

Best Practices in Research Data Management

Hauke Sonnenberg & Michael Rustler

2018-06-28 14:43:01

Contents

Preface	5
1 Introduction	7
1.1 Why RDM ?	7
1.2 For whom ?	7
2 Best Practices	9
2.1 Plan and fund a new project	9
2.2 Data storage	10
2.3 Data collection	20
2.4 Data publishing and sharing	20
3 Case studies	25
3.1 Geogenic salination	25
3.2 LCA modelling	25
3.3 Pilot plants	26
4 Other Projects	29
4.1 Spree2011 (2007)	29
4.2 MIACSO (2009)	29
4.3 KURAS	30
4.4 OGRE	31
4.5 Flusshygiene	31
4.6 DEMOWARE	31
4.7 OPTIWELLS	31
4.8 RWE	31
5 FAQs	33
5.1 Tools for researchers	33
5.2 Writing more robust R code	33
5.3 Learning R on DataCamp	34
5.4 Using Subversion at KWB	34
5.5 How to install KWB R packages?	34
5.6 How to build your own KWB styled R package?	34
5.7 Working with EXCEL	35
5.8 Heterogenous software versions on KWB computers	36
5.9 R package/version dependency of R scripts	36
5.10 Complex R script dependencies	37
5.11 Heterogenous (R-)coding styles	37
5.12 Collaborative version control	38
5.13 Workflow automation	38
5.14 Encoding	38

5.15 Collaborative writing	39
5.16 Exchanging data with colleges/project partners	39
6 Glossary	41
6.1 Reproducibility	41
6.2 Provenance	41
6.3 Techniques	41
6.4 Tools	42

Preface

“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



This document is the outcome of the KWB project FAKIN (Forschungsdatenmanagement an kleinen Instituten = research data management at small institutes).

It defines best practices for research data management. It is mainly based on the personal experiences of

the authors having worked in many different research projects at KWB.

The document is outlined as follows:

- Chapter 1 explains why and for whom research data management is important
- Chapter 2 defines Best Practices for different topics.
- Chapter 3 gives an overview on three case studies with different data related tasks that have been solved within the FAKIN project, while
- Chapter 4 what data challenges occurred in the past in other KWB projects and how they were solved in order to increase our awareness at KWB,
- Chapter 5 provides FAQs and finally
- Chapter 6 is a glossary for explaining commonly used terms within this document.

This document is assumed to be a “living” document. We highly appreciate any comments and suggestions for improvements. What are your experiences with research data management tasks? Can you provide solutions for specific tasks?

The online version of this report is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.



Chapter 1

Introduction

A very good book on how research data management (RDM) improves reproducibility is provided by Kitze et al. (2018) which also is available online at <http://practicereproducibleresearch.org/>.

1.1 Why RDM ?

Doing research data management (RDM) is useful according to Labfolder.com (2018) as it:

- Saves time for research
- Avoids risk of data loss
- Ensures transparency and reproducibility
- Increases data visibility and number of citations
- Fulfill funders' requirements and receive more grants
- Produce new knowledge and make more discoveries just by re-using data
- Archive, retrieve and re-use own data

Recommended literature:

- Introduction to research data management in the geosciences (in German!)
- Handbook research data management (in German!) (Buettner et al., 2011)

1.2 For whom ?

Info on the document – create links between projects and thematic groups

Chapter 2

Best Practices

2.1 Plan and fund a new project

2.1.1 Data management plan

Action 1: Look through a data management planning checklist DMP checklist?

Action 2: Create a data management plan. You can use DMP Online as it has a template or if required, use a funder's format.

RDM checklist before starting with a new project Brauchen wir eine Data Management Plan? See e.g. <https://dmponline.dcc.ac.uk/>

2.1.2 Electronic Lab Notebooks

Electronic Lab Notebooks (ELNs) <https://www.labfolder.com/research-data-management/> ELNs are an essential asset for researchers to fulfil any requirements for data management, and they create direct bridges between scientists and stakeholders. By adopting an ELN, the data lifecycle can proceed smoothly and easily: from creating and collecting data digitally in one place to one-click data archiving, ELNs empower researchers by allowing them to implement their RDM plan without effort and time investment.

An ELN to make your data FAIR? Welches ist das?

2.1.3 Define acronyms

At the start of a research project

- Choose a project acronym and store it in **PROJECTS.txt**.
- Check if the organisations that you expect to get data from are listed in **ORGANISATIONS.txt** and extend this file if necessary.
- Create a subfolder for your project and subfolders for the organisations in the rawdata folder structure.

At the start of a project or if an employee or trainee enters the project

- Give an introduction to our research data management as described in this document.

Regularly during the project

- Check if the folder structure within your project's rawdata subfolder still complies with the rawdata folder structure and clean the structure, if not.

2.2 Data storage

The Ten Simple Rules for Digital Data Storage by Hart et al. (2016) should serve us as a guideline:

Rule 1: Anticipate How Your Data Will Be Used

Rule 2: Know Your Use Case

Rule 3: Keep Raw Data Raw

Rule 4: Store Data in Open Formats

Rule 5: Data Should Be Structured for Analysis

Rule 6: Data Should Be Uniquely Identifiable

Rule 7: Link Relevant Metadata

Rule 8: Adopt the Proper Privacy Protocols

Rule 9: Have a Systematic Backup Scheme

Rule 10: The Location and Method of Data Storage Depend on How Much Data You Have

— (Hart et al., 2016)

2.2.1 Naming

2.2.1.1 Folder and file names

Useful file names are:

- consistent
- meaningful to you and your colleagues
- allow you to find the file easily

It is useful if your department/project agrees on the following elements of a file name:

- Vocabulary – choose a standard vocabulary for file names, so that everyone uses a common language
- Punctuation – decide on conventions for if and when to use punctuation symbols, capitals, hyphens and spaces
- Dates – agree on a logical use of dates so that they display chronologically i.e. YYYY-MM-DD
- Order - confirm which element should go first, so that files on the same theme are listed together and can therefore be found easily
- Numbers – specify the amount of digits that will be used in numbering so that files are listed numerically e.g. 01, 002, etc.

!!!! Good practice: Remove spaces from file names or use punctuation such as underscores and hyphens to separate words e.g. “AHRC_TechnicalApp_Response20120925.docx” or “AHRC-TechnicalApp-Response20120925.docx” rather than “what we got back from funders about the data stuff.docx !!!!

Keine Sonderzeichen, keine Umlaute, keine Leerzeichen

Umlaute ä, ö, ü werden durch **ae**, **oe**, **ue** bzw. **Ae**, **Oe**, **Ue** und der Buchstabe ß durch **ss** ersetzt.

- Vermeidung von Fehlern bei der automatischen Verarbeitung
- Keine Probleme beim internationalen Austausch von Dateien

Keine Leerzeichen, stattdessen **_** Unterstrich oder **-** Bindestrich

- Vermeidung von Fehlern bei der automatischen Verarbeitung
- doppelte Leerzeichen lassen sich besser erkennen

Datum in der Form yyyy-mm-dd (2017-06-30)

- Alphanumerische Sortierung führt automatisch zu Sortierung nach Datum
- Bei Zerlegung des Dateinamens bei Unterstrich bleibt das Datum als Einheit erhalten

Zusammengesetzte Wörter (z.B. Projektnamen) in “CamelCase”

Zusammengesetzte Wörter (z.B. Projektnamen) ohne Unterstrich in CamelCase (`ReliableSewer` vs. `Reliable_Sewer`)

- Bei Zerlegung des Dateinamens bei Unterstrich bleibt Projektname erhalten

Einheitliche Schreibweise von Projektnamen

Einheitliche Sprache: englisch oder deutsch?

Verwendung eines Vokabulars für wichtige Begriffe

z.B. “validiert”, “kalibriert”

- Order - confirm which element should go first, so that files on the same theme are listed together and can therefore be found easily
- Numbers – specify the amount of digits that will be used in numbering so that files are listed numerically e.g. 01, 002, etc.

!!!! Good practice: Remove spaces from file names or use punctuation such as underscores and hyphens to separate words e.g. “AHRC_TechnicalApp_Response20120925.docx” or “AHRC-TechnicalApp-Response20120925.docx” rather than “what we got back from funders about the data stuff.docx !!!!

Keine Sonderzeichen, keine Umlaute, keine Leerzeichen

Umlaute ä, ö, ü werden durch `ae`, `oe`, `ue` bzw. `Ae`, `Oe`, `Ue` und der Buchstabe ß durch `ss` ersetzt.

- Vermeidung von Fehlern bei der automatischen Verarbeitung
- Keine Probleme beim internationalen Austausch von Dateien

Keine Leerzeichen, stattdessen `_` Unterstrich oder `-` Bindestrich

- Vermeidung von Fehlern bei der automatischen Verarbeitung
- doppelte Leerzeichen lassen sich besser erkennen

Datum in der Form yyyy-mm-dd (2017-06-30)

- Alphanumerische Sortierung führt automatisch zu Sortierung nach Datum
- Bei Zerlegung des Dateinamens bei Unterstrich bleibt das Datum als Einheit erhalten

Zusammengesetzte Wörter (z.B. Projektnamen) in “CamelCase”

Zusammengesetzte Wörter (z.B. Projektnamen) ohne Unterstrich in CamelCase (`ReliableSewer` vs. `Reliable_Sewer`)

- Bei Zerlegung des Dateinamens bei Unterstrich bleibt Projektname erhalten

Einheitliche Schreibweise von Projektnamen

Einheitliche Sprache: englisch oder deutsch?

Verwendung eines Vokabulars für wichtige Begriffe

z.B. “validiert”, “kalibriert”



Regeln zur Vermeidung langer Dateipfade

Lange Dateipfade können im Windows-Betriebssystem ein Problem darstellen. Das äußert sich beim Kopieren von Dateien, wenn der Pfad des Zielorts, der sich durch die Kopie ergeben würde, eine bestimmte Länge überschreitet. Die zugehörige Fehlermeldung lautet dann: *Die Dateinamen wären zu lang für den Zielordner. Kürzen Sie die Dateinamen und wiederholen Sie den Vorgang, oder verwenden Sie einen anderen Ort, der einen kürzeren Pfad hat.

2.2.1.2 Acronyms

Acronyms are unique, clear names for objects. They should

- be short but meaningful and easy to remember,
- be all lowercase,
- consist of only alphanumeric letters (a-z, 0-9) or the hyphen (-).

2.2.1.2.1 Projects

- Haben wir eine Regel für die Benennung, Schreibweise von Projektnamen?
- Wir sollten einheitliche, offizielle Schreibweisen definieren
- Sollen wir Acronyme (z.B. drei Zeichen) definieren, um diese z.B. in Dateinamen oder in Übersichten zu verwenden?

At the start of a project we define a project acronym. This acronym is intended to be used in file and folder names.

Whenever we want to indicate the relation to a certain project in a file or folder name, we use the project acronym in exactly the typing that was defined. This is important as we want to distinguish between raw data, processed data and project results in our data workflow.

The project acronyms are defined in a simple text file `PROJECTS.txt` in the `//server/projects$` folder, see Project Folder Structure.

2.2.1.2.2 Organisations

It is very important to know the owners of data. Therefore we define unique acronyms for the owners of data that we use. The acronyms are defined in a special file `ORGANISATIONS.txt`

2.2.2 Metadata

What information would someone need to find/re-use your data? Location, title, creator name, description, date collected

Why? Metadata are required for interpreting raw data and gaining an overview about the available data.



Important metadata for raw data

von wem, wann, über wen und welches Medium erhalten, z.B. “E-Mail von A an B am 25.01.2018” oder “USB-Stick von C an D persönlich am 26.01.2018”

Nutzungseinschränkung, z.B. “nur für Projekt x”, “nur innerhalb des KWB”, “darf auf keinen Fall veröffentlicht werden”

Erläuterungen zum Inhalt und zum Format (Bedeutung von Spalten, Einheiten, Messgeräte, Methoden...)



Important metadata for processed data

von wem (Person), wann, mit welchen Methoden (z.B. Skript) unter welchen Randbedingungen (z.B. Versionen von Software, Paketen) erstellt?

Welche Rohdaten wurden verwendet?

keep metadata in plain text file "readme.txt"

Tools for metadata tracking and data standards are:

- metadata editor, e.g. online editor of GFZ Potsdam

Best Practices:

- 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
- Metadaten zu Rohdaten sind immer abzulegen. Metadaten zu verarbeiteten Daten sind ebenso wichtig. Sie können bei automatisierter Datenverarbeitung aber aus den Skripten geschlossen werden.

Metadatenstandards:

Wir wollen einen Metadatenstandard verwenden. Beispiele für Metadatenstandards sind (in Klammern stehen Institutionen, die unter anderem Daten mit Metadaten in diesem Standard veröffentlichen):

- DataCite Metadata Schema
 - Leibniz-Zentrum für Agrarlandschaftsforschung e. V. (ZALF)
 - Deutsches GeoForschungsZentrum Potsdam (GFZ) (+ ISO + DIF + Dublin Core)
- ISO 19115-2
 - Bundesanstalt für Gewässerkunde (BfG)
 - Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung (AWI)

Die Auswahl erfolgt gemäß einer guten Übertragbarkeit auf die Anforderungen des KWB und gemäß der Verfügbarkeit von Tools.

Tools:

- Metadaten-Editor, z.B. Online-Editor vom GFZ Potsdam

We propose to define some special files that contain metadata related to files and folders. To indicate that these files have a special meaning, the file names are all uppercase.

copied from 10_old_german_chapters.Rmd :

Beispiel:

Wenn wir Zeitreihen von den Berliner Wasserbetrieben bekommen, dann haben diese entweder eine Spalte mit Zeitstempeln oder zwei Spalten mit Zeitstempeln, wobei die erste den Anfang und die zweite das Ende des Zeitraums enthält, auf den sich die Messwerte in der entsprechenden Zeile beziehen.

Wenn nur eine Spalte Zeitstempel enthält, dann ist nicht klar, ob der Zeitstempel den Anfang oder das Ende (oder gar einen anderen Zeitpunkt innerhalb) des Zeitraums angibt.

Bei einer Zeitreihe mit fünfminütigen Messwerten, z.B.

Zeit	Wert
2018-06-05 16:20:00	0
2018-06-05 16:25:00	1
2018-06-05 16:30:00	2
2018-06-05 16:35:00	3
2018-06-05 16:40:00	4

ist nicht klar, ob sich der Wert 1 auf den Zeitraum 16:20 - 16:25 oder auf den Zeitraum 16:25 - 16:30 bezieht.

Dies ist ein Beispiel für eine sehr konkrete Metadatenanforderung. An dieser Stelle gehen wir meistens von Annahmen aus oder hat jemand von Euch schon mal eine Aussage darüber von einem BWB-Mitarbeiter erhalten?

Die Kenntnis über den Bezugspunkt des Zeitstempels ist zum Beispiel wichtig beim

- Erzeugen von Regen-Inputdateien für die Software InfoWorks

In der obigen Beispielzeitreihe ist auch nicht klar, auf welche Zeitzone sich die Zeitstempel beziehen. Die Zeitzone bzw. die Angabe des UTC-Offsets ist eine wichtige Metainformation. Am besten wäre es, wenn alle Zeitstempel diese Information enthielten, z.B.

```
2018-06-05 16:20:00+02
2018-06-05 16:25:00+02
```

Dann bräuchte man diese Information nicht extra zu führen.

Metadaten werden oft erst bei Bedarf erhoben, nämlich genau dann, wenn man sich z.B. die obigen Fragen nach Zeitstempelbezug und Zeitzone stellt. Dann wird nachgefragt und im besten Fall wird eine Antwort darauf gegeben, z.B. in Form einer E-Mail. Die Metadaten sind dann in der E-Mail enthalten.

Fragen:

- Wo und unter welchem Namen legen wir diese E-Mail ab?

Beispiel für Metadaten aus meiner persönlichen Logdatei:

```
Regendaten gelesen; ca. 10 Impulse gegen 14:00; Dannecker geschrubbt, neuer Akku: Batterie defect, alte Batterie drin gelassen, Daten auslesen nicht möglich; H=55cm
```

Das müssen wir irgendwie besser hinkriegen, aber wie?

Die grundlegenden Fragen sind:

- Wer ist für die Ablage von Metadaten verantwortlich?
- Wo werden die Metadaten abgelegt? Direkt bei den Daten oder in einer übergeordneten Datei/Datenbank?
- In welchem Format werden die Metadaten abgelegt? Einfaches Textformat vs. XML-Format, das sich wohl einfach nur mit Hilfe eines Programms editieren lässt.

2.2.2.1 File PROJECTS.txt

This file contains the project acronyms as we want to use them e.g. as top-level folder names in our project folder structure. The projects are grouped by department.

Possible content of PROJECTS.txt:

```
# Department SUW (Surface Water)
dswt: DSWT
flusshygiene: Flusshygiene
```

```

kuras: KURAS
mia-cso: MIACSO
monitor: MONITOR
ogre: OGRE
reliable-sewer: RELIABLE_SEWER
sema: SEMA
sema-berlin: SEMA Berlin
sema-berlin-2: SEMA Berlin 2
spree-2011: SPREE2011
spree-2011-2: SPREE2011 "reloaded"

# Department GRW (Groundwater)
optiwells: OPTIWELLS
optiwells-2: OPTIWELLS 2
wellma: WELLMA

# Department WWT (Wastewater Treatment)
...
```

In the file `PROJECTS.txt` the project acronyms appear in alphabetical order. They map the acronym to a project name or a project title and the year of the start of the project.

Question: Do we already have a place where “official” metadata about projects are stored? If yes, the acronym could be included there. But then, everybody should know about it!

2.2.2.2 File `ORGANISATIONS.txt`

Possible content of `ORGANISATIONS.txt`

```

bwb: Berliner Wasserbetriebe
kwb: Kompetenzzentrum Wasser Berlin
uba: Umweltbundesamt
```

2.2.3 Raw data

Especially in environmental sciences, raw data often cannot be reproduced (e.g. rainfall, river discharge measurements) and are therefore of high value. Thus the following rules apply in case of raw data:

- can be renamed (in case this is documented in a metadata file)
- the content must not be changed
- will be stored as read-only in an secure area (i.e. one folder contains all raw data)

added from: *04_rawdata.Rmd*

As raw data we define data that we receive from a device or from a project partner.

Most of our research results are based on data. We acknowledge the importance of raw data by

- storing then in a special place where it is specially secured
- describing them with metadata

Rawdata are stored in the rawdata folder structure

2.2.4 Data workflow

For each project separation of:

- raw data (i.e. `\\server\\rawdata`)
- data processing (i.e. `\\server\\processing`)
- results (i.e. `\\server\\projekte$`)

Raw data are highly valuable because they if necessary cannot or only with high costs be reproduced. Thus raw data will be compiled on the highest level (i.e. `\\server\\rawdata`) for each project and origin (e.g. KWB, BWB) as shown below:

Raw data folder structure:

```
TestProjekt
  BWB
    Regen
      METADATEN
      regen.xls
    Labor
      METADATEN
      labor.xls

  KWB
    Durchfluss
      METADATEN
      q01.csv
      q02.csv
      q03.csv
```

The data processing is a “playground”, where different approaches or scenarios can be tested and where possibly different versions for a specific approach or scenario are available. The files will be stored here according to the corresponding project and topic and the data processing step, respectively.

Data processing folder structure:

```
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
```

In the results folder only the “reporting relevant” results should be saved. Relevant results files (e.g. plots) should be added according to project and project structure (i.e. per working package) in form of links from the data processing folder structure (see above).

Results folder structure:

```

TestProjekt
  Data-Work Packages
    WP1_Monitoring
    WP2_Modellierung
      sommer.lnk
      winter.lnk

```

2.2.5 Folder Structures**Restrictions/Conventions:**

- Each top-level folder should represent a project, i.e. should be defined in the top level file `PROJECTS.txt`.
- Each possible owner should be defined in the top level file `ORGANISATIONS.txt`.
- The naming convention for the organisations is the same as for projects.

s)

2.2.5.1 Project Folder Structure

We said that we want to concentrate on the folder structures within the project folders. Nevertheless, we would like to give a recommendation on how the project folders could be organised within its top level folder. In this structure, there are no subfolders for the different departments any more.

```

//server/projects$
- PROJECTS.txt
- project-1/
- project-2/
- project-3/

```

In the `projects$` folder

- each subfolder name should appear in the file `PROJECTS.txt`
- there should not be any folder on the top level that does not represent a project.
- there should be no other files on the top level as the files that are described in this documentation.
- there are no subfolders representing departments any more. The mapping of projects to departments is done in the file `PROJECTS.txt`

2.2.5.2 Rawdata Folder Structure

We will create a network folder `//server/rawdata$` in which all files have set the read-only property. We suggest to store raw data by project first and by the organisation that owns (i.e. generated, provided) the data second. This could look like this:

```

//server/rawdata$
- ORGANISATIONS.txt
- PROJECTS.lnk [Symbolic Link to PROJECTS.txt in //server/projects$]
- flusshygiene
  - bwB
  - kwB
  - uba
  - ...
- ogre
  - kwB

```

- bw**b**
- ub**a**
- ...
- ...

2.2.6 Versioning

Versioning or version control is the way by which different versions and drafts of a document (or file or record or dataset or software code) are managed. Versioning involves the process of naming and distinguishing between a series of draft documents that leads to a final or approved version in the end. Versioning allows you to disclose an audit trail for the revision and update of drafts and final versions.

2.2.6.1 Manual

We propose the following workflow:

- It is only allowed to modify the current file (e.g. `filename.pptx`), which contains no version name as postfix
- A version is created by copying the current file. This copy gets a not already defined version name as file name ending (e.g. `filename_v0.1.pptx`)
- The versioned file is moved to a folder ‘VERSIONS’ and set as read-only.
- In the folder `VERSIONS` is a file `VERSIONS.txt`, which contains additional information on the available version within that folder

Drawbacks:

- possibly more time demanding,
- needs time getting used to it and
- requires high level of discipline

Advantages:

- Simple and safe method, which requires not version control software
- Cleaner workspace as old versions are stored in a separate folder
- Sorting by name equals the chronology

Example:

```
BestPractices_Workshop.ppt
VERSIONS/
+ VERSIONS.txt
+ BestPractices_Workshop_v0.1.ppt
+ BestPractices_Workshop_v0.2.ppt
+ BestPractices_Workshop_v1.0.ppt
```

Content of file `VERSIONS.txt`:

```
BestPractices_Workshop.ppt
- v0.1: first draft version, written by Michael Rustler
- v0.2: after additions by Hauke Sonnenberg
- v1.0: first final version, after review by Pascale Rouault
```

2.2.6.2 Automatically

The versioning is done automatically in case a version control software, like Git or Subversion are used.

At KWB we currently use the following version control software:

- Subversion: for internally storing programm code (e.g. R-scripts/packages) we have an Subversion server, which is accessible from the KWB intranet. However, this requires:
 - the installation of the client software TortoiseSVN and a
 - valid user account (for accessing the server) which is currently provided by the IT department on request
- Git: for publishing programm code external on our KWB organisation group on Github. Currently all repositories are public (i.e. are visible for everyone), but also use of private repositories is possible for free as KWB is recognised as non-for-profit company by Github, offering additional benefits for free

Use of version control software is required in case of programming (e.g. in R, Python, and so on) and can be useful in case of tracking changes in small text files (e.g. configuration files that run a specific R script with different parameters for scenario analysis).

Drawbacks:

- Special software (TortoiseSVN), login data for each user on KWB-Server and some basic training are required
- In case of collaborate coding: sticking to ‘best-practices’ for using version control is mandatory, e.g.:
 - timely check in of code changes to the central server,
 - Speaking to each other: so that not two people work at the same time at the same program code in one script as this leads to conflicts that need to be resolved manually, which can be quite time demanding. You are much better off if you avoid this in the upfront by talking to each other

Advantages:

- Only one filename per script (file history and code changes are managed either internally on a KWB server in case of using TortoiseSVN or externally for code hosted on Github)
- Old versions of scripts can be restored easily
- Additional comments during `commit` (i.e. at the time of transferring the code from the local computer to the central version control system about **why** code changes were made and build-in diff-tools for tracking changes improve the reproducibility



Attention: version control software is not designed for versioning of raw data and thus should not be used for it. General thoughts on the topic of ‘data versioning’ are available here: <https://github.com/leeper/data-versioning>



A presentation with different tools for version control is available here: <https://www.fosteropenscience.eu/node/597>

2.2.7 Organising e-mails

“Most people now routinely send and receive lots of messages every day and as a Yesult, their inbox can get very quickly overloaded with hundreds of personal and work-related email. Setting aside some time to organise your emails will ensure information can be found quickly and easily,

and is stored securely. Why should I organise my email? Apart from the obvious frustration and time wasted looking for that email you remember sending to someone last month, email is increasingly used to store important documents and data, often with information related to the attachments within the email itself.”

— University of Cambridge (2018)

How can I ensure my emails remain organised?

Here are some general tips to ensure your email remains organised in the long term University of Cambridge, 2018:

- Delete emails you do not need. Remove any trivial or old messages from your inbox and sent items on a regular (ideally daily) basis.
- Use folders to store messages. Establish a structured file directory by subject, activity or project.
- Separate personal emails. Set up a separate folder for these. Ideally, you should not receive any personal emails to your work email account.
- Limit the use of attachments. Use alternative and more secure methods to exchange data where possible (see ‘data sharing’ for options). If attachments are used, exercise version control and save important attachments to other places, such as a network drive.

2.2.8 Managing references

For managing reference there are plenty of tools available. A detailed overview is provided by (Fenner et al., 2014) in the Chapter Reference Management At KWB we use Endnote, for which an internal guideline document (KWB-EndNote-Guideline-v02.pdf,) was developed.

2.2.9 Data preservation

Data must be retained to support your research findings

Standards: 75 years?

2.3 Data collection

Data are often inconsistent, incomplete, incorrect, or misspelled Data cleaning is essential You may also use OpenRefine <http://openrefine.org/> to clean your messy data Or use the following tools

2.3.1 Logger devices

2.3.2 Spreadsheets

2.4 Data publishing and sharing

“This figure provides a framework for understanding how the “Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing” (Boland et al., 2017) can be translated into easily understood modern life concepts.

Rule 1 is Open-Source Software. The openness is signified by a window to a room filled with algorithms that are represented by gears.

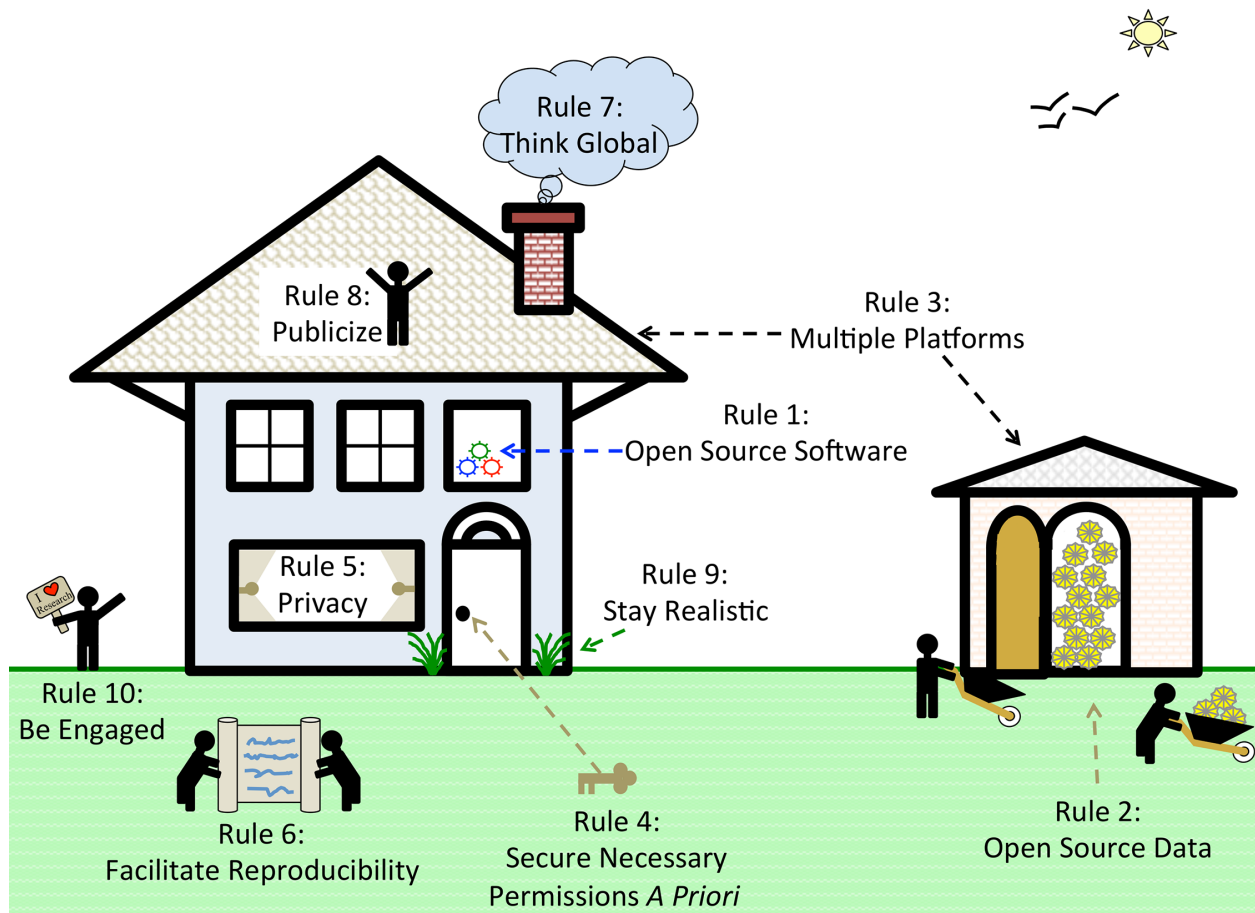


Figure 2.1: Modern life context for the ten simple rules (Boland et al., 2017)

Rule 2 involves making the source data available whenever possible. Source data can be very useful for researchers. However, data are often housed in institutions and are not publicly accessible. These files are often stored externally; therefore, we depict this as a shed or storehouse of data, which, if possible, should be provided to research collaborators.

Rule 3 is to “use multiple platforms to share research products.” This increases the chances that other researchers will find and be able to utilize your research product—this is represented by multiple locations (i.e., shed and house).

Rule 4 involves the need to secure all necessary permissions a priori. Many datasets have data use agreements that restrict usage. These restrictions can sometimes prevent researchers from performing certain types of analyses or publishing in certain journals (e.g., journals that require all data to be openly accessible); therefore, we represent this rule as a key that can lock or unlock the door of your research.

Rule 5 discusses the privacy issues that surround source data. Researchers need to understand what they can and cannot do (i.e., the privacy rules) with their data. Privacy often requires allowing certain users to have access to sections of data while restricting access to other sections of data. Researchers need to understand what can and cannot be revealed about their data (i.e., when to open and close the curtains).

Rule 6 is to facilitate reproducibility whenever possible. Since communication is the forte of reproducibility, we depicted it as two researchers sharing a giant scroll, because data documentation is required and is often substantial.

Rule 7 is to “think global.” We conceptualize this as a cloud. This cloud allows the research property (i.e., the house and shed) to be accessed across large distances.

Rule 8 is to publicize your work. Think of it as “shouting from the rooftops.” Publicizing is critical for enabling other researchers to access your research product.

Rule 9 is to “stay realistic.” It is important for researchers to “stay grounded” and resist the urge to overstate the claims made by their research.

Rule 10 is to be engaged, and this is depicted as a person waving an “I heart research” sign. It is vitally important to stay engaged and enthusiastic about one’s research. This enables you to draw others to care about your research.”

— (Boland et al., 2017)

Recommended literature:

- Ten simple rules to enable multi-site collaborations through data sharing (Boland et al., 2017)
- Guidelines for publishing (PhD) research data (Kaden and Kleineberg, 2018)

2.4.1 Repositories

Repositories for permanently depositing data are for example:

- General
 - Figshare,
 - Zenodo (a joint project between OpenAIRE and CERN),
 - Mendeley data,
 - Dataverse,
 - Dryad
- Focus on environmental and earth sciences

- Pangea
- GFZ Potsdam data services

In addition repositories for publishing program code are for example Github or Gitlab, however these do not offer build-in long term preservation by default.

We are currently using the following repositories and can recommend them for similar usage to others:

- Github: for developing/publishing program code, because it offers the opportunity to automatically get a
- Zenodo DOI for each software released publicly on Github (for details see: <https://guides.github.com/activities/citable-code/>)



A blog post

Bosman and Kramer (2016) provide results of a large survey carried out in 2015 among more than 15000 researchers. Insights can be gained on:

- Which scholarly communications tools are used and
- Are there disciplinary differences in usage?

They finally summarise: “Another surprising finding is the overall low use of Zenodo – a CERN-hosted repository that is the recommended archiving and sharing solution for data from EU-projects and -institutions. The fact that Zenodo is a data-sharing platform that is available to anyone (thus not just for EU project data) might not be widely known yet.”

2.4.2 ORCID

Problem:

”Two large challenges that researchers face today are discovery and evaluation. We are overwhelmed by the volume of new research works, and traditional discovery tools are no longer sufficient. We are spending considerable amounts of time optimizing the impact—and discoverability—of our research work so as to support grant applications and promotions, and the traditional measures for this are not enough. — (Fenner and Haak, 2014)

Solution:

”Open Researcher & Contributor ID (ORCID) is an international, interdisciplinary, open and not-for-profit organization created to solve the researcher name ambiguity problem for the benefit of all stakeholders. ORCID was built with the goal of becoming the universally accepted unique identifier for researchers:

1. ORCID is a community-driven organization
2. ORCID is not limited by discipline, institution, or geography
3. ORCID is an inclusive and transparently governed not-for profit organization
4. ORCID data and source code are available under recognized open licenses
5. the ORCID id is part of institutional, publisher, and funding agency infrastructures.

Furthermore, ORCID recognizes that existing researcher and identifier schemes serve specific communities, and is working to link with, rather than replace, existing infrastructures.”

— (Fenner and Haak, 2014)

2.4.3 Licenses

“In most countries in the world, creative work is protected by copyright laws. International conventions, and primarily the Berne Convention of 1886, protect the copyright of creators even across international borders for 50 years after the death of the creator. This means that copying and using the creative work is limited by conditions set by the creator, or another copyright holder. For example, in many cases musical recordings may not be copied and further distributed without the permission of the musician, or of the production company that has acquired the copyright from the musician. Facts about the universe that are discovered through research are not subject to copyright, but the collection, aggregation, analysis and interpretation of research data may be considered creative work, and could be protected by copyright laws. Thus, the consumption of research publications is governed by copyright law. Furthermore, even data sharing is often governed by copyright laws, because the compilation of data to be shared often requires a creative effort. Another case of research-relevant copyrighted products is software that is developed in the course of research. In all of these cases, if license terms are not explicitly specified, the work is considered to be protected as “all rights reserved”. This means that no one but the creator of the work can use the work unencumbered. For software this means that copying and further distribution of the software is prohibited. Even running the software may be restricted. The exact selection of a license is beyond the scope of this section, but depends on your intentions and goals with regard to the software”

— `citep(manual["Rokem_2018"])`

Recommended literature:

- Intellectual Property and Computational Science (Stodden, 2014)
- Creative Commons Licences (Friesike, 2014)
- <https://choosealicense.com/>

2.4.4 File formats

“Scientific data is saved in a myriad of file formats. A typical file format might include a file header, describing the layout of the data on disk, metadata associated with the data, and the data itself, often stored in binary format. In some cases (e.g., CSV (or comma-separated value) files), data will be stored as text. The danger of proliferation of file formats in scientific data lies in the need to build and maintain separate software tools to read, write and process all these data formats. This makes interoperability between different practitioners more difficult, and limits the value of data sharing, because access to the data in the files remains limited.”

— `citep(manual["Rokem_2018"])`

2.4.5 Data exchange standards

WaterML2:

“...is a new data exchange standard in Hydrology which can basically be used to exchange many kinds of hydro-meteorological observations and measurements. WaterML2 has been initiated and designed over a period of several years by a group of major national and international organizations from public and private sector, such as CSIRO, CUAHSI, USGS, BOM, NOAA, KISTERS and others. WaterML2 has been developed within the OGC Hydrology Domain Working group which has a mandate by the WMO, too.”

— WaterML2

ODM2: is an information model and supporting software ecosystem for feature-based earth observations

Chapter 3

Case studies

The three following case-studies are tested in detail within FAKIN (i.e. proposed best-practices will be applied for this case studies and cross-checked whether their application is useful).

3.1 Geogenic salination

Adapt and test with new folder drive workflow as proposed

- rawdata
- processing
- processing

3.2 LCA modelling

Challenge:

The LCA modelling software Umberto can produce large raw data output files (> 300 MB csv files) that sometimes are even to big for EXCEL 2010 (> 1 millions) but need to be aggregated (e.g. grouped by specific criteria). This was usually performed manually within EXCEL in case that model output data was below EXCEL's 1 million row limit.

Workflow improvement developed within FAKIN:

An open source R package `kwb.umberto` was programmed for automating:

- data import the Umberto model results,
- performing data aggregating to the user needs and finally
- exporting the aggregated results in an **results.xlsx** EXCEL spreadsheet.

This **results.xlsx** EXCEL spreadsheet is referenced by another EXCEL spreadsheet **figures.xlsx** (which contains the figure templates and just links to the **results.xlsx** in order to update the predefined figures).

This workflow now reduces the time consuming and error-prone formerly manually performed data aggregation in EXCEL, whilst still enabling the users to adapt the figures to their needs without coding knowledge.

3.3 Pilot plants

Challenges:

The output of (on-line) monitoring technologies is often difficult to interpret and also inconvenient to handle as the output formats of different devices (in one water treatment scheme) can vary strongly. Furthermore, frequent reporting and documentation of the treatment performance via (on-line) monitoring can be time consuming for the personnel and requires advanced software solutions. An alternative to commercial (and often expensive) software solutions are tools which are based on the open software R (R Core Team, 2017). The free software approach allows any R programmer to produce customized tools for each individual end-user.

Thus an automated reporting tool is developed within the AQUANES project for enabling an integrative assessment of the different monitoring devices and integration with water quality data obtained from analysis in the laboratory for four different pilot plant sites in order to:

- Increase the reliability and reproducibility of handling large amounts of data by reducing the likeliness in human error in complex systems and by increasing the transparency of the data processing.
- Promote the use of customized R tools for different end-user such as utilities, consultants and other research teams.

Therefore the open source R package `aquan.es.report` (Rustler, 2018) was programmed, which is able to:

- import operational and lab data for each pilot site,
- performs temporal aggregation (e.g. 5 min, 1 h, 24h median values),
- visualises raw or aggregated data either interactively in a web browser or by
- creating a standardised report (e.g. monthly) in html, pdf or docx

For the four different pilot plant sites the data (operational and lab data) for being imported into the R tool came from various sources at different temporal resolutions, which are detailed below:

- Haridwar: operational data stored by Autarcon in mySQL database (temporal resolution: ~ 2 -3 minutes, i.e. ~ 0.7 million data points per month), which is accessible from the web and thus could be easily imported into R. Lab data was provided by Autarcon initially in a unstructured format, which was impossible to be automatically imported into R. However, after agreeing on a standardised EXCEL spreadsheet format (e.g. one spreadsheet per site, one sheet per parameter and additional sheets providing metadata for parameters and sites) it was possible to integrate the lab data into the R tool.
- Basel Lange-Erlen: operational data is provided by the water supplier in EXCEL spreadsheets on a weekly basis for each site (i.e. “Rein” and “Wiese”) with a temporal resolution of 5 minutes (i.e. ~ 0.5 million data points per month). Lab data are provided by the water supplier in a single comma separated csv file, which is exported from a database. Thus the structure of the lab data was standardised and could be easily imported into the R tool.
- Berlin-Schönerlinde: operational data from the WEDECO pilot plant are collected using a SCADA system (\sim temporal resolution: seconds, i.e. ~ 10 million data points per month). Lab data are provided by BWB in a single EXCEL spreadsheet. However, its structure often changes in case it is updated by BWB, making an automated importing using R impossible without adapting the import functions. Thus lab data were not integrated in the R tool for this site.
- Berlin-Tiefwerder: operational data from the PENTAIR pilot plant are collected using a SCADA system (\sim temporal resolution: \sim seconds, i.e. ~ 10 million data points per month). Lab data are provided by BWB in a single EXCEL spreadsheet. However, its structure often changes in case it is updated by BWB, making an automated importing using R impossible without adapting the import functions. Thus lab data were not integrated in the R tool for this site.



The high temporal high resolution (\sim seconds) of the operational data for both Berlin pilot plants resulted in large data amounts (\sim 10 million data points per month), which required an large effort to optimise the performance of the R reporting tool in oorder to enable the visualisation of the pilot plant's raw data for its test operation period (\sim 18 months) on computers with limited RAM ressources (\sim 8-12 GB).

The R tool is used by KWB (for the sites Berlin-Schönerlinde and Berlin-Tiefwerder) regularly for assessing the pilot plants' operational performance interactively. In addition for an advanced assessment only the data importing and aggregation routines and combined with R scripts developed by KWB students.

For the other two pilot plant sites Haridwar and Basel Lange-Erlen the AQUANES project partners use the automated R reporting tool in a similar way.

Chapter 4

Other Projects

Here we are just suming up the data workflows and created tools in (old) KWB projects, but – in contrast to the case studies – the workflows and tools were not tested explicitly within the FAKIN project.

The goal of this chapter is to enlarge the internal knowledge base at KWB about what and how data challenges have already been successfully solved in the past.

4.1 Spree2011 (2007)

Used data by source (data formats in parentheses)

- KWB:
 - water level and discharge at one monitoring site (Text/CSV)
 - rain (Text/CSV)
- BWB:
 - pumping rates in the pumping stations (Excel)
 - water levels in the pumping stations (Excel)
 - rain at some gauges near the monitoring site (Excel)

Tasks and methods by topic

- Dry-weather and wet-weather calibration of a sewer network model (Infoworks)
 - InfoWorks: Creating rain input files
 - InfoWorks: Creating RTC input files

Questions that arose:

- Where to store presentations (trainee vs. employee)?
- Where to store the raw data (personal drive of the trainee)?
- How does Infoworks interpret timestamps, how do BWB provide timestamps? -> metadata

4.2 MIACSO (2009)

Monitoring

- sites: one site in the sewer (monitoring container), more sites in the river
- variables: water quantity and quality
- devices: online sensors

Modelling

- Sewerage: Infoworks
- River hydraulics: Hydrax
- River quality: QSim

Data storage

- High amount of data -> extra server: moby
- We put some effort in planning good folder structures for the data. Nevertheless the structure at the end of the project is not as clean as it was planned.
- Data that we received from project partners was stored in `Daten/EXTERN`.
- Raw data was stored in a folder `Daten/RAW` which was write-protected and required a special user-account for storing new data.

`Daten/`

```
ACCESS/ # MS Access databases, containing raw data
EXTERN/ # External data (by organisation)
META/   # MS Access databases, containing metadata
        # (about calibration, maintenance, sites, variables)
RAW/    # Text files containing raw data, from KWB-own devices only, by site
```

Metadata

- many devices in the container -> meta data about device cleaning and maintenance important -> tool: `META_Maintenance.mdb`

Methods and Tools

- We imported most of the data from text files into MS Access databases in -> tool: `MiaCsoRawImport.mdb`
- We calibrated the sensors offline by using SQL queries to provide calibrated data from raw data -> tool: `MiaCsoMetaCalibControl.mdb`
- We used SQL queries to perform data processing -> tool: `MiaCsoStatAnalysis.mdb`
- Data validation (outlier detection) was done in a two step procedure:
 1. Automatic preselection using MS Access tool `MiaCsoStatAnalysis.mdb`
 2. Manual selection using self-developed graphical tool in Origin

Developed Tools:

- MS Access Applications
 - `MetaMaint.mdb`: Monitoring Metadata Management
 - `MiaCsoRawImport.mdb`: Text File Import to MS Access
 - `MiaCsoStatAnalysis.mdb` (project deliverable): Definition and automatic execution of sequences of SQL queries
- Origin extension to interactively select and store outliers graphically
- R packages
 - `kwb.mia.evalCrit02` (project deliverable): graphical evaluation of critical oxygen conditions in the river
 - `kwb.mia.iw`: Calculation of file sizes of InfoWorks result csv-files exported from InfoWorks.
 - `kwb.miacso`: functions used in MIA-CSO, for example for plotting data availabilities.

4.3 KURAS

Developed Tools:

- Frontend for KURAS Database of Rainwater Management Measures: `KURAS_DB_Acc2003_hs.mdb`
- R package `kwb.kuras`: Interface to KURAS database

4.4 OGRE

- Decision to use CUAHSI Community Observations Data Model (ODM)
- R script to import lab data from Excel to MS Access database implementing ODM

Developed Tools:

- R packages
 - `kwb.ogre`
 - `kwb.ogre.model`
 - `kwb.odm`
 - `kwb.odmx`

4.5 Flusshygiene

- Adaptation of free online monitoring data visualisation HydroServerLite
- Reusage of lab data import script developed in OGRE

4.6 DEMOWARE

Developed R packages:

- Groundwater modelling
 - `kwb.hantush` (Rustler, 2016a)
 - `kwb.vs2dh`
 - `kwb.demoware`
- Quantitative microbiological risk assessment: the R package `kwb.qmra` (Rustler, 2016b) is a generic calculation engine assessing the performance of water supply systems. It was successfully applied for the Old Ford wastewater treatment plant (Kraus et al., 2016; Rustler, 2016c). A detailed documentation for the R package and its usage is available online.

4.7 OPTIWELLS

Created R packages:

- `kwb.wtaq` (Sonnenberg and Rustler, 2016): groundwater modelling, e.g. for assessing the impact of production well (e.g. well diameter, pumping rates) and aquifer characteristics (e.g. hydraulic conductivity) on the resulting drawdown. A detailed tutorial is available online,
- `kwb.epanet`: wrapper for (pressurised) pipe network simulation model EPANET

4.8 RWE

(Semi)automated creation of a complex groundwater simulation model with MODFLOW-2005 for the project Maxflow. The developed model consisted of up to three model layers with up to 1000 abstraction wells with temporal-spatial varying pumping rates within the simulation period. For automating the model generation the Python package `flopy` v3.2.5 (Bakker et al., 2016b,a) was used with some minor modifications by KWB

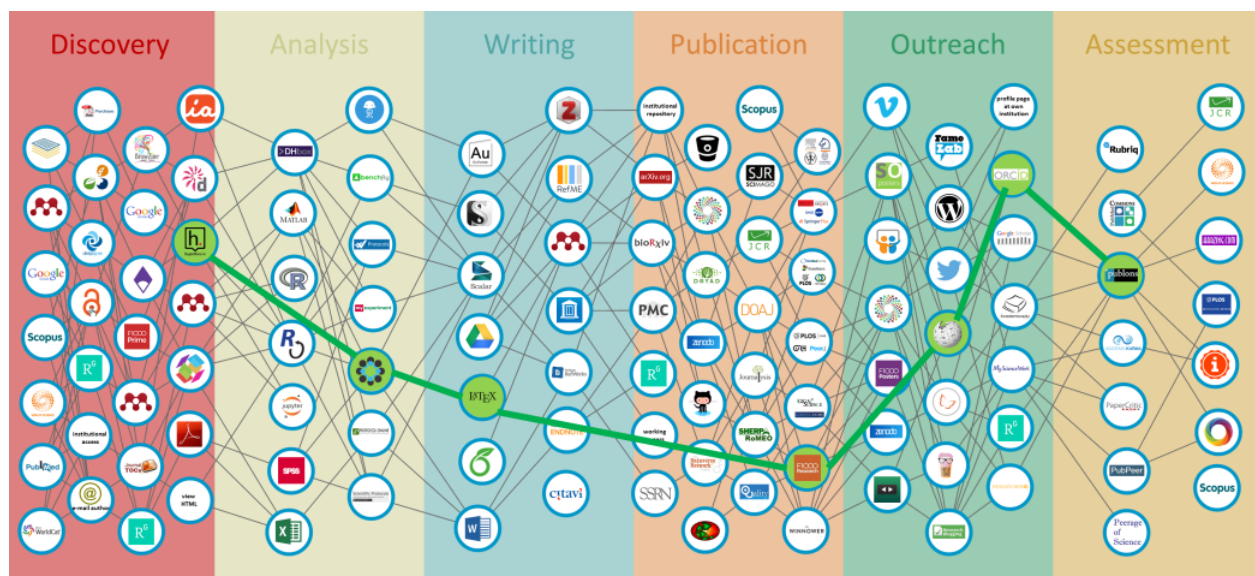
(see [here](#)). Github was used as version control software for tracking changes in the code and each model scenario was stored in its Github branch (for details, see [here](#)).

Chapter 5

FAQs

5.1 Tools for researchers

A very exhaustive overview of tools used in the researcher's workflow are provided by Bosman and Kramer (2018), which are grouped according to different research phases as shown below.



5.2 Writing more robust R code

For details on making R code more robust to work on different computers please read the following tutorial.

Recommended literature:

- Good enough practices in scientific computing (Wilson et al., 2017)
- Ten simple rules for making research software more robust (Taschuk and Wilson, 2017)
- R programming books (freely available online!)
 - R for data science (Grolemund and Wickham, 2017)
 - Advanced R (Wickham, 2015a, 2018)

- R packages (Wickham, 2015b)

5.3 Learning R on DataCamp

Since June 2017 we have a corporate premium group for the online R programming and learning platform DataCamp with five seats, which improved the R learning workflow dramatically.

Before that we already had a free KWB group on DataCamp. However, in case you were interested in doing the paid courses, each user needed to subscribe him/herself and needed to pay the fee on a monthly basis to DataCamp individually, which is now obsolete.

For using DataCamp at KWB the following workflow applies:

1. **Become a member in our free R@KWB DataCamp Gruppe** by clicking on the following link. In case you do not have a DataCamp account yet, please register yourself first by using your KWB email address. The membership in this group serves two purposes: on the one hand you should start on DataCamp with the free courses like for example Introduction to R first. On the other hand this group serves us as a kind of long term memory, to assess who and how frequently the DataCamp courses are used as the amount of users in the free group is – in contrast to our paid premium DataCamp group – not limited.
2. If you want to do a paid DataCamp course, please contact our DataCamp administrator Nicolas Caradot either personally or via email. In case there is a free seat available in our corporate premium group, he will add you so that you can start any of the available premium DataCamp courses.
3. As soon as you either do not want to or you do not have the time for a longer period do one of the paid DataCamp courses, please contact our group administrator Nicolas Caradot either personally or via email, so that he can make your seat available for others at KWB who are interested in learning R.

5.4 Using Subversion at KWB

For details on how to use Subversion at KWB please read the following tutorial

5.5 How to install KWB R packages?

For details please read the following tutorial

5.6 How to build your own KWB styled R package?

For building your own R package from scratch we developed a helper R package `kwb.pkgbuild` (available on Github), which builds a KWB styled skeleton for your future R package.

For more details checkout the tutorial at the package documentation website.

Recommended literature:

- R packages (Wickham, 2015b)
- Advanced R (Wickham, 2015a)

5.7 Working with EXCEL



Excel often crashes in case:

- a formula is applied for a whole column (i.e. 1 million rows)
- a lot of data is processed

Solution: data in own spreadsheet file

Hauke's "best Practices" for EXCEL (unvalidated, to be discussed!)

- Trennung zwischen Eingabe, Verarbeitung und Ausgabe zumindest auf Tabellenblattebene, d.h. ein Tabellenblatt (oder mehrere) für Eingabe, eines (oder mehrere) für Verarbeitung, eines (oder mehrere) für formatierte Ausgabe und / oder Diagramme
- ggf. Aufteilen auf mehrere Dateien. Das hätte den Nachteil, dass nicht mehr alles in einer Datei ist und nicht so leicht übergeben werden kann. In jedem Fall müsste eine Namenskonvention getroffen werden, z.B. `<file_name>_input.xlsx`, `<file_name>_calc.xlsx`, `<file_name>_output.xlsx`
- Verwenden der relativ neuen EXCEL-Features Als **Tabelle** formatieren.

Vorteil: Formeln können auf ganze Tabellenspalten angewendet werden; Spaltennamen anstatt Zellbezüge mit (unsprechenden) Buchstaben und Zahlen. z.B. Formel für Spalte `Volumen_L`: `= Durchfluss[@[Q_L_s]] * 60 * 5`

- Ein Tabellenblatt pro Tabelle
- Genau eine Headerzeile pro Tabelle mit eindeutigen Spaltennamen
- Ein Tabellenblatt, das die Bedeutung der Spaltennamen erläutert mit Spalten **Tabellenblatt**; **Spalte**; **Bedeutung**; **Einheit**; **Formel** Vorteil: Dieses Tabellenblatt sollte ausreichen, um die wesentlichen Berechnungen zu verstehen.

Drawback: needs to be permanently kept up-to-date!

- Hilfsspalten mit (dadurch benannten) Zwischenberechnungen anstatt Wiederholung von langen Ausdrücken in Formeln



A general online workshop on the topic Data Organisation in Spreadsheets is provided for free by the DataCarpentry organisation.

copied from 10_old_german_chapters.Rmd :

Best Practices zum Arbeiten mit Excel:

- z.B. Zellbezüge benennen, dadurch werden Formeln besser lesbar

Ihr werdet es kaum glauben, aber am Anfang meiner Zeit am KWB habe ich noch mit Excel gearbeitet. Ich habe auch komplexe Sachen gemacht und auch Excel-Makros programmiert. Ein Beweis findet sich in meiner persönlichen Logdatei:

Fr, 14.09.07 08:15-19:30 benutzerdefinierte Excelfunktionen zur Modellgüte in Personl.xls,
Modul1; TW-Kalibrierung Wochentag fertig

Ich würde heute nicht mehr empfehlen, Excel-Makros zu programmieren. Es ist umständlich, es gibt keine Bibliotheken und der Quellcode lässt sich nicht unabhängig von der Exceldatei verwalten, so dass keine ordentliche Versionsverwaltung möglich ist. Und wir wollen nicht mehr ohne Versionsverwaltung programmieren!

Wenn wir programmieren, dann sollten wir das einheitlich in R tun. R ist frei, es gibt eine großartige Community und wir haben mittlerweile eine große Expertise erlangt.

5.8 Heterogenous software versions on KWB computers



Unterschiedliche Softwareversionen (z.B. R) können dazu führen, dass Skripte auf verschiedenen Rechnern nicht das gleiche Verhalten zeigen.



Die IT-Abteilung ist in der Lage an bestimmte Nutzergruppen die gleiche Software (z.B. RStudio / R / Miktex) auszurollen. Dies sollte in Zukunft konsequent genutzt werden, indem auf alle Computer an denen potentiell programmiert wird zu dieser Nutzergruppe hinzugefügt werden und somit alle die gleichen Softwareversionen installiert haben.

5.9 R package/version dependency of R scripts

Lösungsvorschlag:

Es ist eine **Mindestdokumentation** der verwendeten R Version und sämtlicher R-Pakete (inklusive ihrer Versionen) zu fordern. Dazu kann in R die Funktion `sessionInfo()` genutzt werden. Die Ausgabe dieser Funktion kann entweder in eine Metadaten-Textdatei `session_info.txt` geschrieben werden oder im Falle der Erzeugung von R-Markdown-Dokumenten direkt am Anfang der Analyse im R-Markdown Dokument ausgegeben werden.

Das Schreiben der Metadaten-Datei `session_info.txt` sollte standardisiert über eine Funktion in einem KWB R-Paket (z.B. `kwb.utils`) implementiert werden.

Direktausgabe in R Console / RMarkdown:

```
sessionInfo()
```

```
## R version 3.5.0 (2017-01-27)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.5 LTS
##
## Matrix products: default
## BLAS: /home/travis/R-bin/lib/R/lib/libRblas.so
## LAPACK: /home/travis/R-bin/lib/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] knitr        citations_1.0.8
```

```
##
## loaded via a namespace (and not attached):
## [1] Rcpp_0.12.17      bookdown_0.7.13  lubridate_1.7.4  digest_0.6.15
## [5] rprojroot_1.3-2   plyr_1.8.4       R6_2.2.2         jsonlite_1.5
## [9] backports_1.1.2   magrittr_1.5     evaluate_0.10.1  highr_0.6
## [13] httr_1.3.1        bibtex_0.4.2     stringi_1.2.2    curl_3.2
## [17] rstudioapi_0.7    xml2_1.2.0       rmarkdown_1.10   tools_3.5.0
## [21] RefManager_1.2.0  stringr_1.3.1    xfun_0.2         yaml_2.1.19
## [25] compiler_3.5.0    htmltools_0.3.6  knitr_1.20
```

Schreiben in standardisierte Metadatei:

```
sink("session_info.txt")
sessionInfo()
sink()
```

Hierzu ist auch noch ein Tutorial zu erstellen!



Komplexere, technische Möglichkeiten zum Paketmanagement werden im gerade vom DFG geförderten Projekt O2R in der Entwicklung befindlichen CRAN task view for computational environments and reproducibility genannt.



Mit dem R Paket packrat lässt sich das Paketmanagement ggf. verbessern.

Als Beispielanwendung dient die Datenanalyse zur Feinstaubbelastung mit Sensebox-Daten, die komplett reproduzierbar sind (mittels der Plattform mybinder) und ohne Installation von Interessierten im Webbrowser ausgeführt werden kann (siehe hier).

5.10 Complex R script dependencies



Dieses Problem tritt insbesondere auf, wenn mehrere verschiedene Nutzer gemeinsam mit den gleichen Skripten arbeiten (wie z.B. im abgeschlossenen Projekt OGRE).



Proposed solution:

Bewusstmachen der Skriptabhängigkeiten

Identifizieren von Optimierungspotentialen -> möglicherweise Elimination von Abhängigkeiten

Workflow dokumentieren und Tutorial, am besten als R-Markdown Dokument, erstellen. Das ist insbesondere wichtig, wenn Skripte häufig verwendet werden.

5.11 Heterogenous (R-)coding styles



Currently there is no established coding style at KWB for in case of programming e.g. R scripts

(R) programmers at KWB will use the tidyverse coding style <http://style.tidyverse.org> as default. This will help increasing both, the readability and reusability of the developed (R-)scripts at KWB (currently: approximately 1000 Rscripts).

Recommended literature:

Clean code (Martin, 2009)

5.12 Collaborative version control



Multiple people developing code together.



1. At KWB we use Subversion (SVN) for version control (see: [How to use subversion?](#))
2. Speak with each other (especially if you are working on code is likely to be changed by others)
3. Regularly perform updates/commits with the version control software SVN



In case the code base (e.g. R scripts) developed in “project A” should be reused/adapted in a new “project B” (e.g. FLUSSHYGIENE) it is not allowed to change the code base in the original “project A”. Instead the code needs to be copied in a folder for the new project by using Subversion’s SVN copy command. In case this procedure is not followed established processes in the source project will be overwritten.

5.13 Workflow automation

Ressources:

- Reproducible Research Automation: This lesson shows how to use automation in R to improve the reproducibility of research by automating tasks.
- useR!2017 Video: Data Carpentry: Open and Reproducible Research (Tutorial)
 - Part1
 - Part2
- Use of an R package to facilitate reproducible research, e.g.:
 - rprkg by rOpenSci
 - rrttools by Ben Marwick

Recommended literature:

- Kitzes et al. (2018)

5.14 Encoding



Umlaute und Sonderzeichen werden falsch angezeigt, wenn R-Skripte in unterschiedlichen Encodings abgespeichert und eingelesen werden.



Vorgabe einer Default Encoding Einstellung in RStudio (z.B. UTF-8)

Alternativ könnten auch alle Umlaute in Unicode dokumentiert werden (siehe folgendes Beispiel). Allerdings ist dies wohl nicht praktikabel, da die Lesbarkeit der Texte erschwert wird und es sollte daher von uns der erste Ansatz (Vorgabe von UTF-8 als Default Encoding) angestrebt werden.

A great blogpost for the topic (How do I write UTF-8 encoded content to a file?) is provided by `citet(manual["Ushey_2018"])`, including background information on how encoding in R works.

For increasing the portability of R script he recommends the for example the following:

“Portable R scripts should use unicode code points, to avoid accidental mis-encoding of string literals.”

— `citet(manual["Ushey_2018"])`

Encoding example in R:

```
### Richtiges Skript Encoding WICHTIG (richtiges Einlesen des Skripts nur wenn
### mit gleichem Encoding eingelesen wie es auch abgespeichert wurde)
print("Ü")
```

```
## [1] "Ü"
```

```
### Skript Encoding EGAL (da Umlaut in Unicode codiert wurde)
print("\u00dc")
```

```
## [1] "Ü"
```

5.15 Collaborative writing



Project proposal with many partners. How can be guaranteed that only the most recently document is used and how can multiple people work at the same time in that document in the “hot phase” of proposal writing?



Ulf made good experiences with Office 365 (provided by KWR at that time). Also writing of up to 10 persons in parallel were not a problem, because the currently processed paragraph automatically gets blocked for others.

However, in case a whole chapter needs to be processed by one person this is not performed automatically. Thus it is recommended to delete the chapter for a short term from the synchronised document and replace it with a placeholder like **Chapter currently in revision by XXXX. Will be put back in this document until YYYY-MM-DD.**

5.16 Exchanging data with colleges/project partners

copied from: “10_old_german_chapters.Rmd”, to be translated:

Es gibt internen Datenaustausch und Austausch mit Projektpartnern. Oft werden Dateien intern per E-Mail verschickt. Das sollten wir nicht tun, da wir unnötig Kopien von Dateien anlegen. Dateien sollten im besten Fall nur an einem Ort abgelegt sein. Dort sollten sie allerdings durch ein Backup-System gesichert sein. Natürlich muss der Zugriff auf diesen Ort gewährleistet sein. Einem Studenten mit eingeschränkten

Rechten eine Datei per mail zuzuschicken, weil sie an dem Ort, an dem sie gespeichert ist, für ihn nicht zugänglich ist, ist meines Erachtens nicht die richtige Lösung. Es muss dann ein Ort geschaffen werden, an dem der gemeinsame Zugriff möglich ist. Laut Bodo ist der Ordner **Exchange** im Projektordner ein solcher Platz. Aus dem Grund der allgemeinen Zugänglichkeit wird er auch in manchen Projekten als allgemeiner Datenablageplatz “misbraucht”. Es kann aber niemandem ein Vorwurf gemacht werden, da ja nirgendwo dokumentiert ist, was in diesem Ort abgelegt werden soll.

Das können wir nun ändern:

Laut Bodo sind Dateien im Exchange-Ordner nur kurzfristig abzulegen, um z.B. Studenten Zugriff darauf zu geben. Die Dateien sollen nach dem Gebrauch wieder aus dem Ordner gelöscht werden. Demnach sollte der Ordner im Normalfall leer sein. Dies ist in einigen Projekten nicht der Fall.

Wir brauchen eine einheitliche Definition der Bedeutungen von Ordnern.

This should be documented in the metadata.

Chapter 6

Glossary

A detailed glossary covering the following topics is provided by Rokem and Chirigati (2018) is used throughout the glossary.

6.1 Reproducibility

”... is a cornerstone of science. Definitions vary greatly across scientific disciplines, but the meaning that we find most prevalent is the ‘calculation of quantitative scientific results by independent scientists using the original datasets and methods’ (Stodden, Leisch, & Peng, 2014). The goals of reproducibility go beyond duplicating someone else’s investigation: it also entails having reproducibility for yourself, defeating self-deception in scientific results (Ioannidis, 2005; Nuzzo, 2015), and extending another researcher’s methods to build your own work. Reproducibility is a matter of degree, not of kind. We say that research is reproducible if reproducibility applies to the results to some extent. That is, some of the corresponding experiments and scientific methods are deemed to be reproducible.

— (Rokem and Chirigati, 2018)

For more details see: Rokem and Chirigati (2018)

6.2 Provenance

“As the volume of digital data increases and the complexity of computational processes that manipulate these data grows, it is becoming increasingly important to manage their provenance. The Oxford English Dictionary defines provenance as the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners. Provenance helps determine the value, accuracy, and authorship of an object.”

— (Rokem and Chirigati, 2018)

For more details see: Rokem and Chirigati (2018)

6.3 Techniques

- Version control

- Literate Programming
- Data Publication
- Data cleaning/munging
- Software Testing
- Continuous Integration
- Workflow Management
- File Format Standards
- Licensing
- Virtualization and Environment Isolation

For details see: Rokem and Chirigati (2018)

6.4 Tools

- Programming Language and Related Tools
- Documentation Generators
- Version Control
- Data Munging and Analysis
- Data Visualization
- Software Testing and Continuous Integration
- Virtualization and Environment Isolation
- Data Sharing and Repositories
- Document Authoring
- File Formats

For details see: Rokem and Chirigati (2018)

Bibliography

- Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J., Starn, J., and Fienen, M. (2016a). Flopy: Python package for creating, running, and post-processing modflow-based models.
- Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., and Fienen, M. N. (2016b). Scripting MODFLOW model development using python and FloPy. *Groundwater*, 54(5):733–739.
- Boland, M. R., Karczewski, K. J., and Tatonetti, N. P. (2017). Ten simple rules to enable multi-site collaborations through data sharing. *PLOS Computational Biology*, 13(1):e1005278.
- Bosman, J. and Kramer, B. (2016). Github and more: sharing data & code. <https://101innovations.wordpress.com/2016/10/09/github-and-more-sharing-data-code/>.
- Bosman, J. and Kramer, B. (2018). Workflows. <https://101innovations.wordpress.com/workflows>.
- Buettner, S., Hobohm, H.-C., and Mueller, L. (2011). *Handbuch Forschungsdatenmanagement*. Bock u. Herchen, Bad Honnef.
- Fenner, M. and Haak, L. (2014). *Unique Identifiers for Researchers*, pages 293–296. Springer International Publishing, Cham.
- Fenner, M., Scheliga, K., and Bartling, S. (2014). *Reference Management*, pages 125–137. Springer International Publishing, Cham.
- Friesike, S. (2014). *Creative Commons Licences*, pages 287–288. Springer International Publishing, Cham.
- Grolemund, G. and Wickham, H., editors (2017). *R for Data Science*. O’Reilly Media, Sebastopol, CA, 1 edition.
- Hart, E. M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., Poisot, T., Woo, K. H., Zimmerman, N. B., and Hollister, J. W. (2016). Ten simple rules for digital data storage. *PLOS Computational Biology*, 12(10):e1005097.
- Kaden, B. and Kleineberg, M. (2018). Guidelines zur veröffentlichung dissertationsbezogener forschungsdaten.
- Kitzes, J., Turek, D., and Deniz, F., editors (2018). *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*. University of California Press, Oakland, CA.
- Kraus, F., Seis, W., Remy, C., Rustler, M., Jubany i Guell, I., Espi, J. J., and Clarens, F. (2016). Deliverable d3.2: Show case of the environmental benefits and risk assessment of reuse schemes. Report, public report in the DEMOWARE project (www.demoware.eu), KWB.
- Martin, R. C. (2009). *Clean Code*. Pearson Education, Inc. ISBN-13 978-0-13-235088-4/978-1498716963.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rokem, A. and Chirigati, F. (2018). *Glossary*. University of California Press, Oakland, CA.

- Rustler, M. (2016a). kwb.hantush (v.0.2.1). Tutorial: <http://kwb-r.github.io/kwb.hantush/>.
- Rustler, M. (2016b). kwb.qmra (v.0.1.1).
- Rustler, M. (2016c). Quantitative microbiological risk assessment for different wastewater reuse options in Old Ford (v.1.0). For performing the quantitative microbial risk assessment the R package kwb.qmra, v.0.1.1 (<https://doi.org/10.5281/zenodo.154111>) is used.
- Rustler, M. (2018). aquanes.report (v.0.5.0).
- Sonnenberg, H. and Rustler, M. (2016). kwb.wtaq (v.0.2.1). Tutorial website: <http://kwb-r.github.io/kwb.wtaq>.
- Stodden, V. (2014). *Intellectual Property and Computational Science*, pages 225–235. Springer International Publishing, Cham.
- Taschuk, M. and Wilson, G. (2017). Ten simple rules for making research software more robust. *PLOS Computational Biology*, 13(4):e1005412.
- Wickham, H., editor (2015a). *Advanced R*. The R Series. Chapman and Hall/CRC, Boca Raton, FL, 1 edition.
- Wickham, H., editor (2015b). *R Packages: Organize, Test, Document, and Share Your Code*. O’Reilly Media, Sebastopol, CA, 1 edition.
- Wickham, H., editor (2018). *Advanced R*. 2 edition.
- Wilson, G., Bryan, J., Cranston, K., Kitjes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6):e1005510.