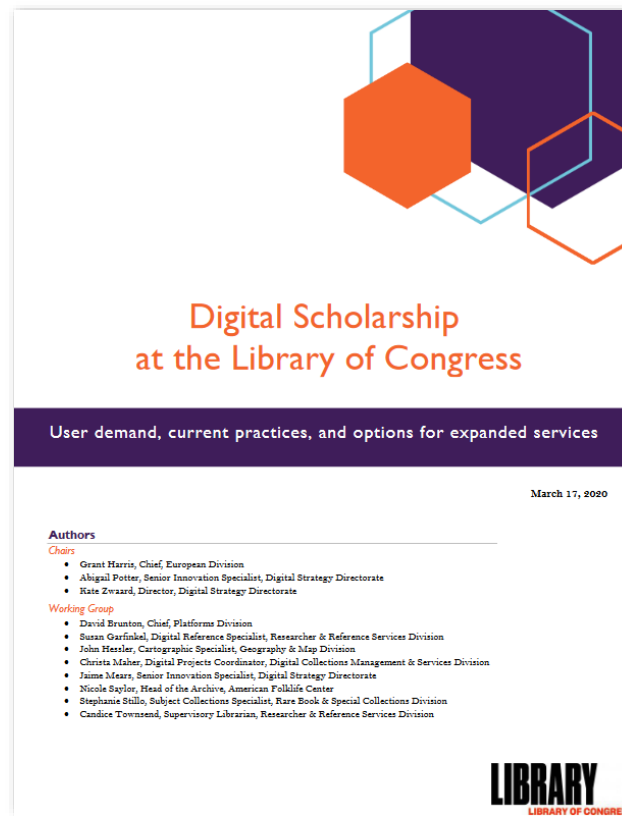
## The Santa Barbara Statement on Collections as Data

*Version 2*

*The Santa Barbara Statement on Collections as Data was written by the Institute of Museum and Library Services supported Always Already Computational: Collections as Data project team. The first version is based on the collaborative work of participants at the first Collections as Data National Forum (UC Santa Barbara, March 1-3 2017). After its release, the team gathered comments from the Hypothesis web annotation tool and sought additional feedback across a series of conversations and workshops (April 2017 - April 2018). The current version of the statement was revised based on that community feedback, especially the close, directed feedback provided by workshop participants at the Digital Library Federation Forum 2017.*

# „Digital Scholarship at the Library of Congress” (2020)

- Prioritize digital collection readiness: enable computational use + provide documentation
- Build institutional capacity: create CoP incl training, partnerships + develop ethical frameworks
- Expand user services: provide tools, services onsite + remote for DS practitioners



Harris, G. u.a. (2020): Digital scholarship at the Library of Congress: User demand, current practices, and options for expanded services.

<https://labs.loc.gov/static/labs/work/reports/DHWWorkingGroupPaper-v1.0.pdf>



# „Mapping the Current Landscape of Research Library Engagement...” (2020)

- Develop data services that work for big data and small data across disciplines
- Provide and sustain machine-actionable collections
- Deliver data science education and consultation

## **Mapping the Current Landscape of Research Library Engagement with Emerging Technologies in Research and Learning: Advancing Digital Scholarship**

By Sarah Lippincott

Edited by Mary Lee Kennedy, Clifford Lynch, and Scout Calvert

July 6, 2020

/ ASSOCIATION  
OF RESEARCH  
LIBRARIES /

**cni**  
Coalition for Networked Information

**born-digital**  
RESEARCH + CONSULTING

**EDUCAUSE**

Lippincott, S. (2020): Mapping the current landscape of research library engagement with emerging technologies in research and learning: Advancing digital scholarship.  
<https://www.arl.org/resources/mapping-the-current-landscape-of-research-library-engagement-with-emerging-technologies-in-research-and-learning/>.

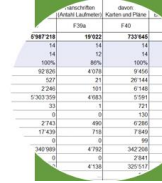
# Digital Scholarship Services

## Praxisbeispiele

- Collections as Data
- Förderung von Digital Literacy
- Spezifische Data Services
- Infrastruktur/Labs

# Collections as Data

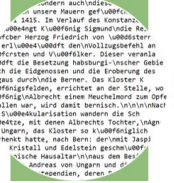
## Bibliothek als Datenprovider



Ländchen	oben	unten	...
(nicht katalogisiert)	Katalog	(nicht katalogisiert)	...
F49	F49	F49	...
5987218	19522	72946	644
14	14	14	14
14	14	14	14
1074	806	1000	1000
5270	4036	9549	12110
527	21	20141	6387
2780	161	6146	56722
5703709	4051	5781	601580
32	1	721	0
0	0	130	84102
2781	490	4780	13202
17430	758	7640	177
0	0	68	177
140389	4792	34220	37
0	0	284	0
0	0	261	0
0	0	50517	0



`"description": null,`  
`"author": "Paul Haupt Bern",`  
`"contributor": null,`  
`"date": "1949",`  
`"type": ["Text", "Journal Article"],`  
`"source": ["Bernese Zeitschrift für Geschichte und Heimatkunde", "280461-x", "0005-9420", "11", "1949"],`  
`"language": null,`  
`"coverage": null,`  
`"format": "application/pdf"`



...den auch hied...  
...unser Herrn geführ...  
...1415. In Verlauf des Konstanz...  
...überlegt Vuböfängig Sigmund...  
...fischer Herzog Friedrich von Vuböföster...  
...bersten und Vuböföfiker. Dieser veranla...  
...überbt die Besetzung habsburgischer Gebie...  
...durch die Eidgenossen und die Eroberung des...  
...Aargaus durch Vuböföfiker. Das Kloster f...  
...Vuböföföfiker, errichtet an der Stelle, wo...  
...Vuböföföfiker einen Heuchelord zum Opf...  
...gelalten war, wird demnach Vuböföföföföf...  
...ter Vuböföföföföföföföföföföföföföföföf...  
...hört hatte, nach dem derinist Jasp...  
...ngarn, das Kloster so Vuböföföföföföföföf...  
...hört hatte, nach dem derinist Jasp...  
...ngarn, das Kloster so Vuböföföföföföföföf...  
...hört hatte, nach dem derinist Jasp...  
...ngarn, das Kloster so Vuböföföföföföföföf...



## Eigene + lizenzierte + frei zugängliche Bestände

- Digitalisate, Volltexte, Metadaten etc. in offenen, spezifischen Formaten
- Detailbeschreibung und Nutzungsbedingungen
- Daten-Dumps + APIs zu Katalog und eigenen Sammlungen
- Dokumentation von Daten-APIs lizenzierter + wichtiger freier Ressourcen

# Collections as Data

## Formen/Gefässe

- Webpages zu DS, TDM, APIs
- interaktive, datenbasierte Bestandspräsentation und Datenexploration:
  - Jupyter Notebooks
  - Katalog/Suchinterface: [Media Monitoring of the Past](#)
  - Webpage: HathiTrust [Bookworm](#), ÖNB [Financial News](#), LNB [Latvian Prose Counter](#)
- Bereitstellung von Ground Truth Data aus eigenen Projekten
- Lizenzierung von TDM-Plattformen: z.B. Nexis Data Lab, Ithaka Constellate

## Text and Data Mining

Information on text and data mining resources available through the Library

Search Library Guide

Search

Home

Newspapers

Text Mining  
Newspapers

Available XML Files  
from ProQuest

Available XML Files  
from East View

Library of Congress  
Newspapers

NewsBank Newspapers

Scholarly Books and  
Journals

Citations and Metadata

Primary Source  
Collections

Online Textual Analysis

Linguistic Corpora

Government

### Text Mining Newspapers

Text mining of newspapers is now available through ProQuest's TDM Studio online service. This offers access to almost all sources in Global Newsstream and ProQuest Historical Newspapers. Our other news sources either don't allow text mining or have conditions on access.

Learn more about TDM Studio on our guide.

- [ProQuest TDM Studio](#)  
by [Greg Fleming](#) Updated Oct 13, 2021

### Available XML Files from ProQuest

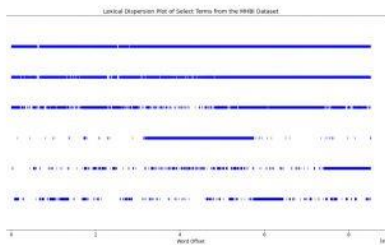
The Library has purchased XML files for the following newspapers from ProQuest. These files can be downloaded for local analysis.

These years are all that are available to us at this time due to publisher restrictions. Use the links to the ProQuest dataset from the Library Catalog in the links below to download the files.

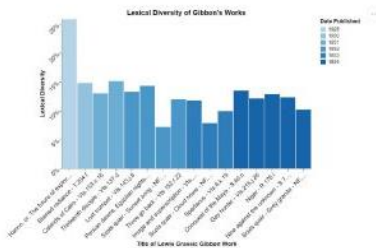
- [Baltimore Afro-American: 1893-1988](#)
- [Boston Globe: 1872-1983](#)
- [Guardian and Observer: 1791-2003](#)
- [Irish Times: 1856-1926](#)
- [Jerusalem Post: 1932-1976](#)

# Jupyter Notebooks

These Jupyter Notebooks provide initial, exploratory analysis of some of the Library's datasets. No prior programming experience is needed to access and use these Notebooks.



Exploring A Medical History of British India



Exploring Lewis Grassie Gibbon First Editions

```

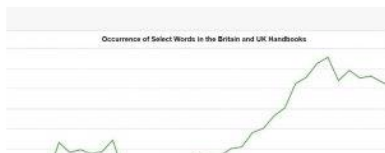
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load the data
df = pd.read_csv('data/19th_dataset.csv')

# Analyze the data
df['word_offset'] = df['word_offset'].astype(int)
df['word_offset'] = df['word_offset'].apply(lambda x: x - 1)

# Plot the data
plt.figure(figsize=(10, 5))
df['word_offset'].hist()
plt.title('Lexical Dispersion Plot of Select Terms from the 19th Dataset')
plt.xlabel('word_offset')
plt.ylabel('Count')
plt.show()
    
```

Exploring The National Bibliography of Scotland (version 1)



SEARCH ARTICLES

SEARCH IMAGES

NGRAMS

Swiss National Library

advertisement

\* 2 search filters ignored.

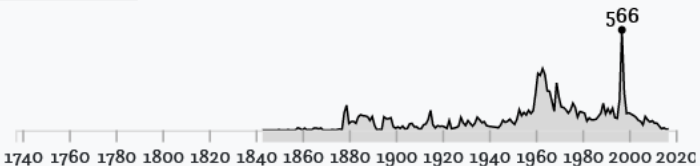
START FRESH, RESET FILTERS!

only results on the front page

PUBLICATION DATE

Number of articles per year

% SUM



ADD NEW DATE FILTER ...

FILTER BY LANGUAGE OF ARTICLES (3 OPTIONS)

check one or more language to filter results

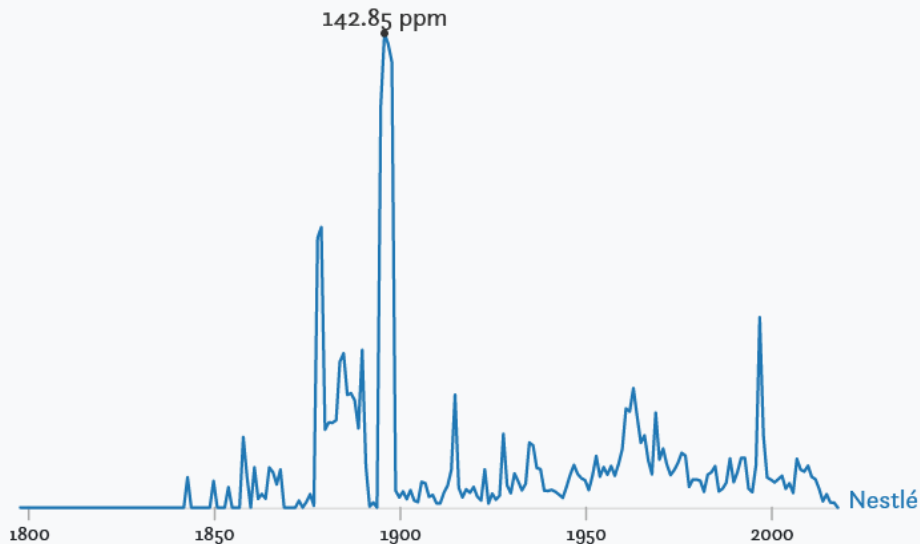
NGRAMS VIEWER

17,403 mentions of "Nestlé" in 10,771 articles - tagged as advertisement

SEE ARTICLES

Enter unigram

Nestlé



## UB Bern

## Collections as Data

## DS Webpage mit...

- TDM-Ressourcen: lizenzierte + frei zugängliche Datenbestände mit Data Sheets
- Dokumentation Anbieter-APIs
- Tools zur Datenarbeit
- Service-Angebot etc.

Universitätsbibliothek

UB Recherche Service Teilbibliotheken Über uns

Ausleihe  
Auskunft  
Open Science  
**Digital Scholarship**  
Kopien und Scans  
Arbeitsplätze  
Kurse und Beratung  
Ausstellungen und Veranstaltungen

## Digital Scholarship

Mit dem Begriff Digital Scholarship wird der Einsatz von digitalen, datenbasierten Methoden beschrieben: Datenanalyse, maschinelles Lernen und Big-Data-Technologien ermöglichen neue Forschungsansätze in allen Disziplinen. Die UB Bern unterstützt mit der Bereitstellung von Datenbeständen und Tools, vermittelt und berät bei deren Nutzung und führt selbst datenbasierte Projekte durch.

**KONTAKT**  
**Digital-Scholarship-Team**  
Telefon +41 31 684 92 02  
ds.ub@unibe.ch

**AKTUELL**

**Aktuell**  
Das Nexis Data Lab ist eine Plattform, die die Medien-inhalte des Anbieters LexisNexis für Text- und Datamining (TDM) verfügbar macht. Der Textbestand umfasst aktuelle und zeitgeschichtliche Berichterstattung von 20'000 Quellen aus über 100 Ländern. Bis zu 100'000 Dokumente können als Korpus gleichzeitig analysiert werden. Die Plattform bietet hierfür online eine Jupyter-Notebook-Umgebung und einfache Skripte für den Einstieg in das TDM. Bei Interesse zu Nutzung des Nexis Data Labs wenden Sie sich gerne an [uns!](#)

**TOOLS**

**Mehr**  
Werkzeuge für die Datenarbeit, digitale Analyse-Methoden und TDM

**DATENQUELLEN FÜR TEXT- UND DATAMINING**

**Mehr**  
Freie und lizenzierte Ressourcen für das Text- und Datamining (TDM)



Datenbank: Times Digital Archive

Provider: Gale Cengage

Times Digital Archive		
Access	Web address, API, Dumps, offline back up copy	<ul style="list-style-type: none"> <li>text-mining drives (includes directories, title manifests, XML files and image files, containing metadata, article segmentation, and page facsimiles (fee, available only for content the UB subscribes to or has purchased)</li> <li>User can create batches of specific issues or titles for bulk download through the Gale Digital Scholar Lab (subscription service)</li> <li>API access is not available</li> </ul>
Documentation	Web address	<a href="https://link.gale.com/apps/TTDA?u=unibern">https://link.gale.com/apps/TTDA?u=unibern</a>
Distribution		<ul style="list-style-type: none"> <li>continuously</li> <li>one volume per year</li> </ul>
Scope	Content Purpose Field of use	<ul style="list-style-type: none"> <li>Times 1785-2014, newspaper archive plus precursors</li> <li>The Daily Universal register (1785-1787)</li> <li>The Times, or, Daily Universal Register (1788)</li> </ul>
Time, Place, Language	temporal, local reference	<ul style="list-style-type: none"> <li>1785-2014</li> <li>UK, universal</li> <li>English</li> </ul>
Data type	What are the basic data types?	<ul style="list-style-type: none"> <li>Facsimiles: TIFF</li> <li>Issue text files with structural mark up (pages, subdivided or zoned into articles): XML</li> <li>bibliographic information: XML, partly within issue text files</li> </ul>
Provenance, dependencies, accompanying material	original data source, manufacturer, data collection procedure, dependencies / links to other data sets / online resources, old versions	A DTD file is provided on the text-mining drives (not online) and the fields are comparable to those found in Dublin Core, MARC and other standard bibliographic standards The definitive dataset is kept in a proprietary XML format, known as the Gale Interchange Format or GIFT, and from this its text-mining and online datasets are derived.
Description Structured text data	Text markup or data structure e.g. TXT, XML, ALTO,	Each XML file contains bibliographic information for the entire issue, automatically zoned during the OCR process, with individual pages and articles are represented as child

		describes the features of the issue, pages, articles <ul style="list-style-type: none"> <li>from 2018: separate issue-level content data (XML)</li> </ul>
Description of databases, tabular data	data tables, existing / recommended data splits (e.g. training / test set)	n/a
Description of image formats	as precisely as possible (e.g. resolution, greyscale / bitonal)	<ul style="list-style-type: none"> <li>to 2007: 300 PPI bitonal TIFFs</li> <li>after 2007: 400 PPI</li> <li>no compression</li> </ul>
Standards, vocabularies	as precisely as possible: standards and vocabularies used	
Data quality: OCR; missing, incorrect, redundant data, noise	For example. OCR error rate, OCR process; different raw data available? Used software?	OCR confidence rating varies across the corpus. About a quarter of the corpus does not have an OCR confidence value associated with it.
Administration, cleanups,	e.g. handling of missing data, cutting, rescaling, NLP preprocessing, used software	Facsimiles: digital restoration was undertaken to reduce the appearance or impact of damaged pages, including manually cropping and cleaning and the insertion of digital titles or page numbers where needed.
Scope /Size	size of data records	1.6 mio pages, 11.8 articles
Metadata	Format/ Standards,	<ul style="list-style-type: none"> <li>bespoke metadata schema developed by Gale</li> <li>hand-keyed issue and article-level metadata</li> <li>until 2017: content + metadata (XML): machine-readable text appears within a single XML file per issue, surrounded by layered metadata that describes the features of the issue, pages, articles</li> <li>metadata fields: article title, article subheadings.</li> </ul>

Ressource	Inhalt	Detailinformationen
OpenAlex <a href="#">Dokumentation</a>	<ul style="list-style-type: none"> <li>• REST API mit Endpoints für 5 Entitäten: works, authors, venues (journals, repositories), institutions, concepts</li> <li>• Metadaten inkl. Zitationsdaten und Volltextlinks</li> <li>• Personendaten</li> <li>• <a href="#">Datenquellen</a> und <a href="#">Daten-Dump</a></li> <li>• <a href="#">R Wrapper</a></li> <li>• <a href="#">Jupyter Notebook</a> in der DS Toolbox</li> </ul>	frei zugänglich; CC0
CrossRef Dokumentation: <a href="#">Metadaten, Event Data</a>	<ul style="list-style-type: none"> <li>• Metadaten, z.T. mit Volltextlinks: diverse APIs, z.B. <a href="#">REST API</a> sowie <a href="#">Dumps</a></li> <li>• Event Data und Funder Data</li> <li>• Datenquellen: Verlage</li> <li>• <a href="#">Wrapper</a> für diverse Sprachen</li> <li>• API-Request für alle für TDM verfügbaren Publikationen (ca. 142'000)</li> <li>• <a href="#">Jupyter Notebook</a> in der DS Toolbox zur Nutzung von CrossRef</li> </ul>	frei zugänglich ( <a href="#">Etiquette</a> ); Metadaten: "free to use", Funder Data: CC0
Elsevier Dokumentation: <a href="#">Start, APIs</a>	<ul style="list-style-type: none"> <li>• Volltexte von Elsevier-Publikationen: Article (Full Text) Retrieval <a href="#">API</a> (max. 6000 records, verschiedene Formate)</li> <li>• Metadaten (inkl. Abstracts, Elsevier-Publikationen oder Scopus)</li> </ul>	gesamter Inhalt für TDM lizenziert (nicht-kommerzielle Nutzung); API Key nötig
Springer Nature Dokumentation: <a href="#">Start, APIs</a>	<ul style="list-style-type: none"> <li>• Volltexte (Springer, Nature, Imprints): via Artikel-Webpages (1 request/sec, XML)</li> <li>• Metadaten (inkl. Abstracts): APIs</li> </ul>	gesamter Inhalt für TDM lizenziert (nicht-kommerzielle Nutzung); User Key für APIs nötig
Wiley <a href="#">Dokumentation</a>	<ul style="list-style-type: none"> <li>• Volltexte und Metadaten (PDF, XML) via Artikel-Webpages</li> </ul>	gesamter Inhalt für TDM lizenziert (nicht-kommerzielle Nutzung); TDM Token nötig
PubMed Central <a href="#">Dokumentation</a>	<ul style="list-style-type: none"> <li>• Open Access Volltexte (XML, TXT, PDF): API, FTP</li> <li>• Accepted Author Manuscripts Volltexte (XML, TXT): FTP</li> </ul>	mind. TDM-Lizenz, CC-Lizenzen, z.T. auch mit kommerzieller Nutzung
Public Library of Science <a href="#">Dokumentation</a>	<ul style="list-style-type: none"> <li>• Metadaten: API</li> </ul>	frei zugänglich; CC BY

# Digital Literacy

## Inhalte

### **Data Literacy**

- Datengewinnung, -manipulation, -analyse und -visualisierung
- mit/ohne Programmierkenntnisse und mittels unterschiedlicher Tools

### **Computer Skills**

- Basics: Command Line, Programmiersprachen, Version Control, Virtualisierung
- Nutzung spezifischer Tools



## Digital Scholarship Resources for Courses

Reach out for help to integrate digital scholarship methodologies and projects into your courses.

Home

Digital Scholarship  
Methodologies

Archiving Code,  
Webpages, and  
Websites

Data Management

Network Analysis

Text Mining

**Web Scraping**

Introduction

Key Audience

Resources for web  
scraping

Open Access  
Resources for Text  
Mining

Public Digital Projects

### Introduction

Do you need to gather an original corpus on the web for your research? Do you have little to no programming expertise? This toolkit is designed for you. Please reach out to us if you need any assistance with any of the workflows.

#### What is web scraping?

Web scraping refers to an automated process of identifying components of a website, and copying (using a programming language) into another file or context. Web scraping is used when an automated process extracts information you need or in a format that you need.

#### A few things to know before you start.

1. There may be an easier, better way to get the data you need. See "Text Mining"
2. If you can't get the online data you need, web scraping is legal, in most cases. You should know that:
  - any data that is publicly available
  - commercial use of scraped data is illegal
  - you cannot scrape sites that require a login (e.g. a library catalog or Facebook)
3. Even if your scraping a site that is legal, you may need to use a "webpages" workflow in Beautiful Soup.

The screenshot shows the University of Washington Libraries website. The top navigation bar includes 'HOME', 'START YOUR RESEARCH', 'USE THE LIBRARIES', 'HELP & SUPPORT', 'ABOUT', and 'LIBRARIES'. A search bar is also present. The main content area is titled 'Research Guides' and features a sidebar with a table of contents. The 'Network Tools' section is highlighted, and the main content displays the 'Digital Scholarship Research Guide: Network Tools' page. This page includes a sub-header 'Popular Network Tools' and a section for 'Gephi', which is described as an open-source program for exploring network graphs. A list of requirements and examples is provided for using Gephi.

Home
Getting Started on Your Project
Copyright Basics
Annotation Tools
Data Visualization
Digital Book Publishing
Map & Timeline Tools
GIS & Data Mapping Tools
<b>Network Tools</b>
Project Presentation
Text and Data Mining
Open Resources for Projects
Research Data Management
Technical Help
Scholarly Publishing and Open Access

HOME / START YOUR RESEARCH / RESEARCH GUIDES / DIGITAL SCHOLARSHIP RESEARCH GUIDE / NETWORK TOOLS

### Digital Scholarship Research Guide: Network Tools

Tools and resources to get you started on your digital scholarship projects

#### Popular Network Tools

**Gephi**  
Gephi website

Gephi is an open source program to explore network graphs and node and link diagrams. The program helps identify clusters, algorithmically arranges the network for readability, and helps visualize change over time. The program also lets you customize the view with colors, labels, grouping, and filtering.

- What you'll need:
  - No programming needed but Java required
- Accepted file formats:
  - There is a data processing wizard, but the underlying data must contain a node table and an edge table that you can identify. There are **instructions** to help.
    - CSV
    - Excel
    - Additional supported graph formats
- Examples/gallery:
  - Gephi Flickr

# UB Bern

## Digital Literacy 1

### DS Toolbox

- Jupyter-Notebook-Tutorials z.B. für OCR, API-Datenzugänge
- Ergebnisse forschungsunterstützender ML-Programmierung
- Kollaboration mit Fachreferenten Historische Bestände + NaWi

ub-unibe-ch / ds-pytools Public

Code Issues Pull requests Actions Projects Wiki Security Insights

main ds-pytools / Literature\_handling / OpenAlex\_API.ipynb

k-woitas Update OpenAlex\_API.ipynb Latest commit 22b206b 10 d

1 contributor

5355 lines (5355 sloc) | 178 KB

### Using the OpenAlex API

- 1 📖 About OpenAlex
- 2 📖 Works endpoint
  - 2.1 Author and affiliation data
  - 2.2 Use the search parameter
  - 2.3 Access multiple Work entities
    - 2.3.1 Filter works by publication type
    - 2.3.2 Filter works by "venue": journal, repository etc
    - 2.3.3 Filter works by author & handle big result sets
    - 2.3.4 Group works due to a certain feature
    - 2.3.5 Combine filters and grouping
- 3 📖 Authors endpoint
  - 3.1 Use the search parameter

## Anlässe mit den Themen

- Data Cleaning in allgemeinem FDM-Kurs
- Einführungen in TDM für Sozialwissenschaften
- Harvesting mit APIs für Digital Humanities
- OpenRefine für UB-MA
- Präsentation TDM-Plattform

Introduction to Research Data Management: Data Cleaning

**Data inspection**  
...with visualization

Bi-/multivariate:

- (Grouped) Box plot
- Heatmap

14

**Data inspection**  
...with pandas-profiling

Interactive HTML [reports](#)

- Type inference, different correlations
- Essentials: unique, missing, most freq values, histograms
- Warnings: zeros, correlations, duplicates
- Quantile and descriptive statistics

Dataset statistics	
Number of variables	21
Number of observations	10300
Missing cells	143
Missing cells (%)	0.1%
Duplicate rows	0
Duplicate rows (%)	+0.1%
Total size in memory	1.7 MB
Average record size in memory	168.8 B

15

**TDM WORKFLOW:**  
4 ANALYSIS

**Supervised learning methods**

- Find the mathematical connection between certain data values to predict outcomes of similar new data

Use Cases:

- *Sentiment analysis:* Recognize the sentiment/mood of a text (think of reviews...)
- *Text classification:* Classify new texts into known groups
- *Named Entity Recognition:* Recognize named entities like places, persons etc.

17

**Start with reading!**

Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling

Plus Müller steuern... Max W. Colquhoun... Yuan Ting Lee... Anne Lippmann... Christian Fritzsche... Jan C. Meye

Energy: Patterns & Stories  
Dietmar Dinkler, 2019, 1000 pages

Visualization: 1950 1960 1970 1980 1990 2000 2010 2019

18

# Spezifische Data Services

Zum Beispiel...

- Datenkorpora: Beratung, Recherche, Erstellung, Aufbereitung von Korpora
- Infrastrukturen als DL: Plattformen/Webpräsenzen für digitale Artefakte/Sammlungen/Projekte
- Vermittlung spezifischer Datenkompetenzen, z.B. GIS, digitale Produktion
- Angebote zur Produktion/Kreation: Residencies/Fellowships, Hackathons



<b>Materialien:</b>	Kinder- und Jugendliteratur aus dem deutschen Sprachraum (D-A-CH), Erscheinungszeitraum 1801-1914
<b>Umfang:</b>	derzeit (11/2021) 146 METS-Dateien; Ziel bis zum Ende des Förderungszeitraums (05/2024): 5.000 METS-Dateien von monographischen Werken sowie in geringem Umfang auch Zeitschriften
<b>Spezifika:</b>	tiefgehende manuell erfasste Strukturdaten; teilweise zusätzliche sachliche Erschließung durch DDC-Notationen; Seiten überwiegend freigestellt
<b>Lizenzen:</b>	Public Domain Mark 1.0
<b>Links:</b>	StaBiKat, Projektseite
<b>Ansprechpartner:</b>	Carola Pohlmann, Sigrun Putjenter

## Deutsches Territorialrecht des 19. Jahrhunderts

<b>Materialien:</b>	Zwischen 1801 und 1900 erschienene Druckwerke zu den Partikularrechten deutschfremdsprachige Titel enthaltend)
<b>Umfang:</b>	10.150 METS-Dateien von monographischen Werken und Zeitschriften; ca. 85-90 % Bestands
<b>Spezifika:</b>	6 Titel fremder Bibliotheken enthalten Sachliche Erschließung durch ARI
<b>Lizenzen:</b>	Public Domain Mark 1.0
<b>Links:</b>	StaBiKat, Projektseite
<b>Ansprechpartner:</b>	Christian Mathieu

Helm, W. u.a. (2019): Distant Viewing-Forschung mit digitalisierten Kinderbüchern: Voraussetzungen, Herausforderungen und Ansätze. In: b.i.t.online: Bibliothek, Information, Technologie, 22. Jg., H. 2, S. 127–134. <https://pub.uni-bielefeld.de/record/2935321>



b  
UNIVERSITÄT  
BERN

## Distant Viewing-Forschung mit digitalisierten Kinderbüchern: Voraussetzungen, Herausforderungen und Ansätze

Wiebke Helm, Thomas Mandl, Sigrun Putjenter, Sebastian Schmideler und David Zellhöfer

### Distant Viewing als Chance für Bibliothek und Forschung

Bibliotheken und Archive haben in den vergangenen Jahren umfangreiche Digitalisierungen ihrer Altbestände durchgeführt und auf diese Weise einen Beitrag zur Sicherung des kulturellen Erbes geleistet. Neben der Bestandserhaltung aus konservatorischen Gesichtspunkten sollte es im Interesse der Beteiligten sein, die digitalisierten Kulturgüter einer breiten Nutzung zuzuführen. Ein Umstand, den auch das neue europäische Urheberrecht für die Anwendbarkeit von Text- und Data-Mining (TDM) insbesondere für Forschungseinrichtungen betont und damit wissenschaftliche Bibliotheken mit Anforderungen konfrontiert, die über das bisher übliche Maß hinausgehen. Denn neben der Online-Bereitstellung der Bestände sollte das Datenmaterial einen Mehrwert für künftige Forschungen bereithalten und die sich wandelnden Forschungsgewohnheiten der Benutzer\*innen einbeziehen, da der Stellenwert der Digital Humanities (DH)-Forschung weiter zunehmen wird. Gemeinsam müssen Möglichkeiten erarbeitet werden, digitalisierte Daten für TDM-Verfahren aufzubereiten und niedrigschwellig zur Verfügung zu stellen. Im vorliegenden Beitrag werden diese Anforderungen exemplarisch an einem Distant Viewing-Forschungsprojekt aus den Digitalen Geisteswissenschaften aufgezeigt, das Illustrationen in historischen Kinder- und Jugendsachbüchern einer automatischen Analyse un-

*Bildererkennung durch KI ist derzeit in aller Munde. Ob im Einsatz großer Player wie Google & Co. oder der medizinischen Diagnostik – die Auswertung visueller Materials spielt eine immer größere Rolle, auch im Bereich der Digital Humanities-Forschung. Nach jahrelanger Fokussierung auf den Text hat dieses Medium beim Digitalisieren von bibliothekarischen Altbeständen bisher vernachlässigt wurde, zeigt der folgende Anwendungsfall, der Desiderata bei der Retrodigitalisierung wie auch Grenzen bestehender Bibliothekskategorien und technischer Systeme in Bezug auf die nachträgliche Inklusion von fachwissenschaftlichen Metadaten und Machine Learning-Prozessen identifiziert. Aus diesen Erkenntnissen werden Aufgaben und Lösungsmöglichkeiten für die zukünftige Forschung abgeleitet.*

*Image recognition by AI is currently on everyone's lips. The evaluation of visual material plays an increasingly significant role – in everyone's use of large players such as Google & Co. or medical diagnostics as well as in the field of digital humanities research. After years of focusing on texts, it has now extended its studies to the image. To what degree this medium has been neglected in the digitisation of library collections is shown by the following application case. Desiderata in retro-digitalisation as well as limits of existing library categories and technical systems will be identified with regard to the inclusion of scientific metadata and machine learning processes at a later point in time. From these findings tasks and options for future research are being derived.*

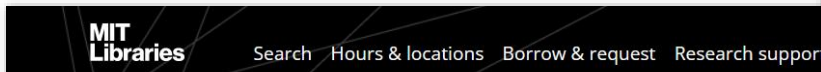
– auch im Hinblick auf zukünftige Nutzungen im Bereich der Digital Humanities.

**Das Projekt: Historische Kinder- und Jugendsachbuchillustrationen aus der DH-Perspektive**

Für das 19. Jahrhundert sind grundlegende Vorände.



# Infrastruktur/Labs



Libraries home » Data Services » GIS & Data Lab

## Data Services

### GIS & Data Lab

The GIS & Data Lab, housed on the 1st floor of [Rotch Library \(7-238\)](#), is available for use during [Rotch's operating hours](#). Computers and in-person help is available for the MIT Community. See [Data Services](#) for more info on how to get help.



Home > Center for Data and Visualization Sciences



About Us

Consulting

Upcoming Workshops

Online Learning

Data Sources

Lab

Blog

Contact Us

[askdata@duke.edu](mailto:askdata@duke.edu)  
Mailing list sign up  
[@dukedata](https://twitter.com/dukedata)

## Center for Data and Visualization Sciences



Attend a workshop



Get expert advice



Access computing

Online office hours or [chat with us](#) during the hours listed below.

Email [askdata@duke.edu](mailto:askdata@duke.edu) to schedule a consultation.

Mon 5/16 1:00pm – 3:00pm  
[Eric Monson](#)

Tue 5/17 1:00pm – 3:00pm  
[Drew Keener](#)

Wed 5/18 1:00pm – 3:00pm  
[John Little](#)

Thu 5/19 12:30pm – 2:30pm  
[Mark Thomas](#)

### SUPPORT AREAS

Data Sources

Data Science

Mapping and GIS

Data Management

Data Visualization

### Contact us

- GIS: [GIS help form](#)
- Data Management: [data-management@mit.edu](mailto:data-management@mit.edu)
- Statistics: [stat-help@mit.edu](mailto:stat-help@mit.edu)

### Our staff

- [Data & Specialized Services staff directory](#)

### Beispiel Korpus-Erstellung von Fachliteratur

- Ziel: mehrere Tsd. Volltexte nach thematischer WoS-Suche
- Vorgehen:
  - Metadatenabzug von WoS + Bereinigung
  - Vervollständigung mit Crossref-API
  - API-Abfragen bzw. Webscraping bei 6 Verlagen
  - Vervollständigung mit OpenAlex-API
- Ergebnis: ca. 4k Volltexte (380 MB), angereichert mit Affiliationen/Ländern

# UB Bern

## Organisation 2022

- Querschnittsthemen erfordern Kollaboration von verschiedenen BB!
- Projektmäßig/AG-mässige Kollaboration:
  - Collection as Data: UB-IT, Historische Bestände
  - Team Digital Toolbox: Fachreferate
  - UB-interne Datenprojekte: Bernensia, GND-Redaktion, IZ-Koordination
  - Projekte mit/für Forschende: Digitalisierung, Fachreferate, UB-IT

# Digital Scholarship + Bibliotheken

## Thesen 1

Datenkompetenzen werden als Schlüsselkompetenzen mittelfristig Bestandteil jedes (Hoch-)Schulbildungsprogramms sein.

UB als Informationsversorger/-vermittler an Universitäten müssen sich in ihren Produktportfolio und Services auf Daten als den zukünftig zentralen «Informationsgrundstoff» einstellen.

# Digital Scholarship + Bibliotheken

## Thesen 2

NLP/TDM-Methoden werden mittelfristig zum Standard-Handwerkszeug von Forschungszweigen werden, die mit Texten arbeiten.

Die hierfür nötigen Daten-Ressourcen, Technologien, Infrastrukturen, Data Services und Literacy-Angebote müssen damit Einzug in HSBen halten.

Analoges kann für Bild-, Audio-, Video-Daten gelten.

Vielen Dank für Ihre Aufmerksamkeit.

Ich freue mich auf Ihre Fragen!

Kathi Woitas, Universitätsbibliothek Bern

[kathi.woitas@unibe.ch](mailto:kathi.woitas@unibe.ch) - [ds.ub@unibe.ch](mailto:ds.ub@unibe.ch) - @library\_pirate

**u<sup>b</sup>**

b  
**UNIVERSITÄT  
BERN**



# Literatur

Borgman, C. L. (2007): Scholarship in the digital age: Information, infrastructure, and the internet. Cambridge, MA.

Greenhall, M. (2019): Digital Scholarship and the role of the research library: RLUK report. <https://www.rluk.ac.uk/digital-scholarship-and-the-role-of-the-research-library-an-rluk-report/>

Harris, G. u.a. (2020): Digital Scholarship at the Library of Congress: User demand, current practices, and options for expanded services. <https://labs.loc.gov/static/labs/work/reports/DHWorkingGroupPaper-v1.0.pdf>

Lippincott, S. (2020): Mapping the current landscape of research library engagement with emerging technologies in research and learning: Advancing digital scholarship. <https://www.arl.org/resources/mapping-the-current-landscape-of-research-library-engagement-with-emerging-technologies-in-research-and-learning/>

Padilla, T. u.a. (2019): Santa Barbara Statement on Collections as Data --- Always already computational: Collections as Data. <https://collectionsasdata.github.io/statement/>

Senseney, M. u.a. (2021): Transforming library services for computational research with text data: Environmental scan, stakeholder perspectives, and recommendations for libraries. <https://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/TransformingLibServices.pdf>