

6. Forschungsdaten akquirieren

Datenakquise: Vorbereitung

1. Methoden und Modelle zur Datenverarbeitung festlegen
2. Datenmanagement planen (Formate, Speicherorte etc.)
3. ggfs. bereits vorhandene Daten lokalisieren und analysieren
4. rechtliche Aspekte der Nutzung und Veröffentlichung klären

Arten der Akquise von Forschungsdaten

Art der Akquise	Vorteile	Nachteile
Erhebung / Erfassung eigener Daten	volle Kontrolle über Art und Umfang	verbunden mit erheblichen Kosten und Risiken
Nachnutzung vorhandener Daten	Kosten / Aufwand der Erhebung wird überwiegend von anderen getragen potenziell größere Reichweite (Metastudien) standardisierte Methoden	Umfang und Format der Daten ist vorgegeben u.U. erheblicher Einarbeitungsaufwand

Mögliche Anlaufstellen

- Fachdatenbanken
 - disziplinäre Forschungsdatenrepositorien
 - Portale oder Index-Services
 - Forschungsdatenzentren
 - Institutionen, z.B. Behörden...
- WWW
 - Webseiten (Inhalt, Domäne, Sprache)
 - Social media (twitter, facebook, blogs...)...
- Assoziierte Projekte
 - Datenträger
 - Zugang zu Shares...

Suche nach passenden Repositorien

- [Re3data.org](#) (Karlsruhe Institute of Technology)
- [OpenDOAR](#) UK
- [Research Pipeline](#) Mediawiki mit Links zu Open Data
- [Repository Finder](#) Suche in Re3Data nach Repositorien, die den Enabling FAIR Data Projekt Prinzipien entsprechen

Suche nach passenden *Daten* in Repositorien

- [DataCite.org](#) NGO / DOI provider
- [EU Open Data Portal](#)
- [B2FIND](#) Recherche in EUDAT FD Sammlungen
- [Research Data Australia](#)
-

Nachnutzung von Daten: Voraussetzungen

- Qualität der Daten muss für das Forschungsvorhaben ausreichend
- Provenienz und Rechte müssen geklärt sein
- Vollständigkeit muss ausreichend sein für das Projektziel
- Interpretierbarkeit (Dokumentation vorhanden?)

Datennachnutzung

Übung (Excel-Tabelle-Datenqualität)

Provenienz/ Dokumentation	Bewerten Sie die Datenqualität:
? Worum geht es bei den Daten	Bewerten Sie die Datenqualität:
? Woher stammen die Daten	? Ist der Datensatz vollständig
? Wann wurden sie erhoben	? Sind die Daten konsistent
? In welchem Format liegen die Daten vor	? Sind die Daten interpretierbar
? Wurden Änderungen versioniert	? Wurden ggfs. Lücken gekennzeichnet

Mögliche Quellen für eigene Erhebungen

- Umfragen
- Aufzeichnungen
- Materialsammlung / Digitalisierung ...
-

Datenakquise: Daten erfassen

1. Methoden und Modelle zur Datenverarbeitung festlegen
2. ggfs. bereits vorhandene Daten lokalisieren
3. Datenmanagement planen (Formate, Speicherorte etc.)
4. rechtliche Aspekte der Nutzung und Veröffentlichung klären
5. Daten erfassen / erheben
6. Metadaten erstellen / Dokumentation

Modellierung - Datenqualität

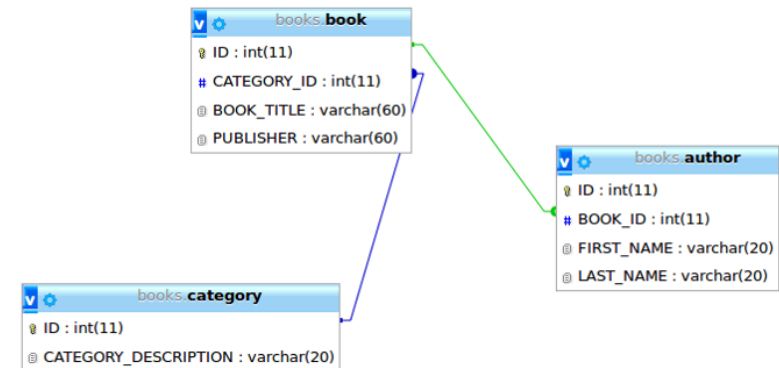
- Erkundigen Sie sich nach möglichen Standards oder Best Practices in Ihrem Feld
- unterstützen Sie die Eingabe von validen Daten durch entsprechende Eingabehilfen
- gestalten Sie die Eingabe nur so flexibel wie nötig (Validierung)

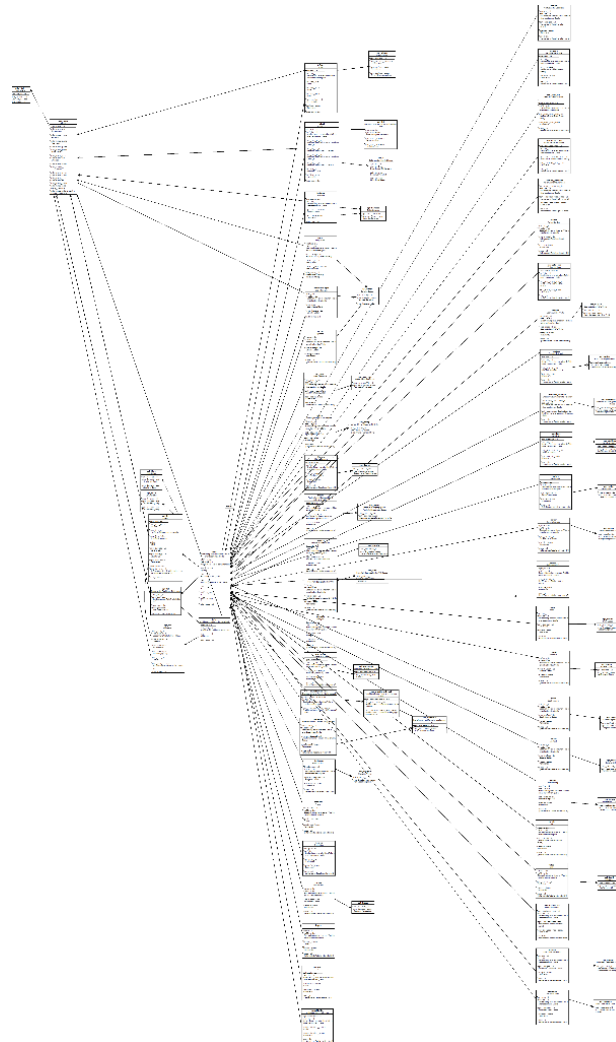
Fokus Datenqualität

- Wertebereiche / Wertelisten helfen bei der Datenqualität
- kontrollierte Vokabulare und Ontologien bzw. Thesauri
- vordefinierte Wertebereiche, Format-vorgaben (z.B. Datum YYYY-MM-DD)
- Validierung (bei Eingabe/Import)
- Aussagekräftige Bezeichnungen
- Überprüfung / Peer Review / 4-Augen-Prinzip

Modellierung von Daten

- Formale Abbildung der Objekte mittels Attributen und Beziehungen
- Eindeutige Spezifikation für die Verwaltung in einem Informationssystem
- Datenmodelle überleben in der Regel damit verbundene Funktionen und Prozesse
- Klare Trennung von Daten und Anwendung wichtig
- erkundigen Sie sich nach Standards in Ihrem Feld





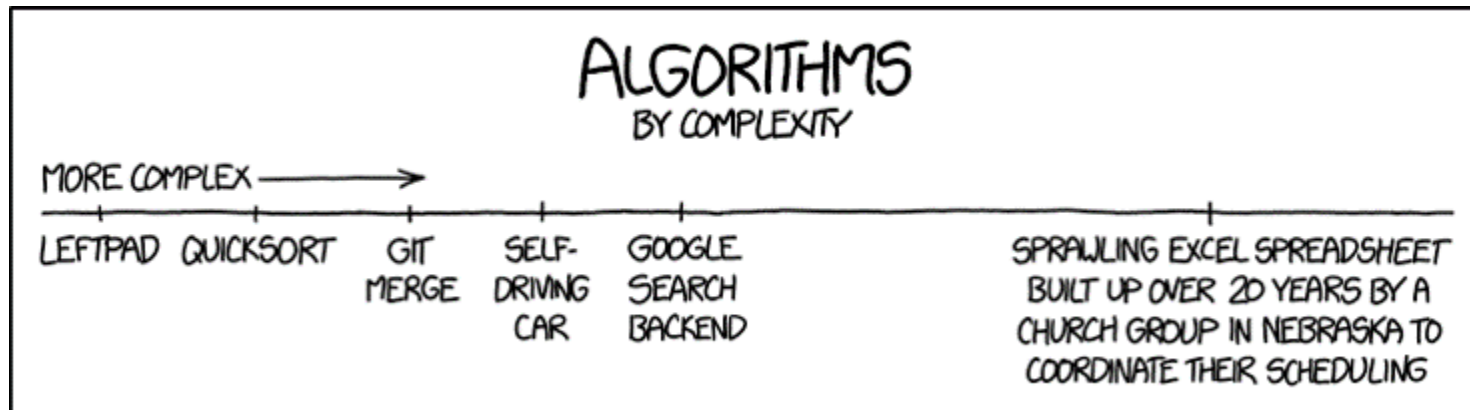
Daten dokumentieren

- Kontext: Projekthistorie, Absicht/ Zielsetzung, Hypothesen ...
- Methoden: Sampling, Umstände der Erhebung, technische Rahmenbedingungen ...
- Datenstrukturen, Beziehungen zwischen Objekten
- Wertebereiche, Qualitätskriterien, Gültigkeit
- Änderungen (!) im Projektverlauf, Versionierung
- Rechtliche Fragen / Zugriff

Publikationen ersetzen nicht die Dokumentation der Daten

Technische Dokumentation

- Namen, Bezeichnungen für Variablen und ihre Werte
- Erklärungen für Codes / Klassifikationsschemata
- Erläuterung der verwendeten Terminologie, ggfs. Definitionen
- Kodierung fehlender Werte / Gründe für fehlende Werte
- Abgeleitete Daten, verwendete Algorithmen, Gewichtungen..
- ..



— Pause —

Daten speichern: Dateiformate

- = die Art und Weise, wie Information digital gespeichert ist
- Dateiendungen (z.B. *.txt*, *.docx*, *.jpg*) geben Hinweise auf Dateiformate und somit die darin enthaltenen Daten

Win und Mac Systeme zeigen Datei-Endungen standardmäßig nicht an. Die Anzeige lässt sich über die Systemeinstellung aktivieren.

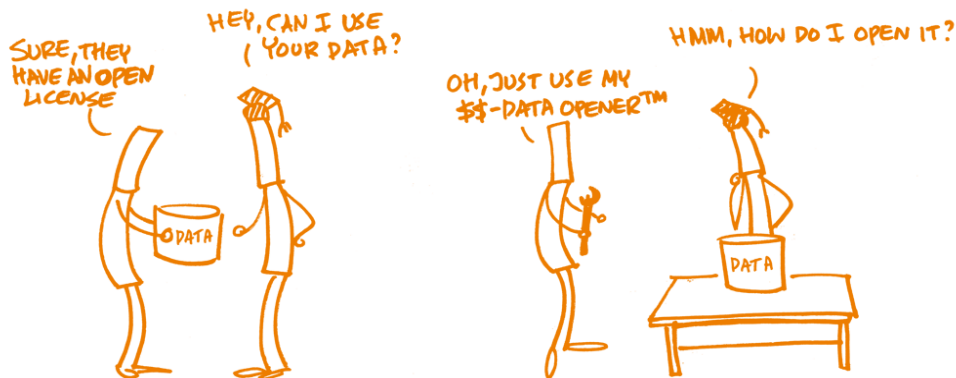
- Der Internet-Media-Type / MIME (Multipurpose Internet Mail Extension) sorgt im WWW dafür, dass den Dateiformaten die richtige Anwendung zugeordnet wird

MIME-Types und File-Extensions

MIME-Typ	Dateiendung(en)	Bedeutung
application/acad	*.dwg	AutoCAD-Dateien (nach NCSA)
application/applefile		AppleFile-Dateien
application/astound	*.asd *.asn	Astound-Dateien
application/dsptype	*.tsp	TSP-Dateien
application/dxf	*.dxf	AutoCAD-Dateien (nach CERN)
application/force-download	*.reg	Registrierungsdateien
application/futuresplash	*.spl	Flash Futuresplash-Dateien
application/gzip	*.gz	GNU Zip-Dateien
application/javascript	*.js	serverseitige JavaScript-Dateien
application/json	*.json	enthält einen String in JavaScript-Objekt-Notation
application/listenup	*.ptlk	Listenup-Dateien
application/mac-binhex40	*.hqx	Macintosh Binärdateien
application/mbedlet	*.mbd	Mbedlet-Dateien
application/mif	*.mif	FrameMaker Interchange Format Dateien
application/msexcel	*.xls *.xla	Microsoft Excel Dateien
application/mshelp	*.hlp *.chm	Microsoft Windows Hilfe Dateien
application/mspowerpoint	*.ppt *.ppz *.pps *.pot	Microsoft Powerpoint Dateien
application/msword	*.doc *.dot	Microsoft Word Dateien
application/octet-stream	*.bin *.file *.com *.class *.ini	Nicht näher spezifizierte Daten, z.B. ausführbare Dateien
application/oda	*.oda	Oda-Dateien

Proprietäre Datei-Formate

- sind meist abhängig von
 - spezifischer Software
 - entsprechenden Software-Lizenzen
 - spezifischen Plattformen / Betriebssystemen
 - sind gefährdet durch 'Software Obsolescence'
 - rasche technologische Weiterentwicklung
 - Kompatibilität zwischen Versionen nicht gewährleistet
 - Marktabhängigkeit
 - Konvertierung / Export in offene Formate meist verlustbehaftet



Dateiformate – offene Formate

Für die Langzeitarchivierung sollten Dateien

- unverschlüsselt sein
- nicht komprimiert sein
- nicht firmeneigen/patentiert sein
- einen offenen/dokumentierten Standard nutzen

Tabelle Dateitypen

Dateityp	Empfehlung	Vermeiden
Tabellendaten	CSV, TSV, SPSS portable	Excel
Text	TXT, ODT, HTML, RTF, PDF/A nur wenn Layouts wichtig sind	Word, PowerPoint
Multimedia	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	QuickTime, H264
Bild	TIFF, JPEG2000, PNG	GIF, JPG
Strukturierte Daten	XML, RDF, JSON	RDBMS

Beispiel CSV

```
ID, Sammlungsname, Band, Röhrchen, Familie, Name lateinisch und Autor, Name deutsch, Name englisch, Date
1, Bredemann & Nieser, 01, 01, Fabaceae, Anthyllis vulneraria L., Wundklee, woundwort, 0001_Anthyllis_vu
2, Bredemann & Nieser, 01, 02, Fabaceae, Astragalus boeticus L., Kaffee-Traganth, swedish coffee, 0002_A
3, Bredemann & Nieser, 01, 03, Fabaceae, Astragalus falcatus Lam., Sicheltraganth, russian milk vetch, 0
4, Bredemann & Nieser, 01, 04, Fabaceae, Astragalus glycyphyllos L., Süßer Tragant, sweet milk vetch, 00
5, Bredemann & Nieser, 01, 05, Fabaceae, Astragalus hamosus L., Angelstragel, yellow milk vetch, 0005_As
6, Bredemann & Nieser, 01, 05a, Fabaceae, Cajanus cajan (L.) Millsp., Taubenerbse, pigeonpea, 0006_Cajan
7, Bredemann & Nieser, 01, 06, Fabaceae, Senna tora (L.) Roxb., Gemüse-Kassie, sicklepod, 0007_Senna_tor
8, Bredemann & Nieser, 01, 07, Fabaceae, Cicer arietinum L., Kichererbse, chickpea, 0008_Cicer_arietinum
9, Bredemann & Nieser, 01, 08, Fabaceae, Coronilla scorpioides (L.) W.D.J.Koch, Skorpionspeitsche, annu
10, Bredemann & Nieser, 01, 09, Fabaceae, Securigera varia (L.) Lassen, Bunte Kronwicke, Crown-vetch, 00
11, Bredemann & Nieser, 01, 10, Fabaceae, Cyamopsis tetragonoloba (L.) Taub., Guar, Guar, 0011_Cyamopsis
12, Bredemann & Nieser, 01, 10a, Fabaceae, Macrotyloma uniflorum (Lam.) Verdc. var. uniflorum, Pferdeh
13, Bredemann & Nieser, 01, 11, Fabaceae, Galega officinalis L., Echte Geisraute, goat's-rue, 0013_Galeg
14, Bredemann & Nieser, 01, 12, Fabaceae, Genista anglica L., Englischer Ginster, petty whin, 0014_Genis
15, Bredemann & Nieser, 01, 13, Fabaceae, Genista germanica L., Deutscher Ginster, german greenweed, 001
16, Bredemann & Nieser, 01, 14, Fabaceae, Genista sagittalis L., Flügelginster, winged greenweed, 0016_G
17, Bredemann & Nieser, 01, 15, Fabaceae, Genista tinctoria L., Färber-Ginster, dyer's greenweed, 0017_G
18, Bredemann & Nieser, 01, 16, Fabaceae, Glycine max (L.) Merr., "Sojabohne, Soja", "soybean, Soya", 00
19, Bredemann & Nieser, 01, 17, Fabaceae, Hedysarum coronarium L., Kronen-Süßklee, French-honeysuckle, 0
```

Beispiel XML

```
<?xml version="1.0" encoding="utf-8"?>
<mycoreobject xmlns:xsi = "http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xlink = "http://www.w3.org/1999/xlink"
  xsi:noNamespaceSchemaLocation = "datamodel-document.xsd"
  ID="ngmcp_document_4376" label="The 4376 NGMCP document">

  <!-- Strukturinformationen fuer MyCoRe-Interna -->
  <structure />

  <!-- Beschreibungsdaten zum Dokument -->
  <metadata>

    <!-- 01 - Titel -->
    <titles class="MCRMetaLangText">
      <title>
        A&#7779;&#7789;avarga
      </title>
    </titles>

    <!-- 02 - Material -->
    <materials class="MCRMetaClassification">
      <material classid="ngmcp_class_00000011" categid="XX" />
    </materials>

    <!-- 03 - Language -->
    <languages class="MCRMetaClassification">
      <language classid="ngmcp_class_00000012" categid="S" />
    </languages>

    <!-- 04 - Script -->
    <scripts class="MCRMetaClassification">
      <script classid="ngmcp_class_00000013" categid="D" />
    </scripts>

    <!-- 06 - Author -->
    <authors class="MCRMetaLinkID" >
      <author xlink:type="locator" xlink:href="ngmcp_author_00000002"
        xlink:label="author" xlink:title=""/>
    </authors>

    <!-- 07 - Subject -->
    <subjects class="MCRMetaClassification" >
      <subject classid="ngmcp_class_00000014" categid="J" />
    </subjects>
  [...]
```

Beispiel JSON

```
{
  "id": 50283,
  "catalogno": "ZMH 22258",
  "collection": "entomology",
  "created_at": "2018-08-16T13:35:45.955Z",
  "updated_at": "2018-09-21T06:04:51.480Z",
  "details": {
    "tax_class": "Insecta",
    "tax_order": "Coleoptera",
    "odesc_name": "Tefflus, meyerlei, Fabricius 1801",
    "tax_domain": "Holometabola",
    "tax_family": "Carabidae",
    "geo_locality": "Côte d'Ivoire, Dimbroko",
    "geo_continent": "Africa",
    "zmh_catalog_no": "ZMH 22258",
    "identification_stage": "Imago",
    "identification_specimen": "pinned"
  },
  "pix": [
    "meyerlei-tefflus-zmh-22258-50mm-1x2-5.6f.jpg",
    "meyerlei-tefflus-zmh-22258-label.jpg"
  ]
}
```

Datenformate - Empfehlungen

- Multimediale Aufnahmen
 - Hohe Qualität
 - Keine verlustbehaftete Kompression
- Texte
 - Unicode-basierte Kodierung
 - Offene und dokumentierte Formate
 - De Facto Standards (UTF-8, XML, TEI, JSON-LD, RDF...)
- Allgemein Empfehlungen
 - Keine proprietären oder exotischen Formate
 - Hohe Qualität
 - Kompression: verbreitete Formate wählen (.zip, .tar.gz, .tar.bz2)

Verlustfreie versus verlustbehaftete Bildkompression

PNG: 1007 Byte



JPEG: 4273 Byte



Von Phrood , <https://commons.wikimedia.org/w/index.php?curid=14906788>



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Datenbereinigung

Werkzeuge zur Bereinigung von Daten

- OpenRefine (ehemals GoogleRefine)
- Python Pandas
- Knime

The screenshot shows the OpenRefine web interface. The browser address bar indicates the project URL: 127.0.0.1:3333/project?project=1586603301677. The page title is "OpenRefine 2017 01 26 Botanik LokiSchmidt DB mit ID csv". The interface includes a "Facet / Filter" sidebar on the left with a "Using facets and filters" section. The main area displays 1124 rows of data. A context menu is open over the "Röhrenchen" column, showing options like "Text facet", "Text filter", "Edit cells", "Edit column", "Transpose", "Sort...", "View", and "Reconcile". The data table has columns: All, ID, Sammlungsname, Band, Röhrenchen, Familie, Name lateinisch und Autor, Name deutsch, Name englisch, Dateinamen in Reihenfolge, and Dateinamen ergänzende Bilder. The first 10 rows of data are visible.

All	ID	Sammlungsname	Band	Röhrenchen	Familie	Name lateinisch und Autor	Name deutsch	Name englisch	Dateinamen in Reihenfolge	Dateinamen ergänzende Bilder
1.	1	Bredemann & Nieser	01				Wundklee	woundwort	0001_Anthyllis_vulneraria	
2.	2	Bredemann & Nieser	01				Kaffee-Tragant	swedish coffee	0002_Astragalus_boeticus	
3.	3	Bredemann & Nieser	01				Sicheltragant	russian milk vetch	0003_Astragalus_falcatus	
4.	4	Bredemann & Nieser	01				Süßer Tragant	sweet milk vetch	0004_Astragalus_glycyphylus	
5.	5	Bredemann & Nieser	01				Angelstragel	yellow milk vetch	0005_Astragalus_hamosus	
6.	6	Bredemann & Nieser	01				Taubenerbse	pigeonpea	0006_Cajanus_cajan	
7.	7	Bredemann & Nieser	01				Gemüse-Kassie	sicklepod	0007_Senna_tora	
8.	8	Bredemann & Nieser	01				Kichererbse	chickpea	0008_Cicer_arietinum	
9.	9	Bredemann & Nieser	01			Coronilla scorpioides (L.) W.D.J. Koch	Skorpionspeitsche	annual scorpion-vetch	0009_Coronilla_scorpioides	0953_Coronilla_scorpioides_Huelsen
10.	10	Bredemann & Nieser	01			Securigera varia (L.) Lassen	Bunte Kronwicke	Crown-vetch	0010_Securigera_varia	0954_Securigera_varia_Huelsen

Herausforderungen

- Redundanz (Dopplungen von Inhalten / Datenstrukturen / Code)
- Lücken (unvollständige Datensätze)
- Inkonsistenzen in den Daten (z.B. Transliteration, Kodierung, Abkürzungen)
- Intransparenz (Dokumentation / Interpretierbarkeit von Werten)
- unklare Herkunft (Expertise der Datenersteller unklar / fehlende Versionierung)



Alles zu spät?!

Welche Daten aufbewahren? (1/2)

- Müssen die Daten aus rechtlichen Gründen aufbewahrt werden, wenn ja, wie lange?
- Können die Daten ohne Informationsverlust erhalten werden?
- Was bedeutet der Informationsverlust für die Daten, sind sie noch brauchbar?
- Gibt es Interessenten an den Daten?
- Sind die Daten Teil einer größeren Sammlung (z.B. Zeitreihe)?
- Können die Daten neu erhoben werden, wenn ja, mit welchem Aufwand?

Welche Daten aufbewahren? (2/2)

- Ist eine Nachnutzung der Daten zu erwarten?
- Welche Art der Nachnutzung ist zu erwarten?
- Sind die Daten von kulturellem / gesellschaftlichem Wert?
- Sind die Daten von genügender Qualität für die Nachnutzung?
- Können die Daten für Lehrzwecke genutzt werden?

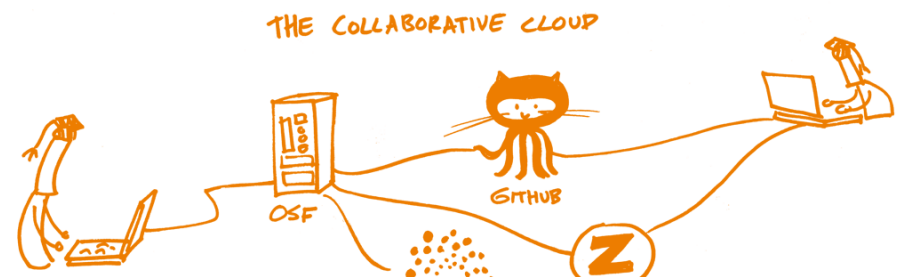
Was kann weg?

- technisch bedingte Strukturen und Daten
- leere Datensätze / Tabellen
- Duplikate
- unvollständige Datensätze (Pflichtfelder?)
- Versionen (welche?)

„Poor data quality can be worse than missing data because it can waste resources and lead to faulty ideas and solutions, or at minimum challenges trust in the results and implications drawn from the data. Improvement in data quality can thus have significant benefits. (Curating for Quality P5)“

Datenorganisation

- Datenmenge wächst sukzessiv während des Projektverlaufs
- Gute Organisation hilft Zeit und Nerven zu sparen
- Konventionen und Abläufe klären für
 - Ablageort (Netzlaufwerken, Shares)
 - Synchronisierung
 - Benennung
 - Versionierung
 - Lebensdauer
 - Sicherung



Daten speichern & teilen ?

- Speichermöglichkeiten – Dropbox, Google, Amazon, Microsoft, Apple usw.
- Google Docs
- Google (u.a. für Suchmaschinen)

Warum nicht all die tollen Tools im Netz?

- Nachhaltigkeit (wie lange wird es den Dienst Gebühren)
- Wie lange werden Daten aufbewahrt (AGB)
- Wird der Dienst irgendwann kostenpflichtig
- Wer hat Zugriff / Rechte an meinen Daten?
- In welchem Format werden meine Daten abgespeichert
- Kann ich sie wieder löschen?



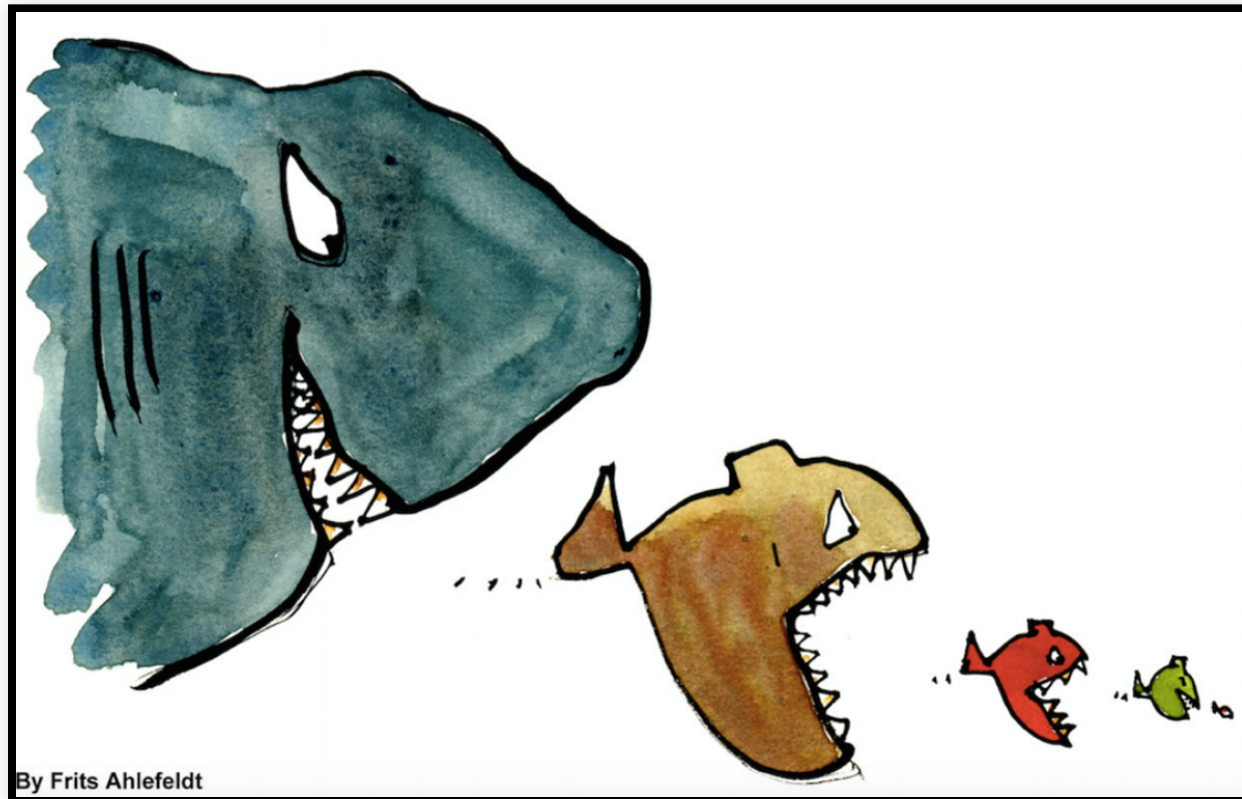
Beispiel Google Drive

Google erhält das Recht Daten weltweit

- zu verwenden
- zu hosten
- zu speichern
- zu vervielfältigen
- zu verändern
- abgeleitete Werke daraus zu erstellen
- zu kommunizieren
- zu veröffentlichen
- zu verteilen

<https://www.heise.de/download/blog/Cloud-Anbieter-Wie-steht-es-um-den-Datenschutz-3658164>

Und nicht zuletzt...



...gestalten wir unsere Welt mit

Dienste am RRZ

UHH-Share

Ordner und Dokumente teilen

UHH Disk

Ordner und Dokumente teilen

OWA- Office Exchange

Kalender und Funktionsmailadressen

Sharepoint

Dateien, Office-Anwendungen, Listen

gitlab

Quellcode / Texte, Wikis, Aufgabenlisten

Wahl der Mittel

- welche Funktion muss der Dienst erfüllen?
- welche Personen(gruppen) sollen darauf Zugriff haben?
- wie sicher ist der Dienst (Backups, Redundanz, Verfügbarkeit)?
- was passiert, wenn ich den Dienst wechseln will / muss?
- welche Nutzungsbedingungen sind an die Nutzung geknüpft?

Datenorganisation - Versionierung

Mit Datei- und Ordnernamen

- Suffix mit laufender Nummer (V1, V2..., final, copy...)
- Verkettung einheitlicher Bestandteile in festgelegter Reihenfolge
- Vermeidung von Sonder- und Leerzeichen
- Nach Möglichkeit jede Änderung kennzeichnen
- Entscheiden Sie sich, wie viele/welche Versionen sie behalten wollen

```
20180301_BohrkernSibirien_ABC_Original  
20180301_Mikrofilm1997_XYZ_005
```

Bleiben Sie konsistent

L_12_cop_22_09.php
L_12_cop_27_10.php
L_12_cop_28_09.php
L_12_cop_30_09.php
L_12.php
L_13_cop_16_12.php
L_13_cop_20_12.php
L_13_cop_mittag.php
L_13.php
L_13_zum_wegnehmen.php
L_6_new.php
L_6.php
L_7-kopie.php

Datenorganisation - Versionierung

Viele Anwendungen bieten integrierte Versionierung

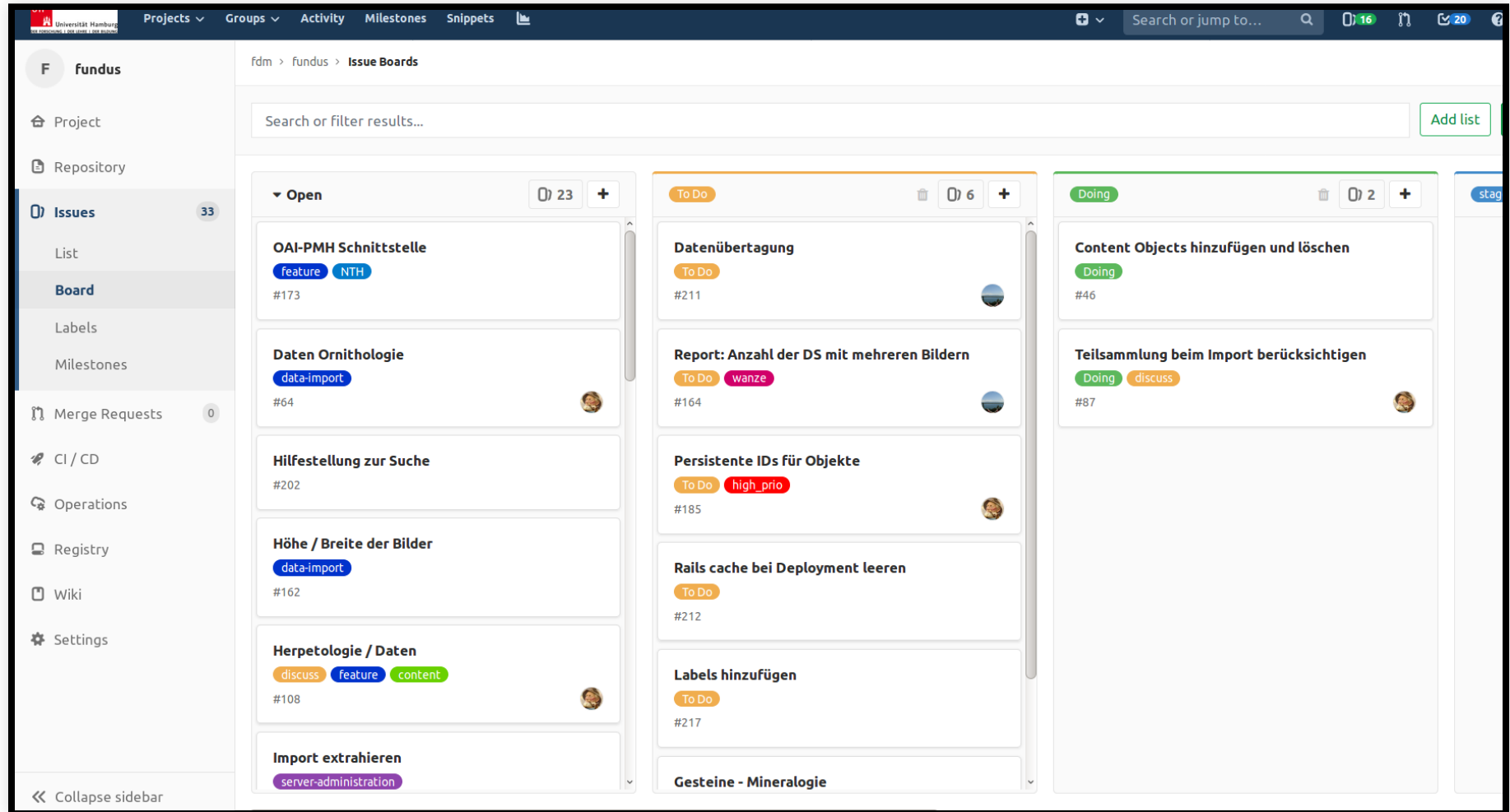
- Versionierung in Office-Programmen, Bildverarbeitung etc. aktivieren
- Persistenz über Sitzungen überprüfen
- Achten Sie dennoch auf regelmäßige Backups
 - automatisiert (TSM-Backup des RRZ)
 - auf Servern
 - auf externen Speichermedien

Versionskontrolle Software / Textbasierte Formate

- Kostenlose Versionskontrollsysteme aus dem Bereich der Softwareentwicklung bieten die Möglichkeit der parallelen Arbeit und kontrollierten Zusammenführung durch Versionierung
- geeignet für alle textbasierten Dateiformate
- *git* ist derzeit de-Facto-Standard, verteiltes Arbeiten
- gitlab ist eine OpenSource Serverlösung für Softwareprojekte

Was bietet Gitlab?

- Intuitive responsive Web-Oberfläche für git-Server
- Datei-Versionierung auf Zeilenebene
- Integriertes 'Tickettool', *kanban*
- Wiki
- Automatische Code-Übersetzung (CI)
- Integration von Kommunikationskanälen



fundus

Projects Groups Activity Milestones Snippets

Search or filter results... [Add list](#)

Open 23 +

- OAI-PMH Schnittstelle** (feature, NTH) #173
- Daten Ornithologie** (data-import) #64
- Hilfestellung zur Suche** #202
- Höhe / Breite der Bilder** (data-import) #162
- Herpetologie / Daten** (discuss, feature, content) #108
- Import extrahieren** (server-administration)

To Do 6 +

- Datenübertragung** (To Do) #211
- Report: Anzahl der DS mit mehreren Bildern** (To Do, wanze) #164
- Persistente IDs für Objekte** (To Do, high_prio) #185
- Rails cache bei Deployment leeren** (To Do) #212
- Labels hinzufügen** (To Do) #217
- Gesteine - Mineralogie**

Doing 2 +

- Content Objects hinzufügen und löschen** (Doing) #46
- Teilsammlung beim Import berücksichtigen** (Doing, discuss) #87

Labels: List, Board, Labels, Milestones

Merge Requests 0

CI / CD

Operations

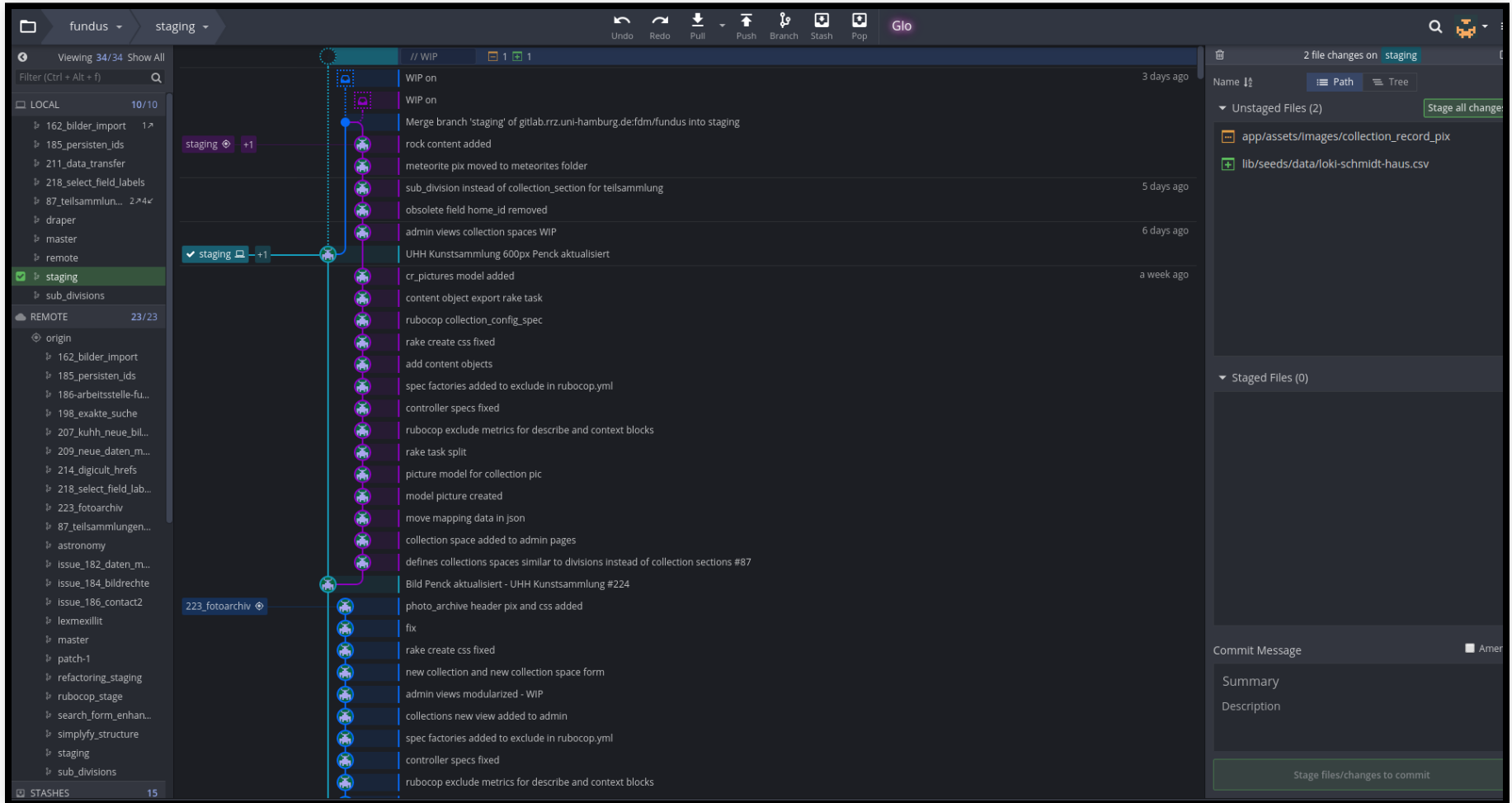
Registry

Wiki

Settings

← Collapse sidebar

Visualisierung der Versionen z.B. durch Gitkraken



The screenshot displays the GitKraken application interface for a repository named 'fundus'. The main window shows a commit history graph with a central vertical line representing the 'staging' branch. Commits are listed on the right, including 'WIP on', 'Merge branch 'staging' of gitlab.rz.uni-hamburg.de:fdm/fundus into staging', 'rock content added', 'meteorite pix moved to meteorites folder', 'sub_division instead of collection_section for teilsammlung', 'obsolete field home_id removed', 'admin views collection spaces WIP', 'UHH Kunstsammlung 600px Penck aktualisiert', 'cr_pictures model added', 'content object export rake task', 'rubocop collection_config_spec', 'rake create css fixed', 'add content objects', 'spec factories added to exclude in rubocop.yml', 'controller specs fixed', 'rubocop exclude metrics for describe and context blocks', 'rake task split', 'picture model for collection pic', 'model picture created', 'move mapping data in json', 'collection space added to admin pages', 'defines collections spaces similar to divisions instead of collection sections #87', 'Bild Penck aktualisiert - UHH Kunstsammlung #224', 'photo_archive header pix and css added', 'fix', 'rake create css fixed', 'new collection and new collection space form', 'admin views modularized - WIP', 'collections new view added to admin', 'spec factories added to exclude in rubocop.yml', 'controller specs fixed', and 'rubocop exclude metrics for describe and context blocks'.

The left sidebar shows the 'LOCAL' and 'REMOTE' branches, with 'staging' selected. The right sidebar shows 'Unstaged Files (2)' and 'Staged Files (0)', with a 'Commit Message' section at the bottom.

Gründe für individuell gehostete Werkzeuge zur Projektorganisation

- Pro
 - Erweiterung des begrenzten Angebot an Lösungen
 - Nutzerverwaltung in Eigenregie
 - z.B. Cloud-Server (Owncloud, Nextcloud..)
 - Datensicherheit, Datenschutz
- Contra
 - Funktionalität und Support bei kostenlosen Versionen i.d.R. begrenzt
 - Updates in Eigenverantwortung
 - Datensicherheit, Datenschutz in Eigenverantwortung

Virtuelle Maschinen am RRZ

- komplette Datenhoheit
- Administrativer Aufwand
- Sicherheit in Eigenverantwortung
- insbes. Eigenentwicklungen ggfs. aufwendig in Administration und Betrieb
- Nachhaltigkeit, Sicherungen, Redundanz
- Standort am RRZ

Kommerzielle Werkzeuge zur Projektorganisation

- meist bieten diese ein Rundum-Sorglos-Paket
- viele OpenSource-Projekte bieten einen Hosted Service an
- teilweise spezialisieren sich Firmen auch auf den Wissenschaftsbereich
- Datensicherheit, Datenschutz, Nutzungsbedingungen beachten
- viele Features (u.a. Workflows, Speichern und Austausch von Daten, Kommunikation, Versionskontrolle, paralleles Arbeiten)
- Zukunftsfähigkeit und Standort der Firma prüfen (Server)

Checkliste virtuelle Forschungsumgebungen

- Speicherung / Austausch von Dateien mit allen Beteiligten ?
- Zugriffskontrolle (Authentifizierung & Authorisierung), wer bekommt Berechtigungen und wie gut lassen sich diese steuern?
- Versionskontrolle - wie oft wird ein backup gemacht?
- Daten - gemeinsames Arbeiten konfliktfrei möglich?
- Kommunikation - welche integrierten Wege gibt es?
- Nachhaltigkeit - welche Exportmöglichkeiten gibt es?
- Transparenz - wo werden die Daten gespeichert / gesichert?
- Import / Export - alle Daten vollständig und korrekt?
- Sicherheit - verfügbar auch kurz vor der Deadline?

Dokumentation - Tools

- Wiki-Software (z.B. Mediawiki, Wiki in gitlab), z.B. für Nutzer-Dokumentation, nicht-technische Dokumentation
- gitbooks (TUHH - Hamburg Open Online University)
- Readme-Dateien (z.B. Readme.md in gitlab)
- Software-Dokumentationstools abhängig von der genutzten Technologie
- ggfs. auch kostenlose hosting-Angebote (z.B. read-the-docs) für Software-Doku (Nachhaltigkeit?)

Es gibt deutliche Unterschiede in der Mächtigkeit der verwendeten Markdown-Sprachen

Dokumentation - Tools

- Wiki-Software (z.B. Mediawiki, Wiki in GitLab), z.B. für Nutzer-Dokumentation, nicht-technische Dokumentation
- gitbooks (TUHH - Hamburg Open Online University)
- Readme-Dateien (z.B. Readme.md in GitLab)
- Software-Dokumentationstools abhängig von der genutzten Technologie
- ggfs. auch kostenlose hosting-Angebote (z.B. read-the-docs) für Software-Doku (Nachhaltigkeit?)

Öffentliche Darstellung und Kommunikation - Angebote

- Fiona CMS als Teil der UHH (Antrag RRZ)
 - Schulungen am RRZ
 - gekoppelt an B-Kennung / Uni-Zugehörigkeit
 - Art der Seite vorstrukturiert (Projekt, Konferenz, Institution...)
- WordPress (Antrag Unikolleg, lokale Kennungen möglich)
 - Webseiten, Design relativ frei
 - Blog posting, Kommentare etc. möglich
 - Hostname xy.blogs.uni-hamburg.de
- Hostnamen (Antrag Abt. Kommunikation und Öffentlichkeitsarbeit /RRZ)
- Mattermost Chat (TUHH)

Empfehlungen Datenorganisation

- Informieren Sie sich eingehend über Dateiformate & Software
- Versuchen Sie sich an Standards und verbreiteten Formaten zu halten
- Strukturieren und dokumentieren Sie Ihre Daten und Dateien so, dass Sie auch ohne Ihr Wissen verständlich sind
- Nutzen Sie zentrale Backup-Speichersysteme und Versionierung zur Datensicherung
- Wählen Sie Werkzeuge passend zu Ihren Anforderungen
- Löschen sie nicht mehr benötigte Daten

Sicherheit

myspace 360 Mio (2008)
Yahoo 500 Mio (2014) 1 Mrd (2013)
LinkedIn 164 Mio (2016)
Adobe 152 Mio (2013)
twitter 32 Mio (2016)
Dropbox 69 Mio (2012)
ebay (145 Mio (2014)
DHL 3 Mio (2016)

Geschäftsmodell Datenklau

About us

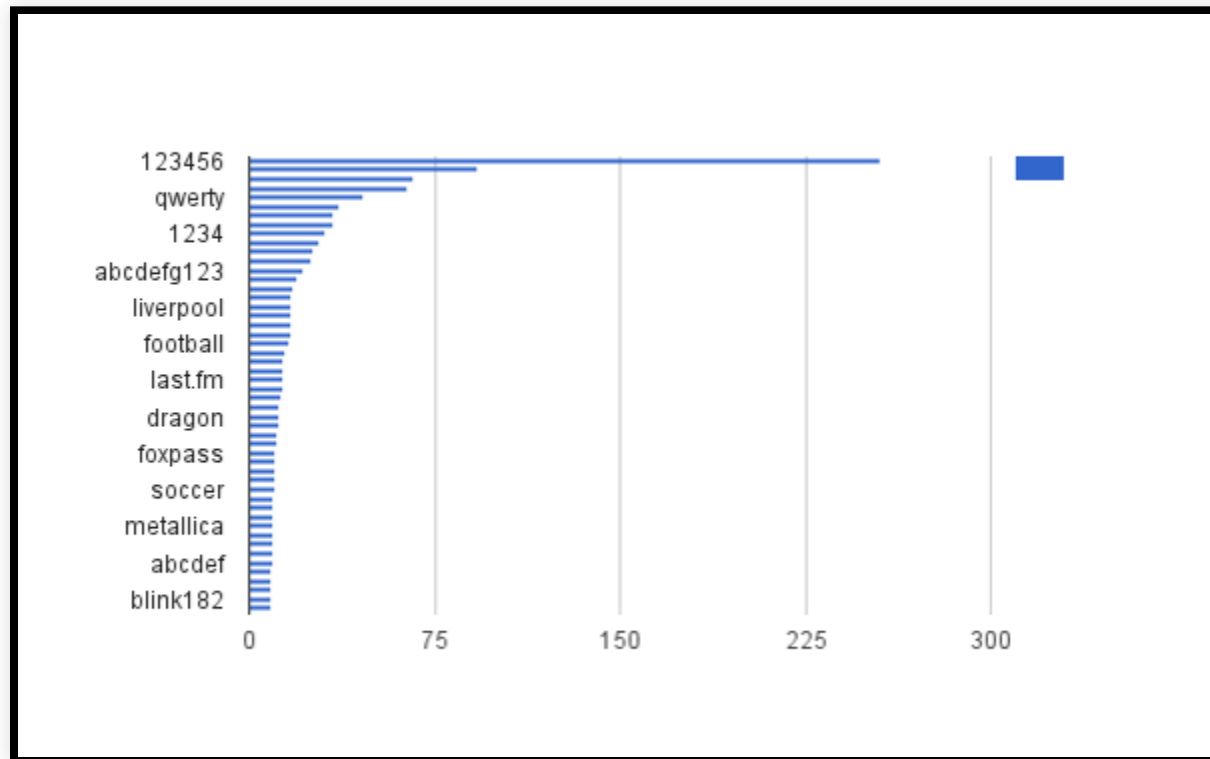
LeakedSource has **exposed** every single mega breach of 2016 including LinkedIn, MySpace, and VK.com but because we are the most effective breach notification service in the world, we're back with more.

We have already helped multiple companies, many of them worth billions of dollars, to secure their users **with our API** which provides cracked plaintext passwords from all 2 billion leaked records. Businesses are then easily able to compare leaked passwords against their own users and if a match is found, they forcibly change the passwords for affected users. If you're interested in using our services, contact us about an API key.

Any journalists that want to get notified about breaches, **DM us on Twitter with your email address**

One final note about the state of affairs on the internet: We have so many databases waiting to be added that if we were to add one per day it would still take multiple years to finish them all. We are also currently processing multiple more mega breaches so stay tuned to our **Twitter**.

Häufigste Passwörter (lastfm)



Passwortsicherheit

- *wichtige* Passwörter nicht wiederverwenden
 - lassen Sie Passwörter automatisch generieren
 - denken Sie sich ein System aus, Varianten von einem Passwort zu erstellen
 - Verwenden Sie einen Passwort-Manager (z.B. LastPass, KeePass)
- Sichere Passwörter enthalten
 - keine Umlaute und Sonderzeichen
 - keine Begriffe aus Wörterbüchern
 - viele Buchstaben / Zahlen / Zeichen
- Achten Sie darauf, wie Ihr Passwort übertragen / gespeichert wird
 - 2-Faktor-Authentifizierung, Verschlüsselung
 - bei Übertragung auf sichere Verbindung ([http](http://)**s**://) achten

Gefahren

- SPAM, Phishing, Trojaner-Mails
- Verkauf im Darknet (insbesondere Streaming-Dienste)
- Einkauf & Umleitung auf Packstation...
- schalten von betrügerischen Anzeigen über Verkaufsplattformen (eBay, Amazon)
- Verlust Ihrer Daten / Ihrer Arbeit

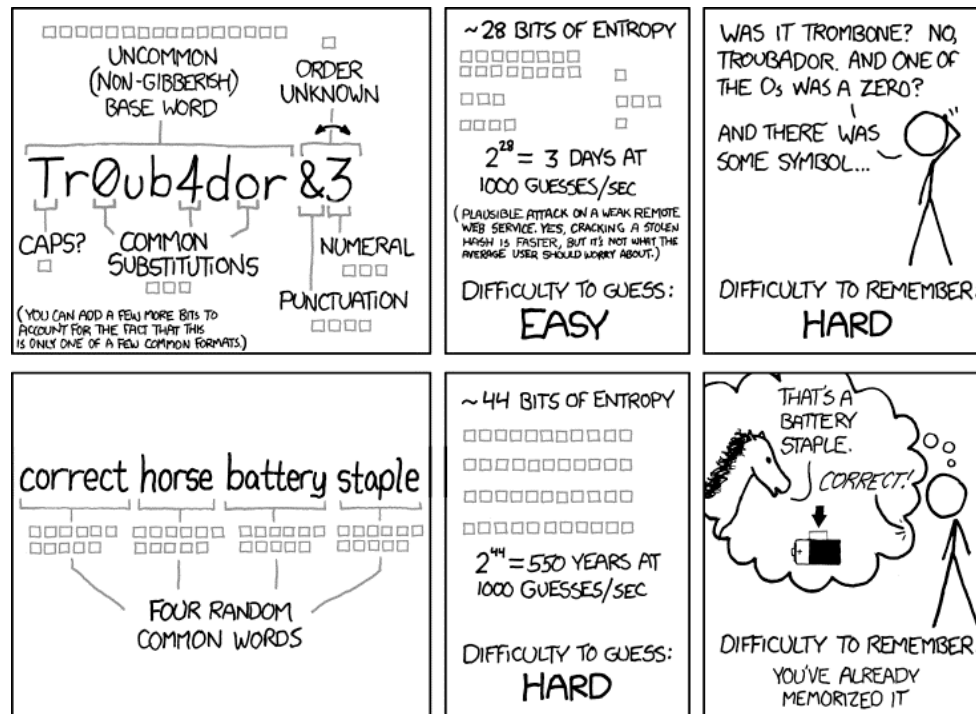
Sicherheit Benutzerkonten

- Überprüfen Sie die Sicherheit Ihrer Konten
 - Unter <https://haveibeenpwned.com/> können Sie herausfinden, ob Ihre Daten offen im Netz kursieren
 - Ändern Sie Dubletten der Login/Passwort-Kombi anderer Konten
- Geizen Sie mit persönlicher Information, persönlichen Details bei Sicherheitsfragen
- Verwenden Sie unterschiedliche E-Mail-Aliase für verschieden Konten
- Trennen Sie berufliche und private Nutzung

Server-Sicherheit

- Datensparsamkeit
- Zugriffsbeschränkungen z.B. für Datenbanken
- SSL / Sichere Verbindung für Kommunikation
- Sichere Authentifizierung
 - Regelung für Passwörter
 - SSH Authentifizierung mit public keys
- Software aktuell halten, Sicherheitslücken sofort schließen
- Regelmäßige Backups

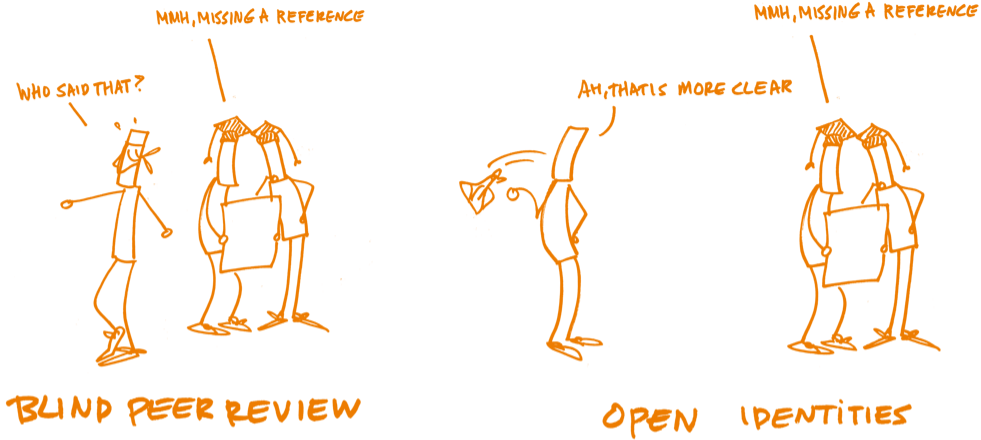
Bleiben Sie auf dem Laufenden



THROUGH 20 YEARS OF EFFORT, WE'VE SUCCESSFULLY TRAINED EVERYONE TO USE PASSWORDS THAT ARE HARD FOR HUMANS TO REMEMBER, BUT EASY FOR COMPUTERS TO GUESS.

6. Datenpublikation

MODES OF PEER REVIEW:



Was sind Repositorien?

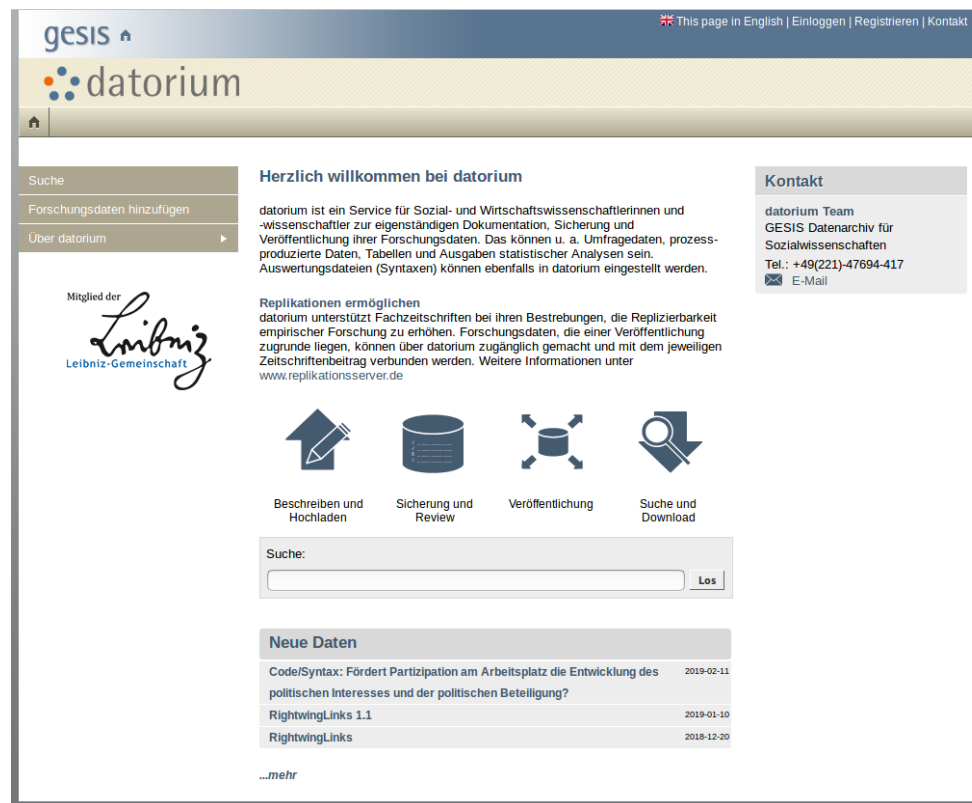
“Ein Repository (lateinisch repositorium, Lager‘) ist ein verwalteter Ort zur Aufbewahrung geordneter Dokumente, die öffentlich oder einem beschränkten Nutzerkreis zugänglich sind. In einem Archiv (lat. Archivum, Aktenschränk‘) werden dagegen ausschließlich historische Dokumente verwaltet.”

“Digitale Forschungsdaten-Repositories sind Informationsinfrastrukturen, die digitale Forschungsdaten möglichst dauerhaft... speichern und organisieren, um die Auffindbarkeit und Zugänglichkeit der Daten zu sichern...”

Quelle: Pampel (2014), Vortrag „Ausgewählte Aspekte digitaler Informationsversorgung“, HU Berlin

Repositorien...

Disziplinspezifische Repositorien sind in der Regel der beste Ort für Ihre Daten:



<https://datorium.gesis.org>

Repositorien mit Dini-Zertifikat

- standardisierte Beurteilung von (Dokumenten-) Servern und erbrachter Dienstleistungen
- Qualitätskontrolle für Dokumenten- und Publikationsservices
- Kriterien u.a.
 - Sichtbarkeit des Gesamtangebotes
 - Betreuung der Autorinnen und Autoren
 - Sicherheit, Authentizität und Integrität des technischen Systems
 - Langzeitverfügbarkeit und Auffindbarkeit der dort archivierten Dokumente
- In Deutschland derzeit ~50 DINI-zertifizierte Repositorien (ausschließlich Dokumenten-Repositorien mit Open Access-Anteil)

Vertrauenswürdige Repositorien: Zertifizierung (Data Seal of Approval)

- Kriterien für DSA (self assessment)
- Daten können im Internet gefunden werden
- Daten sind verfügbar (eindeutige Rechte und Lizenzen)
- Daten sind nachnutzbar (Format)
- Daten sind zuverlässig/vertrauensvoll
- Daten haben PIDs, so dass auf sie verwiesen werden kann

Zertifizierung (Data Seal of Approval)

The data repository ensures the integrity of the digital objects and the metadata. The data repository ensures the authenticity of the digital objects and the metadata. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.



Daten in den Repositorien sind nicht selbstredend

- Der Katalog erläutert nur unzureichend
- Deshalb sind Metadaten unabdingbar
- Detailinformationen auf den Landing Pages und in Zusatzdokumenten
- Landing Pages können langfristig nicht mehr zugänglich sind
- Datenproduzenten stehen für Nachfragen nicht mehr zur Verfügung

Daten sollten nachnutzbar sein!

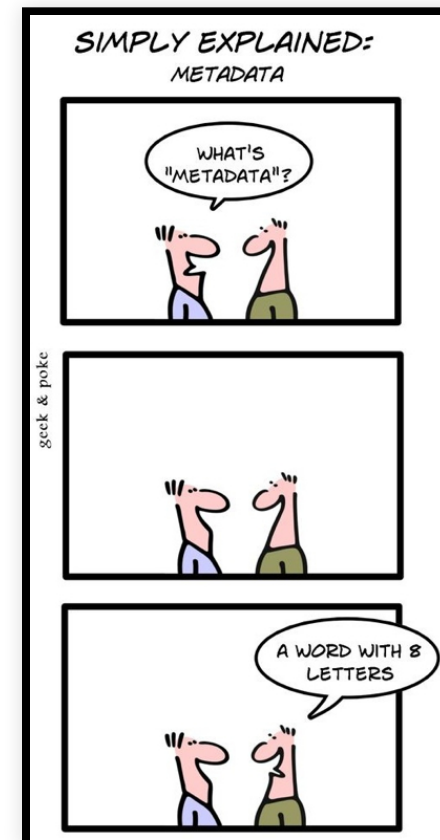
Damit Daten auch zukünftig nutzbar sind, müssen:

- Daten in einem Katalog gefunden werden
- Daten in einem Format sein, das noch geöffnet werden kann
- Formate in einer format-registry beschrieben sein
- alle Daten und Dokumentation der Archivierungsaktionen in einem Informationspaket zusammengefasst sein
- Daten nachvollziehbar sein (Lücken zwischen dem erwarteten Wissen künftiger Nutzer und dem notwendigen Wissen schließen)

Was sind Metadaten?

Metadaten sind „Daten über Daten“ und beinhalten beschreibende Informationen über den Kontext.

- bei der Datengenerierung verwendete Technologie (z.B. Hardware und Software)
- administrative Angaben (Projekt, Beteiligte, Institution usw.)
- Beziehungen zu anderen Daten (Publikationen, Personen, Projekten usw.)



Metadaten-Typen

Metadaten-Typ	Ziel	Beispiel
Beschreibende Metadaten ("descriptive metadata")	Identifizierung und Auffindbarkeit	Titel, Verfasser, Abstract, Schlagwörter...
Verwaltungsmetadaten ("administrative metadata")	Erstellung und Speicherung	Rechtemanagement und Zugriffsrechte, Aufbewahrung...
Strukturelle Metadaten ("structural metadata")	Aufbau der Ressource	Kapitel, Abschnitte, Struktur, Dateien...

Übung „Metadaten“

- Finden Sie sich in zwei Gruppen zusammen
- Suchen Sie sich ein Objekt aus
- Beschreiben Sie die Metadaten (10 min)
- Anschließend stellt jede Gruppe die Ergebnisse kurz vor.

Warum sind Metadaten wichtig?

- Metadaten sichern eine langfristige Datennutzung und Interpretation
- Gut dokumentierte Daten sind auch von für Personen, die nicht an ihrer Erhebung beteiligt waren, verständlich
- Strukturierte Metadaten können maschinell gelesen und verarbeitet werden, was eine schnelle, einfache und verlässliche Suche ermöglicht

Übersicht zu disziplinspezifischen Metadatenstandards

*<http://www.dcc.ac.uk/resources/metadata-standards>
<http://rd-alliance.github.io/metadata-directory/subjects/>*

- Stichworte
 - Data Documentation Initiative (Sozialwissenschaften)
 - GESIS; Datorium oder SSOAR (Social Science Open Access Repository)
 - Dublin Core (bibliographisches Datenformat)
 - Marc21 (bibliographisches Datenformat)

Hilfreiche Links

- Zentrum für FDM (<https://www.fdm.uni-hamburg.de/>)
- IR für GW <https://www.dokserv.gwiss.uni-hamburg.de>
- Forschungsdatenrepositorium <https://fdr.uni-hamburg.de> (coming soon)
- DMP Online <http://dmp.fdm.uni-hamburg.de/> (coming soon)
- www.forschungsdaten.org
- www.forschungsdaten.info

Referenzen - Leseempfehlungen

- Ludwig/Enke, 2013, „Leitfaden zum Forschungsdatenmanagement“
- DFG, 2013, „Sicherung guter wissenschaftlicher Praxis“
- Gesis.org: Anwendung der Datenschutzgesetzgebung auf die empirische Sozialforschung (z.B. Watteler & Kinder-Kurlanda 2015), Sicherheitspraktiken in Organisationen (z.B. Poller et al. 2017), und Datenethik (z.B. Kinder-Kurlanda & Zimmer 2017, Kinder-Kurlanda & Ehrwein 2015). Katsanidou, Alexia, Laurence Horton, and Uwe Jensen. 2016. “Data Policies, Data Management and the Quality of Academic Writing.” International Studies Perspectives online first.



