Forschungsdaten-management

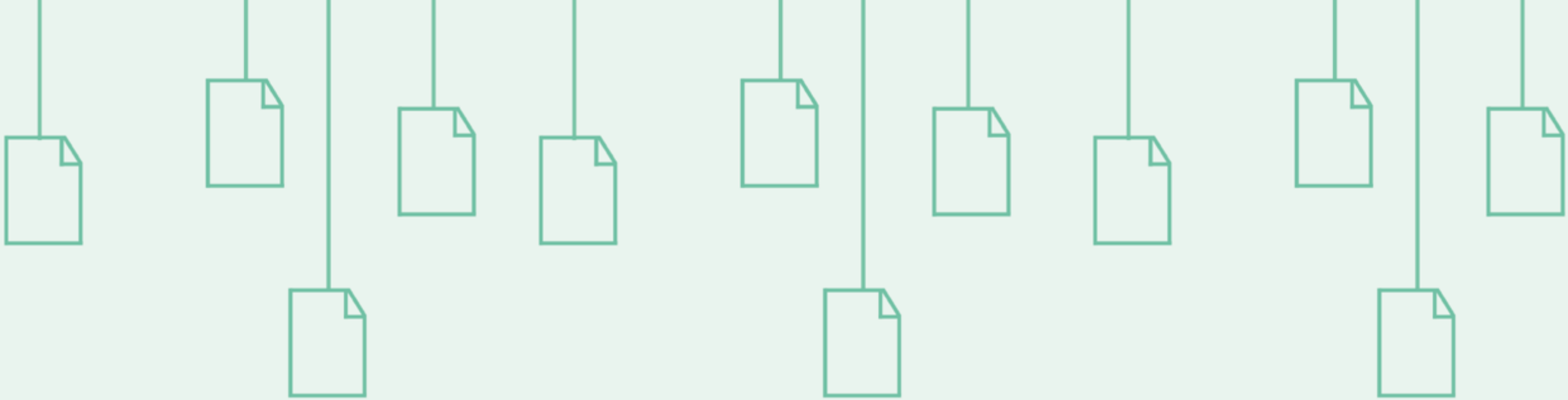
in den Geisteswissenschaften

15. Juni 2026

Referent: Dr. Patrick Helling

DCH

Universität zu Köln



Agenda

- Datentypen und –formate in den Geisteswissenschaften
- Repositorien und Infrastrukturen
- Umgang mit lebenden Systemen

Forschungsdaten?

„Research data is any information that has been collected, observed, generated or created to validate original research findings.“

https://library.leeds.ac.uk/info/14062/research-data-management/61/research-data-management-explained

Typen von Forschungsdaten

- Dokumente, Tabellen
- Feldforschungsnotizen
- Umfragen
- Audio, Video
- Präsentationsfolien
- Sammlungen
- Modelle
- Skripte
- …

DCH

Universität zu Köln

Forschungsdatenmanagement?

Backup

Datenorganisation durch spezifische Ordnerstruktur

Definition einer Lizenz für ein Github-Repositorium

Bachelorarbeit_vs_final_final.docx

Daten teilen mit Dropbox

DCH

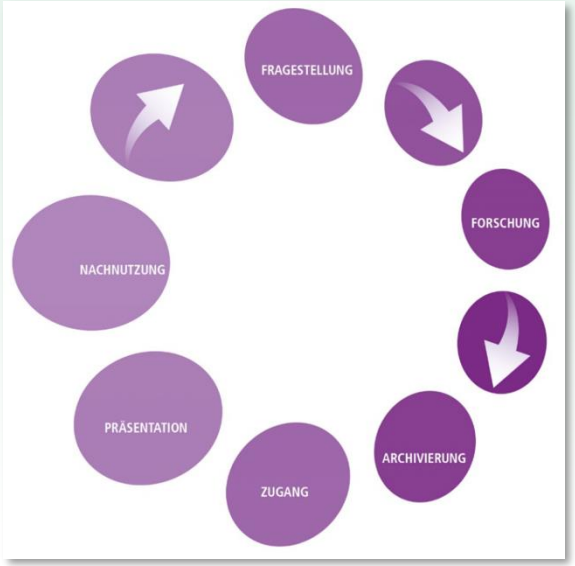
Universität zu Köln

Forschungsdatenmanagement?

„Research data management (RDM) refers to the organisation, storage and preservation of data created during a research project. It covers initial planning, day-to-day processes and long-term archiving and sharing.“

https://www.sheffield.ac.uk/library/rdm/whatisrdm

Umgang mit Forschungsdaten über den gesamten Forschungsdatenlebenszyklus hinweg.

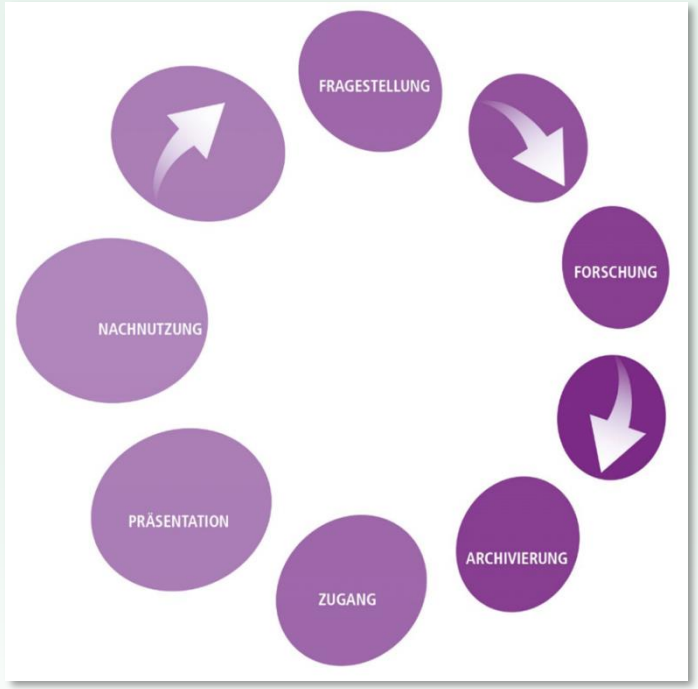


DCH

Universität zu Köln

Forschungsdatenmanagement?

Forschende
(Nachnutzung Forschungsdaten)



Forschende
(erzeugen Forschungsdaten)

Datenarchive/ Repositorien /
Infrastrukturen

DCH

Universität zu Köln

Ein typischer Montag an der Universität…

Ich lese einen Artikel in einem Fachjournal. Das Thema ist sehr interessant für meine eigene BA-Arbeit und ich würde gerne Zugang zu den Forschungsdaten erhalten, die die Grundlage für die im Artikel beschriebenen Ergebnisse waren, um diese nachnutzen zu können. Allerdings kann ich die Daten online nicht finden (und sie sind auch nicht dem Artikel angehängt)

→ Ich muss die Hauptautorin kontaktieren

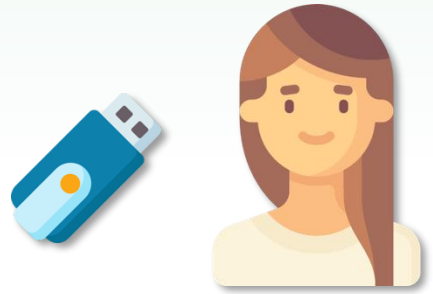
DCH

Universität zu Köln

Ein typischer Montag an der Universität…

„Die Daten wurden auf einem USB-Stick gespeichert. Ich bin gerade in ein neues Haus umgezogen und wohne aktuell noch aus den Umzugskartons. Der Stick sollte hier irgendwo in den Kartons sein. Sobald ich ihn finde, melde ich mich wieder.“

→ Ich muss geduldig sein



DCH

Universität zu Köln

6 MONTHS
LATER...

Ein typischer Montag an der Universität…

Endlich hab ich den USB-Stick per Post zugeschickt bekommen und ich kann auf die Forschungsdaten zugreifen. Allerdings sind die Daten in einem speziellen Format (.glp ??) abgelegt, das ich nicht kenne. Ich kenne auch kein Programm, mit dem ich das Datenformat öffnen könnte.

→ Ich muss die Hauptautorin noch einmal kontaktieren und sie darum bitten mir die Daten in einem anderen Format zur Verfügung zu stellen

DCH

Universität zu Köln

Ein typischer Montag an der Universität…

Bedauerlicherweise sind die Forschungsdaten allerdings nur in dem speziellen Format verfügbar. Die Autor:innen haben für die Generierung und Analyse der Daten eine kommerzielle Software genutzt. Das Unternehmen, dass die Software entwickelt hat, ist vor vier Jahren pleite gegangen. Es gibt keinen Support mehr, die Software ist nicht mehr verfügbar.

Glücklicherweise kenne ich noch eine alte und weise Professorin an meinem Institut, die noch eine Kopie der Software hat und einen alten Commodore 64, auf dem sie läuft. Endlich kann ich die Datei öffnen und mir genauer anschauen. Es handelt sich um eine Datenorganisation in Form einer Tabelle. Allerdings kann ich jetzt die Logik, nach der die Daten strukturiert sind, nicht so richtig verstehen.

→ Ich muss schon wieder die Hauptautorin kontaktieren und sie fragen, ob sie mir erklären kann, wofür die unterschiedlichen Akronyme und Beschreibungen, die die Daten kategorisieren und strukturieren, stehen. Es ist mittlerweile das dritte Mal, dass ich sie kontaktieren muss und sie ist so langsam recht genervt von mir...

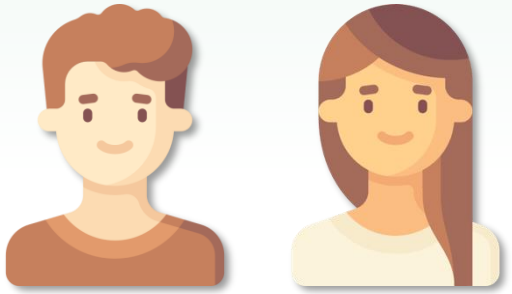
DCH

Universität zu Köln

Ein typischer Montag an der Universität…

Die Hauptautorin kann mir noch bei einigen meiner Fragen weiterhelfen, aber sie kann sich an die meisten Beschreibungsmuster nicht mehr erinnern. Es gibt auch keine weitere Dokumentation. Allerdings könnte mir einer der Co-Autoren des Artikels, der für das Datenmanagement in dem Projekt verantwortlich war, bei meinen Fragen weiterhelfen. Er war PhD-Student und Mitarbeiter in dem Projekt. Er hat seinen PhD mittlerweile beendet, hat aufgehört in der Wissenschaft zu arbeiten und ist zurück in die USA gezogen.

→ Ich würde selbstverständlich gerne den Co-Autoren kontaktieren, um die mir vorliegenden Daten endlich zu verstehen



DCH

Universität zu Köln

Ein typischer Montag an der Universität…

Die Hauptautorin steht mit dem Co-Autoren allerdings nicht mehr in Kontakt. Sie weiß nur noch, dass der nicht mehr an der Universität tätig ist, zurück in die USA gezogen ist und dort mittlerweile als Farmer arbeitet. Sein Name ist John Doe…

→ …???

DCH

Universität zu Köln

Was sind hier die Probleme…

Findability

Interoperability

Accessibility

Reusability

DCH

Universität zu Köln

Forschungsdaten müssen gesichert werden

- langfristige Archivierung von Forschungsdaten
  - Bitstream-Preservation → nachhaltige Sicherung *bit für bit*
  - Sicherstellung, dass man Forschungsdaten immer wiederherstellen kann

- Nutzung von Infrastruktur, die von verlässlichen Institutionen betrieben werden
  - organisatorisch, strukturell und finanziell

DCH

Universität zu Köln

Forschungsdaten müssen auffindbar sein

- langfristige Ablage in Repositorien o.Ä.

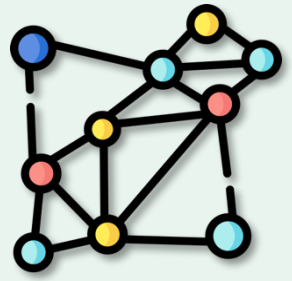
- abrufbar über standardisierte Protokolle, bspw. HTTPS
- referenzierbar über persistente Identifier, bspw. Digital Object Identifier (DOI)
- Beschreibung mit reichhaltigen (vorzugsweise) fachspezifischen Metadaten

    - Administrative Metadaten → bspw. Datenerstellung, Datenanalyse, technische Aspekte
    - Inhaltsbezogene Metadaten → bspw. Inhalt der Daten, Kontext der Daten
    - Bibliographische Metadaten → bspw. Autor:innen, Erscheinungsjahr, Projektname

Forschungsdaten müssen zugänglich sein

- So zugänglich wie möglich, so restriktiv wie nötig
- Klar kommunizierte und standardisierte Zugangsregelungen
  - Offen, embargo, restriktiv (Definition von Bedingungen)
- Sofern es Schranken in der Zugänglichkeit von Forschungsdaten gibt (bspw. durch Datenschutz oder Urheberrecht), sollten mindestens die Metadaten der Forschungsdaten zugänglich sein

Forschungsdaten müssen interoperabel sein



- Nutzung von möglichst offenen und standardisierten Datenformaten, die nicht von kommerzieller Software sind, bspw. .txt, .csv, .svg, .wav, .mpeg

- Nutzung von Open Source Software

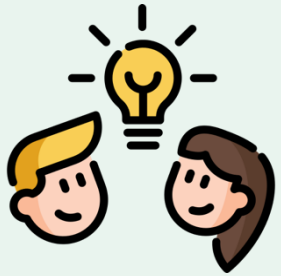
DCH

Universität zu Köln

Forschungsdaten müssen reusable sein

- Klar kommunizierte rechtliche Rahmenbedingungen bezüglich der Nachnutzung von Forschungsdaten
  - Nutzung von Lizenzen (sofern ein Urheberrecht vorhanden ist), e.g. Creative Commons Lizenzen
- Dokumentation von Daten und Software

Forschungsdatenmanagement – Warum?

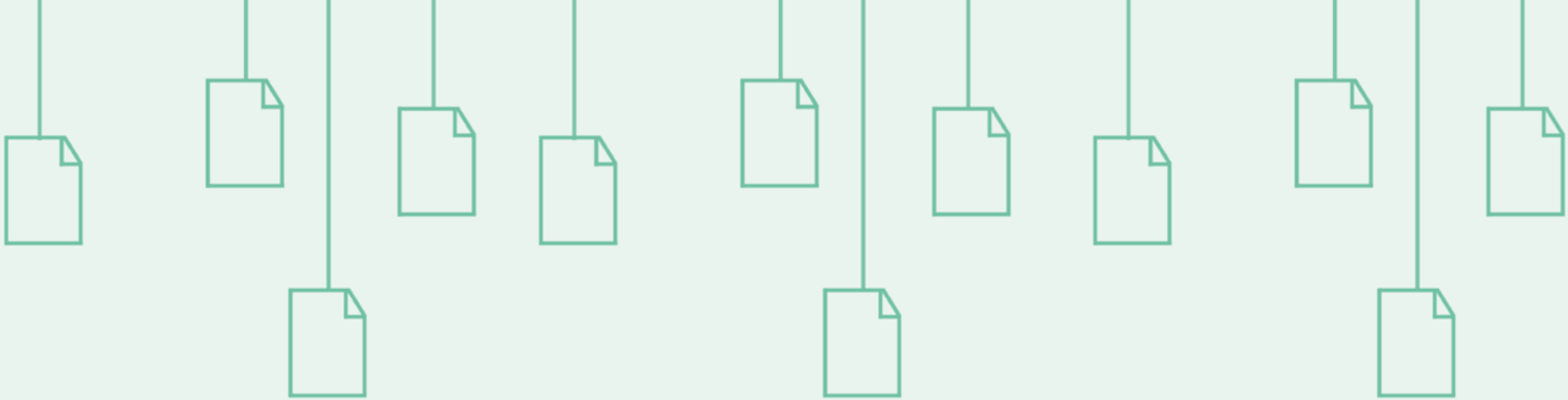


- Dokumentation der eigenen Forschung
- Verbesserung der eigenen Sichtbarkeit/Reputation
- Vergleichbarkeit von wissenschaftlichen Ergebnissen → Überprüfbarkeit
- Wir müssen das Rad nicht immer wieder neu erfinden
- Entwicklung neuer Forschungsfragen
- Potenzial für Kooperationen
- Überleben der eigenen Forschung

→ Beförderung wissenschaftlichen Fortschritts!

DCH

Universität zu Köln



Forschungsdatenmanagement in den Geisteswissenschaften

DCH

Universität zu Köln

Fachbereiche

- Alte Kulturen
- Geschichtswissenschaften
- Kunstgeschichte, Musikwissenschaften, Theater- und Medienwissenschaften
- Linguistik
- Literaturwissenschaften
- Sozial- und Kulturanthropologie, Nicht-Europäische Kulturen, Judaistik, Religionswissenschaften
- Philosophie
- Bildungsforschung
- Psychologie
- Sozialwissenschaften
- Wirtschaftswissenschaften
- Jura

+ 53
Sub-Bereiche

DCH

Universität zu Köln

icon: flaticon.com

Was Forschende in den Geisteswissenschaften tun…

graben

anschauen

beobachten

zuhören

aufnehmen

modellieren

annotieren

lesen

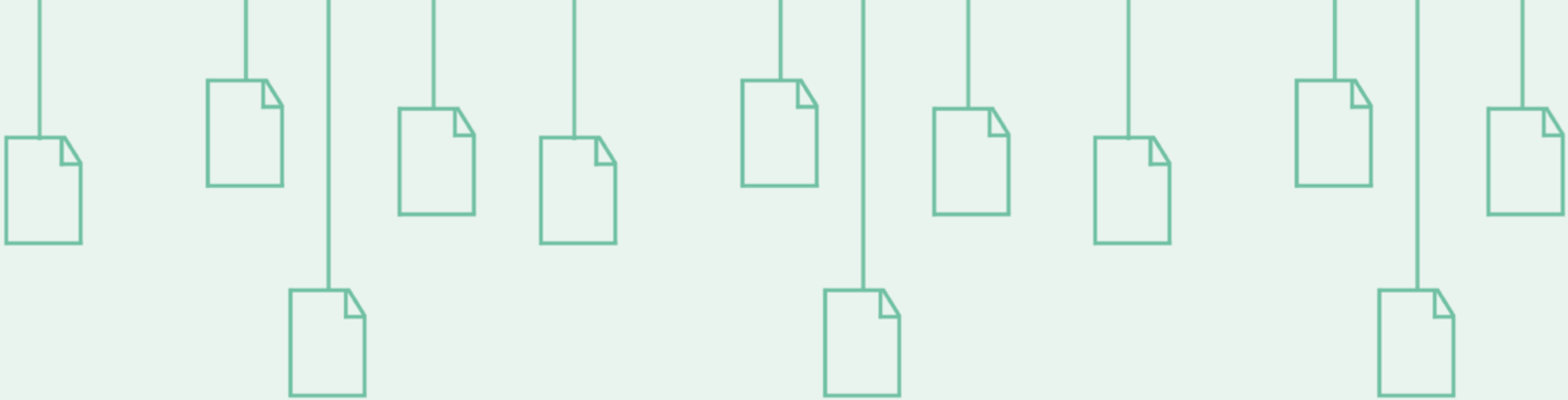
fotografieren

aufzeichnen

DCH

Universität zu Köln

icon: flaticon.com and smashicons.com



Datentypen und –formate in den Geisteswissenschaften

DCH

Universität zu Köln

Kurz und knapp - Daten, Datenmodelle und Schichten der Repräsentation

- Daten sind die Grundlage für wissenschaftliche Ergebnisse
- Rohdaten existieren nicht, da Daten immer in einem Kontext existieren
- Daten werden immer vorverarbeitet/prozessiert, um Forschungsfragen zu beantworten
- (Vor)verarbeitete Daten sind Modelle von Daten (= Datenmodelle), die die Basis von Forschung darstellen
- Daten können über unterschiedliche Schichten und Formate repräsentiert werden

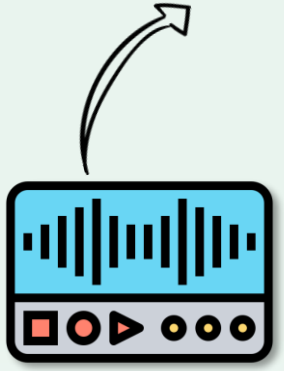
DCH

Universität zu Köln

icon: flaticon.com

Abbildungen

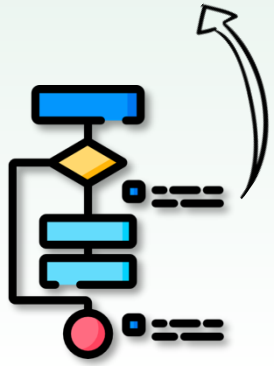
Audio

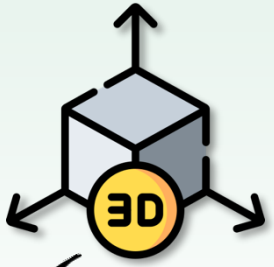


Musik-
partituren

Varianten von
Text

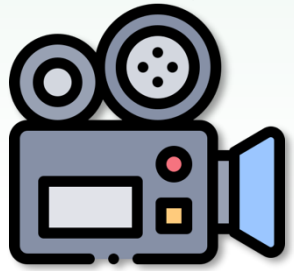
Algorithmen

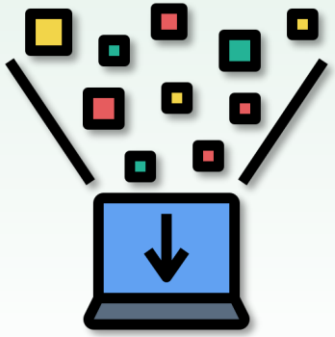




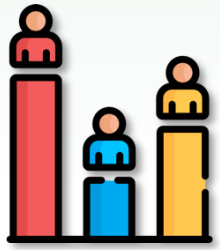
3D-Modelle

Video





Sammlungen



Umfragen

DCH

Universität zu Köln

icon: flaticon.com

Text

DCH

Universität zu Köln

.csv/.tsv

.html

OCR Formate

word-formats

.xml

.jpeg

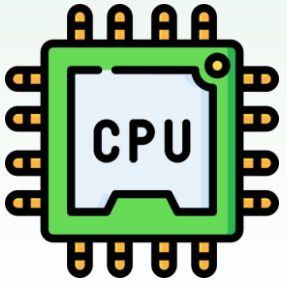
.rtf

.pdf/A

.txt

markdown

.json



DCH

Universität zu Köln

icon: flaticon.com

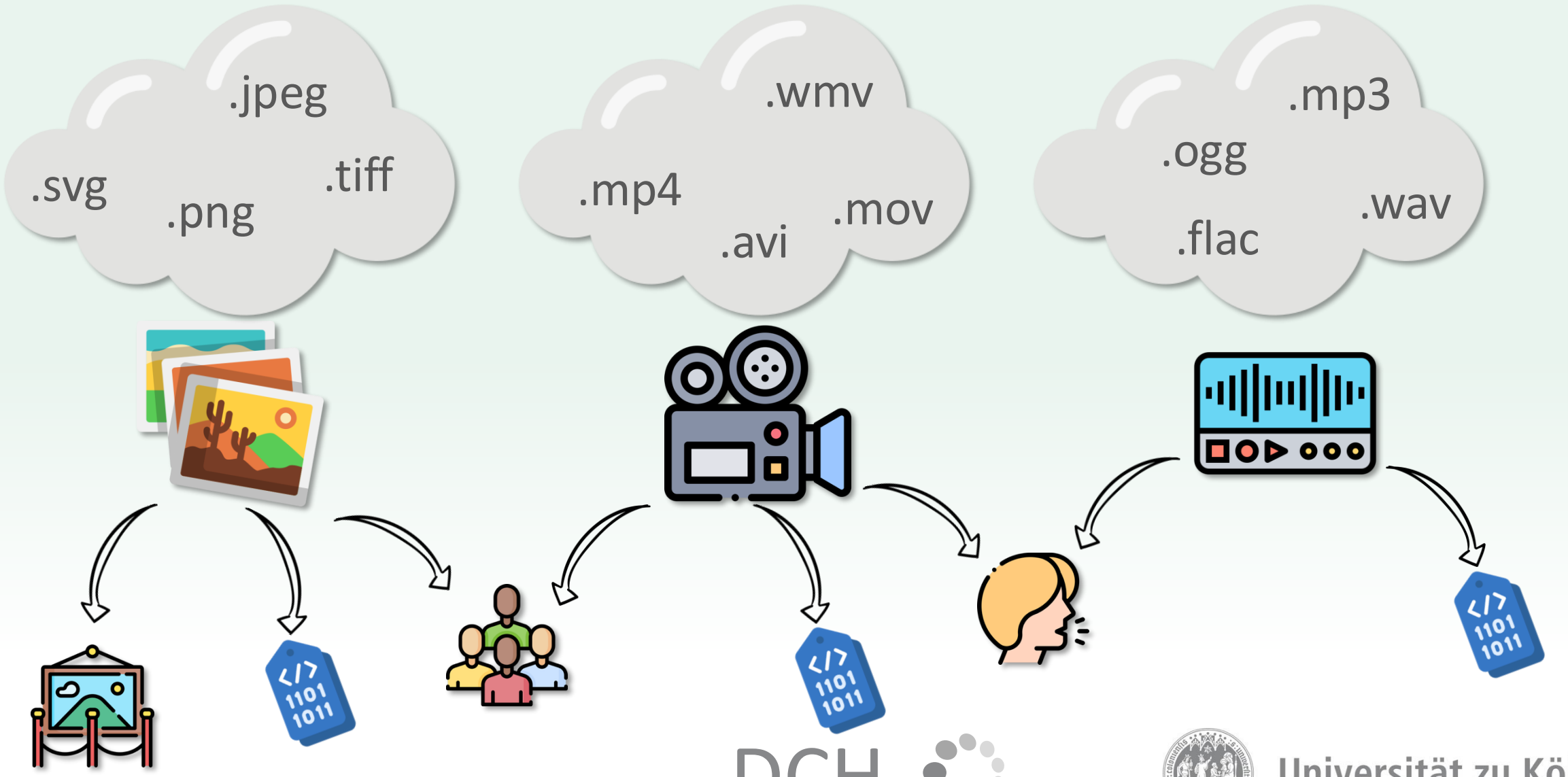
Spezifika von Text…

- Repräsentation eines Datentyps durch einen anderen Datentyp → digitales Scan/Bild eines gedruckten Textes
- Verschiedene Datentypen repräsentiert durch ein Datenformat, bspw. Bilder kombiniert mit Text in einem .pdf
- Repräsentation von Daten durch Markup-Sprachen, bspw. .xml
- Verschiedene Level von Analysierbarkeit, bspw. .csv vs. word-Formate

Bilder/Video/Audio

DCH

Universität zu Köln



DCH

Universität zu Köln

icon: flaticon.com

Spezifika von Bildern/Videos/Audios…

- Datenschutz
- Urheberrecht
- „unsichtbare“ Metadaten

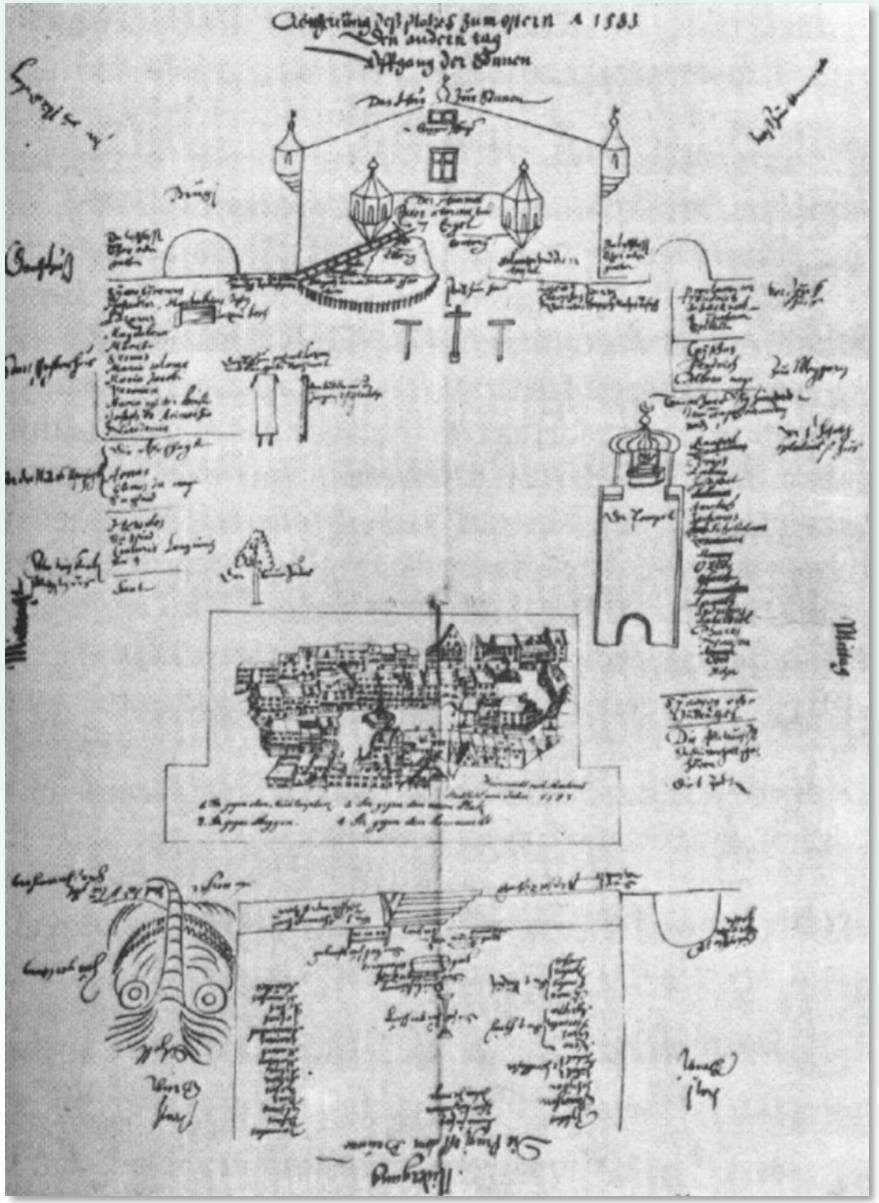
3D-Modelle

DCH

Universität zu Köln

.jpeg
.svg
.png

Plan der Passionsspiele in Luzern, 2. Tag, Renward Cysat (1583)



https://de.wikisource.org/wiki/Passionsspiele

DCH

Universität zu Köln

.jpeg
.svg
.png

Rekonstruktion der Bühne der Passionsspiele in Luzern, August am Rhym (1929).

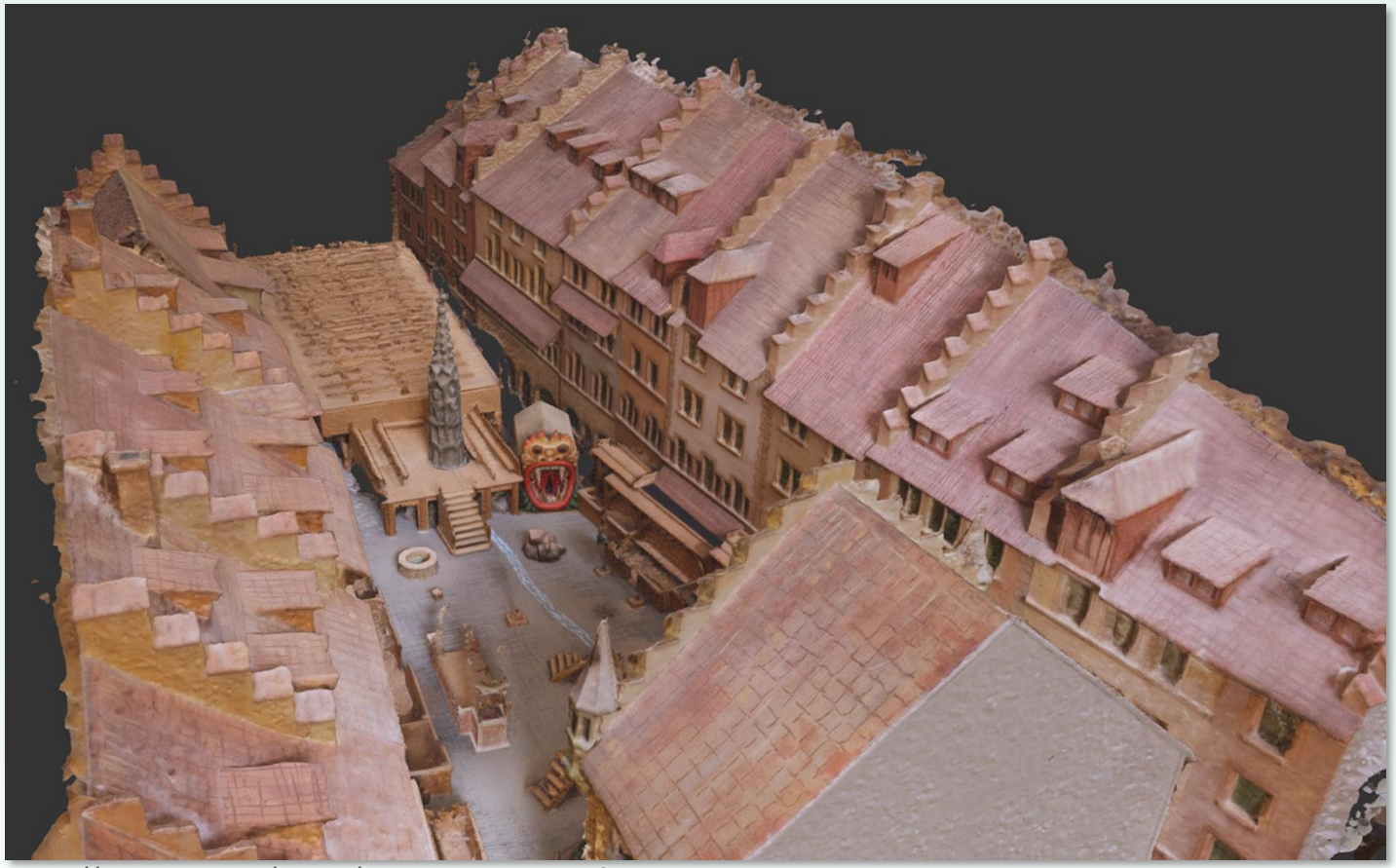
https://www.lu-wahlen.ch/uploads/media/4_Cysat_Theater.pdf

DCH

Universität zu Köln

.3ds
.stl
.obj

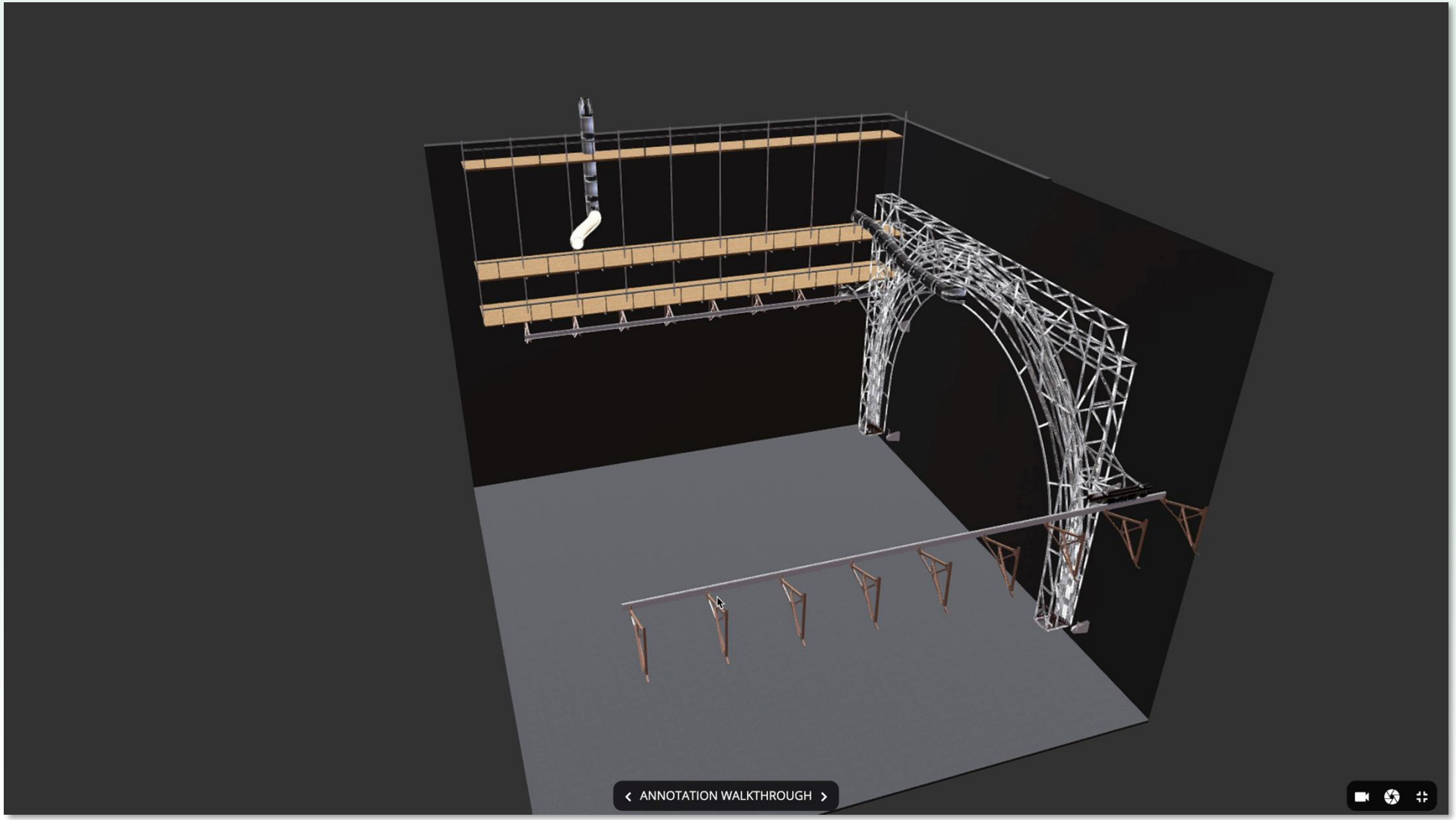
3D-Modell einer Rekonstruktion der Bühne der Passionsspiele in Luzern, Enes Türkoglu (Theaterwissenschaftliche Sammlung, Universität zu Köln)



https://kompakkt.de/entity/5e4bca0123be011fdc1d3613

DCH

Universität zu Köln



DCH

Universität zu Köln

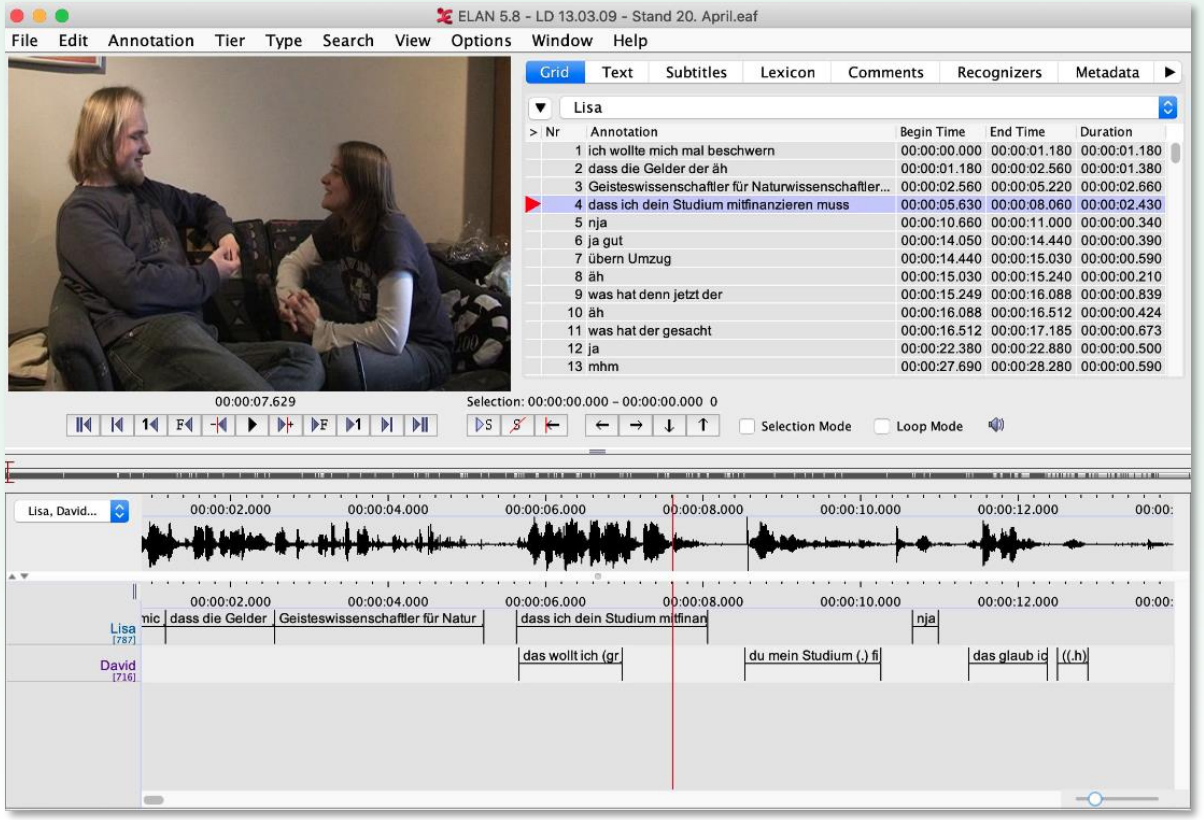
Spezifika von 3D-Modellen…

- Verschiedene Quellen, bspw. analoge Karten, Modele, Objekte
- Verschiedene Präsentationslayer
- Bedarf großer Speichermengen

Und grundsätzlich…

Datenmodelle

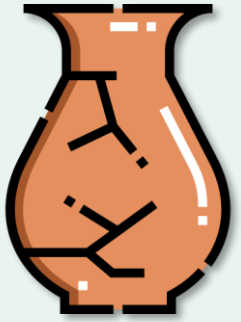
- *Ausgangspunkt:* Video (.mpeg)
- Audioaufnahme (.wav)
- Waveform (Visualisierung)
- Sprachsegmentierung (.eaf)
- Transkription (.eaf)



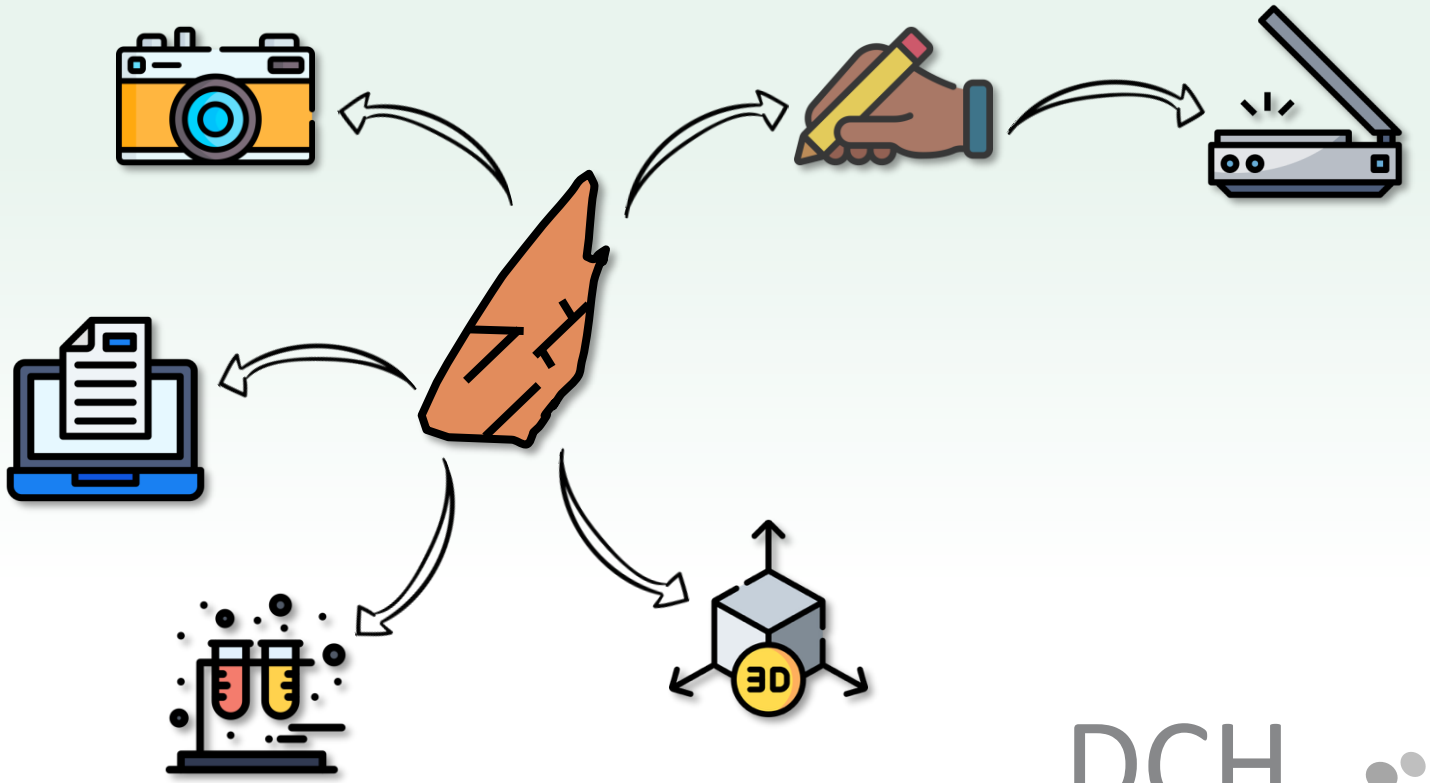
DCH

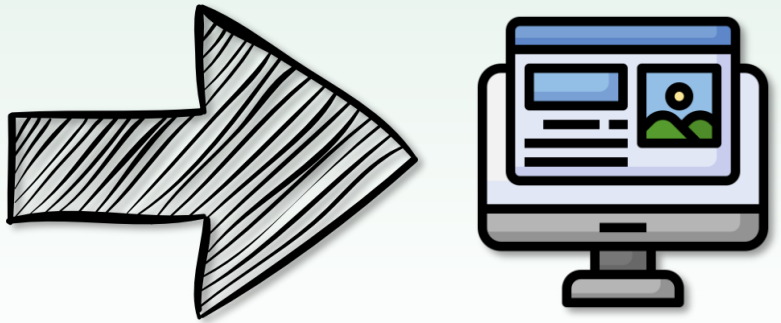
Universität zu Köln

Und grundsätzlich…



- Verschiedene Datentypen (und -formate) beschreiben den gleichen Forschungsgegenstand, bspw. ein Stück einer antiken Vase in einem Museum



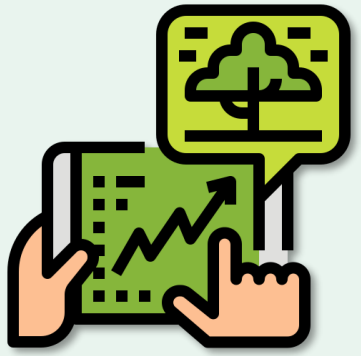


DCH

Universität zu Köln

icon: flaticon.com

Nachhaltigkeit von Datenformaten



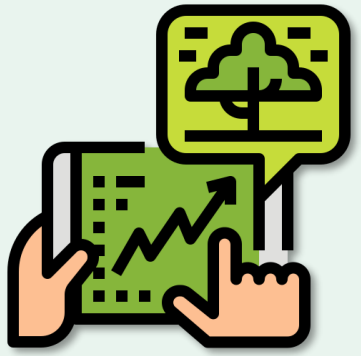
- **Offenheit:** standardisiert, patentiert, Lesbarkeit
- **Umsetzbarkeit:** viele Implementierungen
- **Komplexität:** menschenlesbar
- **Kopierschutz:** technische Beschränkungen in Nutzung und Lesbarkeit
- **Dokumentation:** reichhaltige Metadaten
- **Abhängkeiten:** Hardware, OS, oder Anwendungsabhängkeiten

DCH

Universität zu Köln

icon: flaticon.com

Nachhaltigkeit von Datenformaten



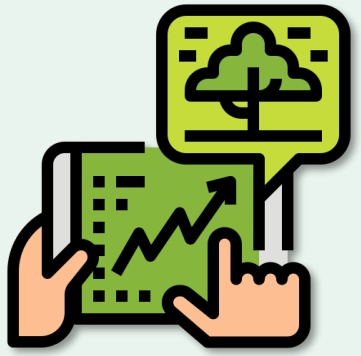
		✓	✗
Standard		.pdf/A (ISO)	.rtf
Öffentliche Documentation		.mp3	.doc
Verbreitung	.docx		.wrf
Menschenlesbar		.xml	.doc
Verlustfreiheit		.wav	.mp3
Kein dynamischer Inhalt	.csv		.pdf

DCH

Universität zu Köln

icon: flaticon.com

Nachhaltigkeit von Datenformaten

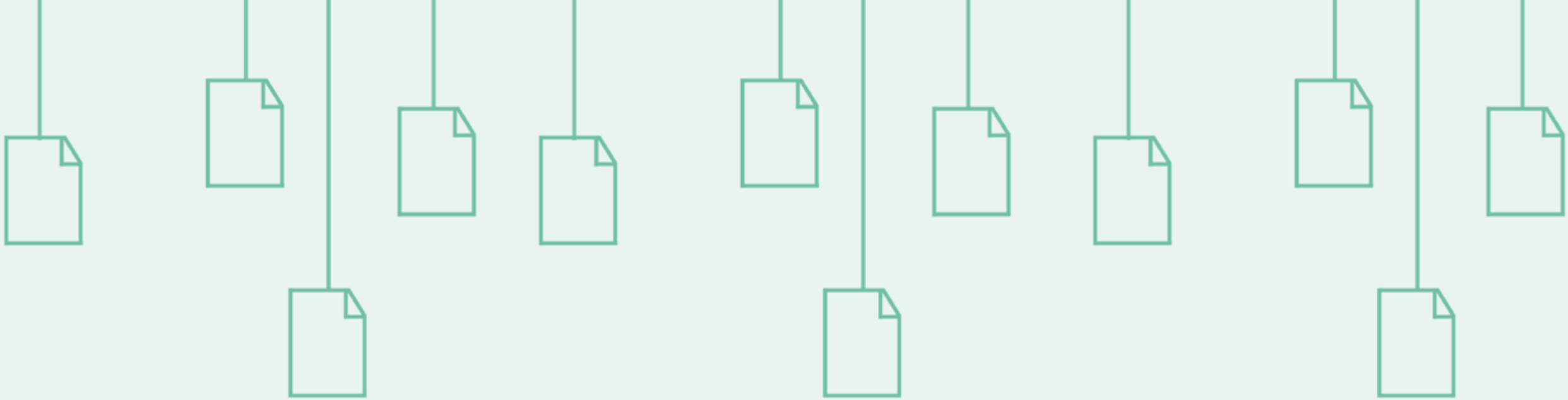


- Text → .txt, .xml, .pdf/a
- Tabellen → .csv, .tsv
- Bilder → .svg (vector), .png (raster)
- Audio → .wav, .flac
- Video → .mp4

DCH

Universität zu Köln

icon: flaticon.com



Repositorien und Infrastrukturen in den Geisteswissenschaften

DCH

Universität zu Köln

Reposi… what?

- Infrastrukturen, die
  - Große Datenmengen halten können
  - Forschungsdaten auffindbar und zugänglich machen
  - persistente Identifier (PID) Systeme nutzen

- Infrastrukturen, die ggf.
  - Authentifizierungssysteme verwenden
  - die Versionierung von Forschungsdaten ermöglichen

- Infrastrukturen, die
  - von stabilen und auf Dauer ausgelegten Institutionen betrieben werden
  - die Archivierung von Forschungsdaten ermöglichen

Reposi… what?

- Publikationsserver
- Generische Repositorien
- Institutionelle Repositorien
- Fachspezifische Repositorien
- Dark Archives

DCH

Universität zu Köln

Publikationsserver

- Publikation von elektronischen Texten, bspw. Dissertationen

→ KUPS, https://kups.ub.uni-koeln.de

Kein Service für Forschungsdaten!

DCH

Universität zu Köln

Generische Repositorien

- Keine fachspezifischen Einschränkungen
- Speicherung von (nahezu) allen Forschungsdatentypen und –formaten aller Fachrichtungen
- Eingeschänkte Findbarkeit aufgrund fehlender, fachspezifischer Dokumentation und Beschreibung durch Metadaten

→ Zenodo, https://zenodo.org

Institutionelle Repositorien

- Infrastruktur für Mitglieder einer spezifischen Einrichtung, bspw. einer Universität
- Normalerweise ähnlich zu generischen Repositorien
- Antwort auf Forschungsdatenmanagement-Policies und –Empfehlungen

→ Refubium, https://refubium.fu-berlin.de

Fachspezifische Repositorien

- Infrastruktur für fachspezifische / inhaltsspezifische / datenspezifische Foschungsdaten
- Service i.d.R. für die gesamte Forschungscommunity (weltweit)
- Fachspezifischer Support, Kuration und Beschreibung/Dokumentation von Forschungsdaten

→ Language Archive Cologne, https://lac.uni-koeln.de

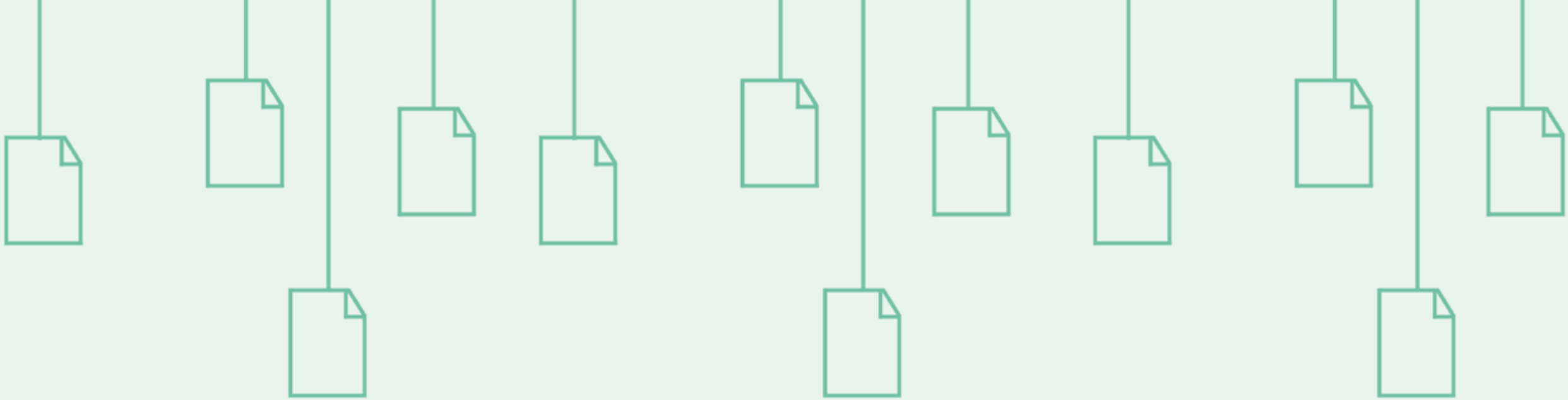
Dark Archives

- Keine Zugänglichkeit von Forschungsdaten
- Forschungsdaten sind archiviert und max. über Metadaten auffindbar

→ TSM Bandspeicherarchiv der UzK, https://rrzk.uni-koeln.de/daten-speicher-und-teilen/backup-system-tsm

DCH

Universität zu Köln



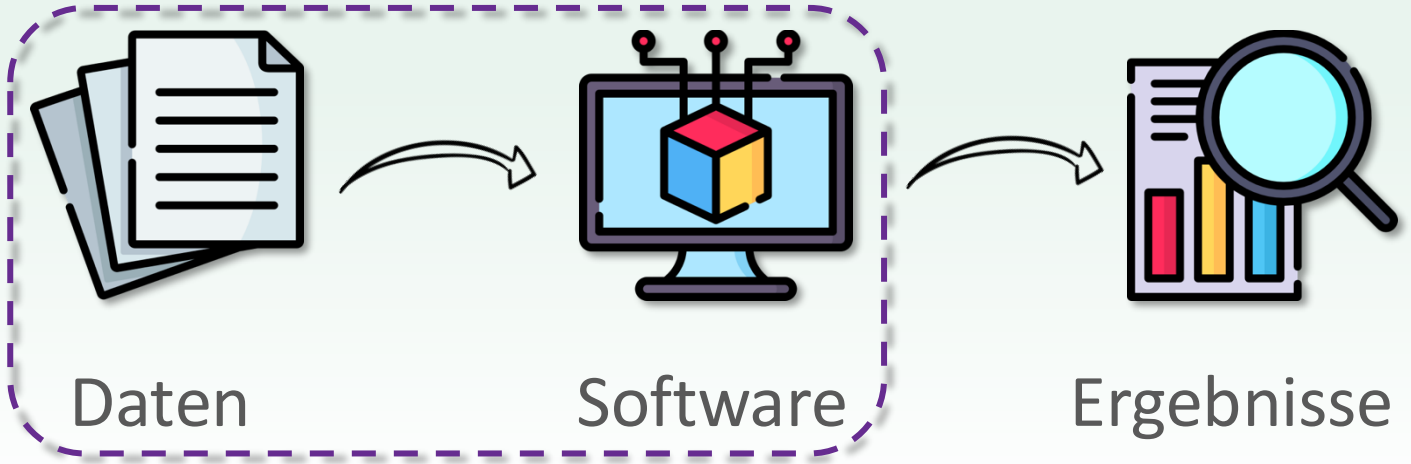
Umgang mit lebenden Systemen

DCH

Universität zu Köln

Software und Forschungsdaten

- Code (Code als Text)
- Software
- Häufige Komponente: **Datenbanken**



DCH

Universität zu Köln

icon: flaticon.com

Spezifische Software und Tools in den Geisteswissenschaften

- Digitale Editionen
- Text Korpora
- Text mit Markup
- Websites mit Datenbanken
- Thematische Datensammlungen
- Forschungsdaten mit Analysen und Annotationen

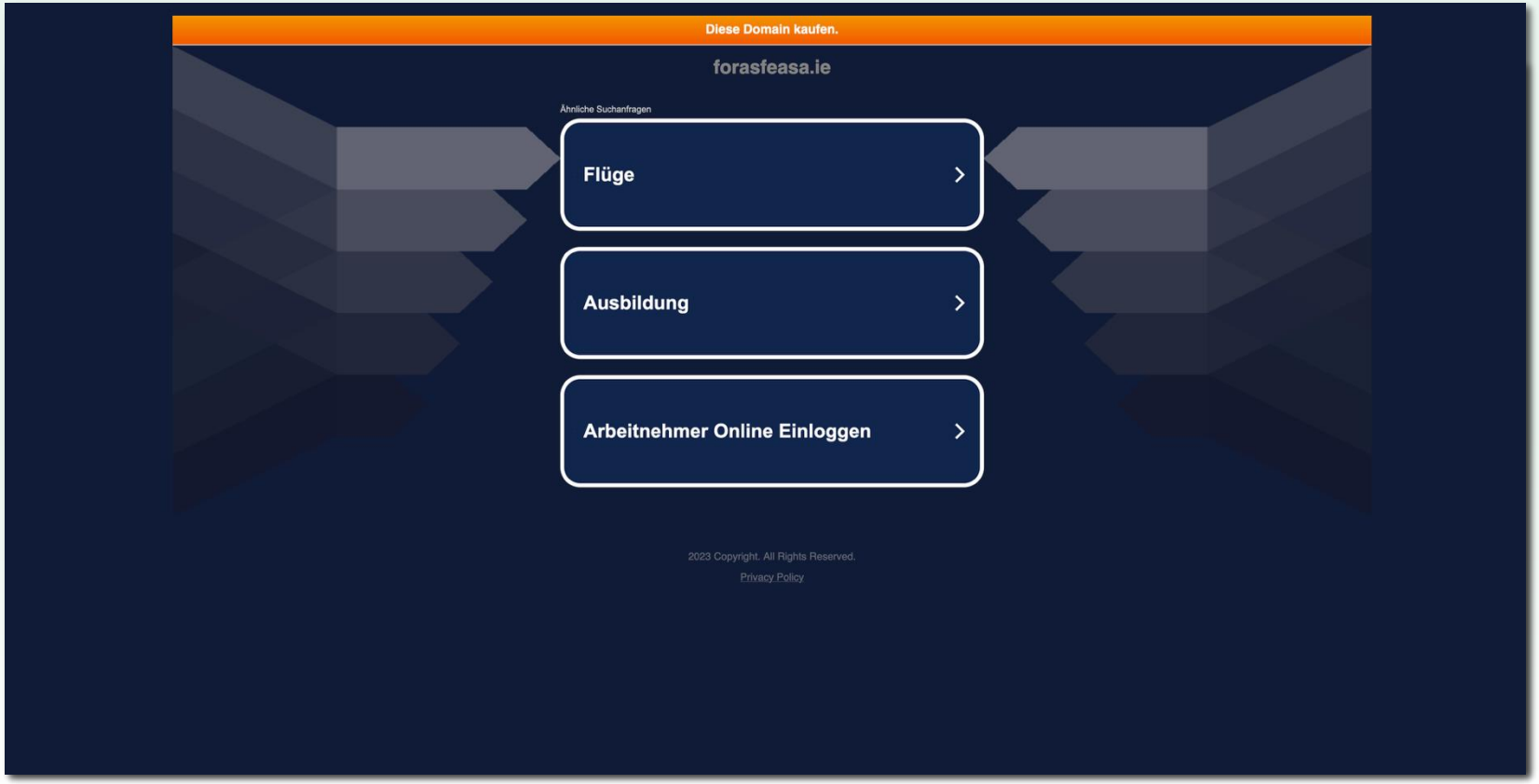
DCH

Universität zu Köln

Schicksal

- Digitale Ressourcen ziehen um (Broken Links)
- Updates von Technologiestacks
- Veränderte Abhängigkeiten
- Organisatorische/Finanzielle Veränderungen
- …

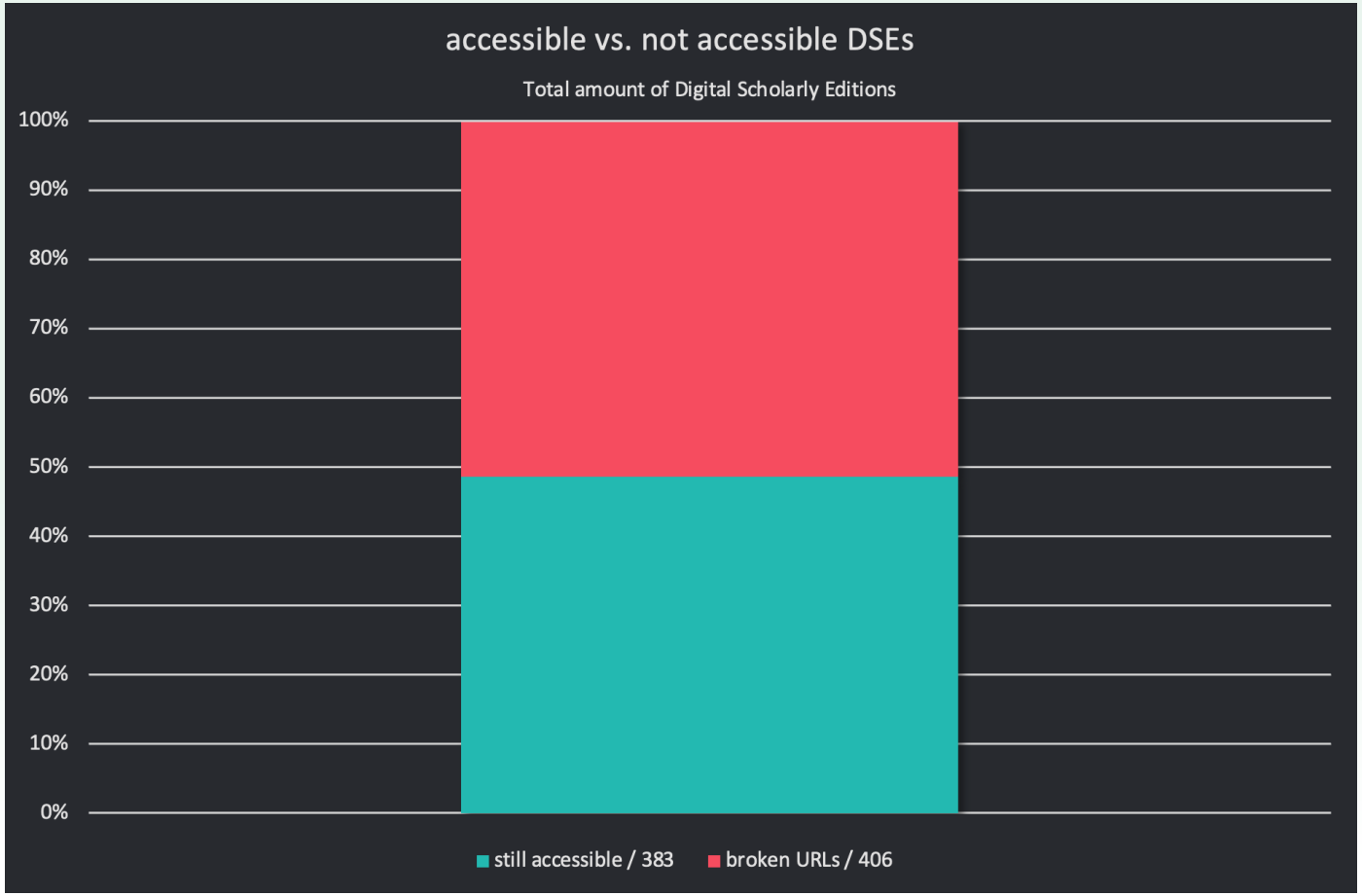
The Alcalá Account Book



DCH

Universität zu Köln

Catalog of Digital Scholarly Editions (Sahle)

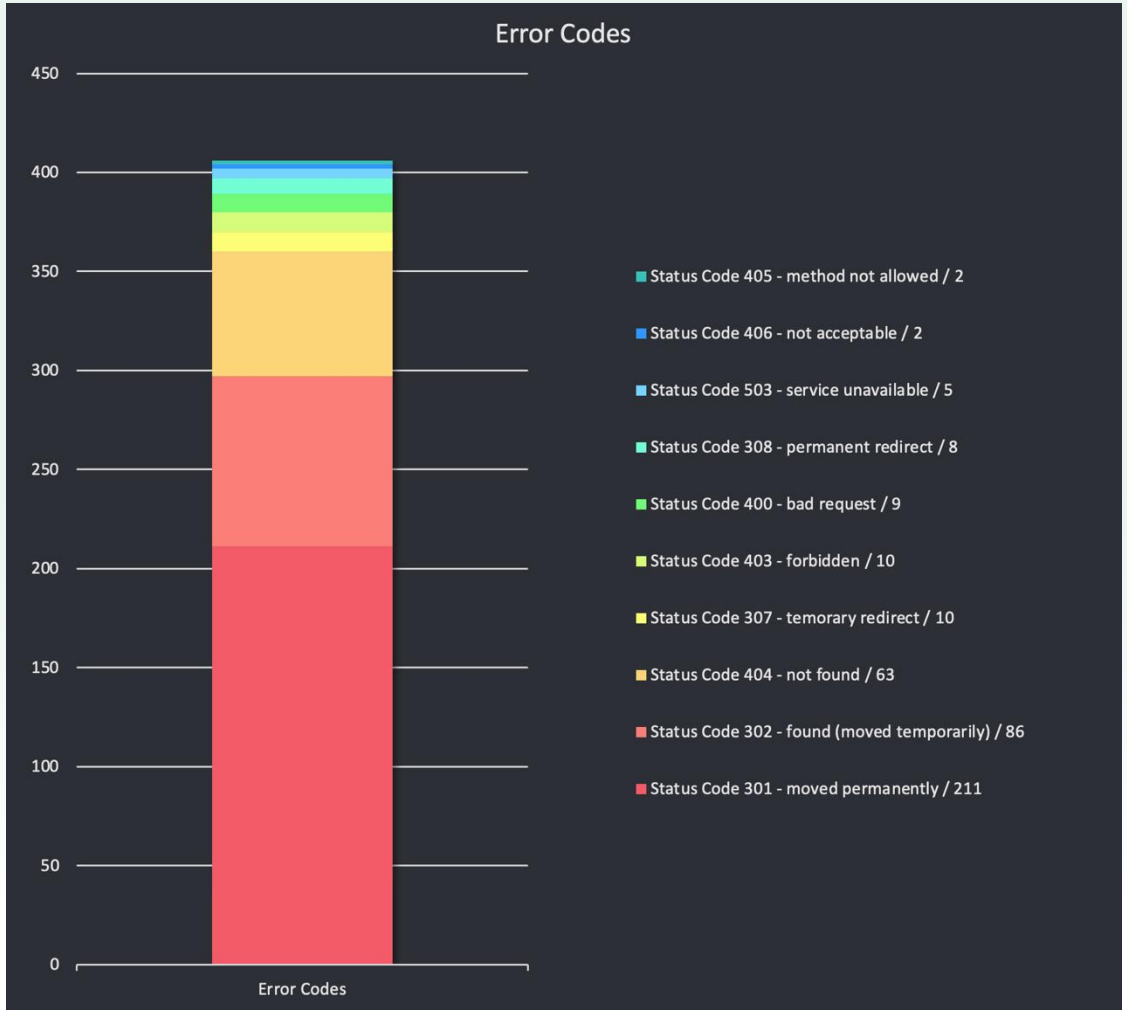


N=789

DCH

Universität zu Köln

Catalog of Digital Scholarly Editions (Sahle)



N=406

DCH

Universität zu Köln

Lösungen?!

Ausreichend personelle und finanzielle Ausstattung

- Skaliert nicht mit einer wachsenden Anzahl an lebenden Systemen
- Wer bezahlt das?

Einschränkung von Technologie-Stacks bei der Entwicklung lebender Systeme

- Nicht ausreichend definierte Standards von Forschenden
- Widerspruch zur Freiheit der Wissenschaft

Virtualisierung, bspw. durch “Docker“

- Bedarf spezifischer Infrastruktur
- Bedarf qualifizierter Mitarbeiter:innen

Hosting in statischer Form, bspw. HTML

- Nicht immer eine mögliche Strategie
- Reduktion von bereits entwickelten, dynamischen Funktionalitäten

Ein reales Beispiel

Digitale Datenbank

- Professor Emeritus einer Universität in Deutschland
- Audioaufnahmen (Interviews) mit Sprecher:innen einer, gesammelt seit 1991
- Insgesamt 700 Interviews, ca. 2 Stunden (300 GB Daten, hauptsächlich .wav und .mp3) organisiert in einer digitalen Datenbank
- Daten waren über eine Website zugänglich (+ weiterer Software, bspw. Lame, MP3splt, FFmpeg)
- Hosting auf Windows Server an der Universität des Professors

Digitale Datenbank

- Konkreter: Datenbank und Website wurden ursprünglich auf einem Server des Instituts gehostet
- Nachdem der Professor die Universität verlassen hatte, konnte sich niemand mehr am Institut um die Ressource kümmern
- Der Professor dachte, er hätte das Überleben seiner Ressource gesichert, indem er sie an die Bibliothek seiner Universität übergeben hat

Die Jahre zogen ins Land…

- Die Bibliothek konnte irgendwann den Betrieb der Ressource nicht mehr sicherstellen (und genau genommen gab es nie dezidiertes eine Finanzierung oder Personal, um die Ressouce zu hosten)

Digitale Datenbank

Die Bibliothek informierte den Professor, dass sie die Ressource innerhalb eines Montas abschalten werden…

… und die Panik begann (überall)

DCH

Universität zu Köln

Lösungen!?

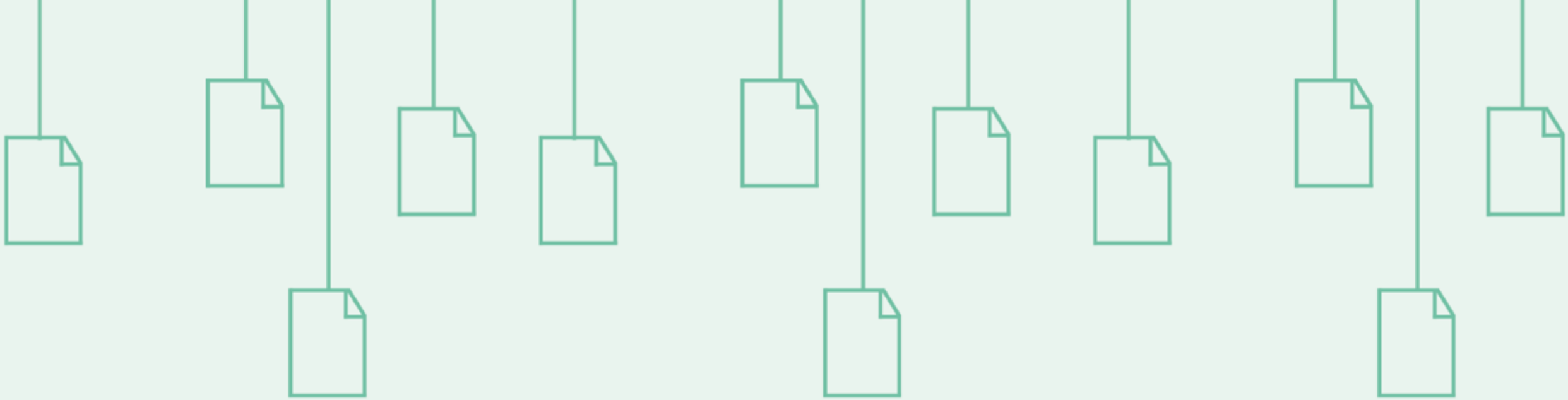
- **Lobbyarbeit**
  - Überzeugung von verschiedenen Stakeholdern, bspw. die Universität, das lokale Rechenzentrum, die Leitung des DCH
- **Migration**
  - der Ressource auf Linux Webserver
- **Reduktion**
  - von dynamischen Funktionalitäten
- **Implementierung**
  - einer einfachen Suchfunktion
- **Upload und Tests**
  - Ablage der Ressource ink. Datenbank auf lokaler Infrastruktur
- **Anpassungen**
  - … work in progress

DCH

Universität zu Köln

Lessons learned?

- Möglichst standardisierte, weit verbreitete Technologiestacks
- Möglichst wenig dynamische Inhalte
- Less fancy, more solid
- Nutzung von Forschungsdaten unabhängig von lebendem System



Zusammenfassung

Take home messages

- Die Geisteswissenschaften weisen in Bezug auf Fachbereiche, Forschungsfragen, Methoden, Datentypen und –formate, Forschungsergebnisse und entsprechend auch Forschungsdatenmanagement Bedarfe eine **große Heterogenität** auf

- Es gibt bereits **einige fachspezifische Lösungen**, bspw. Repositorien und Infrastrukturen, de-facto Standards bei Datentypen und –formaten etc. (allerdings nicht für die Geisteswissenschaften im Allgemeinen, sondern lediglich für einige geisteswissenschaftliche Fachbereiche)

- Standardisierte und offene Datenformate
- Standardisierte und offene Software
- Daten unabhängig von Strukturen und Systemen
- Berücksichtigung fachspezifischer Best Practices

DCH

Universität zu Köln

icon: flaticon.com

Vielen Dank!

Fragen?

Präsentationsfolien:

Data Center for the Humanities
Kölner Datenzentrum
für die Geisteswissenschaften

Data Center for the Humanities (DCH)
Universität zu Köln
Tel.: +49 221 470 1307
E-Mail: patrick.helling@uni-koeln.de | info-dch@uni-koeln.de

DCH

Universität zu Köln