UNIVERSITÄT BERN

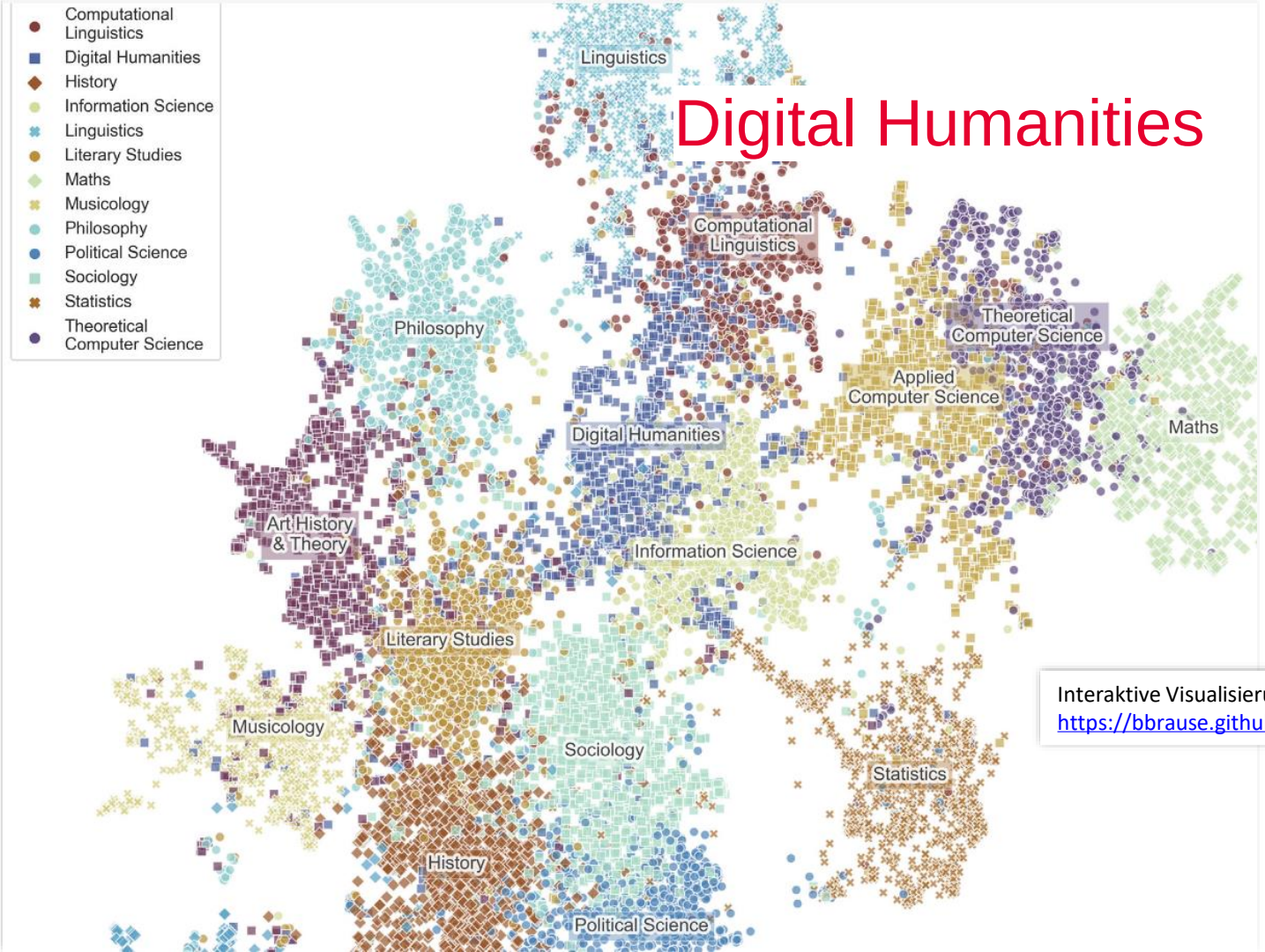
Digital Scholarship Services

Neue Dienstleistungen von wissenschaftlichen Bibliotheken für die datenbasierte Forschung

Kathi Woitas, Universitätsbibliothek Bern

01.06.2022 8. Bibliothekskongress Leipzig

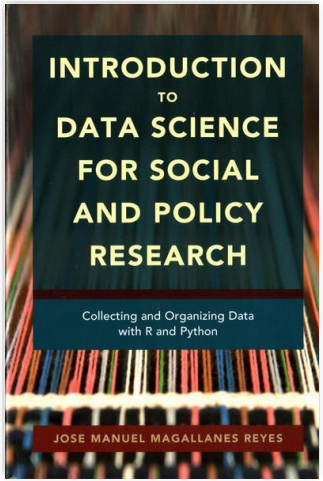
Digital Humanities

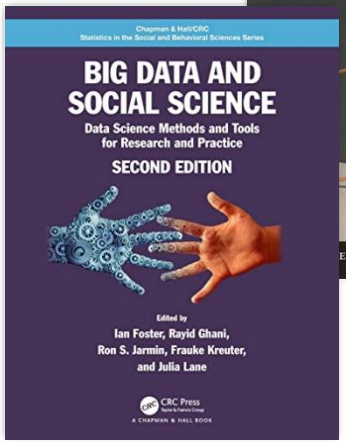


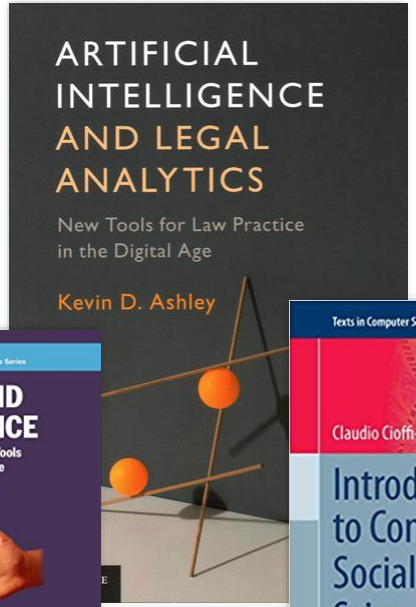
Luhmann, J., und M. Burghardt (2021): „Digital Humanities—A discipline in its own right? An analysis of the role and position of Digital Humanities in the academic landscape“. *Journal of the Association for Information Science and Technology*, S. 1-24. https://doi.org/10.1002/asi.24533

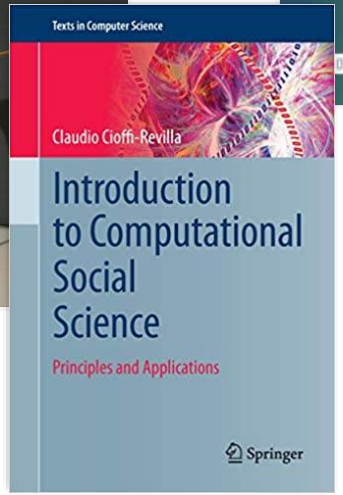
Interaktive Visualisierung: https://bbrause.github.io/dh_academic_landscape/

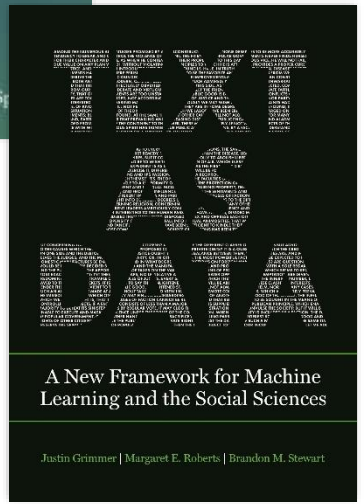
Gesellschaftswissenschaften











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 *et al.* [47]	Frequency of specific search terms on Google	Google Trends	Life satisfaction Emotional well-being	Longitudinal trends within one nation (United States)
Carlquist *et al.* [16]	Closed vocabulary (self-constructed lexicon)	Newspaper articles	Emotional well-being	Longitudinal trends within one nation (Norway)
Collins *et al.* [12]	Online activity; closed vocabulary (LIWC)	Facebook status updates and likes	Life satisfaction	Individual
Curini *et al.* [45]	Open vocabulary	Twitter	Emotional well-being	Italian provinces
Doré *et al.* [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 *et al.* [24]	Open vocabulary	Weibo posts	Emotional well-being	Individual
Hung *et al.* [34]	Mobile phone usage	Calling states, app usage	Emotional well-being	Individual
Jones *et al.* [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)

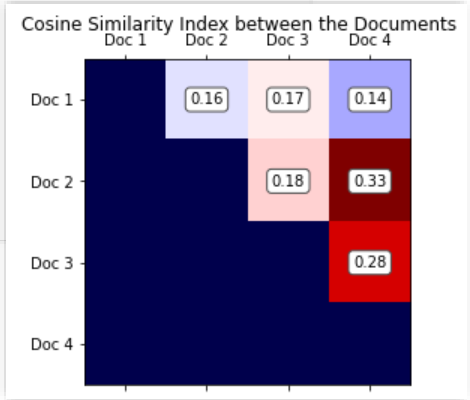
Computational Linguistics

UNIVERSITÄT BERN

```
# 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?

UNIVERSITÄT BERN

*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.*

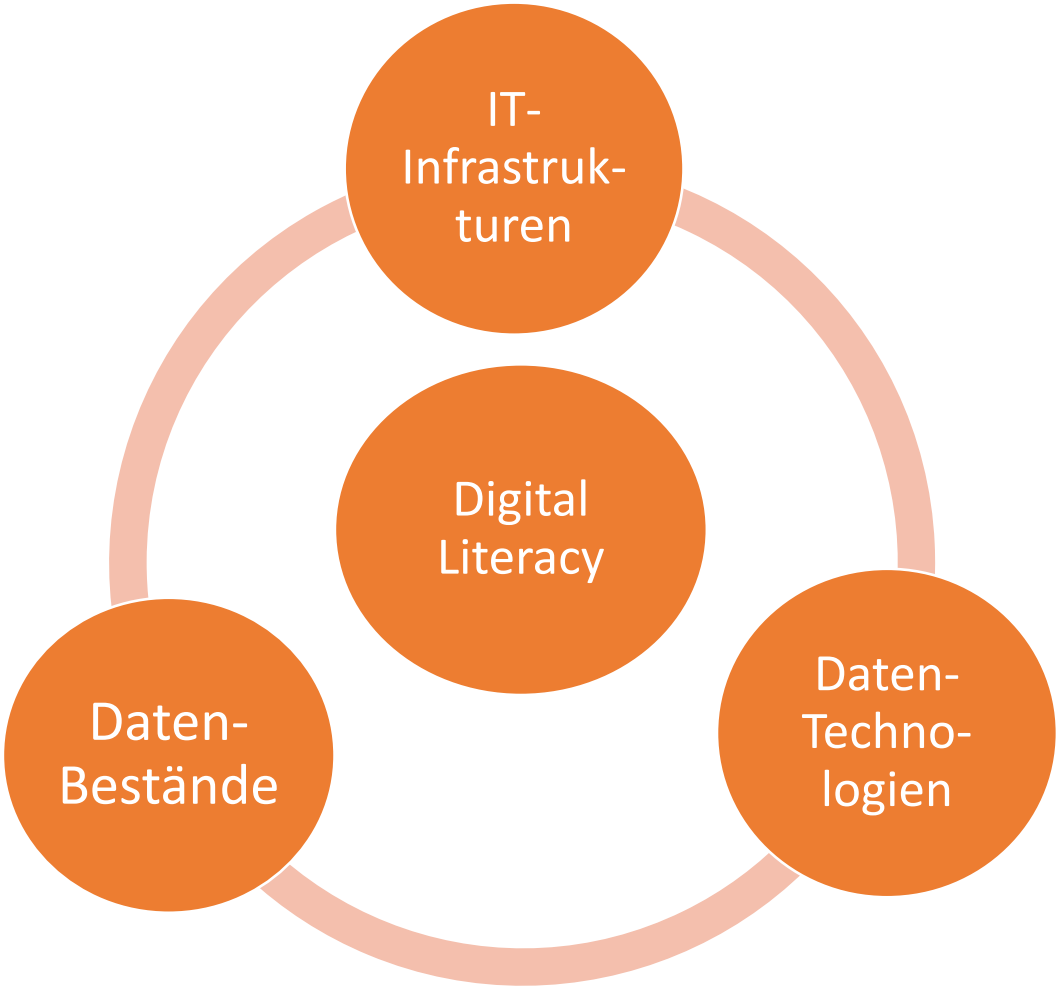
Senseney, M. u.a. (2021): Transforming library services for computational research with text data: Environmental scan, stakeholder perspectives, and recommendations for libraries. Glossary. https://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/TransformingLibServices.pdf

Digital Scholarship *Services?*

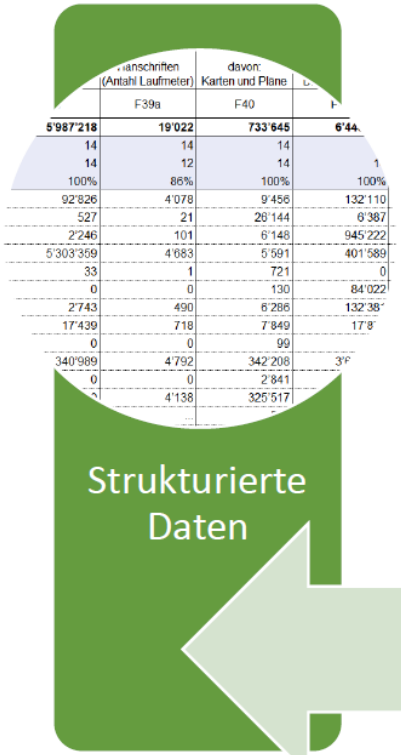
UNIVERSITÄT BERN

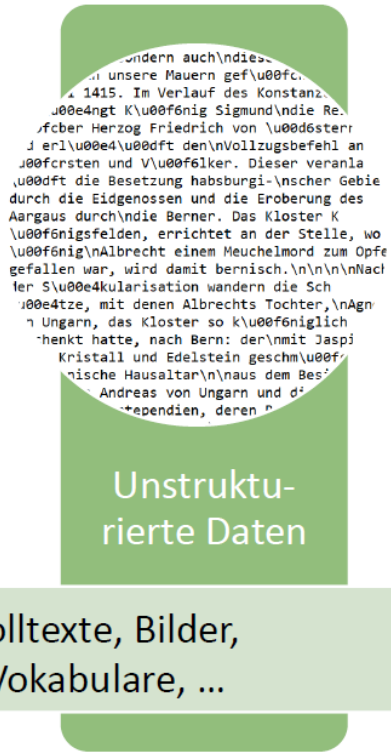
*This increasing emphasis on **data- and computationally intensive research methods** creates opportunities **for libraries** to contribute to the **education, tools, infrastructure, and communities** that sustain and expand these practices. Given the complexity of big-data analysis and the specialized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

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/. S. 18


IT-
Infrastruk-
turen
Digital
Literacy
Daten-
Bestände
Daten-
Techno-
logien

Datenbestände





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

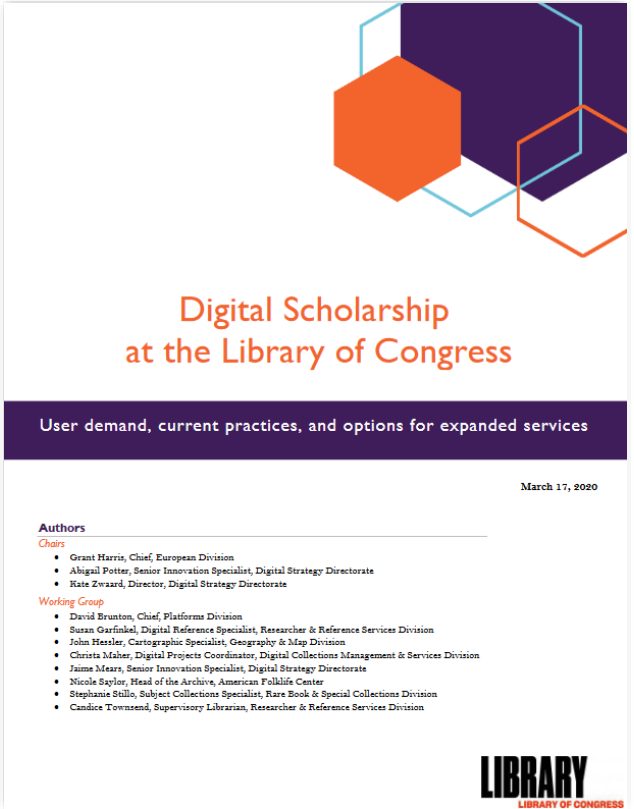
- **verteilte** Anwendungen für Datenhaltung, Datenintegration, Datenverarbeitung, -abfrage
- hoch skalierbar, ausfallsicher, performant

Digital Scholarship Services

Empfehlungen

„Digital Scholarship at the Library of Congress" (2020)

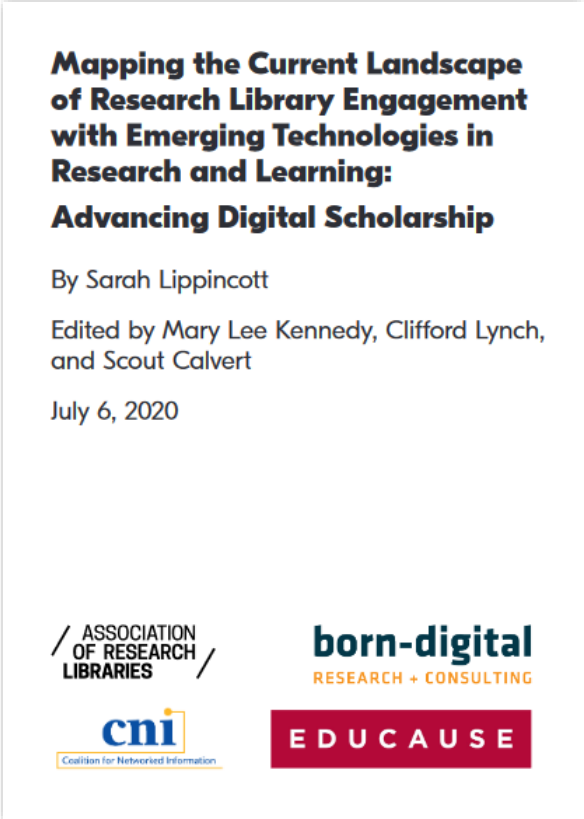
- Prioritize digital collection readiness: enable computational use + provide documentation
- Build institutional capacity: create CoP incl training, partnerships + develop ethical frameworks
- Expand user services: provide tools, services onsite + remote for DS practioners



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.

„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



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

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

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

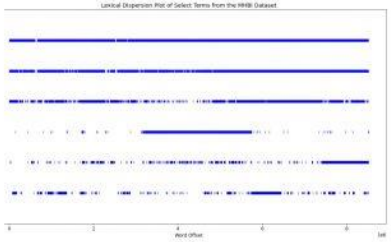
Available XML Files from ProQuest

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.

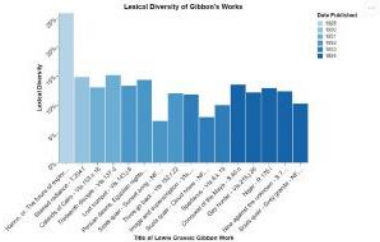
- Baltimore Afro-American: 1893-1988
- Boston Globe: 1872-1983
- Guardian and Observer: 1791-2003
- Irish Times: 1856-1926
- Jerusalem Post: 1932-1976

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.

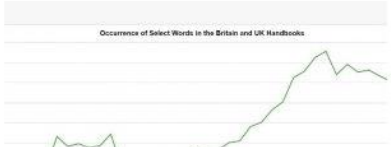


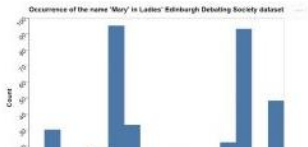
Exploring A Medical History of British India



Exploring Lewis Grassic Gibbon First Editions

Exploring The National Bibliography of Scotland (version 1)





SEARCH ARTICLES | SEARCH IMAGES | NGRAMS

Swiss National Library | advertisement

* 2 search filters ignored.

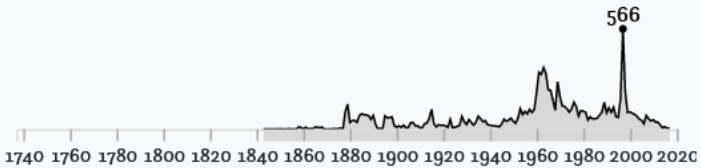
START FRESH, RESET FILTERS!

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)

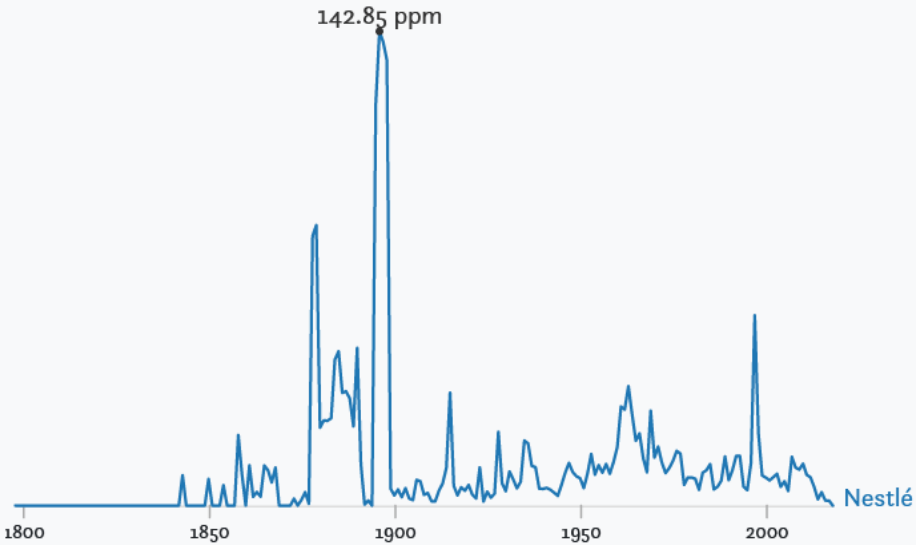
NGRAMS VIEWER

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

SEE ARTICLES

Enter unigram

Nestlé ×

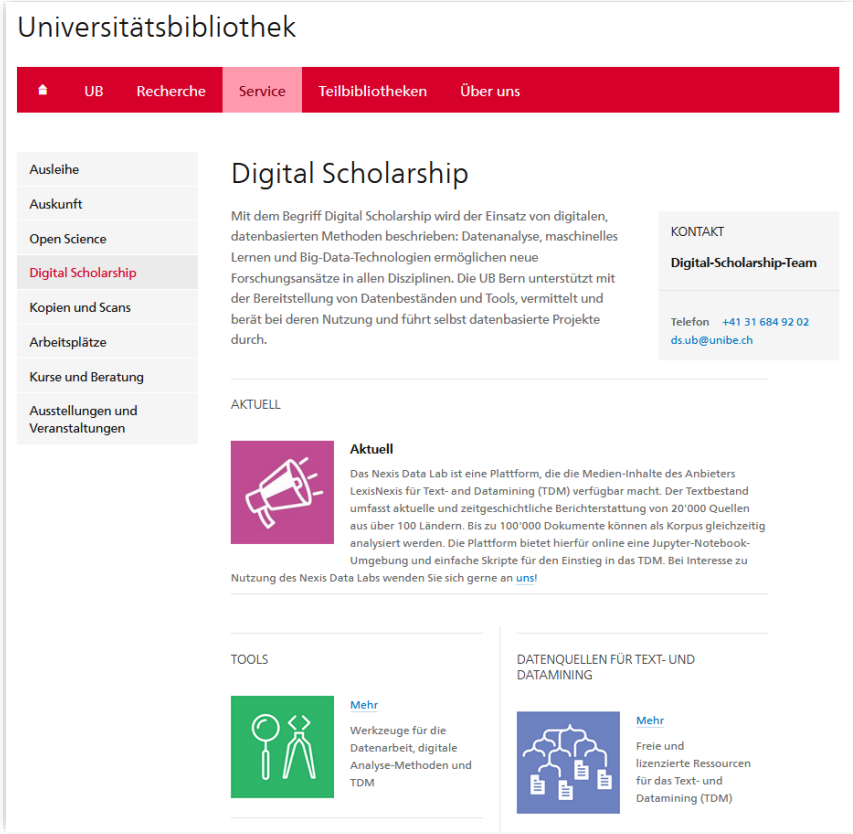


UB Bern

Collections as Data

DS Webpage mit…

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



Datenbank: Times Digital Archive

Provider: Gale Cengage

		Times Digital Archive
Access	Web address, API, Dumps, offline back up copy	• 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) • User can create batches of specific issues or titles for bulk download through the Gale Digital Scholar Lab (subscription service) • API access is not available
Documentation	Web address	https://link.gale.com/apps/TTDA?u=unibern
Distribution		• continously • one volume per year
Scope	Content Purpose Field of use	• Times 1785-2014, newspaper archive plus precursors • The Daily Universal register (1785-1787) • The Times, or, Daily Universal Register (1788)
Time, Place, Language	temporal, local reference	• 1785-2014 • UK, universal • English
Data type	What are the basic data types?	• Facsimiles: TIFF • Issue text files with structural mark up (pages, subdivided or zoned into articles): XML • bibliographic information: XML, partly within issue text files
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 • from 2018: separate issue-level content data (XML)
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)	• to 2007: 300 PPI bitonal TIFFs • after 2007: 400 PPI • no compression
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,	• bespoke metadata schema developed by Gale • hand-keyed issue and article-level metadata • 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 • metadata fields: article title, article subheadings,

Wissenschaftliche Publikationen und Metadaten | Ressourcen anderer Anbieter (Nationalbibliotheken, usw.)

Ressource	Inhalt	Detailinformationen
OpenAlex Dokumentation	• REST API mit Endpoints für 5 Entitäten: works, authors, venues (journals, repositories), institutions, concepts • Metadaten inkl. Zitationsdaten und Volltextlinks • Personendaten • Datenquellen und Daten-Dump • R Wrapper • Jupyter Notebook in der DS Toolbox	frei zugänglich; CC0
CrossRef Dokumentation: Metadaten, Event Data	• Metadaten, z.T. mit Volltextlinks: diverse APIs, z.B. REST API sowie Dumps • Event Data und Funder Data • Datenquellen: Verlage • Wrapper für diverse Sprachen • API-Request für alle für TDM verfügbaren Publikationen (ca. 142'000) • Jupyter Notebook in der DS Toolbox zur Nutzung von CrossRef	frei zugänglich (Etiquette); Metadaten: "free to use", Funder Data: CC0
Elsevier Dokumentation: Start, APIs	• Volltexte von Elsevier-Publikationen: Article (Full Text) Retrieval API (max. 6000 records, verschiedene Formate) • Metadaten (inkl. Abstracts, Elsevier-Publikationen oder Scopus)	gesamter Inhalt für TDM lizenziert (nicht-kommerzielle Nutzung); API Key nötig
Springer Nature Dokumentation: Start, APIs	• Volltexte (Springer, Nature, Imprints): via Artikel-Webpages (1 request/sec, XML) • Metadaten (inkl. Abstracts): APIs	gesamter Inhalt für TDM lizenziert (nicht-kommerzielle Nutzung); User Key für APIs nötig
Wiley Dokumentation	• Volltexte und Metadaten (PDF, XML) via Artikel-Webpages	gesamter Inhalt für TDM lizenziert (nicht-kommerzielle Nutzung); TDM Token nötig
PubMed Central Dokumentation	• Open Access Volltexte (XML, TXT, PDF): API, FTP • Accepted Author Manuscripts Volltexte (XML, TXT): FTP	mind. TDM-Lizenz, CC-Lizenzen, z.T. auch mit kommerzieller Nutzung
Public Library of Science	• Metadaten: API	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


BROWN UNIVERSITY LIBRARY
Search BruKnow
Search
All Guides/ Digital Scholarship Resources for Courses
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 proc
identifying components of a website, and
programming language) into another file o
contexts.Web scraping is used when an A
information you need or in a format that yo
A few things to know before you start.
1. There may be an easier, better way
Text Mining"
2. If you can't get the online data you n
scraping is legal, in most cases. You
any data that is publicly availab
commercial use of scraped dat
you cannot scrape sites that re
catalog or Facebook)
3. Even if your scraping a site that is le
webpages" workflow in Beautiful Sou

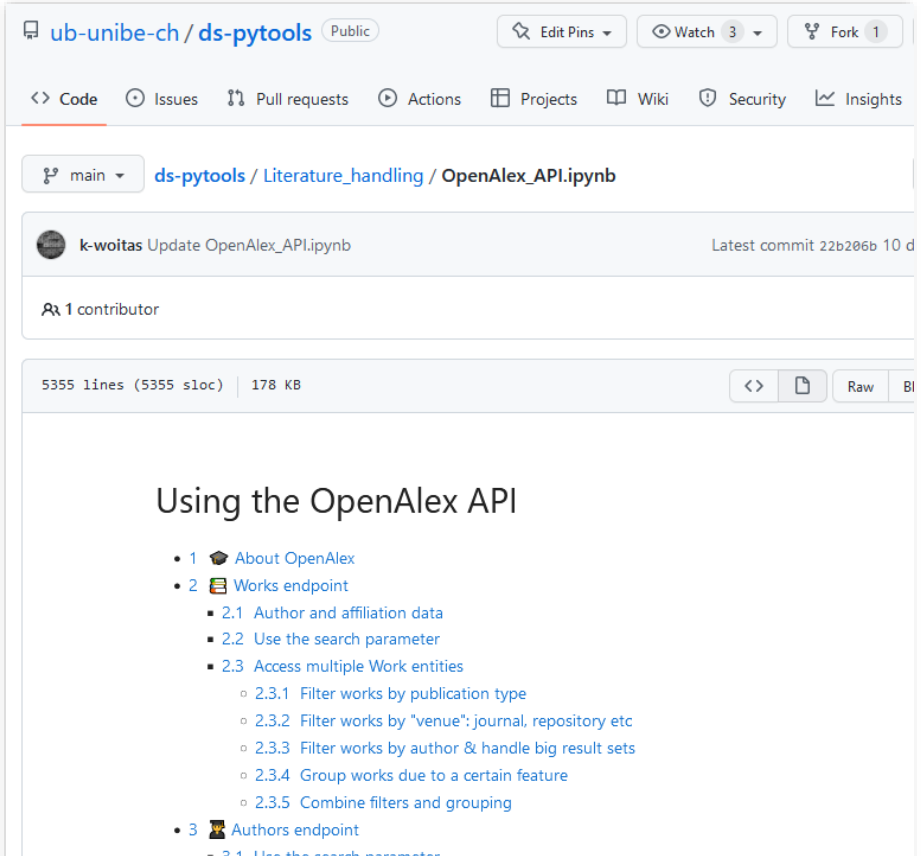
UNIVERSITY LIBRARIES
UNIVERSITY of WASHINGTON
My UW
Calendar
Maps
Directories
HOME
START YOUR RESEARCH
USE THE LIBRARIES
HELP & SUPPORT
ABOUT
LIBRARIES
ask us!
Research Guides
Home
Getting Started on Your Project
Copyright Basics
Annotation Tools
Data Visualization
Digital Book Publishing
Map & Timeline Tools
GIS & Data Mapping Tools
Network Tools
Project Presentation
Text and Data Mining
Open Resources for Projects
Research Data Management
Technical Help
Scholarly Publishing and
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 Flicker

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

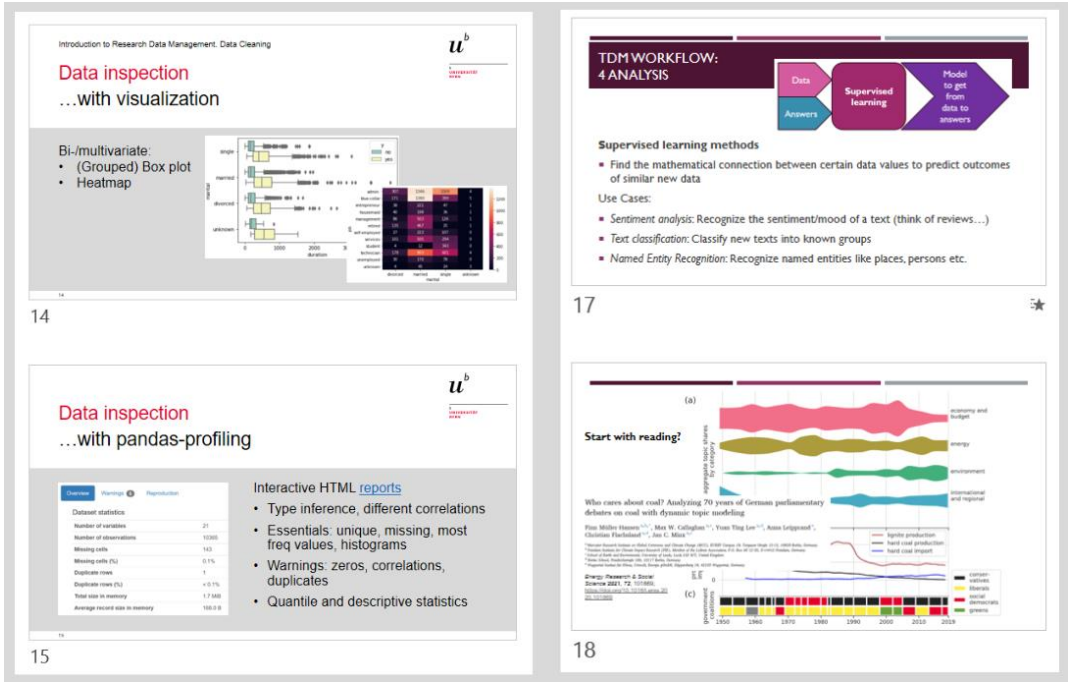


UB Bern

Digital Literacy 2

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



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

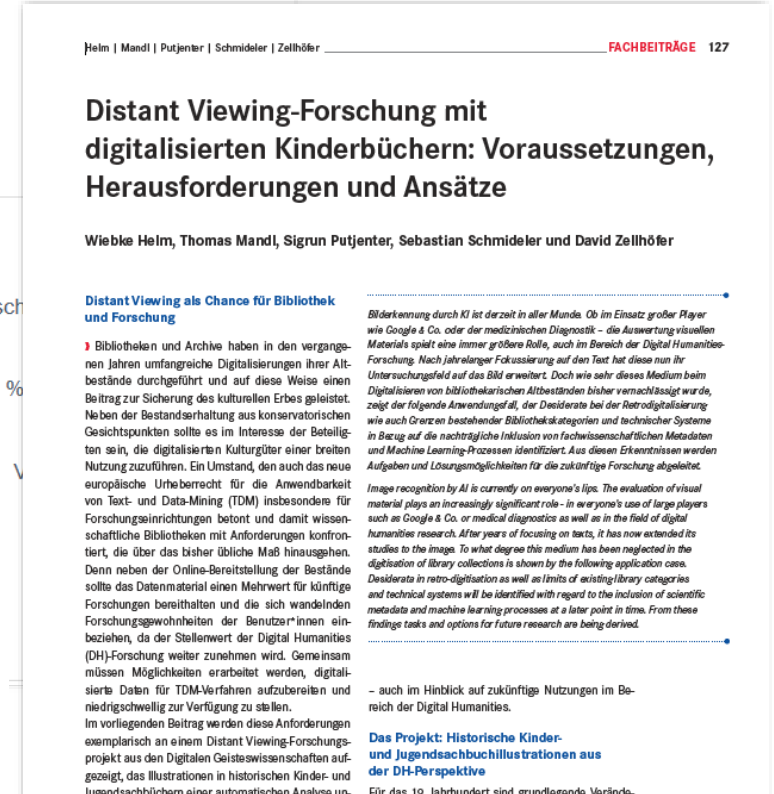
Colibri

Materialien:	Kinder- und Jugendliteratur aus dem deutschen Sprachraum (D-A-CH), Erscheinungszeitraum 1801-1914
Umfang:	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
Spezifika:	tiefgehende manuell erfasste Strukturdaten; teilweise zusätzliche sachliche Erschließung durch DDC-Notationen; Seiten überwiegend freigestellt
Lizenzen:	Public Domain Mark 1.0
Links:	StaBiKat, Projektseite
Ansprechpartner:	Carola Pohlmann, Sigrun Putjenter

Deutsches Territorialrecht des 19. Jahrhunderts

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



Helm | Mandl | Putjenter | Schmideler | Zellhöfer FACHBEITRÄGE 127

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

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

Bilderkennung durch KI ist derzeit in aller Munde. Ob im Einsatz großer Player wie Google & Co. oder der medizinischen Diagnostik – die Auswertung visuellen Materials spielt eine immer größere Rolle, auch im Bereich der Digital Humanities-Forschung. Nach jahrelanger Fokussierung auf den Text hat diese nun ihr Untersuchungsfeld auf das Bild erweitert. Doch wie sehr dieses Medium beim Digitalisieren von bibliothekarischen Altbeständen bisher vernachlässigt wurde, zeigt der folgende Anwendungsfall, der Desiderate 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-digitisation 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.

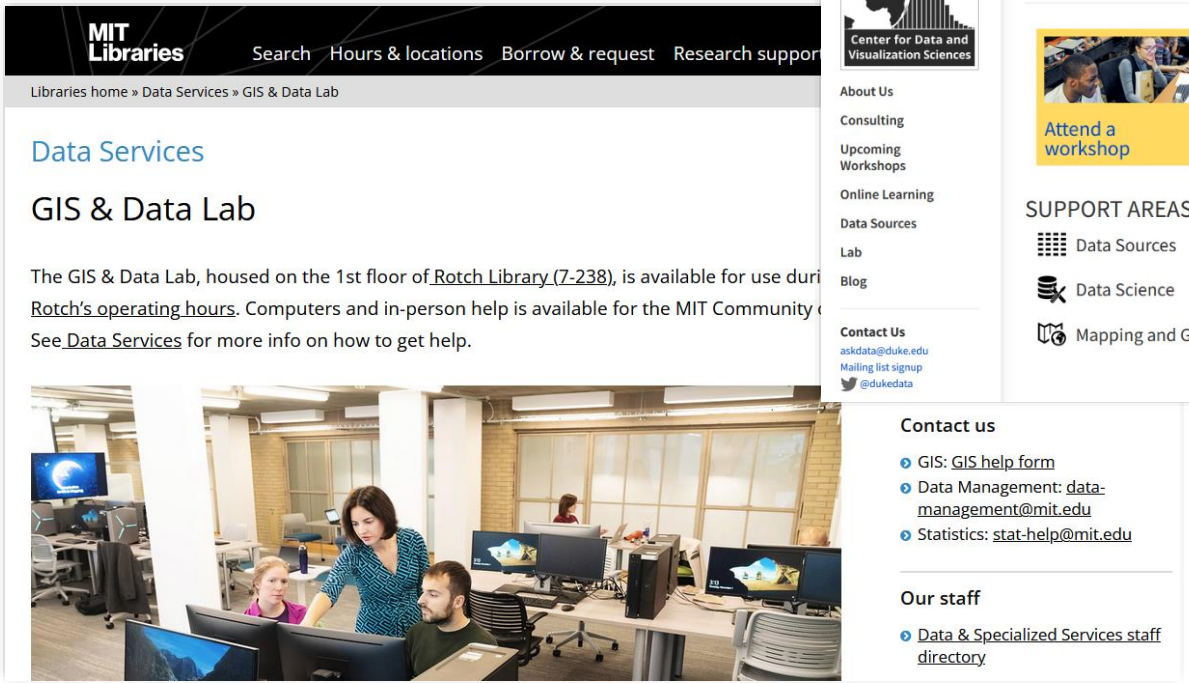
Distant Viewing als Chance für Bibliothek und Forschung

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

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

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

Infrastruktur/Labs



UB Bern

Data Services

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 nötigten 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 - ds.ub@unibe.ch - @library_pirate

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