Best Practices Workshop

(**F**orschungsdatenmanagement **a**n **k**leinen **In**stituten)

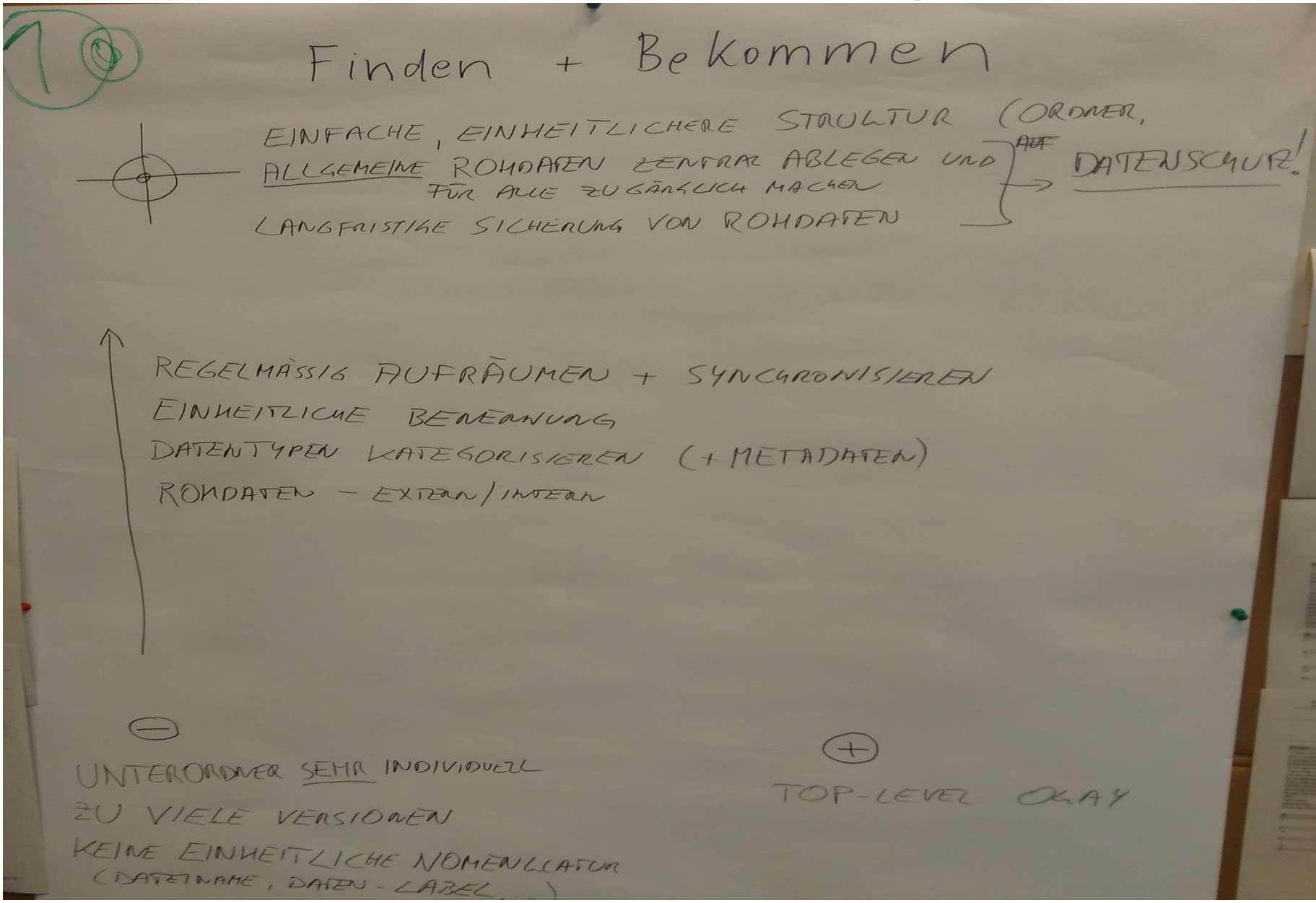
Gefördert durch: Bundesministerium für Bildung und Forschung

KOMPETENZZENTRUM Wasser Berlin

Agenda

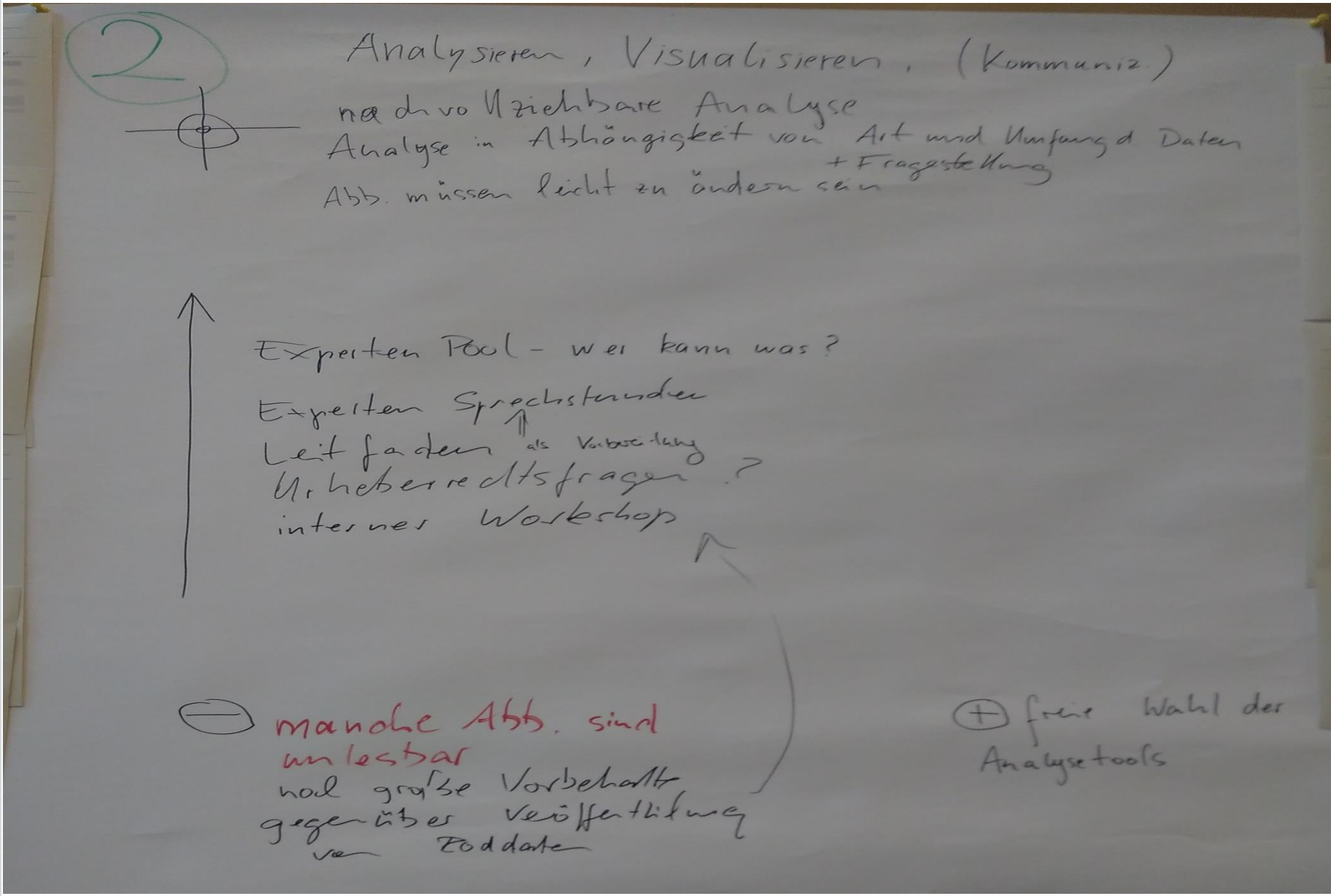
13:00	Rückblick Auftaktworkshop (Michael)
13:10	Zusammenfassung der Projektleiterinterviews (Hauke)
13:30	Best Practices für „Daten finden und bekommen“ (Hauke)
13:40	Konzept für neuen Daten-Workflow (Hauke)
13:50	Diskussion / Feedback
14:05	Pause
14:10	Best Practices für „Datenanalyse“ (Michael)
14:20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

Rückblick: Auftaktworkshop



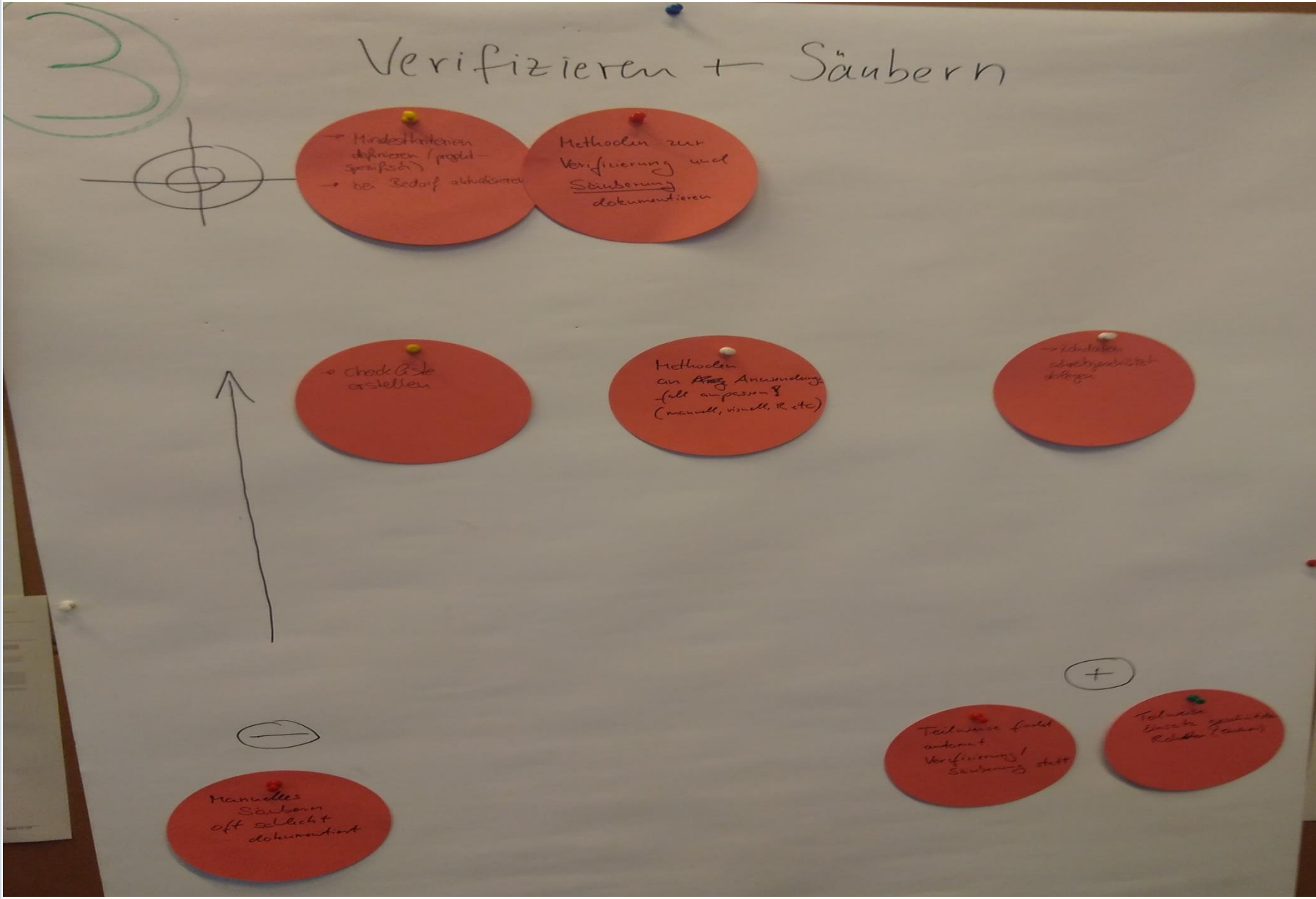
KOMPETENZZENTRUM Wasser Berlin

Rückblick: Auftaktworkshop



KOMPETENZZENTRUM Wasser Berlin

Rückblick: Auftaktworkshop



KOMPETENZZENTRUM Wasser Berlin

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

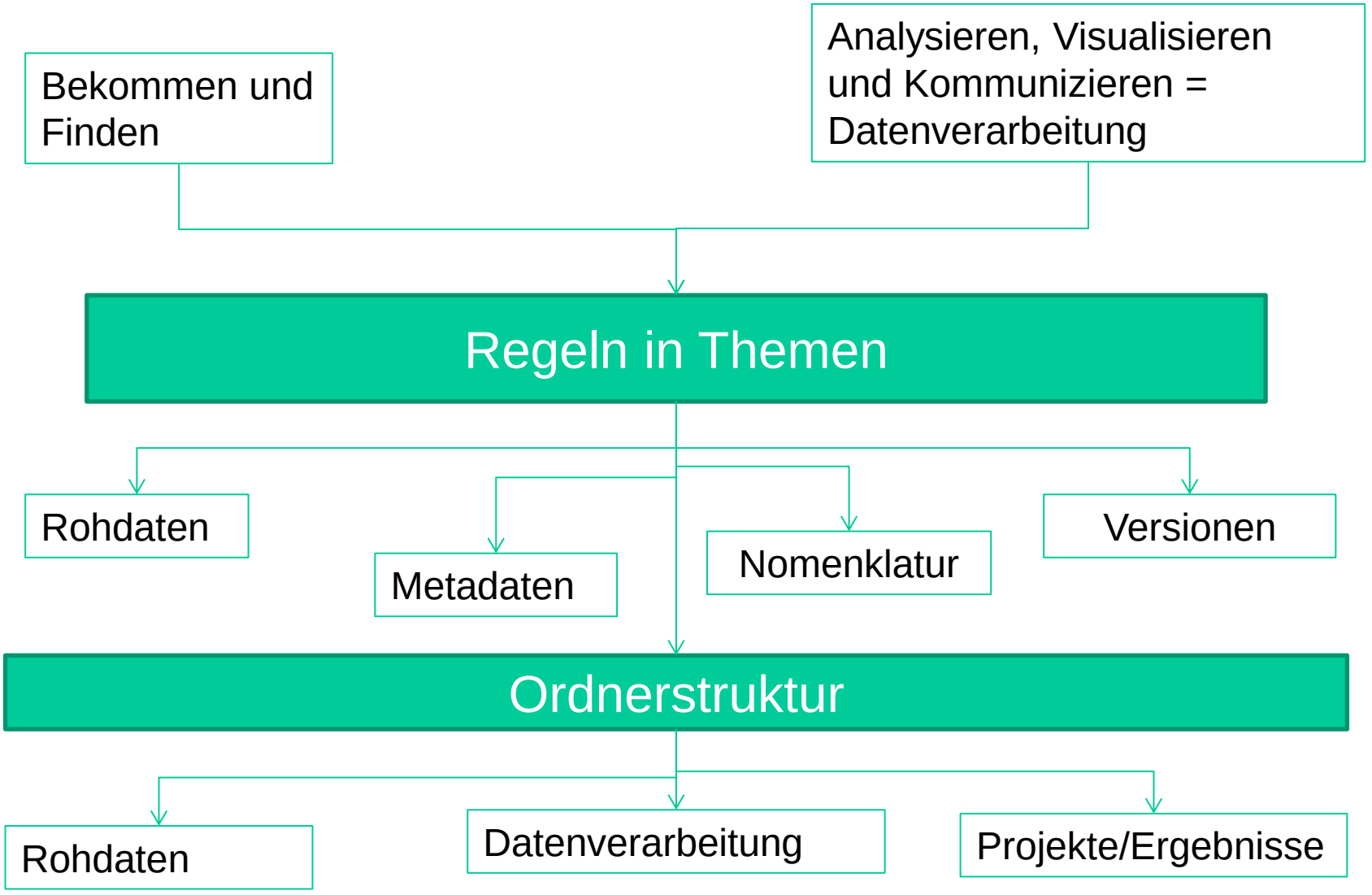
Source: elifesciences.org (2018)

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

Rohdaten

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)

Metadaten

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

Nomenklatur

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

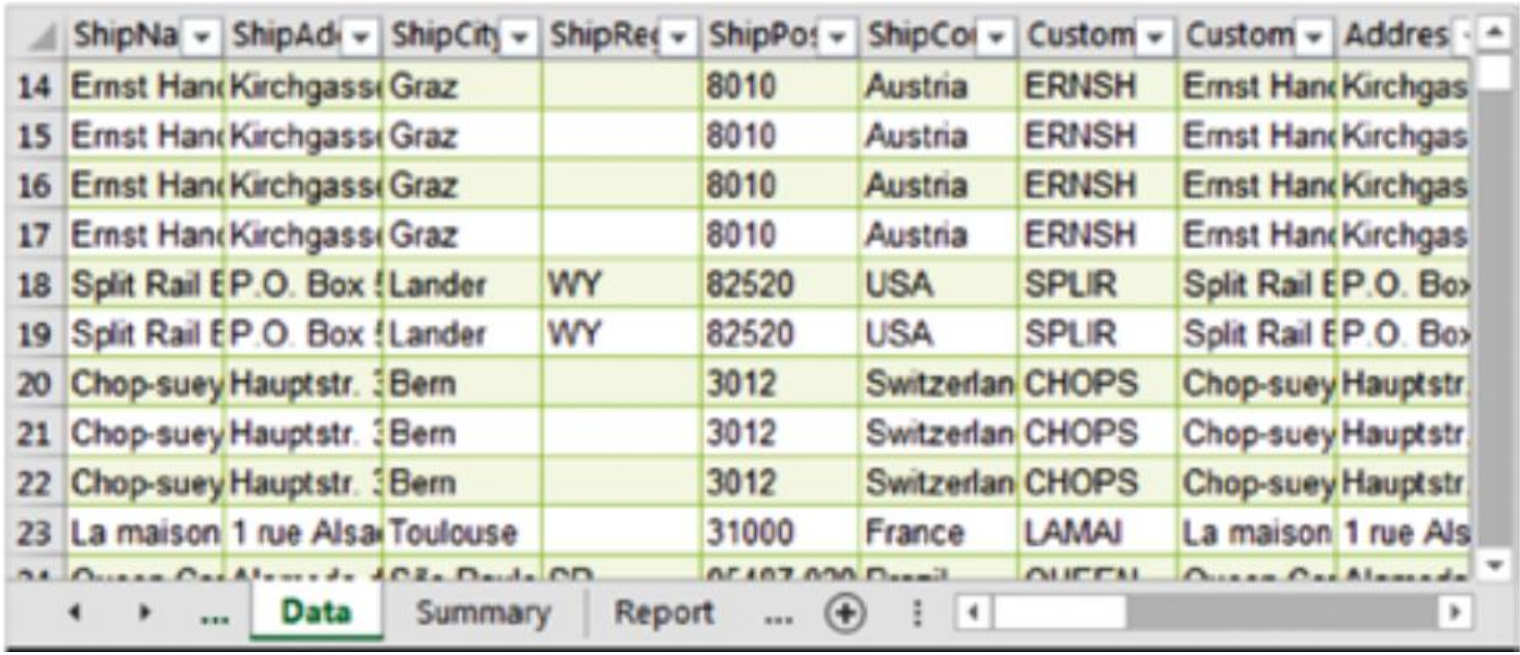
Verweis auf existierende Best Practices, z.B.

- Data Carpentry „Data organisation in spreadsheets“ (DataCarpentry, 2018)
- Twenty principles for good speadsheet 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	ShipRe	ShipPo	ShipCo	Custom	Custom	Addres
14	Ernst Han	Kirchgass	Graz		8010	Austria	ERNSH	Ernst Han	Kirchgas
15	Ernst Han	Kirchgass	Graz		8010	Austria	ERNSH	Ernst Han	Kirchgas
16	Ernst Han	Kirchgass	Graz		8010	Austria	ERNSH	Ernst Han	Kirchgas
17	Ernst Han	Kirchgass	Graz		8010	Austria	ERNSH	Ernst Han	Kirchgas
18	Split Rail E	P.O. Box !	Lander	WY	82520	USA	SPLIR	Split Rail E	P.O. Box
19	Split Rail E	P.O. Box !	Lander	WY	82520	USA	SPLIR	Split Rail E	P.O. Box
20	Chop-suey	Hauptstr.	Bern		3012	Switzerlan	CHOPS	Chop-suey	Hauptstr
21	Chop-suey	Hauptstr.	Bern		3012	Switzerlan	CHOPS	Chop-suey	Hauptstr
22	Chop-suey	Hauptstr.	Bern		3012	Switzerlan	CHOPS	Chop-suey	Hauptstr
23	La maison	1 rue Alsa	Toulouse		31000	France	LAMAI	La maison	1 rue Als

... Data Summary Report ...

Quelle: ICAEW (2015)

Datenanalyse mit R

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) Einhält 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)

Datenanalyse mit R

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 (Aggregierung der Modellergebnisse)

Ein Monitoringprojekt (Aquanes oder Flusshygiene ???)

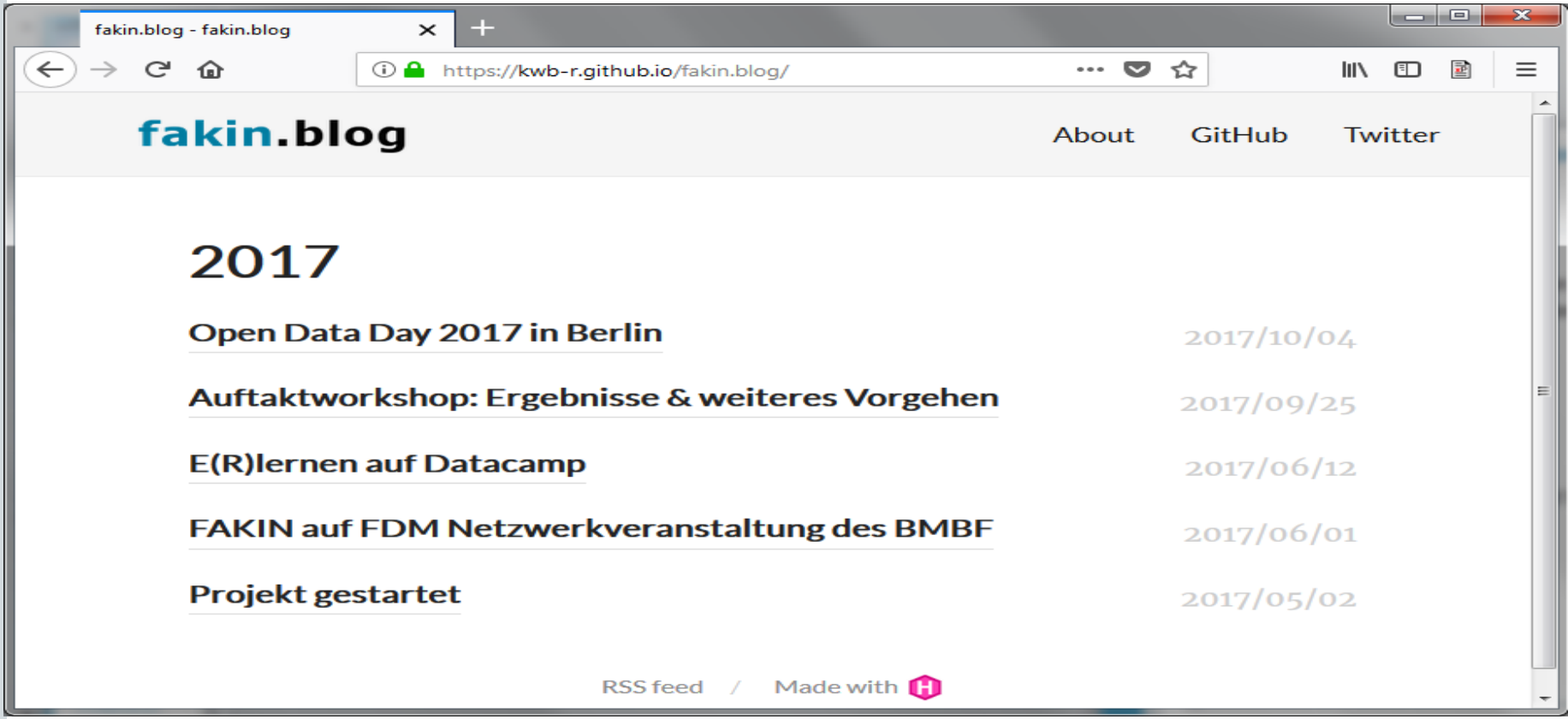
- Bearbeitung durch viele Personen
- Viele Partner
- Viele Rohdaten:
  - Aquanes: Loggerdaten für Berliner Versuchsstandorte: ~ 10 Mio. Datenpunkte pro Monat
  - Flusshygiene: 5min Regendaten, ….

Kommunikation

Kommunikation: FAKIN

Blog

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

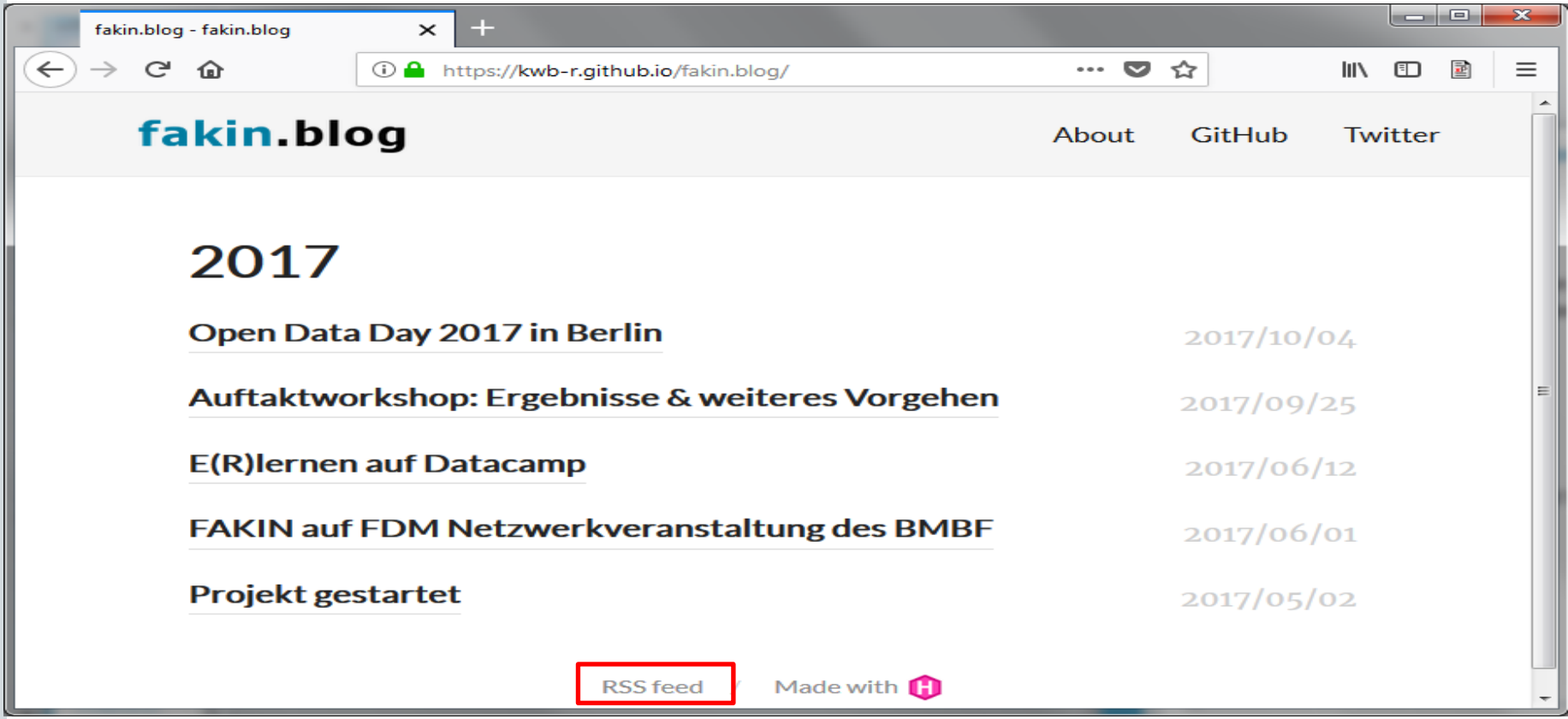


KOMPETENZZENTRUM Wasser Berlin

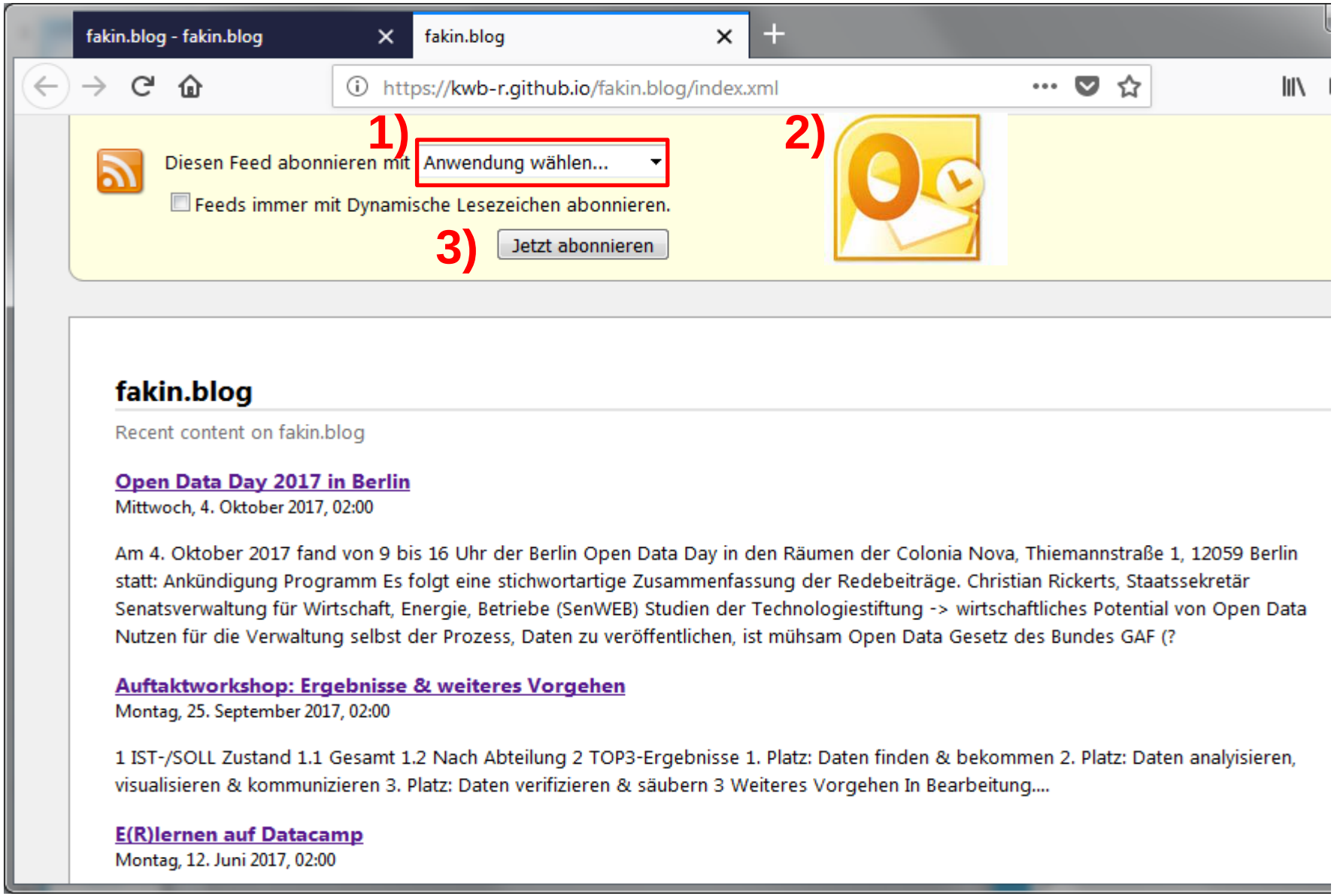
Kommunikation: FAKIN

Blog

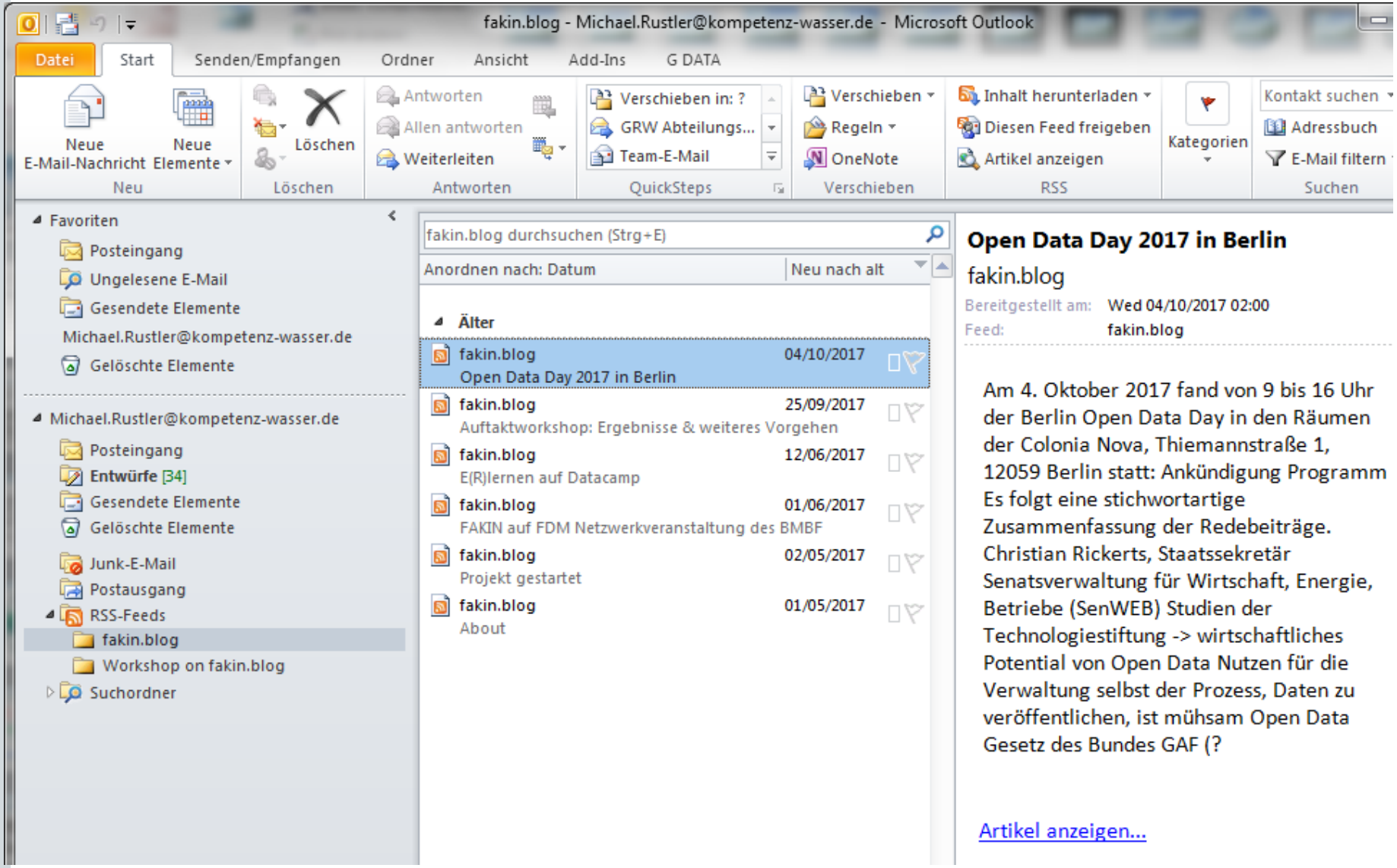
- Themen des Forschungsdatenmanagements
- Kann von Interessierten in Outlook als RSS-Feed abonniert werden



Blog: RSS Feed mit Outlook abonnieren



Blog: RSS Feed mit Outlook abonnieren

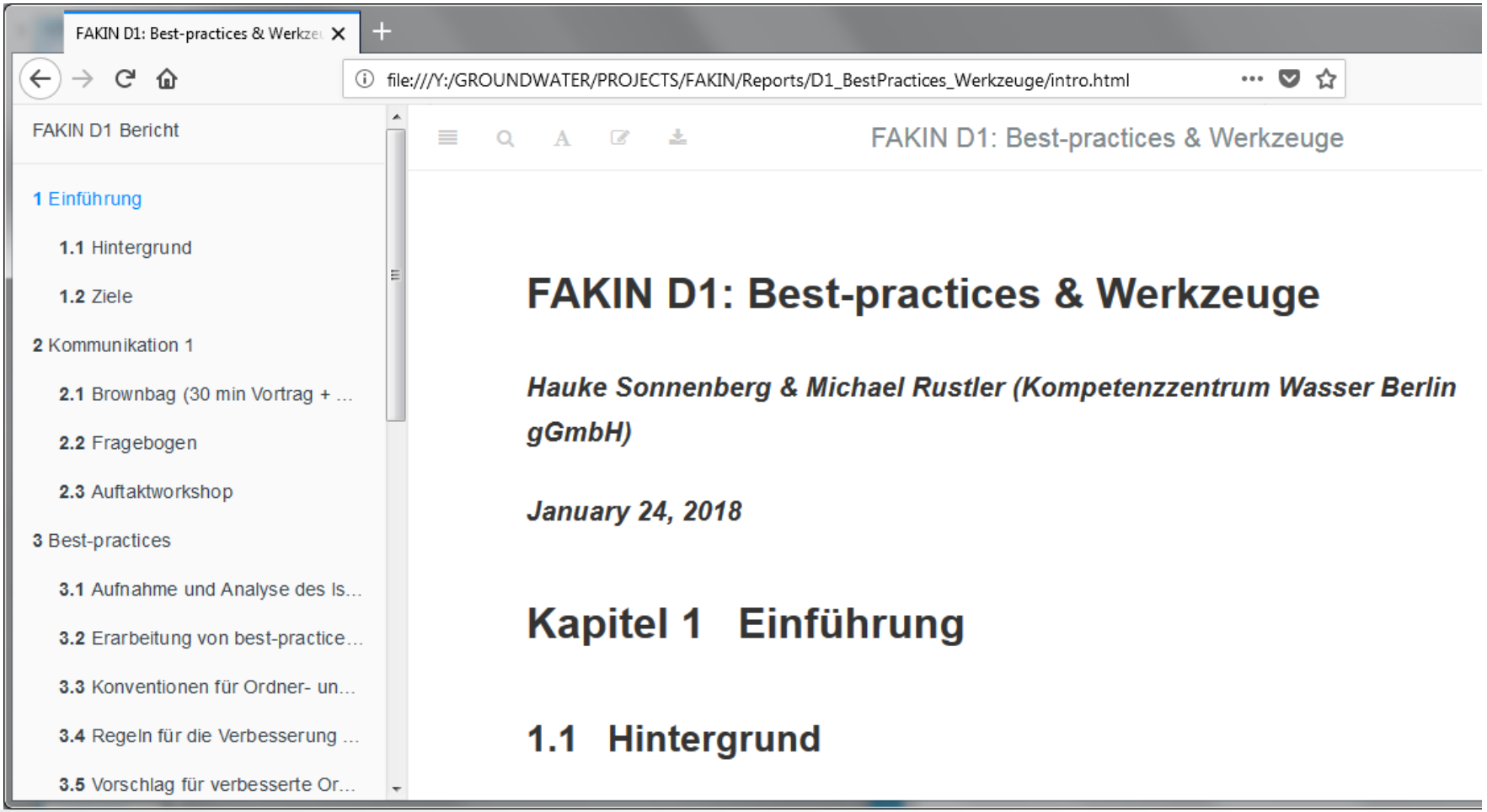


KOMPETENZZENTRUM Wasser Berlin

Kommunikation: FAKIN

Best Practices Bericht:

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



KOMPETENZZENTRUM Wasser Berlin

Zusammenfassung