

Working with Research Data

Markus Stocker

September 12, 2017

Outline

- Accessing and reusing research data
- Computational environments for data processing
- Curating and storing data, from files to databases
- Research data versioning and backup

Data Access

- It's complicated but it is improving
- Drivers for better access
 - ▶ Open Data imperative
 - ▶ Credit for publishing data
 - ▶ Increase return on investment in scientific research
 - ▶ Funders requiring data to be published
- Correspondingly, supporting infrastructures is
 - ▶ Increasing in number and quality
 - ▶ Adopting principles, guidelines, standards

Data Access

- You know how to access *your* data
- More difficult is access to data authored by others
- Presumes others have published their data
- Then you may be able to
 - ▶ Find their data
 - ▶ Retrieve the data
 - ▶ Reuse the data

Find Data

- Useful data can be found in a lot of places
- Online or offline, e.g. printed books
- In data repositories or as files on a server
- You could try a Google search
- Or ask your supervisor and fellow students
- The authors of papers you read may cite data and/or sources

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

 Search

re3data.org Reaches a Milestone and Begins Offering Badges

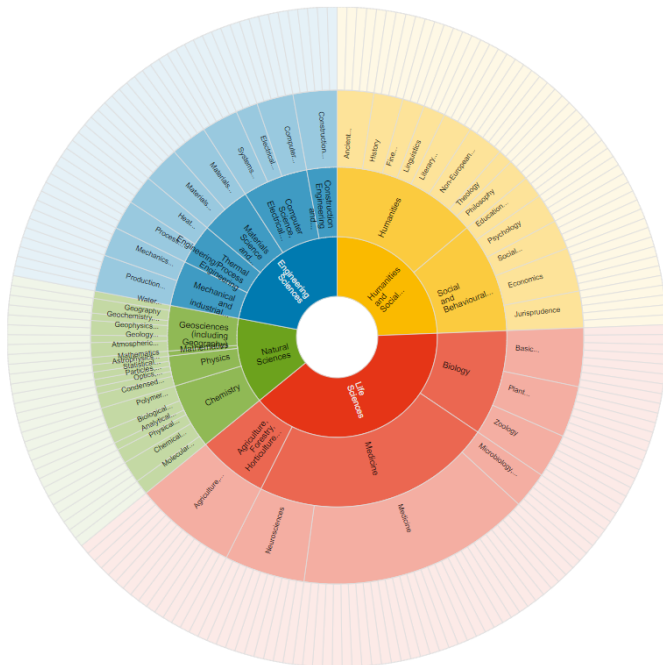
re3data.org has reached a milestone of identifying and listing 1,500 research data repositories, making it the largest and most

Enhancements to creating and updating re3data

We are happy to announce a new feature that enables users to more easily suggest corrections and enhancements of

New re3data.org Schema and Search Functionality















We are pleased to announce the publication of version 3.0 of the "Metadata Schema for the Description of Research Data Repositories" (Rücknagel et al., 2015).



Retrieve Data

- Typically download of one or more files
- An API for programmatic retrieval may be available
- Data repositories generally support search
- Often data are retrieved as they were deposited (original format)
- Repository may standardize data during ingestion

Parameter(s):

#	Name	Short Name	Unit
1	DEPTH, sediment/rock 	Depth	m
2	AGE 	Age	ka BP
3	Sample code/label 	Sample label	
4	Duration 	Duration	ka
5	Biozone 	Biozone	
6	Temperature, coldest month 	CMT	°C
7	Temperature, coldest month 	CMT	°C
8	Temperature, coldest month 	CMT	°C
9	Sigma 	Sigma	
10	Sigma 	Sigma	
11	Temperature, warmest month 	WMT	°C
12	Temperature, warmest month 	WMT	°C
13	Temperature, warmest month 	WMT	°C
14	Covariance 	Cov	

License: Creative Commons Attribution 3.0 U

Size: 960 data points

Data

[Download dataset as tab-delimited text \(use the fo](#)

1	2	3	4	5
Depth [m]	Age [ka BP]	Sample label	Duration [ka]	Biozon
19.100	115.125 80		10.875 E7	
19.300	115.325 79		10.675 E7	
19.500	115.525 78		10.475 E7	
19.700	115.790 77		10.210 E7	
19.900	116.060 76		9.940 E7	
20.100	116.310 75		9.690 E7	

<https://doi.pangaea.de/10.1594/PANGAEA.548373?format=textfile>

Bispingen tempera t ^

The screenshot shows an Excel spreadsheet titled "Bispingen_temperature - Excel". The spreadsheet is organized into rows and columns. The first row (row 25) contains the following data:

Depth [m]	Age [ka BP]	Sample la	Duration [Biozone]	CMT [°C]	CMT [°C]	CMT [°C]	Sigma (of)	Sigma (of)	WMT [°C]	WMT [°C]	WMT [°C]	WMT [°C]
19.1	115.125	80	10.875	5	-5.6	-16.2	6.4	2.5	19.6	15.6	15.6	15.6
19.3	115.325	79	10.675	-1.3	-7.8	-14.4	4	1.9	18	14.7	14.7	14.7
19.5	115.525	78	10.475	-1.3	-7.8	-14.4	4	1.9	18	14.7	14.7	14.7
19.7	115.75	77	10.21	0	-5.5	-11	3.3	1.7	18.6	15.9	15.9	15.9
19.9	116.06	76	9.94	4.5	-2.9	-10.3	4.5	1.9	19.9	16.7	16.7	16.7
20.1	116.31	75	9.69	6.4	-1.5	-9.4	4.8	2.4	20.9	17	17	17

The spreadsheet also includes a "License" section with Creative Commons Attribution 3.0 Unported (CC-BY) and a "Size" section with 960 data points.

[Alle anzeigen](#)

Reuse Data

- Complicated!
- Generally substantial processing needed to make reuse possible
- Even if accessible, data are generally not interoperable
- Lack syntactic interoperability due to different formats
- Lack semantic interoperability due to different terminology
- Data and metadata quality may not be adequate
- Different resolution, gaps, outliers, insufficient information, ...
- Data need to be integrated: common syntax and semantics
- A lot of time required to prepare for reuse

Data Processing

- Assume integrated data
- Your next step is to process them for your purpose
- Staggering amount of methods
- Programming (scripting) languages
- Computational environments and other tools
- Processing results in new (derived) data
- Various kinds of processing, transformation, interpretation, ...

Curating and Storing Data

- Data need to be identified, described, quality controlled, etc.
- Curated data are stored and possibly preserved
- How you curate and store data depends on various factors, e.g.
 - ▶ Longevity: from temporary to preserved data
 - ▶ Sharing: with yourself or a community
 - ▶ Dynamism: from static files to queriable databases

Databases

- Many kinds but relational most common
- Help structuring data, be consistent with datatypes
- Flexible data retrieval with declarative language (SQL)
- Access to data from programming languages (Python, R, Java, ...)
- Processing of large data quantities
- Access management and security
- Backup



Databases

```
create table data (  
  id integer primary key,  
  time timestamp,  
  latitude double precision,  
  longitude double precision,  
  temperature double precision  
)
```

Databases

```
insert into data values(  
    1,  
    timestamp '2016-07-26 00:00:00',  
    -70.650000,  
    -8.250000,  
    -19.3  
)
```

Databases

```
select
  id, time, latitude, longitude, temperature
from data
```

```
select * from data where temperature < -15
```

id	time	latitude	longitude	temperature
-----+-----+-----+-----				
1	2016-07-26 00:00:00	-70.65	-8.25	-19.3

From Python

```
import pyodbc

con = pyodbc.connect('DRIVER=...;SERVER=...;PORT=...;
                     DATABASE=...;UID=...;PWD=...')
cur = con.cursor()

cur.execute('select latitude, longitude, temperature
            from data where temperature < -15')

for row in cur.fetchall():
    print('Latitude: {}; Longitude: {}; Temperature: {}'.
          .format(row[0], row[1], row[2]))

cur.close()
con.close()
```

Backup and Versioning

- Your (Word) manuscripts are data
- Backup your other data as you backup your manuscripts
- Avoid memory sticks as backup media
- Test your backup strategy: can you recover lost files?
- Make sure to version your data
- Including if you collaborate only with yourself
- Remote version control can act as backup

Take aways