

# Best Practices Workshop

## (Forschungsdatenmanagement an kleinen Instituten)



Gefördert durch:



Bundesministerium  
für Bildung  
und Forschung

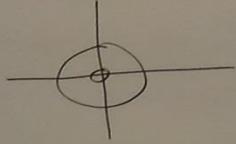
# Agenda

- 13:00 Rückblick Auftaktworkshop (Michael)
- 13:10 Zusammenfassung der Projektleiterinterviews (Hauke)
- 13:30 Best Practices für „Daten finden und bekommen“ (Hauke)
- 13:40 Konzept für neuen Daten-Workflow (Hauke)
- 13:50 Diskussion / Feedback
- 14:05 Pause
- 14:10 Best Practices für „Datenanalyse“ (Michael)
- 14:20 Festlegen der beiden Testprojekte (Michael)
- 14:25 Abstimmung / Diskussion / Feedback (alle)
- 14:40 Kommunikation (Michael)
- 14:55 Zusammenfassung (Michael)

# Rückblick: Auftaktworkshop

10

## Finden + Bekommen



EINFACHE, EINHEITLICHERE STRUKTUR (ORDNER,  
ALLGEMEINE ROHDATA ZENTRAL ABLEGEN UND  
FÜR ALLE ZUGÄNLICH MACHEN } <sup>AUF</sup> DATENSCHUTZ!  
LANGFRISTIGE SICHERUNG VON ROHDATA

↑  
REGELMÄSSIG AUFRÄUMEN + SYNCHRONISIEREN  
EINHEITLICHE BENENNUNG  
DATENTYPEN KATEGORISIEREN (+ METADATEN)  
ROHDATA - EXTERN/INTERN

⊖

UNTERORDNER SEHR INDIVIDUELL  
ZU VIELE VERSIONEN  
KEINE EINHEITLICHE NOMENKLATUR  
(DATEINAME, DATEN-LABEL, ...)

⊕

TOP-LEVEL OLAY

# Rückblick: Auftaktworkshop

2



Analysieren, Visualisieren, (Kommuniz.)  
nachvollziehbare Analyse  
Analyse in Abhängigkeit von Art und Umfang d. Daten  
+ Fragestellung  
Abb. müssen leicht zu ändern sein



Experten Pool - wer kann was?

Experten Sprechstunden

Leitfaden als Vorbereitung  
Urheberrechtsfragen?

interner Workshop

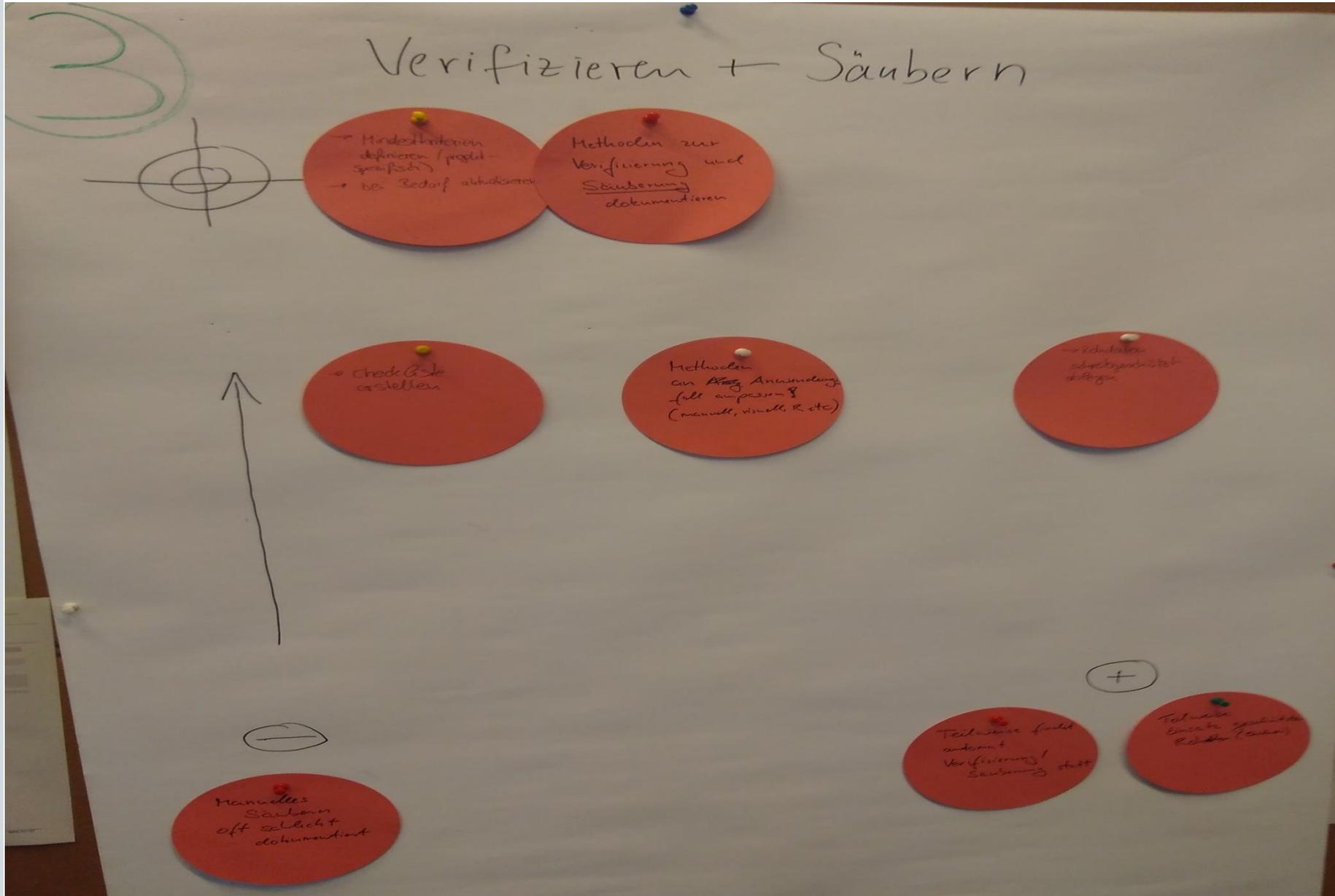


manche Abb. sind  
unlesbar  
noch große Vorbehalte  
gegenüber Veröffentlichung  
von Rohdaten



freie Wahl der  
Analysetools

# Rückblick: Auftaktworkshop



# Thema heute: Best Practices

## Problem (nicht nur bei uns!!!):

“

These days, data trails are often a morass of separate data and results and code files in which no one knows which results were derived from which raw data using which code files.”

— Professor Charles Randy Gallistel, Rutgers University

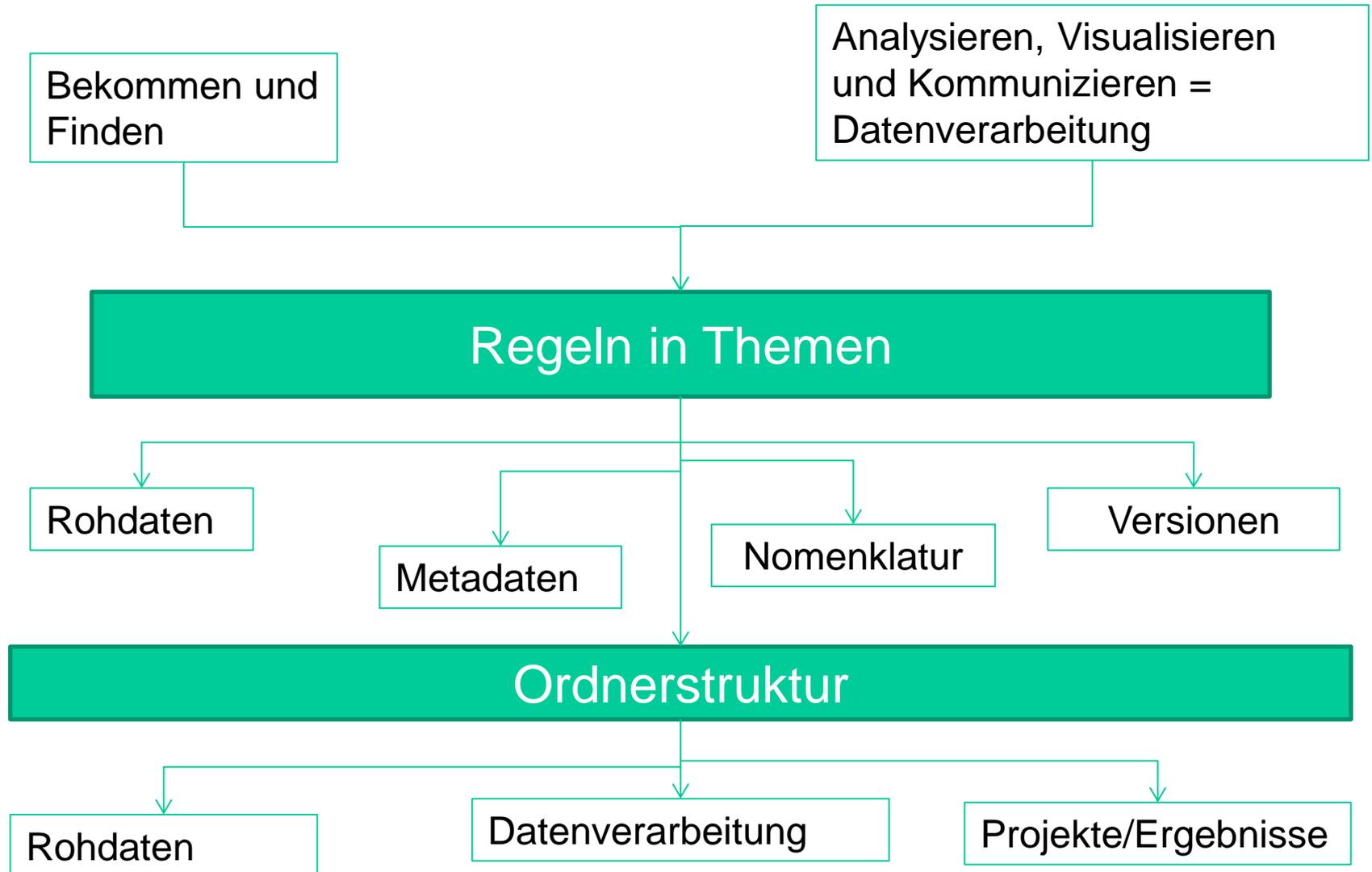
[Source: elifesciences.org \(2018\)](https://elifesciences.org)

## Unser Ansatz in FAKIN:

Entwicklung von **Best Practices** und deren **Anwendung in zwei KWB-Projekten** mit dem **Ziel einer transparenten, nachvollziehbaren Datenverarbeitung** (von Rohdaten bis zur Abbildung im Endbericht)

## Zusammenfassung der Projektleiterinterviews

# Von der Aktion über das Thema zur Bearbeitungs- und Ordnerstruktur



## Best Practices zu den Themen

## **Definition:**

Alles was in erster Version nicht von uns selbst erstellt wurde, z.B.:

- Loggerdaten von einem Messgerät
- Daten die von externen Partnern geliefert werden

## **Best Practices für Rohdaten-Dateien**

- dürfen umbenannt werden, dies muss in den Metadaten dokumentiert werden
- dürfen inhaltlich nicht verändert werden
- werden schreibgeschützt abgelegt
- werden in einem eigenen Bereich abgelegt (ein Ordner enthält alle Rohdaten)

## **Definition:**

Daten, die andere Daten beschreiben

## **Best Practices Metadaten**

- Mindestanforderungen definieren (unterschieden nach Roh- und verarbeiteten Daten)
- Metadatenstandards prüfen, z. B. DataCite (siehe u.a. ZALF, GFZ Potsdam)
- Werden wir am konkreten Anwendungsfall in den Testprojekten entwickeln

## Best Practices Ordner- und Dateinamen

- Keine Sonderzeichen, keine Umlaute, keine Leerzeichen
- Datum in der Form yyyy-mm-dd (2017-06-30)
- Zusammengesetzte Wörter (z.B. Projektnamen) in „CamelCase“
- Einheitliche Schreibweise von Projektnamen
- Einheitliche Sprache: englisch oder deutsch?
- Verwendung eines Vokabulars für wichtige Begriffe, z.B. „validiert“, „kalibriert“

# Versionierung

## **Option 1: Manuell**

- Wir machen einen Vorschlag

## **Option 2: mit Versionsverwaltungssoftware, z.B. Subversion**

- Verpflichtend für Programmcode und ggf. kleinerer Textdateien
- Aber: nicht geeignet für Rohdatenversionsverwaltung

## **Best Practices zur Ordnerstruktur und Konzept für neuen Daten-Workflow**

# Konzept für neuen Daten-Workflow

Für jedes Projekt Trennung von

## Rohdaten

\\server\rohdaten\$

\TestProjekt\...

**Wertvoll**  
(ggf. nicht  
reproduzierbar!)

*Nach  
Herkunft*

## Datenverarbeitung

\\server\datenverarbeitung\$

\TestProjekt\...

**„Spielwiese“**  
(viele Varianten,  
Versionen!)

*Nach Thema bzw.  
Bearbeitungs-  
schritt*

## Ergebnisse

\\server\projekte\$

\TestProjekt\...

**Nur „berichts-  
relevante“  
Ergebnisse**

*Nach Projekt-  
struktur*

## Best Practices Daten analysieren

- Auszuwertende Rohdaten werden zuerst in ein standardisiertes Format gebracht
- Die Auswertung beginnt bei den standardisierten Daten
- Vorteil: formale Änderungen werden nur einmal gemacht und nicht in jeder Auswertung erneut und ggf. verschieden
- Standards werden je nach Art der Rohdaten und der Weiterverarbeitung definiert
- CSV-Dateien werden in ein einheitliches Format gebracht
- Excel-Tabellenblätter, die (auch) automatisiert verarbeitet werden sollen, werden nach CSV exportiert und in das einheitliche CSV-Format gebracht
  
- Die Datenverarbeitung erfolgt getrennt von den Rohdaten.

# Konzept für neuen Daten-Workflow

Für jedes Projekt Trennung von

## Rohdaten

\\server\rohdaten\$

TestProjekt  
  BWB  
    Regen  
      METADATEN  
      regen.xls  
    Labor  
      METADATEN  
      labor.xls  
  
  KWB  
    Durchfluss  
      METADATEN  
      q01.csv  
      q02.csv  
      q03.csv

## Datenverarbeitung

\\server\datenverarbeitung\$

TestProjekt  
  01\_Bereinigung  
    METADATEN  
    regen\_roh.csv  
    regen.csv  
    qualitaet.csv  
    durchfluss.csv  
  02\_Modellierung  
    sommer  
    winter  
    VERSIONEN  
      v0.1  
      v1.0  
      sommer  
      winter  
  Software

## Ergebnisse

\\server\projekte\$

TestProjekt  
  Data-Work Packages  
    WP1\_Monitoring  
    WP2\_Modellierung  
      sommer.lnk  
      winter.lnk

# Best Practices zur Datenanalyse

# Datenanalyse mit EXCEL

## Verweis auf existierende Best Practices, z.B.

- Data Carpentry „Data organisation in spreadsheets“ ([DataCarpentry, 2018](#))
- Twenty principles for good spreadsheet practice ([ICAEW, 2015](#))

## Beispiel:

### 10. Separate and clearly identify inputs, workings and outputs

A properly structured spreadsheet will be easier to understand and to maintain. If pivot tables are used, it may be possible to relax this principle, but clarity remains crucial. Design to ensure that any input should be entered only once.

	ShipNa	ShipAd	ShipCity	ShipReg	ShipPo	ShipCo	Custom	Custom	Address
14	Ernst Hanc	Kirchgasse	Graz		8010	Austria	ERNSH	Ernst Hanc	Kirchgasse
15	Ernst Hanc	Kirchgasse	Graz		8010	Austria	ERNSH	Ernst Hanc	Kirchgasse
16	Ernst Hanc	Kirchgasse	Graz		8010	Austria	ERNSH	Ernst Hanc	Kirchgasse
17	Ernst Hanc	Kirchgasse	Graz		8010	Austria	ERNSH	Ernst Hanc	Kirchgasse
18	Split Rail	EP.O. Box	Lander	WY	82520	USA	SPLIR	Split Rail	EP.O. Box
19	Split Rail	EP.O. Box	Lander	WY	82520	USA	SPLIR	Split Rail	EP.O. Box
20	Chop-suey	Hauptstr.	Bern		3012	Switzerland	CHOPS	Chop-suey	Hauptstr.
21	Chop-suey	Hauptstr.	Bern		3012	Switzerland	CHOPS	Chop-suey	Hauptstr.
22	Chop-suey	Hauptstr.	Bern		3012	Switzerland	CHOPS	Chop-suey	Hauptstr.
23	La maison	1 rue Als	Toulouse		31000	France	LAMAI	La maison	1 rue Als
24	Queen Car	Alameda	San Francisco	CA	94107-0200	USA	QUEEN	Queen Car	Alameda

Quelle:  
[ICAEW](#)  
(2015)

**Häufiges Problem:** Mein R-Skript funktioniert nur auf meinem Rechner aber nicht bzw. „anders“ auf dem Rechner eines Kollegen?

## **Mögliche Ursachen:**

- 1) Verwendest du **die aktuelle Version des R-Skriptes**?
- 2) Enthält dein R-Skript „**hart**“ **codierte Pfade** (z.B. „C:\Users\meinName\....“) **zu Dateien auf deinem lokalem Rechner** ?  
Falls ja: ersetze diese konsequent durch **allgemeingültige Pfade** (z.B. [\\server\](#))
- 3) Habt ihr **die gleichen Versionen** installiert von:
  - **R und ggf. weiterer abhängiger Software** (Miktex, Pandoc, Stan?)
  - **R Paketen** (die von deinem Skript verwendet werden?). *Mit der R Funktion `sessionInfo()` kannst du das prüfen!*
- 4) Verwendet ihr die **gleichen Region- und Ländereinstellungen** (oftmals wichtig beim Import von CSV Dateien, da in R „Defaultwerte“ hierüber gesetzt werden)

## Vorgaben für gemeinsames Programmieren:

Nutzung des Versionsverwaltungssystems Subversion für R Code sowie **Befolgen der dazugehörigen Best Practices**, d.h.:

- **Regelmäßige „Commits“** (Einchecken eigener Änderungen) mit einer kurzen Nachricht „Warum“ Modifikation nötig war
- **Regelmäßige „Updates“** (d.h. Abholen von Codeänderungen durch Kollegen)

**R-Skripte sind so zu programmieren**, dass sie nicht nur auf dem eigenen Rechner funktionieren sondern auch auf anderen, unter der Voraussetzung dass auf diesen die benötigte Software (R/Rstudio, Miktex, Pandoc) in den gleichen Versionen installiert ist

**Tutorials für Best Practices werden in den Testprojekten erarbeitet und getestet !**

## Festlegen der beiden Testprojekte

# Anwendung der Best Practices für zwei KWB Projekte

## Unser Vorschlag:

### Ein Modellierungsprojekt: LCA „Umberto“ (Fabian)

- Projektbearbeitung durch eine Person
- Keine „Rohdaten“
- Aber: große Dateien (Modellkonfiguration: > 200 MB, exportierte CSV/EXCEL Ergebnisdateien: > 100000 Zeilen)
- Zeitaufwändige Datenverarbeitung (Aggregation der Modellergebnisse)

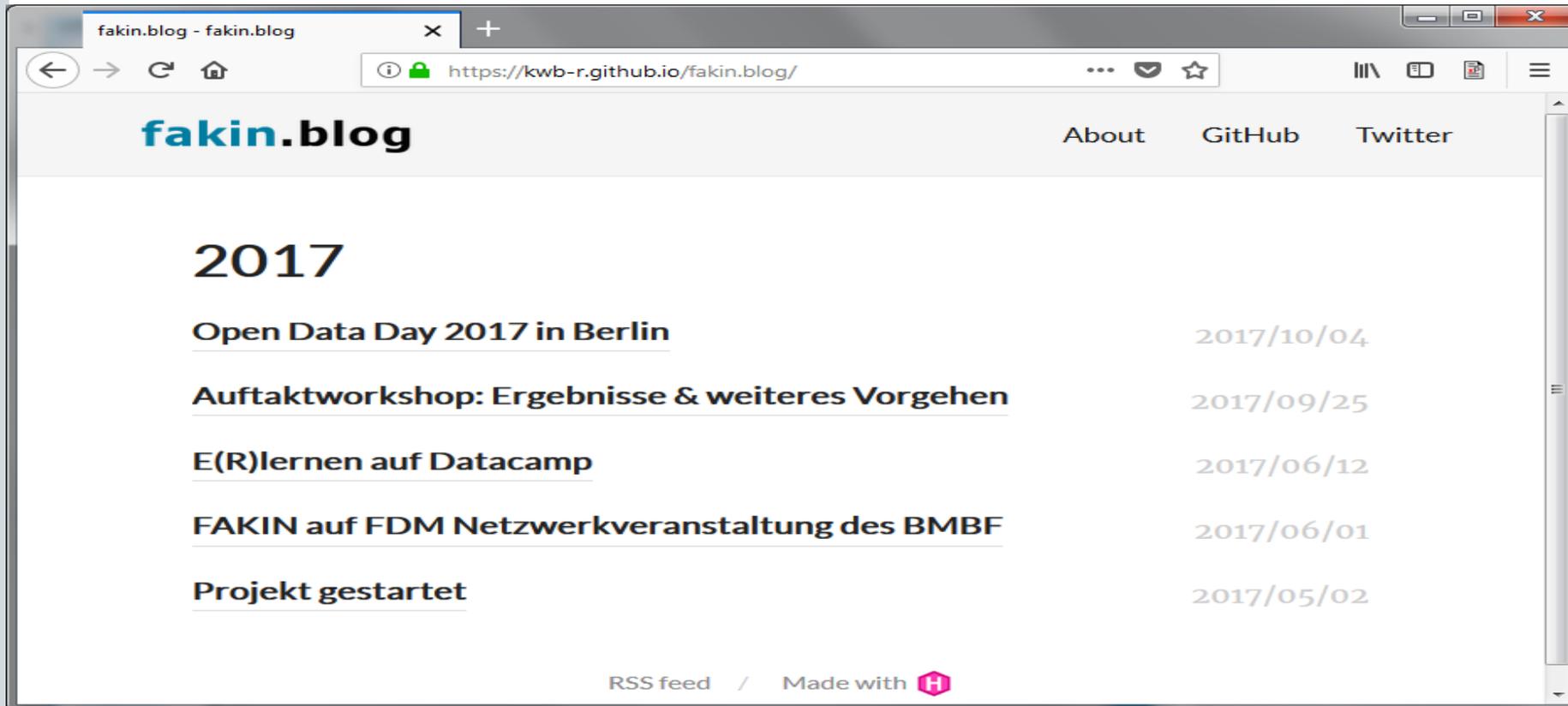
### Ein Monitoringprojekt (Aquanet oder Flusshygiene ???)

- Bearbeitung durch viele Personen
- Viele Partner
- Viele Rohdaten:
  - Aquanet: Loggerdaten für Berliner Versuchsstandorte: ~ 10 Mio. Datenpunkte pro Monat
  - Flusshygiene: 5min Regendaten, ....

# Kommunikation

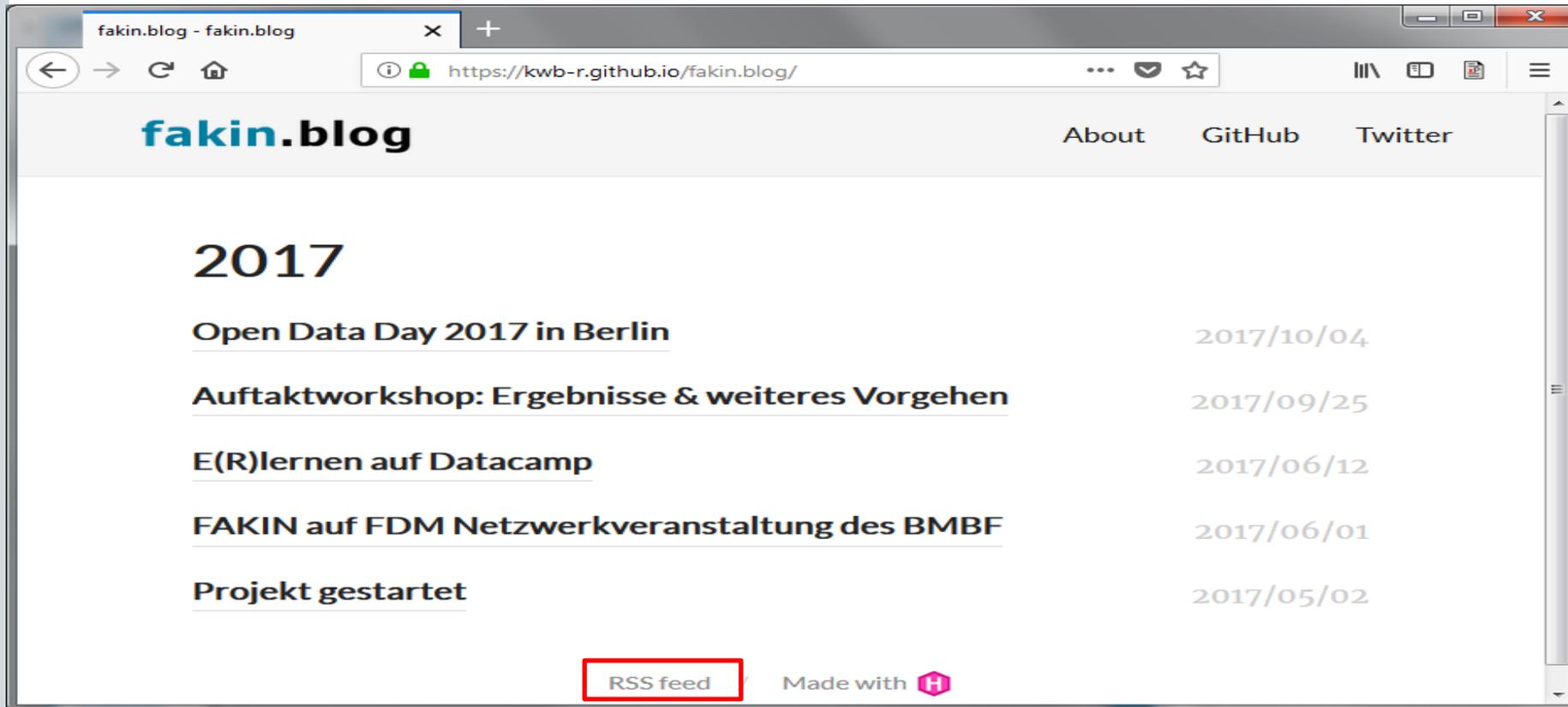
## Blog

- Themen des Forschungsdatenmanagements
- Kann von Interessierten in Outlook als RSS-Feed abonniert werden



## Blog

- Themen des Forschungsdatenmanagements
- Kann von Interessierten in Outlook als RSS-Feed abonniert werden



# Blog: RSS Feed mit Outlook abonnieren

fakin.blog - fakin.blog x fakin.blog x +

https://kwb-r.github.io/fakin.blog/index.xml

1) Diesen Feed abonnieren mit Anwendung wählen...

2) 

3)  Feeds immer mit Dynamische Lesezeichen abonnieren.

---

## fakin.blog

Recent content on fakin.blog

**[Open Data Day 2017 in Berlin](#)**  
Mittwoch, 4. Oktober 2017, 02:00

Am 4. Oktober 2017 fand von 9 bis 16 Uhr der Berlin Open Data Day in den Räumen der Colonia Nova, Thiemannstraße 1, 12059 Berlin statt: Ankündigung Programm Es folgt eine stichwortartige Zusammenfassung der Redebeiträge. Christian Rickerts, Staatssekretär Senatsverwaltung für Wirtschaft, Energie, Betriebe (SenWEB) Studien der Technologiestiftung -> wirtschaftliches Potential von Open Data Nutzen für die Verwaltung selbst der Prozess, Daten zu veröffentlichen, ist mühsam Open Data Gesetz des Bundes GAF (?)

**[Auftaktworkshop: Ergebnisse & weiteres Vorgehen](#)**  
Montag, 25. September 2017, 02:00

1 IST-/SOLL Zustand 1.1 Gesamt 1.2 Nach Abteilung 2 TOP3-Ergebnisse 1. Platz: Daten finden & bekommen 2. Platz: Daten analysieren, visualisieren & kommunizieren 3. Platz: Daten verifizieren & säubern 3 Weiteres Vorgehen In Bearbeitung....

**[E\(R\)lernen auf Datacamp](#)**  
Montag, 12. Juni 2017, 02:00

# Blog: RSS Feed mit Outlook abonnieren

fakin.blog - Michael.Rustler@kompetenz-wasser.de - Microsoft Outlook

Datei Start Senden/Empfangen Ordner Ansicht Add-Ins G DATA

Neue E-Mail-Nachricht Elemente Neu Löschen Antworten Allen antworten Weiterleiten Verschieben in: ? GRW Abteilungs... Team-E-Mail Verschieben Regeln OneNote Inhalt herunterladen Diesen Feed freigeben Artikel anzeigen Kontakt suchen Adressbuch E-Mail filtern Suchen

**Favoriten**

- Posteingang
- Ungelesene E-Mail
- Gesendete Elemente
- Michael.Rustler@kompetenz-wasser.de
- Gelöschte Elemente

**Michael.Rustler@kompetenz-wasser.de**

- Posteingang
- Entwürfe [34]
- Gesendete Elemente
- Gelöschte Elemente
- Junk-E-Mail
- Postausgang
- RSS-Feeds**
  - fakin.blog**
  - Workshop on fakin.blog
- Suchordner

fakin.blog durchsuchen (Strg+E)

Anordnen nach: Datum Neu nach alt

Älter	Datum	Icon
fakin.blog Open Data Day 2017 in Berlin	04/10/2017	🔍
fakin.blog Auftaktworkshop: Ergebnisse & weiteres Vorgehen	25/09/2017	🔍
fakin.blog E(R)lernen auf Datacamp	12/06/2017	🔍
fakin.blog FAKIN auf FDM Netzwerkveranstaltung des BMBF	01/06/2017	🔍
fakin.blog Projekt gestartet	02/05/2017	🔍
fakin.blog About	01/05/2017	🔍

## Open Data Day 2017 in Berlin

fakin.blog

Bereitgestellt am: Wed 04/10/2017 02:00  
Feed: fakin.blog

Am 4. Oktober 2017 fand von 9 bis 16 Uhr der Berlin Open Data Day in den Räumen der Colonia Nova, Thiemannstraße 1, 12059 Berlin statt: Ankündigung Programm

Es folgt eine stichwortartige Zusammenfassung der Redebeiträge. Christian Rickerts, Staatssekretär Senatsverwaltung für Wirtschaft, Energie, Betriebe (SenWEB) Studien der Technologiestiftung -> wirtschaftliches Potential von Open Data Nutzen für die Verwaltung selbst der Prozess, Daten zu veröffentlichen, ist mühsam Open Data Gesetz des Bundes GAF (?)

[Artikel anzeigen...](#)

## Best Practices Bericht:

- KWB-intern auf Server als HTML, PDF, DOCX verfügbar
- Fertigstellung der ersten Version: Mitte Februar 2018

FAKIN D1: Best-practices & Werkzeuge

file:///Y:/GROUNDWATER/PROJECTS/FAKIN/Reports/D1\_BestPractices\_Werkzeuge/intro.html

FAKIN D1 Bericht

1 Einführung

- 1.1 Hintergrund
- 1.2 Ziele

2 Kommunikation 1

- 2.1 Brownbag (30 min Vortrag + ...)
- 2.2 Fragebogen
- 2.3 Auftaktworkshop

3 Best-practices

- 3.1 Aufnahme und Analyse des Is...
- 3.2 Erarbeitung von best-practice...
- 3.3 Konventionen für Ordner- un...
- 3.4 Regeln für die Verbesserung ...
- 3.5 Vorschlag für verbesserte Or...

## FAKIN D1: Best-practices & Werkzeuge

*Hauke Sonnenberg & Michael Rustler (Kompetenzzentrum Wasser Berlin gGmbH)*

*January 24, 2018*

### Kapitel 1 Einführung

#### 1.1 Hintergrund

# Zusammenfassung