



Open Data: Organisation und Veröffentlichung von Forschungsdaten

Vortragender: Kevin Lang
Donnerstag, 05.06.2020

Agenda

Vorstellung

Warum Open Data?

Drei Modelle zum Umgang mit Forschungsdaten

- Five S Data Model
- 3-2-1 Backup-Regel
- FAIR Data Prinzipien

Diskussion

Vorstellung

Zur eigenen Person

Kevin Lang, Master of Science

- **2011 – 2016:** Bachelor-Studium, Medieninformatik
- **2016 – 2018:** Master-Studium, Computer Science and Media
- **seit 2015:** Arbeit bei Webis
 - Schwerpunkte: Natural Language Processing, Machine Learning, Artificial Intelligence und Big Data
- **seit 2018:** Servicestelle für Forschungsdaten und Mitglied des Thüringer Kompetenznetzwerk Forschungsdatenmanagement (TKFDM)

TKFDM: Services

- **Beratung**
 - Datenmanagementplan, rechtliche Grundlagen, Software, Portale, Formate, ...
- **Schulung**
 - thüringenweite Schulungen, Workshops, Train-the-Trainer und Informationsveranstaltungen
 - auch auf Anfrage möglich
- **Vernetzung**
 - Zwischen den Hochschulen, Forschungseinrichtungen, NFDI-Konsortien und anderen Interessierten



TKFDM: Portal



The screenshot shows the top portion of the TKFDM website. The header is a dark green bar with a network diagram background. It contains navigation links: [AKTUELLES](#), [VERANSTALTUNGEN](#), [MATERIAL](#), [UNSER NETZWERK](#), and [KONTAKT](#). In the top right corner, there is a language selector: [Deutsch](#) | [English](#). Below the header is a white circular logo featuring a stylized green and grey leaf-like shape with the text "FDM THÜRINGEN" underneath. The main content area is white and contains the following text:

Willkommen auf den Seiten des Thüringer Kompetenznetzwerks Forschungsdatenmanagement (TKFDM)

Das TKFDM ist Ansprechpartner für Forschende aller Thüringer Hochschulen im Bereich Forschungsdatenmanagement. Dies umfasst alle Aspekte des Umgangs mit Daten, die während des Forschungsprozesses entstehen.

Das TKFDM vernetzt die an den Thüringer Universitäten (Universität Erfurt, Technische Universität Ilmenau, Friedrich-Schiller-Universität Jena, Bauhaus-Universität Weimar) angesiedelten Beratungsstellen zum Forschungsdatenmanagement und stellt seine Services **allen Thüringer Hochschulen** zur Verfügung. Wir bieten Beratung und Schulungen zu allen Fragen rund um das Thema Forschungsdaten. **Kontaktieren Sie uns!**

www.forschungsdaten-thueringen.de

TKFDM: Materialien und Veranstaltungen

- Verschiedene Informationsflyer
- Handreichungen & Best-Practices
 - Forschungsförderung und Anforderungen, Forschungsdatenrepositorien, Virtuelle Forschungsumgebungen, ...
 - Qualitätskontrolle, eLabFTW, GitLab, LaTeX, Großprojekte, ...
- 23 Dinge zu Forschungsdatenmanagement
- ScaryTales
 - über 50 Geschichten zu schlechten Datenmanagement



Warum Open Data?

Gründe für Open Data



Zugang fördern und erleichtern



Transparenz



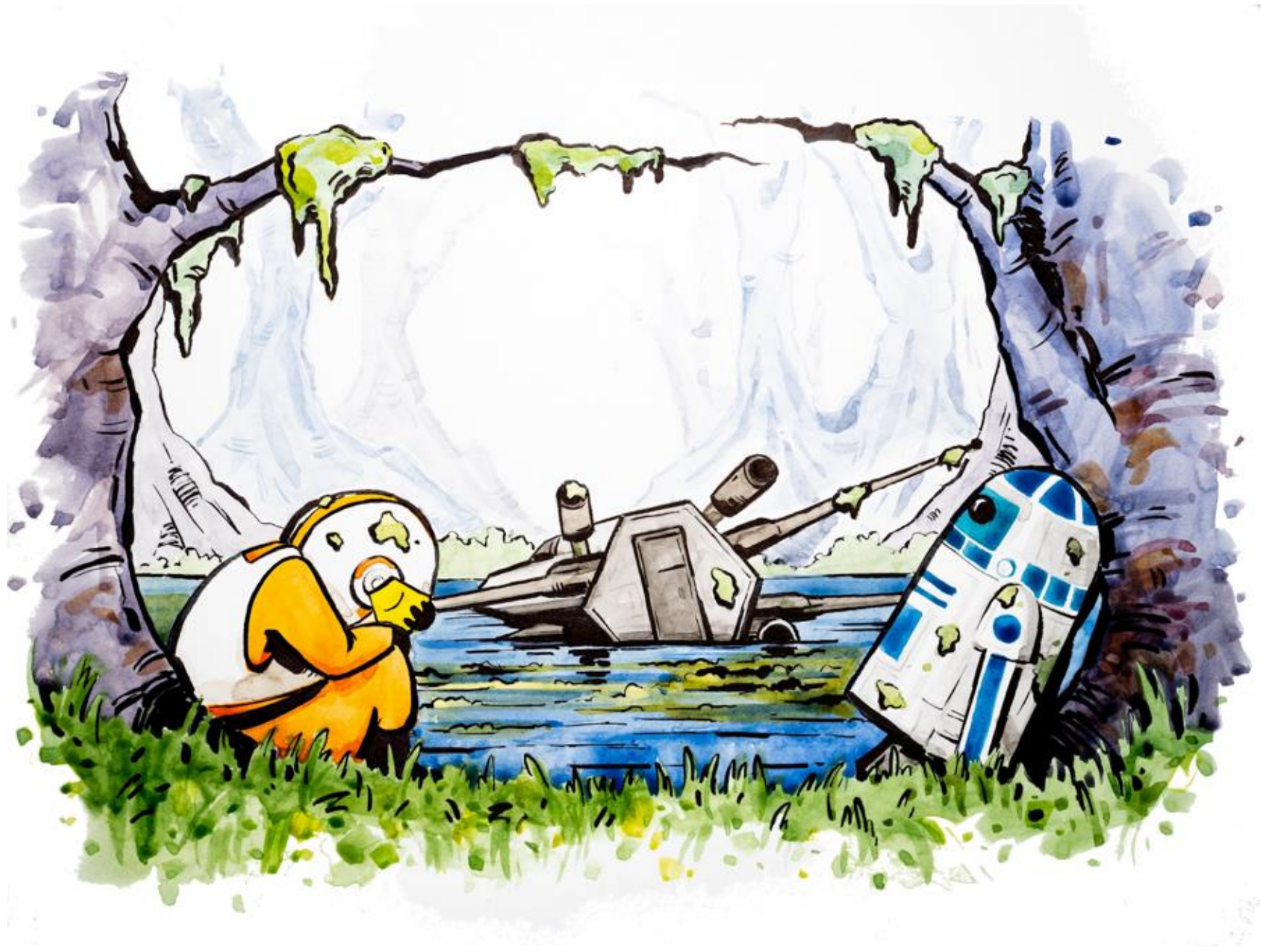
Datenverlust reduzieren



Forschungsförderung

- **Problem...**

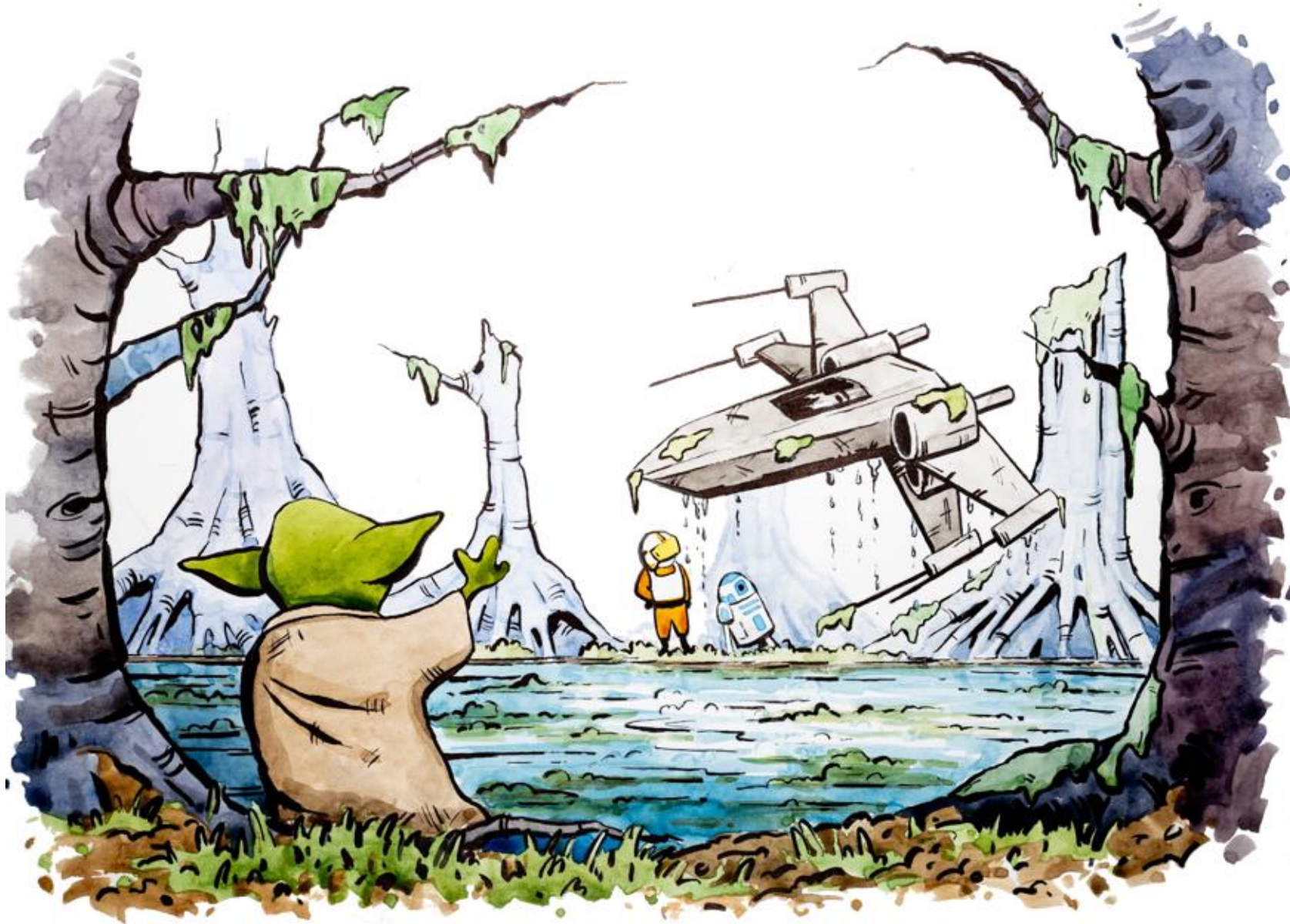
- Fehlende Kenntnis zu Open Science
- Probleme mit dem Umgang von Daten (keine "Digital Natives")





Problembeispiele in der Forschung:

- Datensatz nicht kompatibel
- Unorganisierter oder fehlerhafter Programmiercode
- Fehlende Zeit am Ende des Projekts





Wo Open Data die Forschung verbessert:

- Open Source Programme und Sprachen
- Kollaborative Zusammenarbeit
- Teilen von Präsentationen oder Vorlesungen





Es gibt nicht den Datenexperten für alles...

- Planung, Organisation und Rechtliches
- Skripte und Automatisierung
- Analyse und Kooperation
- Präsentation, Werbung und Veröffentlichung
- Speicherung und Archivierung

Open Data...

Forschungsdaten- management!



Modelle zum Umgang mit Forschungsdaten

Übersicht

- **Five S Data Model**
 - Organisation von Forschungsdaten
- **3-2-1 Backup Regeln**
 - Sicherung von Forschungsdaten
- **FAIR-Prinzipien**
 - Veröffentlichung von Forschungsdaten

Five S Data: Ursprung

- Five S stammt aus dem Konzept von „Kaizen“ aus dem Toyota-Produktionssystem, welches soviel bedeutet wie: „*Die Handlung, Schlechtes besser zu machen*“

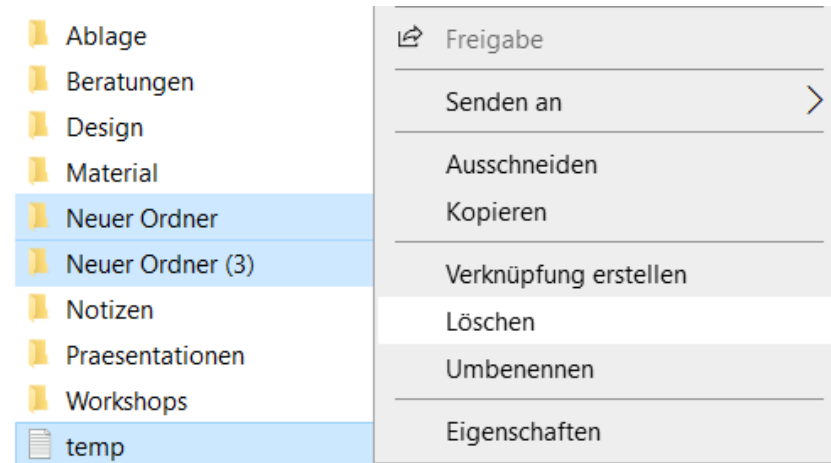
改善
“kai” “zen”

1	Seiri	Sort	Aussortieren
2	Seiton	Set in Order	Aufräumen
3	Seiso	Shine	Arbeitsplatzsauberkeit
4	Seiketsu	Standardize	Anordnung zur Regel machen
5	Shitsuke	Sustain	Alle Punkte einhalten und verbessern

- bekannt bei herstellungsorientierten Unternehmen wie Toyota, Boeing oder Hewlett-Packard
- Bezug zu Forschungsdaten von RDA geschaffen

Five S: 1) Sort

- Ziel:
 - Zeitgewinnung bei der Suche
 - Gewinnung von mehr Datenspeicher
- Umsetzung:
 - unnötige Dateien/Ordner löschen, oder für geplante Löschung markieren
 - Vor temporären Dateien sauber halten

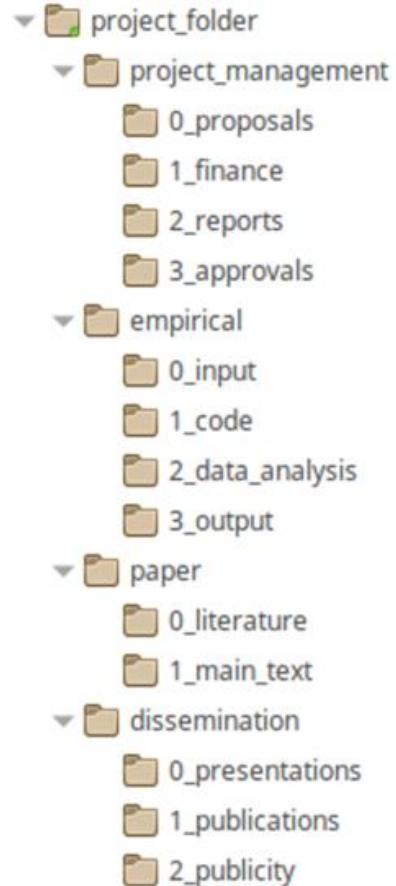


Five S: 2) Set in Order

- Ziel:
 - Ordnerstrukturen aufbauen, um Arbeitsvorgänge zu vereinfachen
- Umsetzung:
 - Sinnvolle Ordnerstrukturen
 - Namenskonventionen
 - Readme für Erklärungen von Ausnahmen



Gutes Beispiel:



Schlechtes Beispiel:

A screenshot of a file explorer showing a flat, unorganized folder structure. The root folder is 'project_folder', which contains a single subfolder 'project_management'. Inside 'project_management', there are numerous subfolders, including '1_PROJEKTIT', '2_PEOPLE', '3_MANUSCRIPTS', '4_APPLICATIONS AND GRANTS', '5_SEMINARS and MEETINGS', '6_PERSONAL', '7_ADMIN', '8REFEREE TASKS', '9ABSTRACTS AND PRESENTATIONS', '10LUENNOT JA OPETUS', '11POPULAR SCIENCE', 'Articles', 'Bernasconi_thesis', 'elokuvia', 'Jonna's documents_old', 'Photos', 'photos_from_anytrans', 'R', 'Sample lists', and 'Team speciant'. The folders are not organized into a clear hierarchy, and the names are inconsistent and uninformative.

Name	Date Modified	Size	Kind
▶ 1_PROJEKTIT	9 May 2019 at 11.39	--	Folder
▶ 2_PEOPLE	25 Mar 2019 at 9.10	--	Folder
▶ 3_MANUSCRIPTS	25 Apr 2019 at 13.05	--	Folder
▶ 4_APPLICATIONS AND GRANTS	Yesterday at 23.46	--	Folder
▶ 5_SEMINARS and MEETINGS	12 Mar 2019 at 20.54	--	Folder
▶ 6_PERSONAL	21 Dec 2018 at 12.44	--	Folder
▶ 7_ADMIN	28 Mar 2019 at 9.48	--	Folder
▶ 8REFEREE TASKS	7 May 2019 at 13.27	--	Folder
▶ 9ABSTRACTS AND PRESENTATIONS	28 Feb 2019 at 10.52	--	Folder
▶ 10LUENNOT JA OPETUS	8 May 2019 at 21.38	--	Folder
▶ 11POPULAR SCIENCE	17 Jan 2019 at 21.34	--	Folder
▶ Articles	21 Dec 2018 at 12.27	--	Folder
▶ Bernasconi_thesis	6 Jul 2016 at 14.48	--	Folder
▶ elokuvia	18 Dec 2017 at 4.44	--	Folder
▶ Jonna's documents_old	Today at 13.08	--	Folder
▶ Photos	14 May 2019 at 0.15	--	Folder
▶ photos_from_anytrans	15 Nov 2017 at 23.12	--	Folder
▶ R	31 Jan 2019 at 9.03	--	Folder
▶ Sample lists	20 Jul 2015 at 12.03	--	Folder
▶ Team speciant	12 Mar 2018 at 13.53	--	Folder

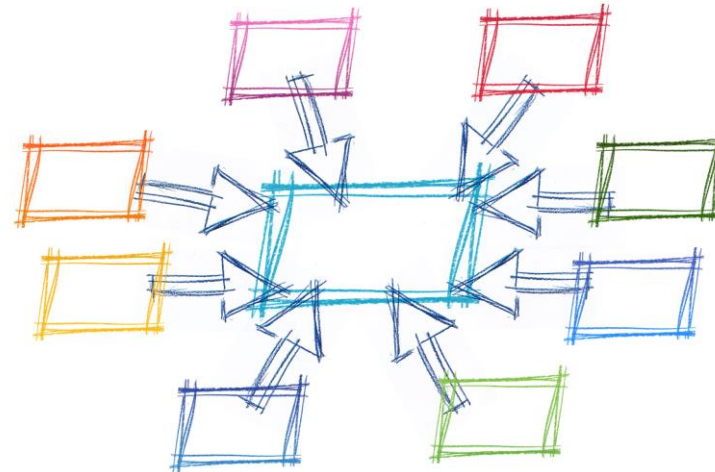
Five S: 3) Shine

- Ziel:
 - Qualität erhalten und ggf. anpassen
 - dokumentieren und für Mitarbeiter verständlich halten
- Umsetzung:
 - Vorgänge kontrollieren und selber durchsetzen
 - regelmäßige Routinen



Five S: 4) Standardize

- Ziel:
 - Prozesse und Termine etablieren, um erste 3 "S" zu ermöglichen
- Umsetzung:
 - Best-Practices, Leitlinien und Regeln dokumentieren
 - Diskussion, Verantwortlichkeiten klären



Five S: 5) Sustain

- Ziel:
 - mit Selbstdisziplin und Gewohnheit entwickelte Prozesse erhalten
- Umsetzung:
 - Praktiken und Organisation erhalten
 - Übungseinheiten mit Gruppe und neue Mitglieder einweisen
 - Verbesserungen einarbeiten



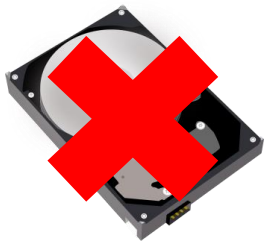
3-2-1 Backup Regeln: Ursprung

- historische goldene Regel in Unternehmen zur Organisation von Backups
- 2009 als Regel 3-2-1 zum ersten mal formuliert von Fotograf Peter Krogh
 - 3 Kopien
 - 2 verschiedene Technologien
 - 1 externer Standort
- erweitert als 3-2-1-0 oder 3-2-1=0 Backup Regel
 - 0 Fehler bei Wiederherstellung

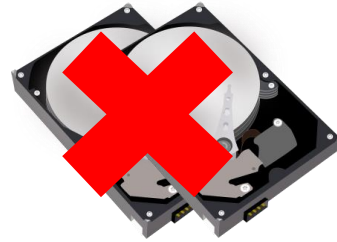


3-2-1 Backup Regeln: 3 Kopien

- 1 Arbeitskopie auf dem Arbeitsplatzrechner, 2 Backup-Kopien
- Warum 3? **Stochastik!**



**1 zu
100**



**1 zu
10 000!**



**1 zu
1000 000!!!**

- im privaten Gebrauch meist 2 Kopien ausreichend
- bei Institutionen und Gewerbe 3 oder mehr Kopien

3-2-1 Backup Regeln: 2 verschiedene Technologien

- unterschiedliche Fehleranfälligkeiten bei verschiedenen Speichertechnologien und Medientypen
- Unterschiede in:
 - Anzahl von Schreib-/Lesezyklen
 - Haltbarkeit des Materials
 - Schnittstellen
 - Sicherheitsaspekte
- durch Medienbruch lässt sich Datenverlust verringern



3-2-1 Backup Regeln: 1 externer Standort

- Was passiert, wenn alle Kopien an einem Ort sind?
- Mögliche Probleme:
 - Brand
 - Überschwemmung
 - Stromausfall
 - Hacker-Angriffe
 - Netzwerkbefehle/
Softwareupdates
- Getrennte Räume und Technik



3-2-1 Backup Regeln: 0 Fehler

- 0 Fehlertoleranz bei der Wiederherstellung von Dateien
- Welche Probleme können passieren?
 - Vorgang/Kontakt
 - Backup/Versionierung deaktiviert (Default Settings)
 - zu große Intervalle
 - Art und Stand der gesicherten Dateien
- Immer System testen und richtig konfigurieren



FAIR-Prinzipien: Ursprung

- Erste Veröffentlichung im März 2016
 - **F**indable (Auffindbarkeit)
 - **A**ccessible (Zugänglichkeit)
 - **I**nteroperable (Anwendbarkeit)
 - **R**eusable (Wiederverwendbarkeit)
- Hohe Akzeptanz in vielen Organisationen (G20, GO FAIR, CODATA, RDA, DFG, ...)
- Versuche zur Erweiterung:
 - ausführlichere Erklärungen zur Umsetzung
 - CARE-Prinzipien



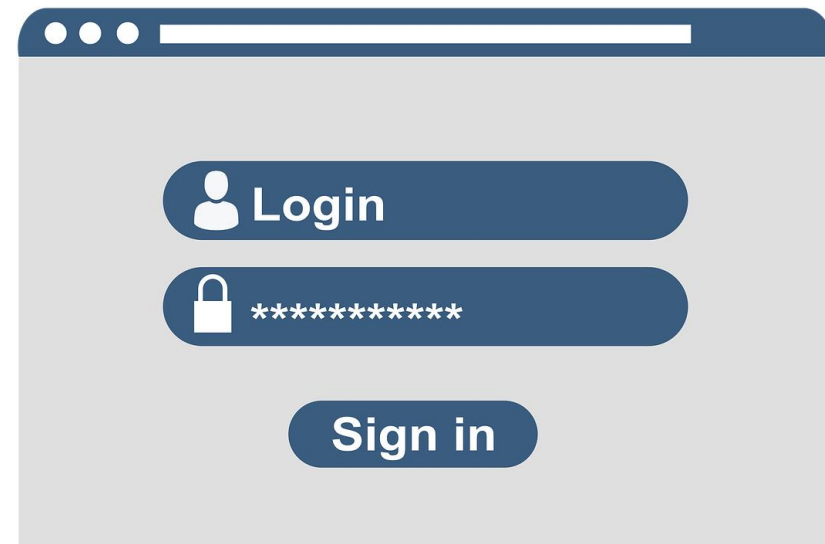
FAIR-Prinzipien: Findable

- Ziel:
 - Die Daten sollten mit viel Kontext auffindbar sein.
- Umsetzung:
 - Persistenter Identifikator
 - viele Metadaten
 - Repository (→ re3data.org)
 - Auffindbar in Suchmaschinen oder Register



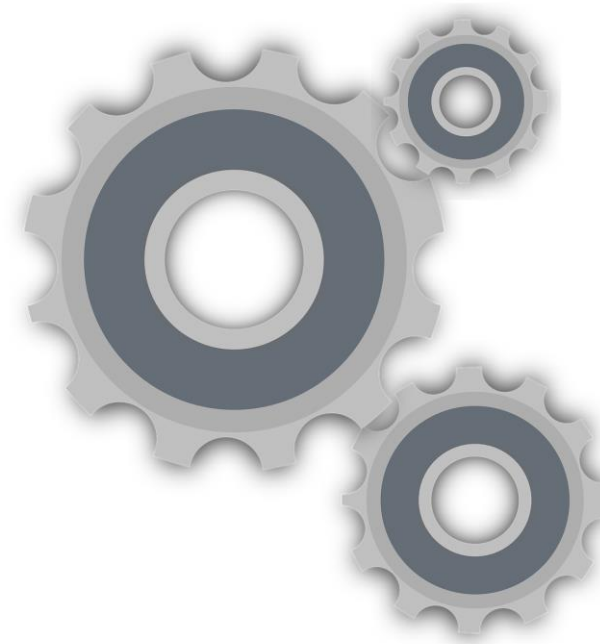
FAIR-Prinzipien: Accessible

- Ziel:
 - klar definierter Zugang für die Daten
- Umsetzung:
 - Downloadmöglichkeiten oder API
 - Authentifizierungs- bzw. Autorisierungsschritte
 - Bedingungen
 - bestehende Metadaten



FAIR-Prinzipien: Interoperable

- Ziel:
 - Die Datenformate und Metadaten entsprechen üblichen Standards
- Umsetzung:
 - Verwendung offener Formate
 - Vokabular nach Standards
 - Referenzen zu verwandten (Meta-)Daten



FAIR-Prinzipien: Reusable

- Ziel:
 - Dokumentation und Verwendungsmöglichkeiten
- Umsetzung:
 - Datennutzungslizenz vergeben
 - Beschreibung der Erstellung/Herkunft und Veränderung
 - Verweis auf benötigte Anwendungen



Offene Diskussion

Vielen Dank für Ihre Aufmerksamkeit.

Quellen

- Thüringer Kompetenznetzwerk Forschungsdatenmanagement:
 - [Portal von TKFDM](#)
 - [TKFDM Community auf Zenodo](#)
 - [Research Data Scarytales](#)
- Motivation:
 - <https://www.labfolder.com/why-we-need-open-data-access/>
 - <https://www.openscapes.org/media/>
 - Star Wars Artworks von [Allison Horst](#)
- Stockimages mit CC0 Lizenz von:
 - [pixabay.com](#)
 - [unsplash.com](#)

Quellen

- Five S DATA:
 - <https://www.helsinki.fi/en/research/organizing-data-folders-with-5sdata-method>
 - <https://kanbanize.com/de/lean-management-de/wert-verschwendung/was-sind-die-5s-in-lean>
- 3-2-1 Backup:
 - <https://www.storage-insider.de/was-ist-die-3-2-1-backup-regel-a-782641/>
 - <https://spanning.com/blog/simplifying-the-3-2-1-rule-for-data-protection-in-the-age-of-the-cloud/>
- FAIR Principles:
 - [The FAIR Guiding Principles for scientific data management and stewardship](#)
 - [FAIR Principles auf go-fair.org](#)
 - [How FAIR are your data?](#)
 - <https://www.gida-global.org/care>