Saving and Organizing data

- File Formats
- File Organization
- Workflows
- Versioning and Backups
- Securing Information

File Formats

Save the data: File formats

- = the way in which information is stored digitally
- File extensions (e.g. .txt, .docx, .jpg) provide information about file formats and thus the data they contain

Win and Mac systems do not display file extensions by default. The display can be activated via the system settings.

The Internet-Media-Type / MIME (Multipurpose Internet Mail Extension) ensures in the WWW that the correct application is assigned to the file formats

Media Type (MIME-Type) 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

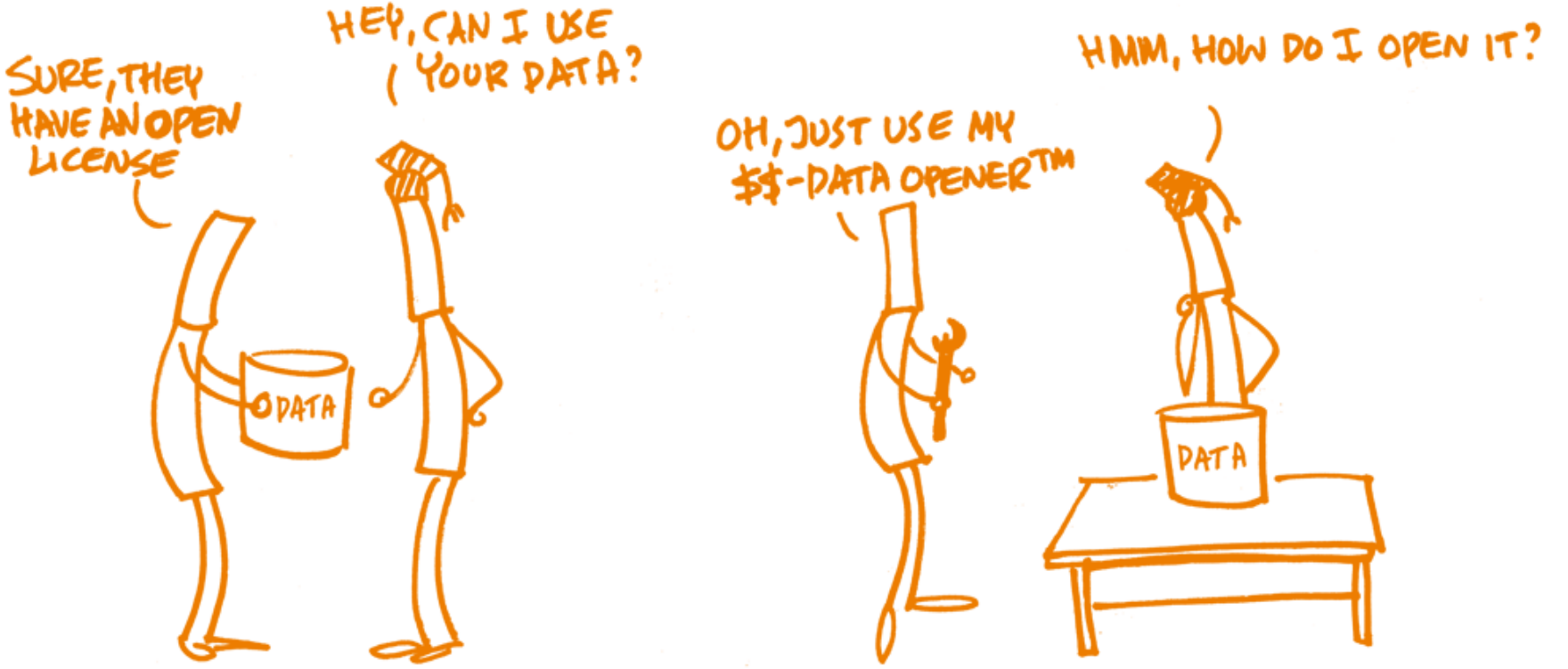
Open <-> proprietary file formats

Proprietary file formats

- Are mostly dependent on
  - specific software
  - corresponding software licenses
  - specific platforms / operating systems
- Are endangered by
  - software obsolescence (dependency on the market)
  - rapid technological development (incompatability between versions)
- conversion / export into open formats mostly lossy

Open file formats

- can be viewed without independant of specific
- are defined by a publicly available specifiction
- are mostly maintained by standards organizations (DIN, ANSI, JISC…)
- are not encumbered by any copyrights, patents, trademarks or other restrictions
- can be used without any costs and for any purpose


HEY, CAN I USE YOUR DATA?
SURE, THEY HAVE AN OPEN LICENSE
DATA
OH, JUST USE MY $$-DATA OPENER™
HMM, HOW DO I OPEN IT?
DATA

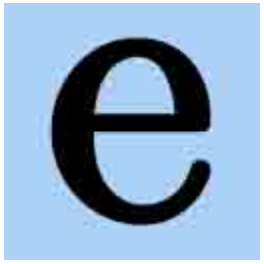
File Types Table

Filetype	Recommended	Avoid
Tabular data	CSV, TSV, SPSS portable	Excel
Text	TXT, ODT, HTML, RTF, PDF/A only if layouts are important	Word, PowerPoint
Multimedia	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	QuickTime, H264
Image	TIFF, JPEG2000, PNG	GIF, JPG
Structured data (Databases)	XML, RDF, JSON, CSV	RDBMS
Container	TAR, GZIP, ZIP	WAR

Lossless versus lossy image compression

PNG: 1007 Byte

JPEG: 4273 Byte



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

Example 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,Pferdeb
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
```

Example 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"
  ]
}
```

Summary - Recommendations

- General Recommendations
  - try to avoid branded or exotic formats
  - high quality
  - compression: select common formats (.zip, .tar.gz, .tar.bz2)
- Multimedia recordings
  - high quality
  - no lossy compression
- Texts
  - Unicode-based encoding
  - Open and documented formats
  - De Facto Standards (UTF-8, XML, TEI, JSON-LD, RDF…)

Organizing of Files and Folders

Strategies for file organization

- File types: Documents, Pictures, Videos…
- Task Type: legal, finances, analysis, consulting, documentation, forms
- Technology: Adobe PDFs, MS-Word-Files, CAD..
- Project names: project1, project2
- Folder templates: e.g. fixed subfolder structure for each project

Data Organization - Dos and Don'ts

Do	**don't**
choose your own system	settle on the system default
organize your files by purpose	organize your files by file type
use folders and subfolders	save everything on the desktop
choose filenames wisely	mistake it with a novel (technical maximum 260 character)
keep private things seperately	mix private files with professional files

Data Organization - Tricks and Tipps

Tipp	Advantage
save new files in a kind of inbox	collect everything before tidying up
seperate data from the application	easier to organize and backup the data; won't be overwritten by new installation
keep your data on a seperate hard drive /partition	easier to access and recover data after a system failure
use shortcuts / bookmarks for active projects	avoids copies and involuntarily diverged version
use archive-folder(s)	inactive data can clutter the view on your file system
tidy up regularly	saves space and nerves

File names

Name files and folders consistently

- find your own system and stick to it
- think about, what your typical search case would be if you misplaced a file
- Chain uniform components in a fixed sequence (e.g. project_type_date)
- Avoid of special characters and spaces
- Versioning
  - Decide on a versioning system
  - Decide how many/ what versions you want to keep

```
20180301_BohrkernSibirien_ABC_Original
20180301_Mikrofilm1997_XYZ_005
```

Stay consistent

```
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
```

Summary: Recommendations data organization

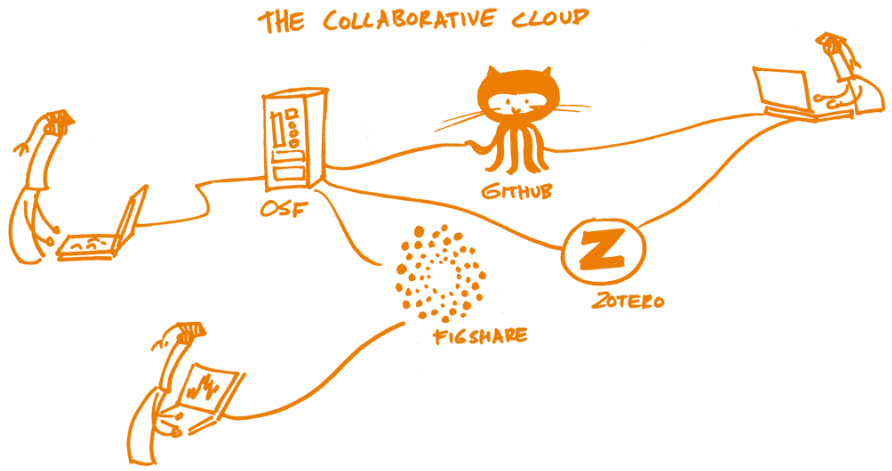
- Get detailed information about file formats & software
- Try to adhere to standards and common formats
- Structure and document your data and files in such a way that they are understandable even without your knowledge
- Use central backup storage systems and versioning for data backup
- Select tools to suit your requirements
- Delete data you no longer need

Ready to clean up?

Organizing Data - Team & Workflow

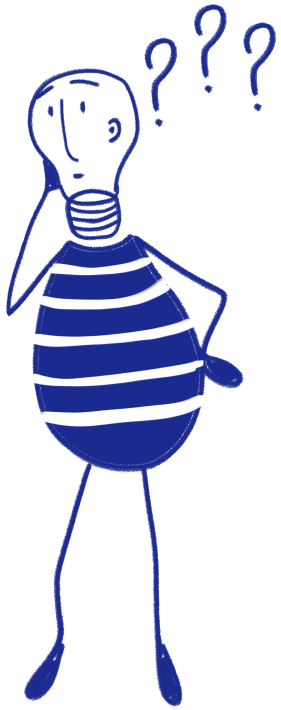
Working in Teams

- Clarify conventions and procedures for
  - storage location
  - synchronization & access
  - versioning
  - lifespan
  - security



Commercial solutions:

Dropbox, Google & Co?



Commercial storage solutions

- Sustainability
- How long data is kept (AGB)
- Does the service eventually become costly?
- Who has access / rights to my data?
- In which format will my data be saved?
- Can I delete it again?

Services @ RRZ

Service	Folders & Files	Calendar	Office Applications	Access	Sync
UHH-Share	+	-	-	UHH	Desktop
UHH-Disk	+	-	-	UHH	Desktop / Cifs
Sharepoint	+	+	+	UHH	-
Nextcloud (2020?)	+	+	+	UHH	Desktop

Choice of means

- Which function must the service fulfil?
- Which persons (groups) should have access to it (external project partners)?
- How secure is the service (backups, redundancy, availability)?
- What happens if I want / have to change the service?
- Which terms of use are linked to the use?
- Are there contractual / legal obligations to use specific storage (funding agency)

Workflow tools

Synchronization and access

Commercial project organisation tools

- Mostly they offer an all-round carefree package
- Many open source projects offer a hosted service
- Some companies also specialize in the scientific field
- Observe data security, data protection and terms of use
- Many features (e.g. workflows, storage and exchange of data, communication, version control, parallel working)
- Check future viability and location of the company (server)

Virtual Research Environments Checklist

- Access control
- Data integrety
- Data - silmutaneaous work
- Communication
- Sustainability - Import / Export
- Stability

Host your own web application

Pros	Cons
free choice of tools & apps	often limited functionality
your own research environment	administrative effort
updates according to your schedule	responsibility for security, data protection and privacy issues

Documentation - Tools

- Wiki software (e.g. Mediawiki, Wiki in VREs), e.g. for user documentation
- Gitbooks (TUHH - Hamburg Open Online University)
- Readme files
- Software documentation tools for technical documentation

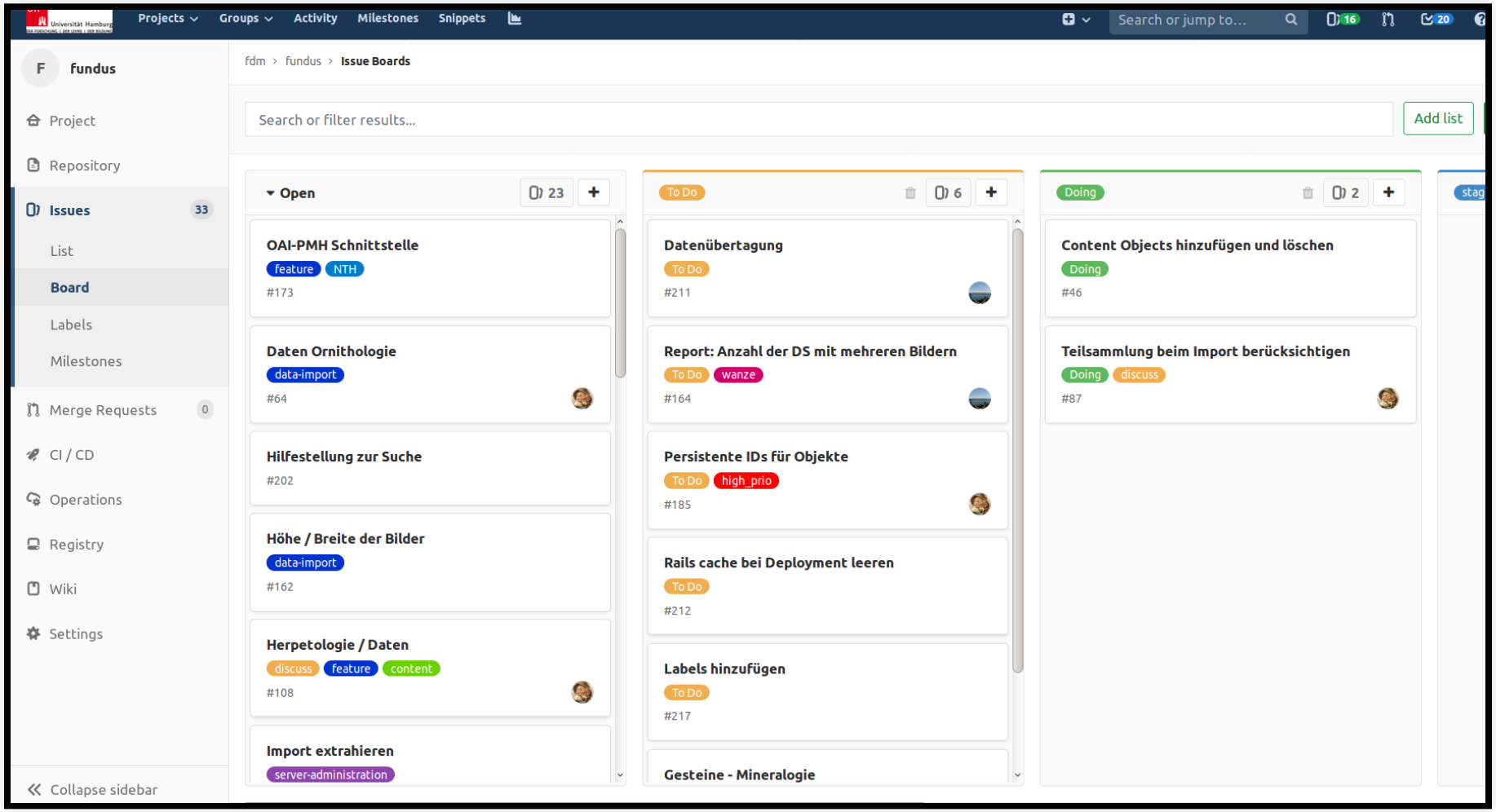
Versioning and backup

Life span of Data

- magnetic storage media: 10~20 years
- CDs / DVDs : shelf life 5~100 years, CD-RW 2~3 years
- flash drives: 50~100 years
- usb flash drives: 5-75 years

Gitlab @ RRZ

- DevOps platform (open source version of github)
  - versioning of source code / text files
  - rights management on the basis of projects
    - Repository (folders and files)
    - Issue Tracking (Issues and Assignees), *kanban*
    - Metrics
    - Wiki
    - CI uvm…
  - UHH & external partners (register via ServiceLine)


Projects
Groups
Activity
Milestones
Snippets
Search or jump to...
fundus
Project
Repository
Issues 33
List
Board
Labels
Milestones
Merge Requests 0
CI / CD
Operations
Registry
Wiki
Settings
Collapse sidebar
fdm › fundus › Issue Boards
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
Datenübertagung
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

Versioning non-textual files

- Activate versioning in Office programs, image processing, etc.
- Checking persistence across sessions
- Be sure to make regular backups anyway
  - automated TSM-Backup @ RRZ
  - on servers
  - on external storage media

Take-away message: 3-2-1 Rule of backup

- keep 3 separate copies of your data
- at least on 2 different storage media devices (e.g. laptop, share/cloud *one copy not synchronized*)
- 1 copy must be physically separated from the others (in case of natural disaster)

CASH REWARD

for returning my lost backpack

- Black [AK] Burton Rucksack
- Lost on Friday 15. July at 8 pm in the Panton Arms pub 43, Panton St. Cambridge
- Containing a laptop (white MacBook), a black external hard drive and scientific research documents

The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

If you found it, I would be extremely grateful if you could return it to the Panton Arms or contact me on: [redacted]

Thank you!!

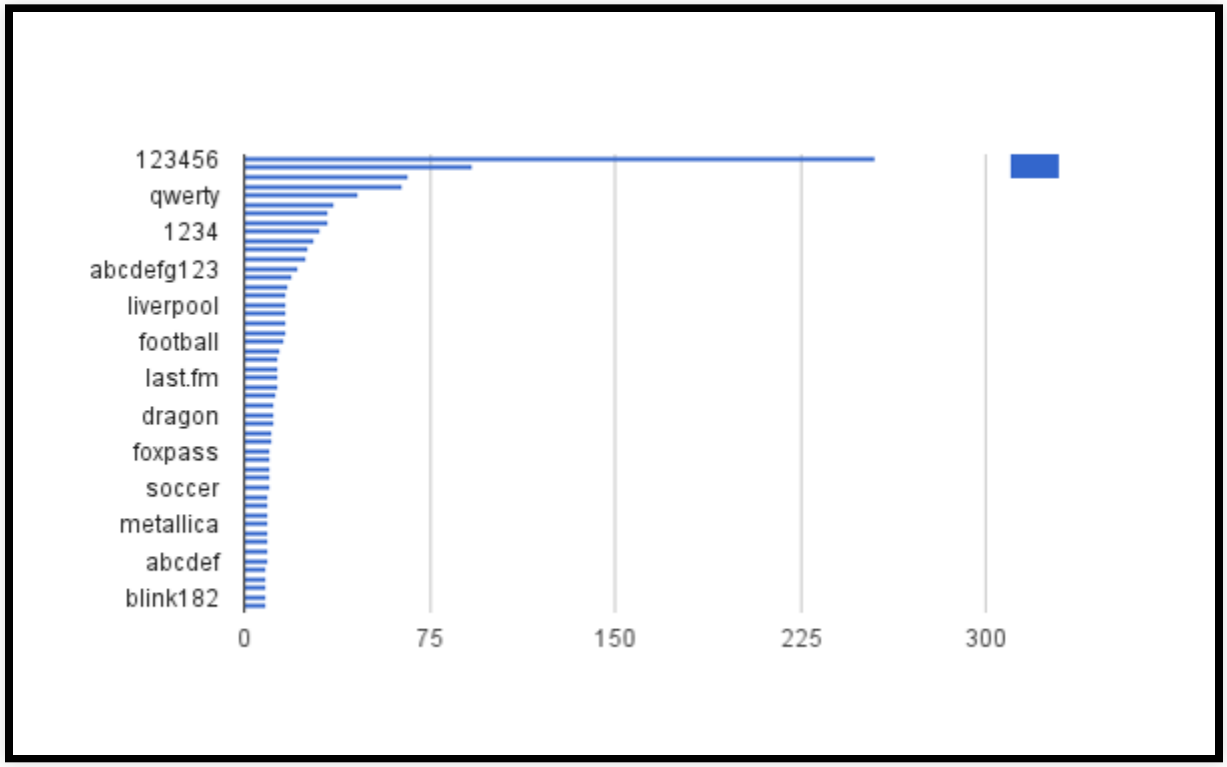
Securing your data

Data breach examples

Organisation	No of records	Reason	year
Facebook	540 Mio	because of poor security	2019
Yahoo	500 Mio	hacked	2014
Yahoo	3 Billion	hacked	2013
Microsoft	250 Mio	misconfiguration	2019
Google Plus	500,000	poor security	2018
Capital One	106 Mio	unsecured S3 bucket	2019
University of Utah Hospital & Clinics	2.2 Mio	lost / stolen media	2008

https://en.wikipedia.org/wiki/List_of_data_breaches

Most common passwords (lastfm)



Password security

- Do not reuse *important* passwords
  - generate passwords automatically
  - think up a system to create variants of a password
- Secure passwords contain
  - no terms from dictionaries
  - many letters / numbers / charactersgit status
- Pay attention to how your password is transferred / stored
  - ensure a secure connection during transmission (https://)
- 2-factor authentication, encryption
  - biometrical data
  - fido2 usb stick
  - smartphone & PC

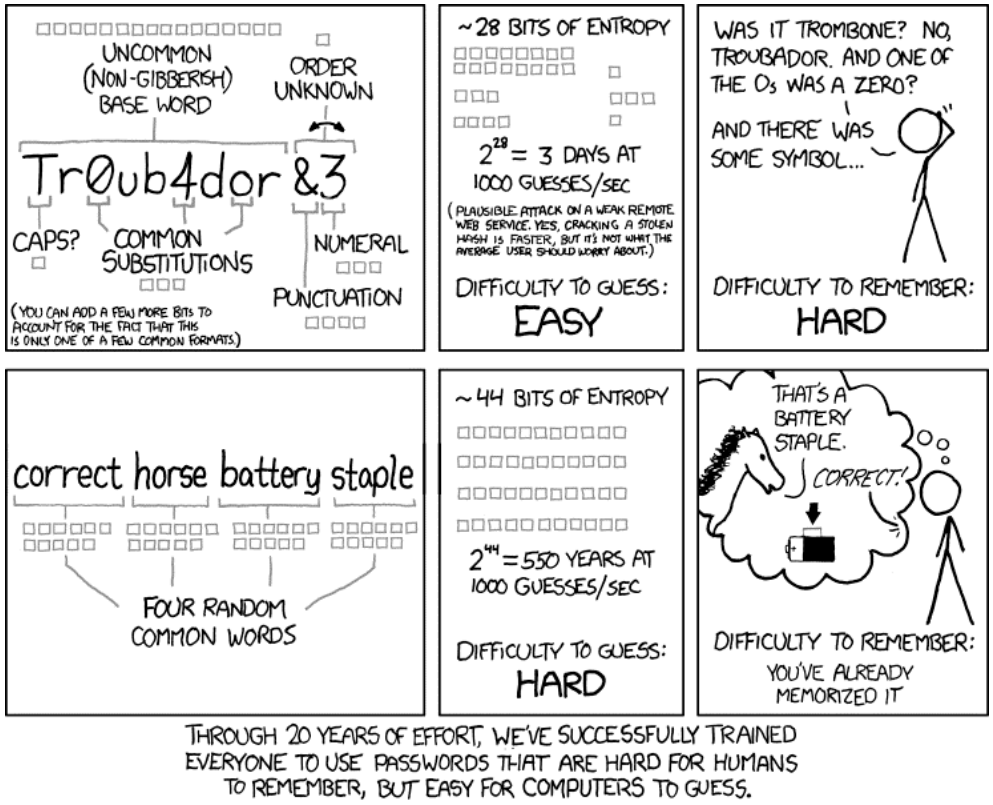
Risks

- SPAM, Phishing, Trojanen mails
- Sale in Darknet (especially streaming services)
- Purchasing & diversion to Packstation…
- Place fraudulent advertisements via sales platforms (eBay, Amazon)
- **Loss of your data / your work**

User Accounts

- Check the security of your accounts & password
  - at https://haveibeenpwned.com/ you can find out whether your data is circulating openly in the network
  - change duplicates of the login/password combination of other accounts
- Be sparing with personal information, personal details on security issues
- Use different email aliases for different accounts
- Separate professional and private use

Stay up to date



https://xkcd.com/

References - Reading recommendations

- Ludwig/Enke, 2013, „Leitfaden zum Forschungsdatenmanagement“
- Gesis.org: Anwendung der Datenschutzgesetzgebung auf die empirische Sozialforschun
  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 Mana
  Writing.” International Studies Perspectives online first.
- Safeguarding Good Scientific Practice:
  https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wi