Göttingen/eResearch Alliance

Research Data Management
What‘s in it for me?

GGNB, 06.09.2017

Timo Gnadt

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

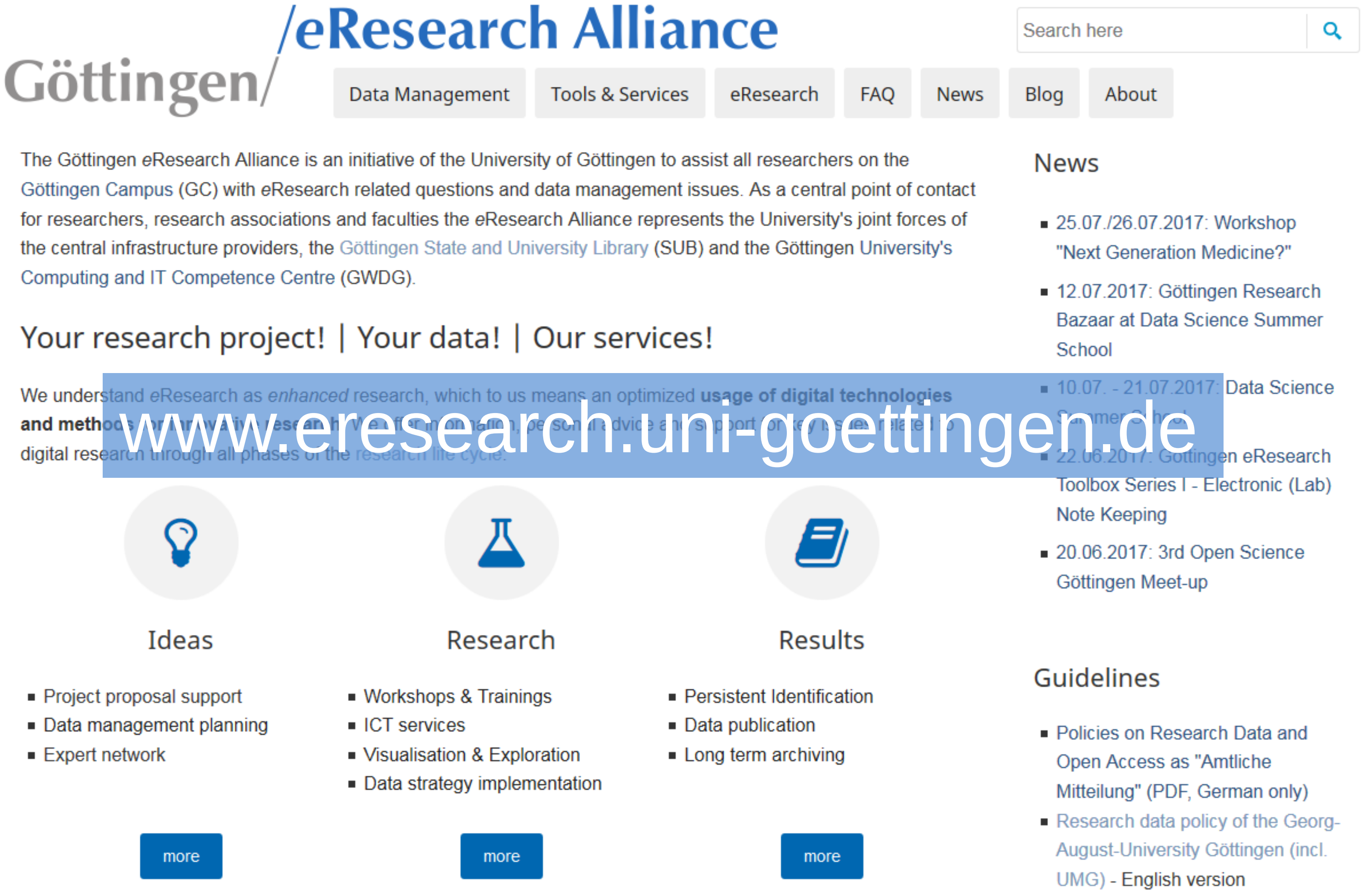
NIEDERSÄCHSISCHE STAATS- UND
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

SUB

GWDG

Outline

- Introduction
- Research / Data / Management
- Data Management Planning
- Backup & Storage
- Organization & Documentation
- Data sharing and legal aspects



The Göttingen eResearch Alliance is an initiative of the University of Göttingen to assist all researchers on the Göttingen Campus (GC) with eResearch related questions and data management issues. As a central point of contact for researchers, research associations and faculties the eResearch Alliance represents the University's joint forces of the central infrastructure providers, the Göttingen State and University Library (SUB) and the Göttingen University's Computing and IT Competence Centre (GWDG).

Your research project! | Your data! | Our services!

We understand eResearch as *enhanced* research, which to us means an optimized **usage of digital technologies and methods for innovative research**. We offer information, personal advice and support for key issues related to digital research through all phases of the research life cycle.

Ideas

- Project proposal support
- Data management planning
- Expert network

more

Research

- Workshops & Trainings
- ICT services
- Visualisation & Exploration
- Data strategy implementation

more

Results

- Persistent Identification
- Data publication
- Long term archiving

more

News

- 25.07./26.07.2017: Workshop "Next Generation Medicine?"
- 12.07.2017: Göttingen Research Bazaar at Data Science Summer School
- 10.07. - 21.07.2017: Data Science Summer School
- 22.06.2017: Göttingen eResearch Toolbox Series I - Electronic (Lab) Note Keeping
- 20.06.2017: 3rd Open Science Göttingen Meet-up

Guidelines

- Policies on Research Data and Open Access as "Amtliche Mitteilung" (PDF, German only)
- Research data policy of the Georg-August-University Göttingen (incl. UMG) - English version

www.eresearch.uni-goettingen.de

Göttingen eResearch Alliance (eRA)

- diverse backgrounds
  - mainly in natural sciences, humanities, computer science

Dr Philipp Wieder
Degree in Electrical Engineering, PhD in Computer Science

Dr Jan Brase
Degree in Mathematics, PhD in Computer Science

Fatih Berber
Computer scientist

Timo Gnadt
Computer scientist

Péter Király
Software developer and historian

Christopher Menke
Computer scientist

Lena Steilen
Historian and cultural anthropologist

- run mutually by

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

SUB | NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

- extensive expertise on e-research related topics

→ *we are not experts in your discipline, but we can relate to your data management requirements*

What eRA can do for you

- Consultations / support
  - Research Data Management
  - Publication strategies
  - Digital methods, software and technologies to enhance a research project
  - Information hub for experts & expertise on the whole campus
- Training
  - (like right here & now)
  - Information material / knowledge base
- Collaboration
  - Liaising project partnership
  - Project as a service

RDM WS GGNB
Research / Data / Management

06.09.2017

Research Data Management

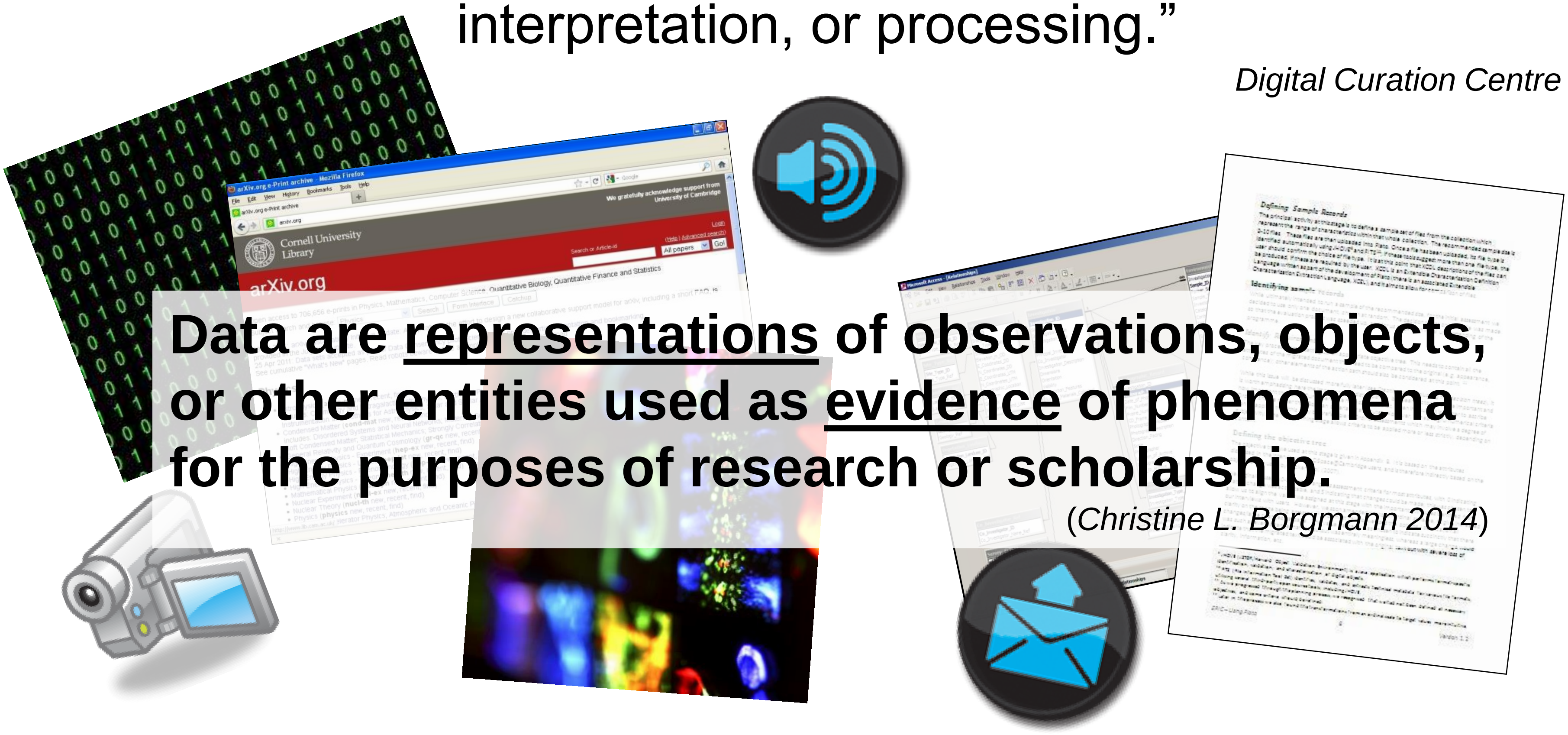
Surely you know what that is…

… and how to do it. RIGHT?

What is ‘data’?

“A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.”

Digital Curation Centre



Data are <u>representations</u> of observations, objects, or other entities used as <u>evidence</u> of phenomena for the purposes of research or scholarship.

(Christine L. Borgmann 2014)

Adapted from: Open Access Teaching Materials for Digital Preservation, Produced by Anna Collins (2012) for the JISC-funded PrePARe project

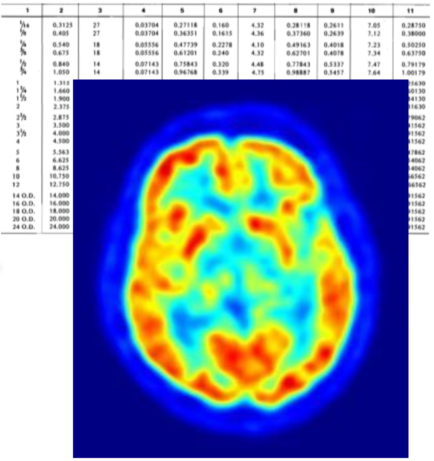
What is Research Data?

Any information you use in your research:

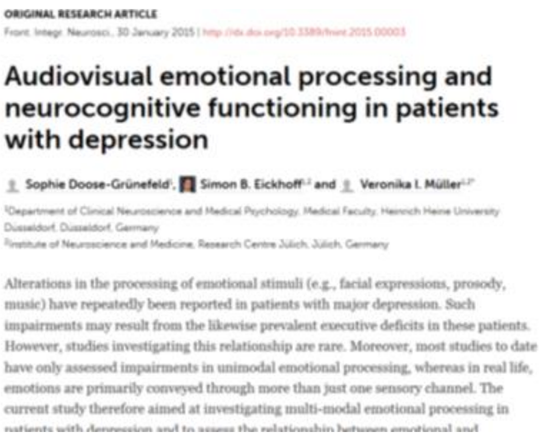
statistics, interviews, simulations, measurement data from experiments, observational data from instruments, text with semantic annotations, 3D scans, model drawings, numerical representations, ...

In many forms:

Video, audio, images, spreadsheets, paper documents, binary data, software, text files, lab notebooks, ...



1	2	3	4	5	6	7	8	9	10	11
1/16	0.3125	27	0.03704	0.27118	0.160	4.32	0.28118	0.2611	7.05	0.28750
1/8	0.405	27	0.03704	0.36351	0.1615	4.36	0.37360	0.2639	7.12	0.38000
1/4	0.540	18	0.05556	0.47739	0.2278	4.10	0.49163	0.4018	7.23	0.50250
3/8	0.675	18	0.05556	0.61201	0.240	4.32	0.62701	0.4078	7.34	0.63750
1/2	0.840	14	0.07143	0.75843	0.320	4.48	0.77843	0.5337	7.47	0.79179
3/4	1.050	14	0.07143	0.96768	0.339	4.75	0.98887	0.5457	7.64	1.00179
1	1.315									5630
1 1/4	1.660									0130
1 1/2	1.900									4130
2	2.375									1630
2 1/2	2.875									9062
3	3.500									1562
3 1/2	4.000									1562
4	4.500									1562
5	5.563									7862
6	6.625									4062
8	8.625									4062
10	10.750									6562
12	12.750									6562
14 O.D.	14.000									1562
16 O.D.	16.000									1562
18 O.D.	18.000									1562
20 O.D.	20.000									1562
24 O.D.	24.000									1562



ORIGINAL RESEARCH ARTICLE

Front. Integr. Neurosci., 30 January 2015 | http://dx.doi.org/10.3389/fnint.2015.00003

Audiovisual emotional processing and neurocognitive functioning in patients with depression

Sophie Doose-Grünefeld[1], Simon B. Eickhoff[1,2] and Veronika I. Müller[1,2*]

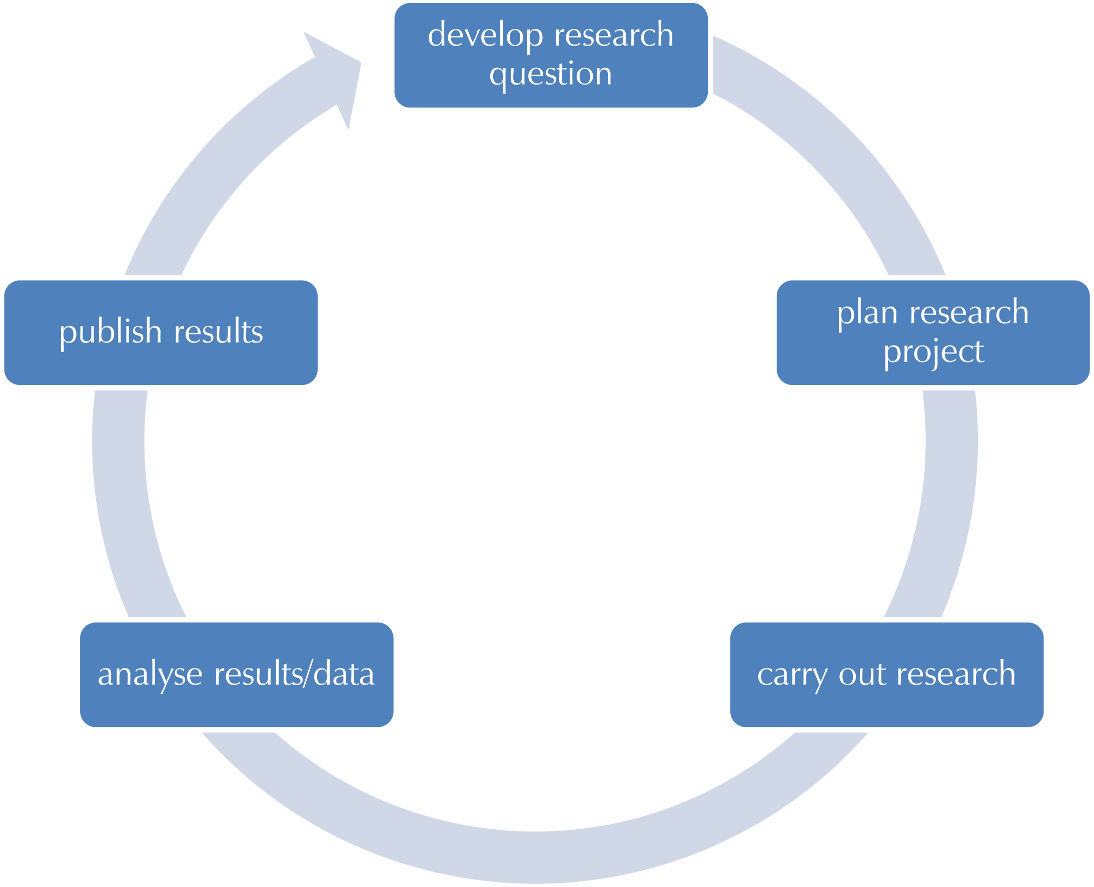
[1]Department of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Institute of Neuroscience and Medicine, Research Centre Jülich, Jülich, Germany

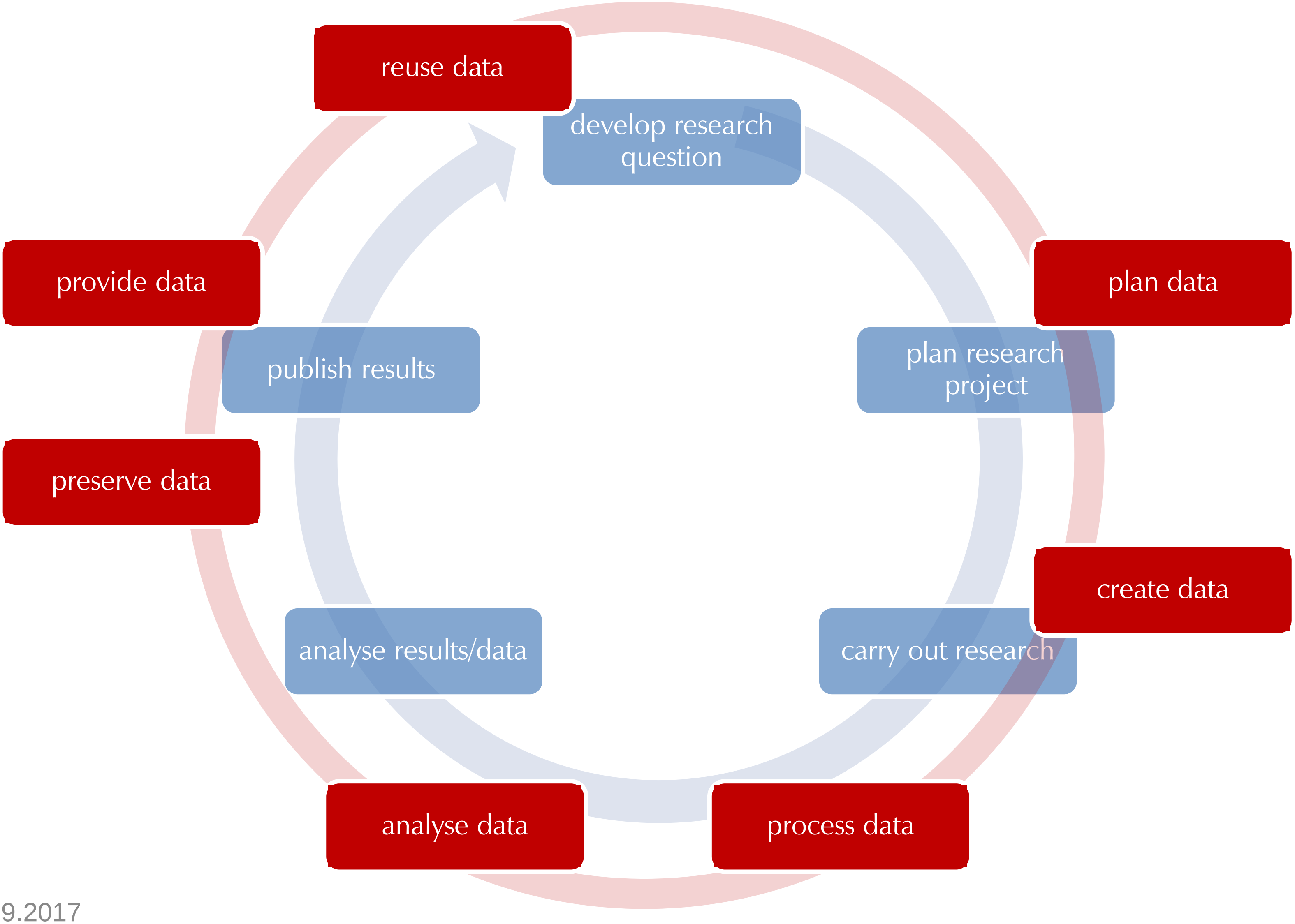
Alterations in the processing of emotional stimuli (e.g., facial expressions, prosody, music) have repeatedly been reported in patients with major depression. Such impairments may result from the likewise prevalent executive deficits in these patients. However, studies investigating this relationship are rare. Moreover, most studies to date have only assessed impairments in unimodal emotional processing, whereas in real life, emotions are primarily conveyed through more than just one sensory channel. The current study therefore aimed at investigating multi-modal emotional processing in patients with depression and to assess the relationship between emotional and

research object → **research data** → **result/ publication**

Research lifecycle



Research data lifecycle



Research data – a valuable investment

Source: European Space Agency: Rosetta and Philae at comet, on flickr. CC-BY-SA-2.0

Rosetta & Philae

Duration:

- >10 years preparation
- 10 years from start to data

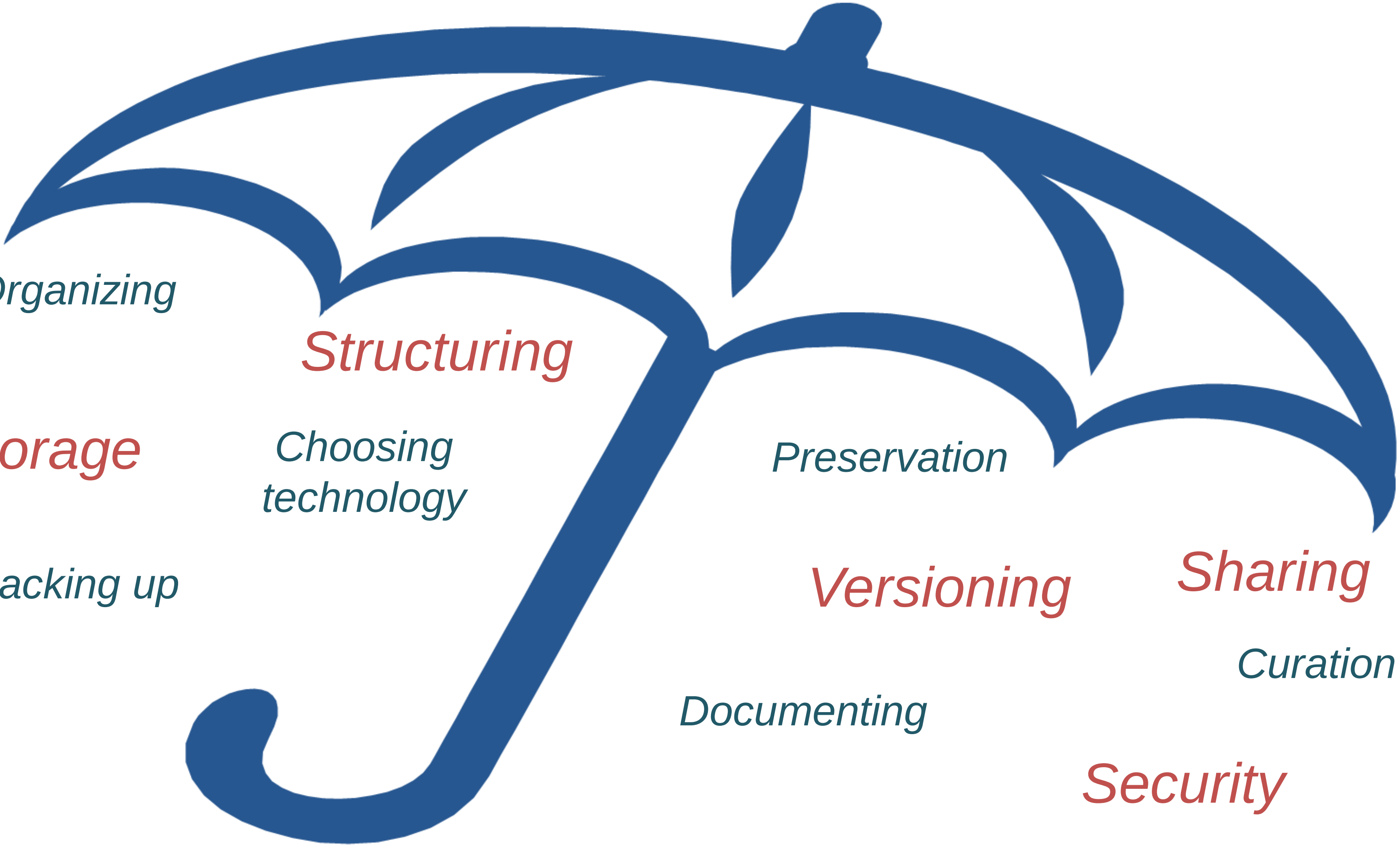
Costs:

- over € 1.000.000.000

Outcome:

- some cool photos
- lots of data
- *a radically new theory on the origin of the universe?*

What is Research Data Management?



Organizing

Structuring

Storage

Choosing technology

Backing up

Preservation

Versioning

Sharing

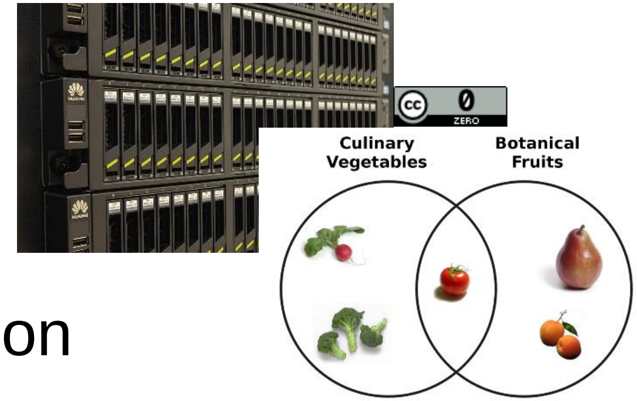
Curation

Documenting

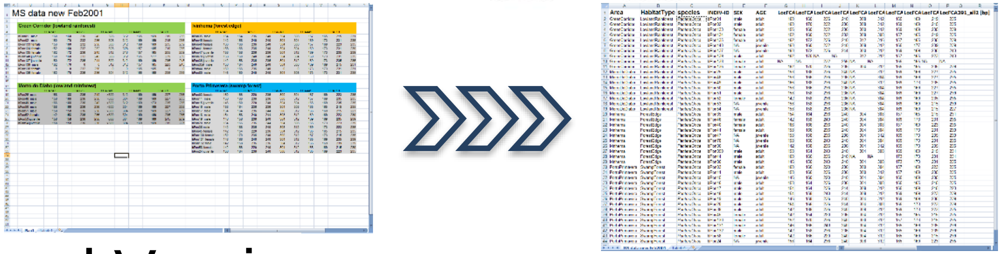
Security

Adapted from “Research data management – a very brief introduction”, the Research Support team, IT Services, University of Oxford

What is Research Data Management?



- Backup and Storage
- Metadata and Documentation
- Data Quality
- File Names, Identifier and Versions
- Ethics, Rights and Licenses


MS data new Feb2001

doi:10.10.1038/nphys1170

Thesis_final_v13b_revised.docx

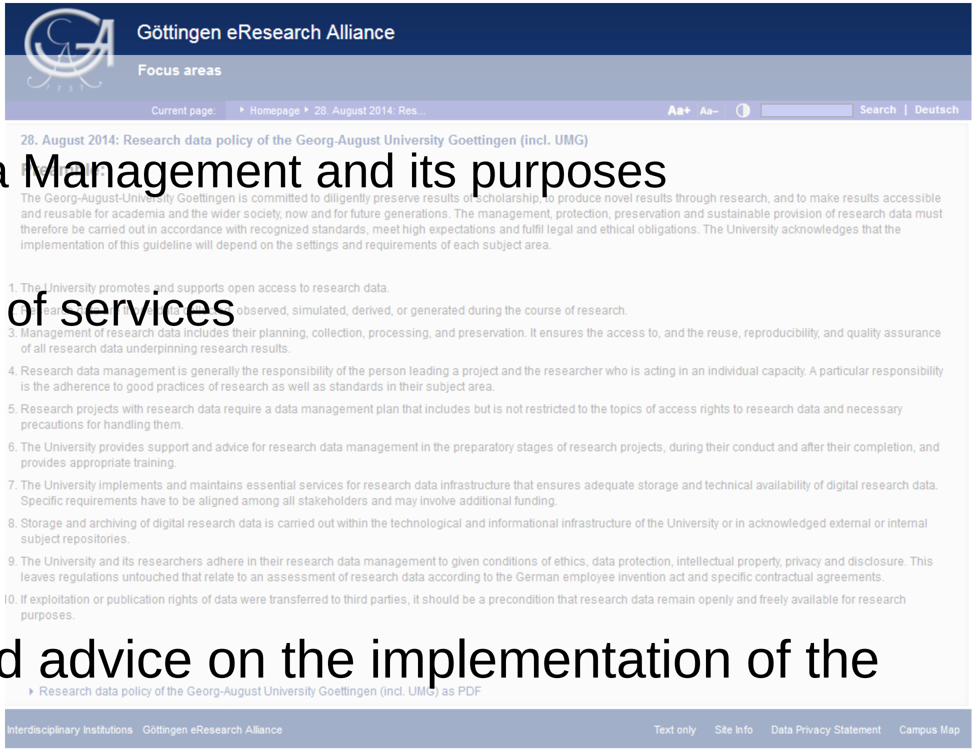
BY NC SA

Research Data Policy of the Georg-August Universität Göttingen

- Officially issued on 28th August 2014
- One of the first German universities with such a policy

- Topics addressed:
  - Research Data, Research Data Management and its purposes
  - Data Management Plans
  - Support, training and provision of services
  - Storage location
  - Ethical and legal standards
  - Open Access

- eResearch Alliance: support and advice on the implementation of the RDP for the Göttingen Campus


Göttingen eResearch Alliance
Focus areas
Current page: Homepage 28. August 2014: Res...
Aa+ Aa– Search | Deutsch

28. August 2014: Research data policy of the Georg-August University Goettingen (incl. UMG)

Preamble:

The Georg-August-University Goettingen is committed to diligently preserve results of scholarship, to produce novel results through research, and to make results accessible and reusable for academia and the wider society, now and for future generations. The management, protection, preservation and sustainable provision of research data must therefore be carried out in accordance with recognized standards, meet high expectations and fulfil legal and ethical obligations. The University acknowledges that the implementation of this guideline will depend on the settings and requirements of each subject area.

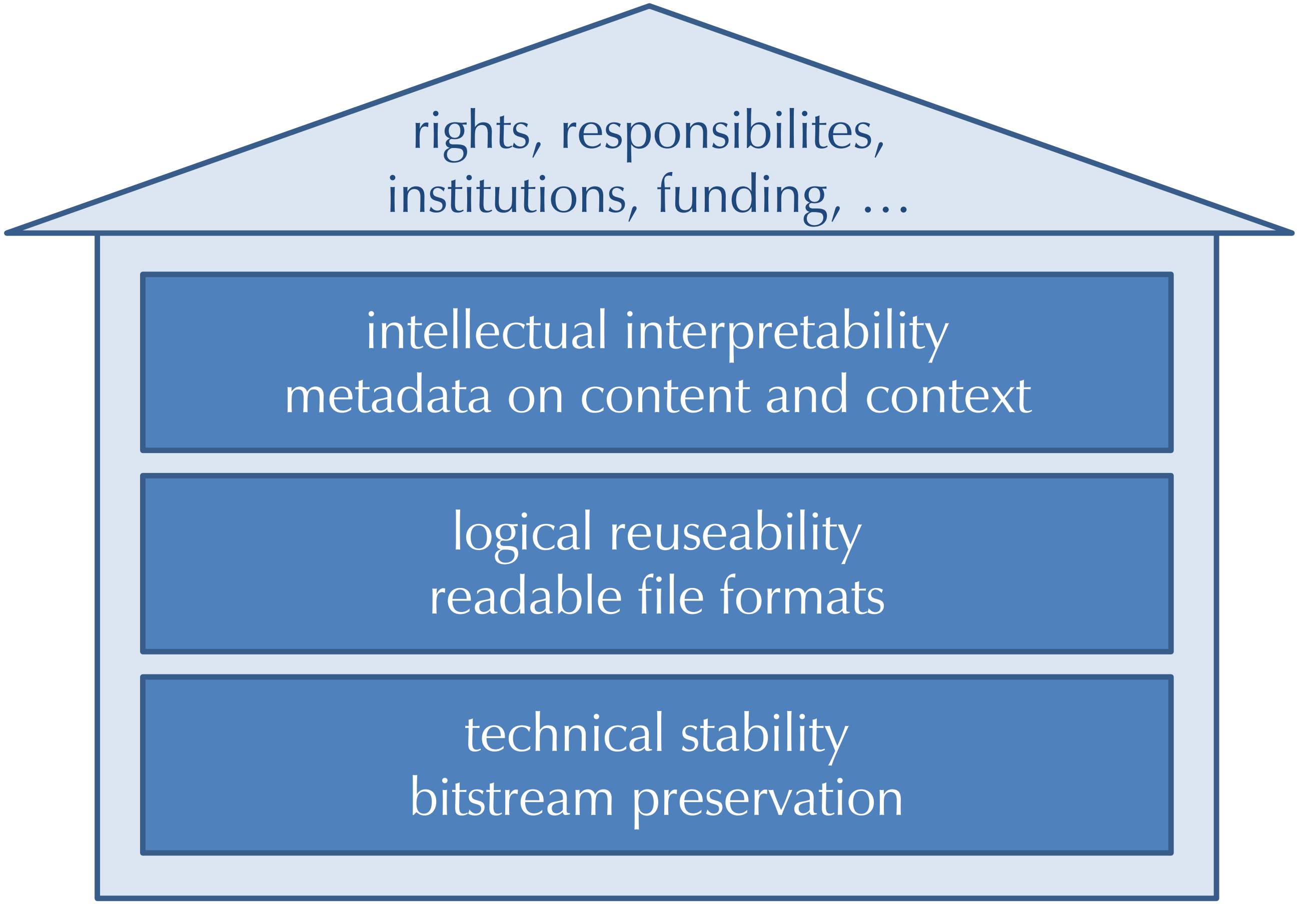
1. The University promotes and supports open access to research data.
2. Research data are those data collected, observed, simulated, derived, or generated during the course of research.
3. Management of research data includes their planning, collection, processing, and preservation. It ensures the access to, and the reuse, reproducibility, and quality assurance of all research data underpinning research results.
4. Research data management is generally the responsibility of the person leading a project and the researcher who is acting in an individual capacity. A particular responsibility is the adherence to good practices of research as well as standards in their subject area.
5. Research projects with research data require a data management plan that includes but is not restricted to the topics of access rights to research data and necessary precautions for handling them.
6. The University provides support and advice for research data management in the preparatory stages of research projects, during their conduct and after their completion, and provides appropriate training.
7. The University implements and maintains essential services for research data infrastructure that ensures adequate storage and technical availability of digital research data. Specific requirements have to be aligned among all stakeholders and may involve additional funding.
8. Storage and archiving of digital research data is carried out within the technological and informational infrastructure of the University or in acknowledged external or internal subject repositories.
9. The University and its researchers adhere in their research data management to given conditions of ethics, data protection, intellectual property, privacy and disclosure. This leaves regulations untouched that relate to an assessment of research data according to the German employee invention act and specific contractual agreements.
10. If exploitation or publication rights of data were transferred to third parties, it should be a precondition that research data remain openly and freely available for research purposes.

Research data policy of the Georg-August University Goettingen (incl. UMG) as PDF

Interdisciplinary Institutions Göttingen eResearch Alliance
Text only Site Info Data Privacy Statement Campus Map

Source: http://www.uni-goettingen.de/en/488918.html

Levels of data preservation



rights, responsibilites,
institutions, funding, …

intellectual interpretability
metadata on content and context

logical reuseability
readable file formats

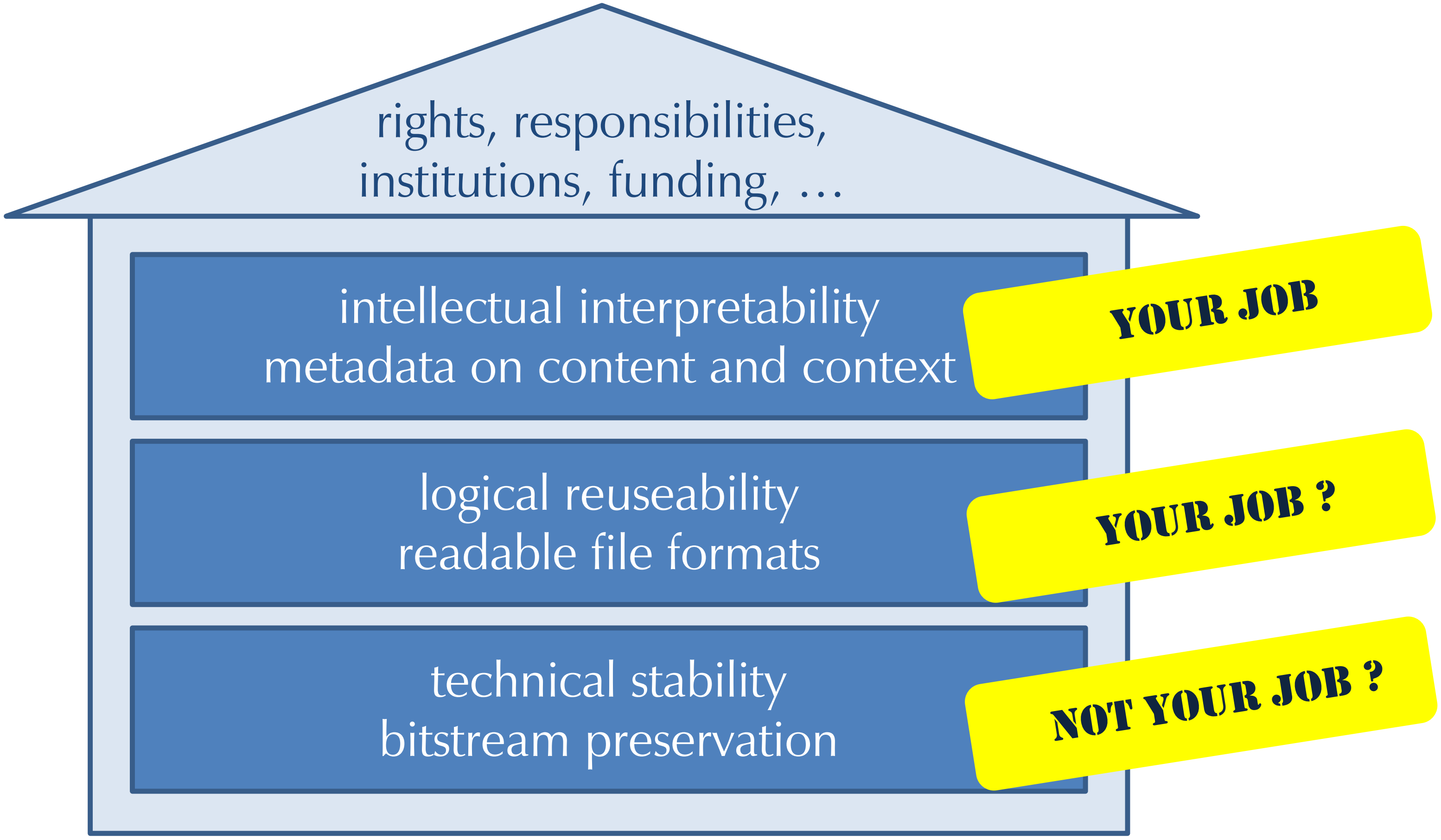
technical stability
bitstream preservation

Data preservation motivation

Video:
„Data Management SNAFU in 3 short acts“
By NYU Health Sciences Library

https://www.youtube.com/watch?v=66oNv_DJuPc

Levels of data preservation


rights, responsibilities,
institutions, funding, …

intellectual interpretability
metadata on content and context

logical reuseability
readable file formats

technical stability
bitstream preservation

Why Research Data Management?

1. Improve your research

- prevent data loss
- prevent unnecessary work
- better data quality

2. Good Scientific Practice

- reproducibility, accountability and compliance
- "Primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin." (DFG, Proposals for safeguarding good scientific practice, 1998)
- Requirement from DFG: every new project proposal has to explain how it will deal with research data and whether it will be shared.

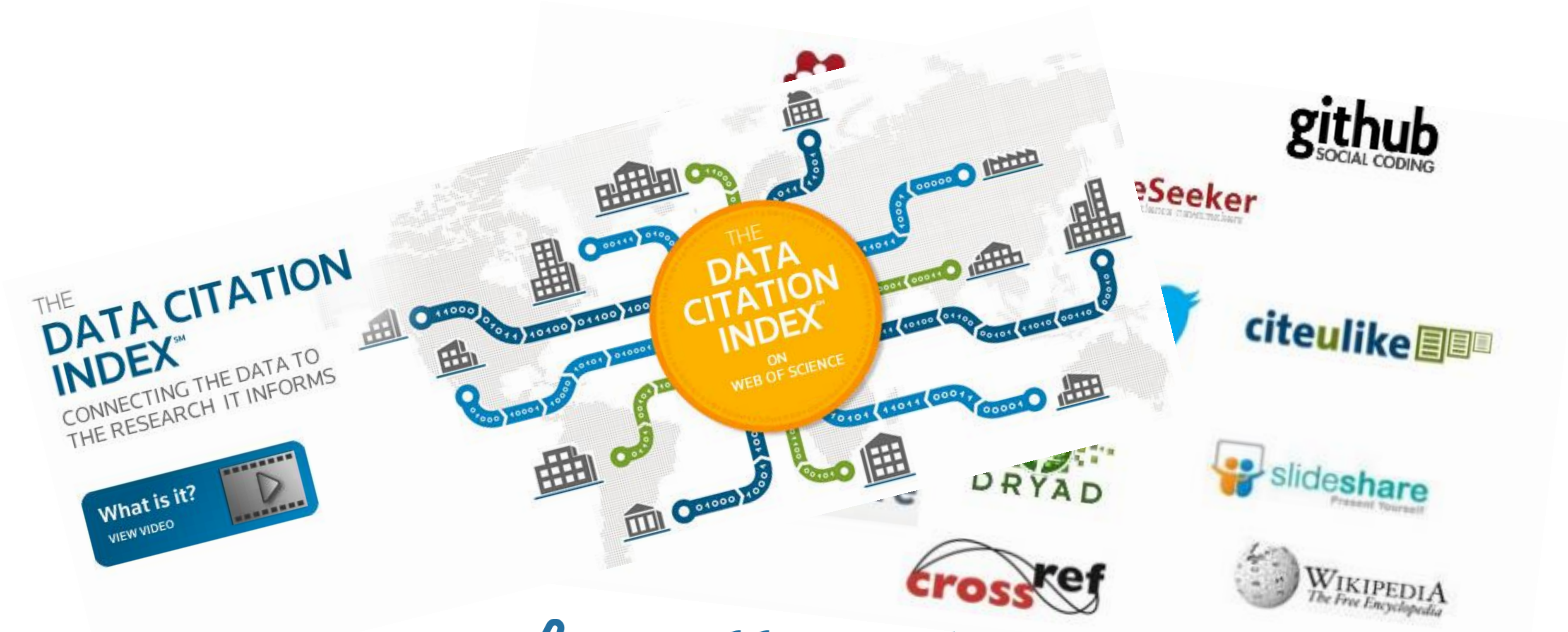
3. Data Sharing with Colleagues

- Research can be *very* expensive and the only result of long research journeys may be data.
- Data management costs are small in comparison to data creation costs.
- Productive data sharing is simply a matter of efficiency.

Why Research Data Management?

nature.com
Publications A-Z index
Browse by subject
SCIENTIFIC DATA
Home
Archive
Collections
RCSB PDB
PDB PROTEIN DATA BANK
An Information Portal to 121654 Biological Macromolecular Structures
Plant Genomics and Phenomics Research Data Repository
e!DAL electronic data archive library
Last Published: 2017-01-12
Version: 2.4.8
Welcome
Download & Source Code
Running Repositories
Demos and Samples
Paper - Talks - Poster
Development
Run
License and disclaimer
Contact
Imprint / Data Protection Policy
Acknowledgment / Public Grants
Project Information
Built by maven
Copyright © 2017 Leibniz Institute of Plant Genetics and Crop Plant Research
DSDR Data Sharing for Demographic Research
ANNOUNCEMENTS
ABOUT US
FIND DATA
GEO Gene Expression Omnibus
NCBI
Public holdings
Series
Platforms
Samples
Organisms
History
Series type
Count
Expression profiling by array
Expression profiling by genome tiling array
Expression profiling by high throughput sequencing
Methylation profiling by array
Protein profiling by protein array
SNP genotyping by SNP array
Other
Third-party reanalysis
Search
DATASETS 1 TO 25 OF 25,376
knb
ABOUT
DATA
SHARE
TOOLS
SIGN IN
Earth System Science Data
REVISE SEARCH

Why Research Data Management?



Usage
Views and
downloads

Mentions
Blogged, mentioned
in Wikipedia

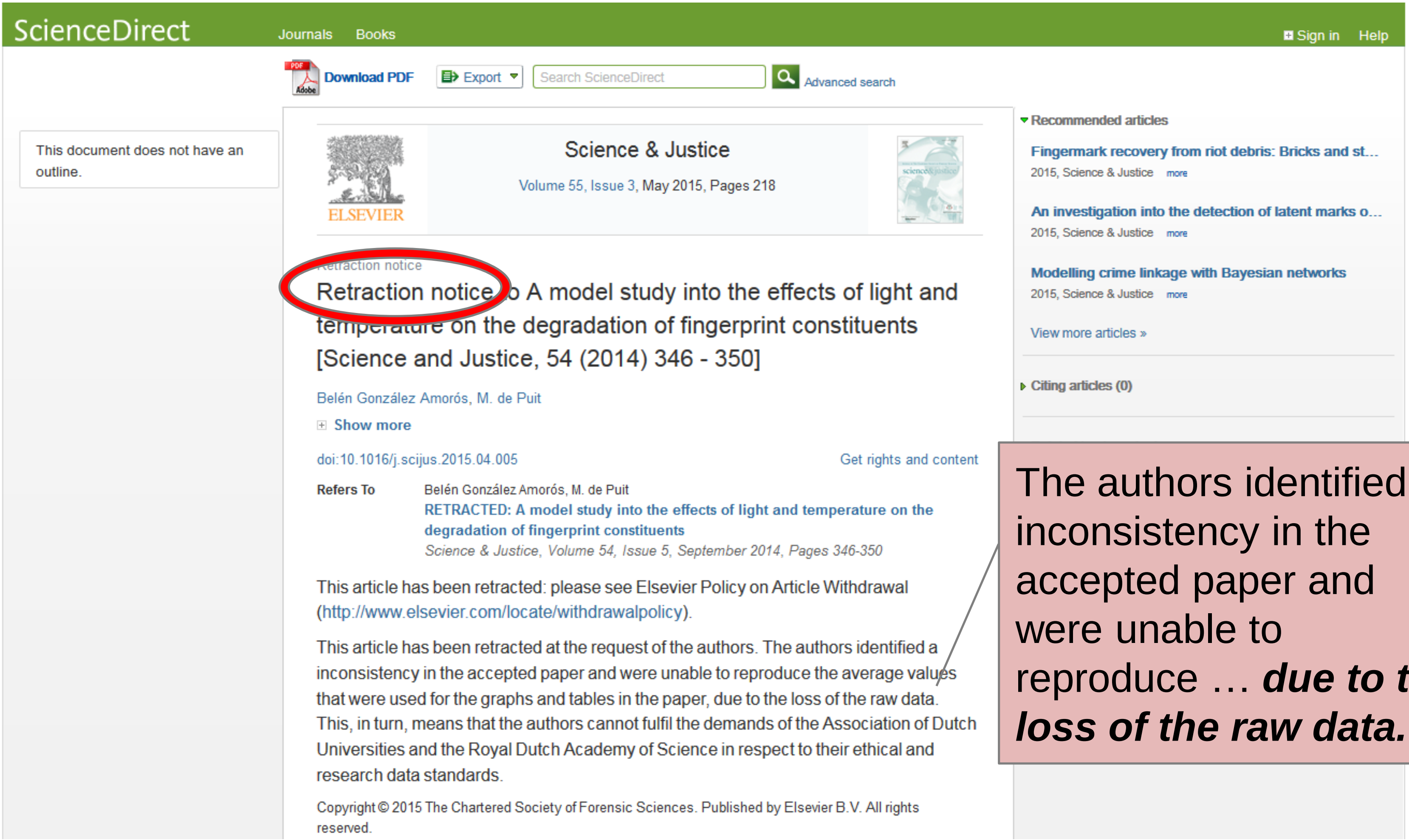
Social Media
“Likes” on Facebook,
tweets on Twitter

Cites
Web of Science,
PubMed, Scopus

Captures/ Shares
Bookmarks on CiteULike,
shares on Mendeley

Altmetric

Why Research Data Management?



This document does not have an outline.

Science & Justice

Volume 55, Issue 3, May 2015, Pages 218

Retraction notice

Retraction notice to A model study into the effects of light and temperature on the degradation of fingerprint constituents [Science and Justice, 54 (2014) 346 - 350]

Belén González Amorós, M. de Puit

Show more

doi:10.1016/j.scijus.2015.04.005

Get rights and content

Refers To Belén González Amorós, M. de Puit
RETRACTED: A model study into the effects of light and temperature on the degradation of fingerprint constituents
Science & Justice, Volume 54, Issue 5, September 2014, Pages 346-350

This article has been retracted: please see Elsevier Policy on Article Withdrawal (http://www.elsevier.com/locate/withdrawalpolicy).

This article has been retracted at the request of the authors. The authors identified a inconsistency in the accepted paper and were unable to reproduce the average values that were used for the graphs and tables in the paper, due to the loss of the raw data. This, in turn, means that the authors cannot fulfil the demands of the Association of Dutch Universities and the Royal Dutch Academy of Science in respect to their ethical and research data standards.

Copyright © 2015 The Chartered Society of Forensic Sciences. Published by Elsevier B.V. All rights reserved.

Recommended articles

Fingermark recovery from riot debris: Bricks and st...
2015, Science & Justice more

An investigation into the detection of latent marks o...
2015, Science & Justice more

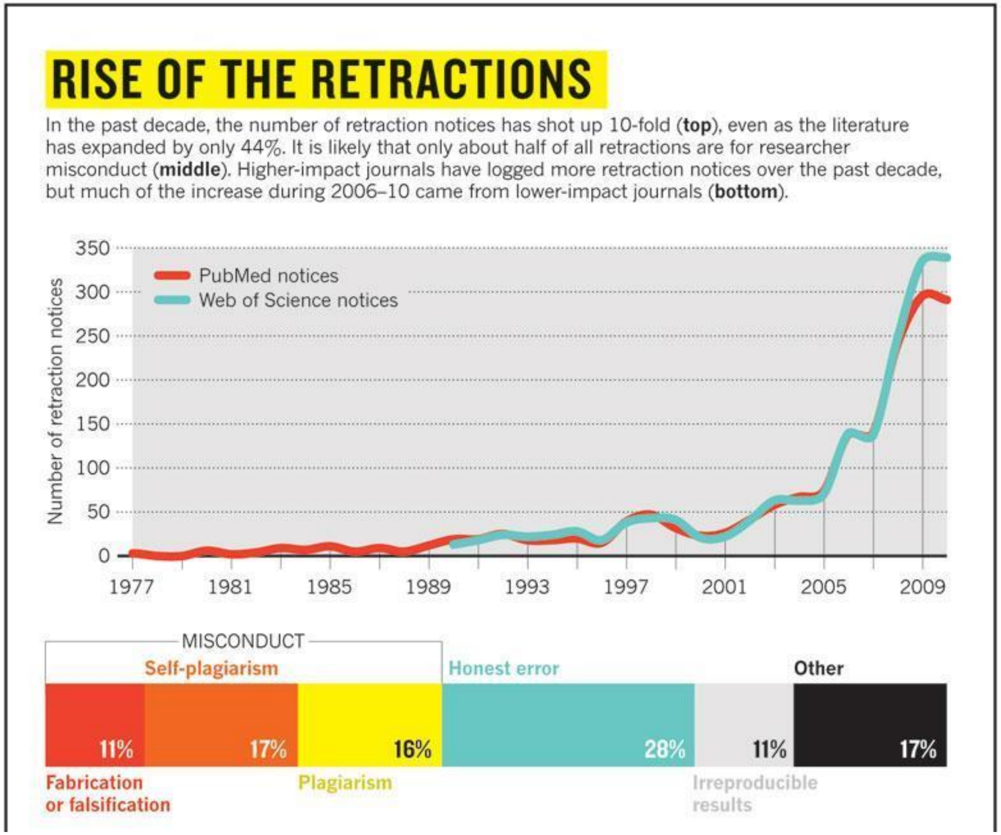
Modelling crime linkage with Bayesian networks
2015, Science & Justice more

View more articles »

Citing articles (0)

The authors identified a inconsistency in the accepted paper and were unable to reproduce … ***due to the loss of the raw data.***

Why Research Data Management?



RISE OF THE RETRACTIONS

In the past decade, the number of retraction notices has shot up 10-fold (**top**), even as the literature has expanded by only 44%. It is likely that only about half of all retractions are for researcher misconduct (**middle**). Higher-impact journals have logged more retraction notices over the past decade, but much of the increase during 2006–10 came from lower-impact journals (**bottom**).

Why Research Data Management?

1. Improve your research

2. Good Scientific Practice

3. Data Sharing with Colleagues

4. Data Publication
   - Required by increasing number of journals
   - Get credit for your data!

5. Enable new kinds of research
   - Feedback loops between empirical and modeling approaches
   - Initiating research questions in completely different fields

Publications are arguments made by authors, and data are the evidence used to support the arguments.

umann, 2014)

HOW MUCH STRONGER WILL YOUR ARGUMENTS BE WHEN ANYONE CAN VERIFY THE EVIDENCE?

The deeper meaning of Research Data Management

Source: cmhughes on pgfplots, CC-BY 2.5

RDM WS GGNB
Data Management Planning

06.09.2017

Why plan your Data Management?

1. Become aware of problems before they arise

- Like you plan your thesis or research project, you should plan your data management
- Identify roles, responsibilities, resources and solutions *before* data are generated

2. Prevent double work and time pressure

- Keep data management problems to a minimum during hot research phases
- Rest assured knowing that your (intermediate) research results are well-managed

3. Requirement by funders

- DFG requires comprehensive description of how data is dealt with
- BMBF asks: ”*Please provide a concrete data utilisation and data management concept as annex*”
- In the rest of the world, especially US and UK, DMPs are mandatory for quite some time already

What is a Data Management Plan?

- A formal document specifying how data is being handled during and after a research project (i.e., across the full data lifecycle)
- A measure to ensure and document how research data can be kept ***FAIR***
- A reference for workflows, procedures, responsibilities regarding data management
- An opportunity to comprehensively address data-related issues in a project
- **NOT** just a static document to be delivered with a project proposal
- **NOT** a checkmark, yes/no or multiple-choice questionnaire
  - Can be based on a template, on guidelines, or completely free from scratch

THE FOCUS IS ON THE PLANNING, NOT ON THE PLAN

Do I *need* a Data Management Plan?

- **No**, not yet, but more and more funders are moving towards requiring one.
- **No**, since you know already all about what can, will and should happen to your data and who will be responsible if something goes wrong with it, and you can explain and justify this to your supervisor and your funder. RIGHT?
- **No**, since you have an IT department or data representative who takes care of everything concerning your data - NOT

➢ **in essence: No, but it’s still a good idea to start creating one**

What does a Data Management Plan look like?

- It’s up to you
- You can find examples and guidelines, e.g. here:
    - https://www.lib.ncsu.edu/data-management/dmp_examples
    - http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples
    - http://www.ands.org.au/working-with-data/data-management/data-management-plans
    - https://www.openaire.eu/opendatapilot-dmp
- Or tools / checklists to create a DMP:
    - http://www.dcc.ac.uk/dmponline
    - https://dmptool.org/
    - http://data.uni-bielefeld.de/de/data-management-plan
    - http://rdmorganiser.github.io/

What should be in a Data Management Plan?

Try to answer the following questions when writing your DMP:

- What data (types, formats, amounts) will be *created?*
- What *policies* (funding, institutional, ethical, and legal) will apply to the data?
- What data management *practices* (backups, access control, preservation and archiving) will be used?
- How are *ownership*, *data access* and protection of *intellectual property* settled and managed?
- How are *data sensitivity issues* addressed and managed?
- How will the data be *described* and possibly *shared* and/or *reused*?
- What *facilities* and *equipment* (hard-disk space, backup server, repository) will be required and used?
- Who will be *responsible* for each of these activities?

➢ **Don't worry if you don't know all the answers yet!**

THE FOCUS IS ON THE PLANNING, NOT ON THE PLAN

Göttingen/eResearch Alliance

RDM @ GGNB
Backup & Storage

06.09.2017

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

SUB

GWDG

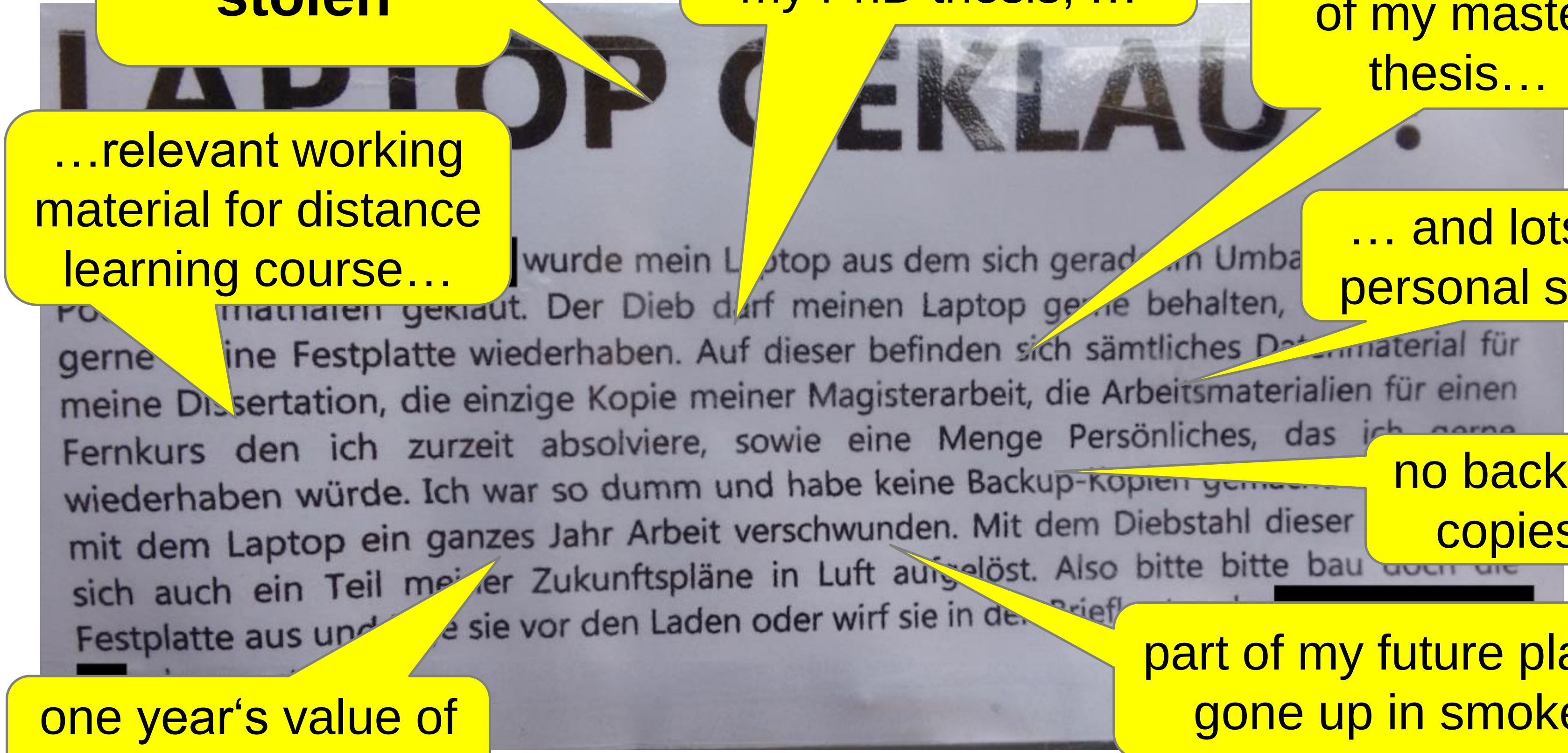
Discussion: Backup

Check for yourself:

- Do you backup your research data? How?
- How often do you do it?
- Have you ever tried to recover a deleted file?
- Can you return to a previous version of a file?
- Who is responsible for Backup and Storage services at your institute, in your research group or project?

Why Backup?

N top Ju tin



LAPTOP GEKLAU

[redacted] wurde mein L ptop aus dem sich gerad n Umba
Po mathalen geklaut. Der Dieb d rf meinen Laptop ge e behalten,
gerne ine Festplatte wiederhaben. Auf dieser befinden sich sämtliches D material für
meine Di sertation, die einzige Kopie meiner Magisterarbeit, die Arbeitsmaterialien für einen
Fernkurs den ich zurzeit absolviere, sowie eine Menge Persönliches, das ich gerne
wiederhaben würde. Ich war so dumm und habe keine Backup-Kopien gemacht
mit dem Laptop ein ganzes Jahr Arbeit verschwunden. Mit dem Diebstahl dieser
sich auch ein Teil me er Zukunftspläne in Luft au elöst. Also bitte bitte bau doch die
Festplatte aus un e sie vor den Laden oder wirf sie in de Briefl [redacted]
[redacted]

Laptop stolen

Contains all data for my PhD thesis, …

… the only copy of my master thesis…

…relevant working material for distance learning course…

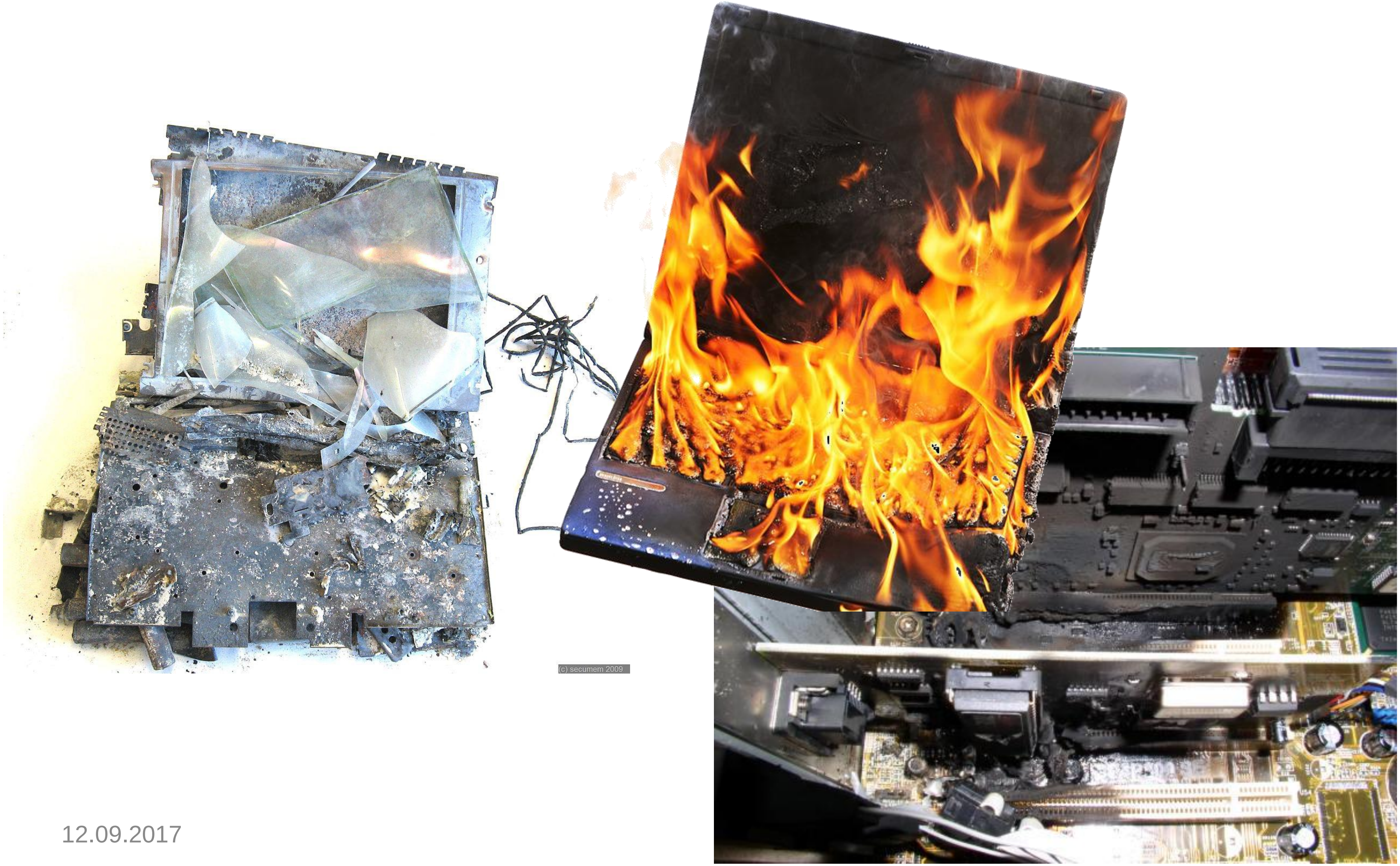
… and lots of personal stuff.

no backup copies

one year‘s value of work disappeared

part of my future plans gone up in smoke

Why Backup?



(c) securmem 2009

Why Backup?

Source: University of Southampton, School of Electronics and Computer Science, 2005

Why Backup?

...because:

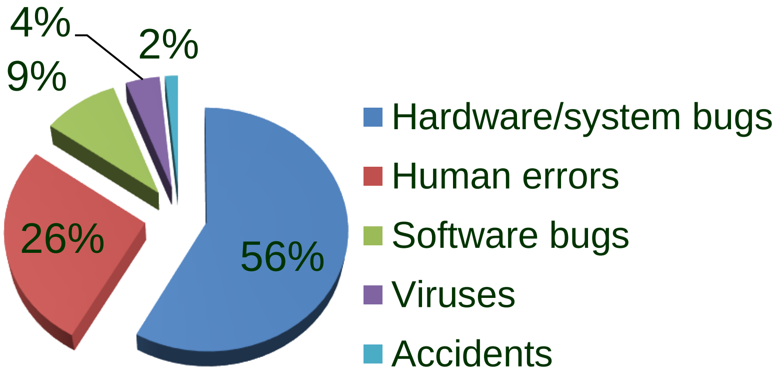
- ○ Don‘t wait until data loss happens to your best friend. It might happen to you first!
- ○ NOBODY is safe from data loss. But EVERYBODY can minimize the risk at a relatively low prize and effort.
- ○ Once it‘s become a habit, you will hardly notice the required effort.

Source: University of Southampton, School of Electronics and Computer Science, 2005

Sources of data loss

- Malware / Theft / Destruction
- Software failures
  - Program errors / bugs / software updates
  - Features
    - (e.g.: Dropbox overwriting on synchronization)
- Hardware failures
  - Bad design / cheap parts / defects
  - Age
  - Dropped laptops / HDDs
  - Liquids (water, coffee, coke)
  - Lightning strikes / electric pulses
- Human errors
  - Accidental deletion
  - Missing knowledge

Source: a man working at home while eating breakfast by Socialeurope via flickr: https://www.flickr.com/photos/socialeurope/4303391587, CC-BY-NC-SA 2.0



Source: Kroll Ontrack, 2007, Robin Harris, http://www.zdnet.com/blog/storage/how-data-gets-lost/167

Further reading: disasters and tales of data loss, statistics on how data gets lost

Sources of data loss

How much of your work can you afford to loose?

- **an accidentally deleted file?**
- **a complete hard drive?**

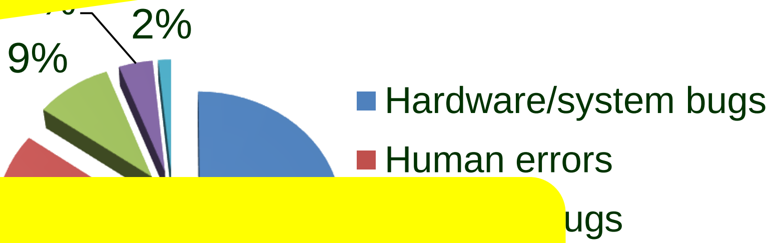
- Mal
- Software failures
  - Program errors / bugs / software upda
  - Features
    - (e.g.: Dropbox overwriting on synchronization)
- Hardware failures
  - Bad design / cheap parts /
  - A
  -
  -
  - kes / electric pulses
- Human errors
  -
  -

breakfast by

03391587,

When can you afford these kinds of loss?

- **at the beginning of your research project?**
- **one month before your thesis submission?**



Let's minimize the risks as far as possible.

Source: Kroll Ontrack, 2007, Robin Harris, http://www.zdnet.com/blog/storage/how-data-gets-lost/167

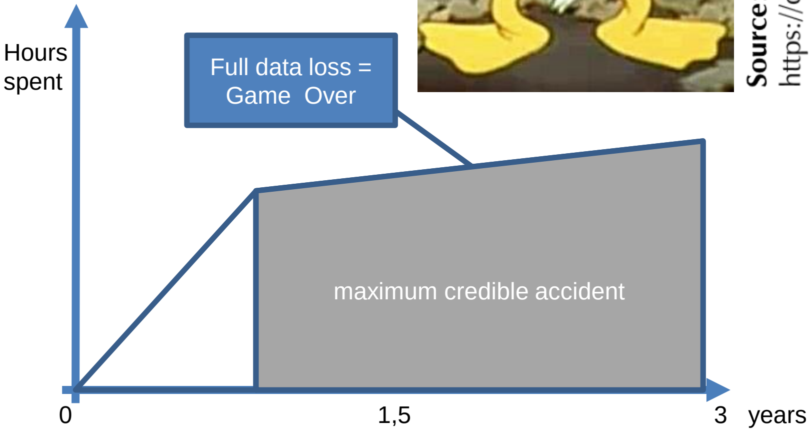
Further reading: disasters and tales of data loss, statistics on how data gets lost

Costs of data loss

Is backing up really worth the effort?

- PhD or postdoc salary costs for employer:
  over € 60.000 / year *
- Estimated costs for losing data of one year's work:
  usually higher

➢ **Besides, you can lose a lot of time … and possibly your nerves**



Source: https://de.wikipedia.org/wiki/Dagobert_Duck

Required investments:

- External hard drives start at € 50,-
- Backup Software is included in most modern operating systems

➢ **When will you start backing up? When will you be required to?**

* DFG staff appropriation rates for 2016: http://www.dfg.de/formulare/60_12/60_12.pdf

Backup: Types, Methods & Media

Backup Types:

- manually vs. automated

Backup Methods:

- full vs. incremental vs. differential

Backup Media:

- USB Sticks: cheap, small (also in storage), *but:* not very reliable
- USB HDD: sufficient storage, affordable, *but:* not shock resistant
- USB SSD: mostly very resilient, *but:* more expensive, often not recoverable
- NAS: safer, more features, *but:* even more expensive, more complex
- Cloud Services (Dropbox, Skydrive, FigShare etc.):
  - File safety is not covered by service terms, several cases of data loss in the past
  - not suitable for personal or sensitive data (since Snowden: no excuses anymore)
  - Internet access can be bottleneck when doing a full restore
- Central Network drives at University institutes / MPIs
  - Mostly rely on professional hardware
  - Should be one central part in your backup strategy
  - *BUT:* *Check their backup policy*
  - *AND:* *Can you access your backup when you need it?*

Backup principles

- Create multiple backups

3-2-1

- 3 copies
- 2 different media
- 1 remote

- Expect human errors (keep older versions)

BACKUP: NOT IN BACKPACK

- Do not use backup drives for sharing files
- Store backups physically separate from your PC / laptop
- Check your backups regularly

ONCE / MONTH

- Practice the worst case and make a full recovery dry-run

ONCE / YEAR

- Discuss the topic with friends to learn their best-practices
- Include your mobile devices in your planning

Backup: Example strategy

- Use an institutional backup solution (e.g. Active Directory)
- Have external harddisks available for backup
  - at your office

  AND
  - at home
- Backup daily to the office harddisk
  - Ideally before you go home
- Backup weekly at home
  - Identify a consistent time slot
- Test both backups at least once a month
  - restore a random number of files or folders and verify their content
- Replace both harddisks after 3-4 years
  - Allow some overlap time

JUST DO IT. REGULARLY.

Backup: Example Strategy (paranoia version)

- One Apple MacBook and one Windows 8 Desktop PC
- 4 USB HDD - 2 for every computer (2 Windows – 2 MacOS)
  - 1 pair located at office (fast access to files from backup)
  - 1 pair located at home (if office burns down, drowns or is robbed)
  - The pairs are swapped every two weeks and stored in lockers
- Google-Calendar Event to get a reminder E-Mail every week
- Automatic backup once a week when attaching the drive to PC
  - Apple OSX: Time machine backup
  - Windows: File Recovery
- Check file system of USB HDD after every backup

➔ Files are stored 3 times per computer

- Replace HDD after getting errors or at least every two years
- Cost: 240 Euro -> 120 Euro per year -> 10 Euro per month

Backup software

Operating system	Integrated Backup SW	Comments
Windows 7	File Recovery	• Needs adjustment to copy other folders than the local libraries • Can create bootable image
Windows 8 & 10	File History	• Only backs up local libraries • Can be adjusted by creating custom libraries and ***excluding*** folders • Cannot create bootable image
Mac OS	Time Machine	• Backs up **everything** except for what is *excluded* • Can use encryption • Can even be used to recover a not-bootable Mac
Ubuntu	Déjà Dup	• Uses encryption, compression • Can use cloud storage

Operating system	Free Third Party Backup SW
Windows	Personal Backup, PureSync, Paragon Backup&Recovery, Robocopy, …
Mac OS	Carbon Copy Cloner, SuperDuper, …
Ubuntu	Rsync, Back in Time

GWDG solutions

Name	Backup	Sharing	Comment
Fileservice / Active Directory	Yes	Maybe	Network drives, e.g. P:, but maybe more Automatic backup
IBM Tivoli Storage Manager (TSM)	Yes	No	Offer to institutes fro centralized backup of all local working machines
CrashPlanProE	Yes	No	Individual Backup solution GWDG license: €26,- per year
CloudShare	Yes	Yes	Free: 10 / 50 GB
ownCloud	Yes	Yes	Free: 10 / 50 GB
CryptShare	No	Yes	Only for MPG
HSM	No	No	For archival of data from closed project
GitLab	No	Yes	Versioning; not for large data amounts

Yes, we store – what for?

	Backup	Archival	Depositing
Storage Purpose	**Ability to restore data** in case of data loss or error propagation	Enable validation by peers through **persistent storage** of data used for research results / publication	Enable verification, citation & reuse of datasets (**data sharing**)
Data Characteristics	Duplication of **current work data** & intermediate work results	Archive format (e.g. zip) containing **all related & relevant data** / files (ideally incl. metadata)	Format specified by repository; **discipline-specific metadata** standards
Process Regularity	Regularly **during work phase** or project runtime	Once for each relevant dataset, usually **at the end of or after work phase**	Once for each selected dataset, **either during or after work phase**
Effort	Depends – e.g.: set up once, verify regularly	Establish predefined procedure with data archive (e.g. data center)	Process documented, sometimes guided by repository

Göttingen/ /eResearch Alliance

RDM @ GGNB
Organization & Documentation

06.09.2017

Photo by candyschwartz
licensed under CC BY-NC-ND 2.0

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

SUB

GWDG

Why organize?

By austinevan on flickr:
http://www.flickr.com/photos/austinevan/1225274637/

Organize your files so that you and others can find and access things when you need them

Source: twechy on flickr :
http://www.flickr.com/photos/twechy/6829994084/

Why organize?

Organize your files so that you and others can find

Source: twechy on flickr :
http://www.flickr.com/photos/twechy/6829994084/

...because:

○ you need to stop working on A and work on B for 2 weeks

○ you get sick & your colleagues need to finish your joint publication

○ your supervisor wants your results from 4 months ago, in 4 minutes

○ you need to eat & sleep from time to time

File naming conventions

To stay organized, you should define:

- A self-describing folder structure or tagging scheme
- What information should be in filenames
- How filenames should be structured
- How to refer to files

… especially when working in a team!

USE WHAT WORKS FOR YOU

AND STICK TO IT !

Self-speaking file name:

`Presentation_GGNB_20170906_V42.pptx`

vs. short file name:

~~`GGNB_final.pptx`~~

Avoid special characters

~~„ “, ‘ ´ ` {} < > : ;
/ \ ? ! $ & ~ *~~

Original file name:

`PICT7639.jpg`

Custom file name:

`20161103_exp01_prb03_001.jpg`

Versioning

`Presentation_GGNB_20170906_V13.pptx`

~~`Presentation_GGNB_20170906_V13final.pptx`~~

~~`Presentation_GGNB_20170906_V13new-final.pptx`~~

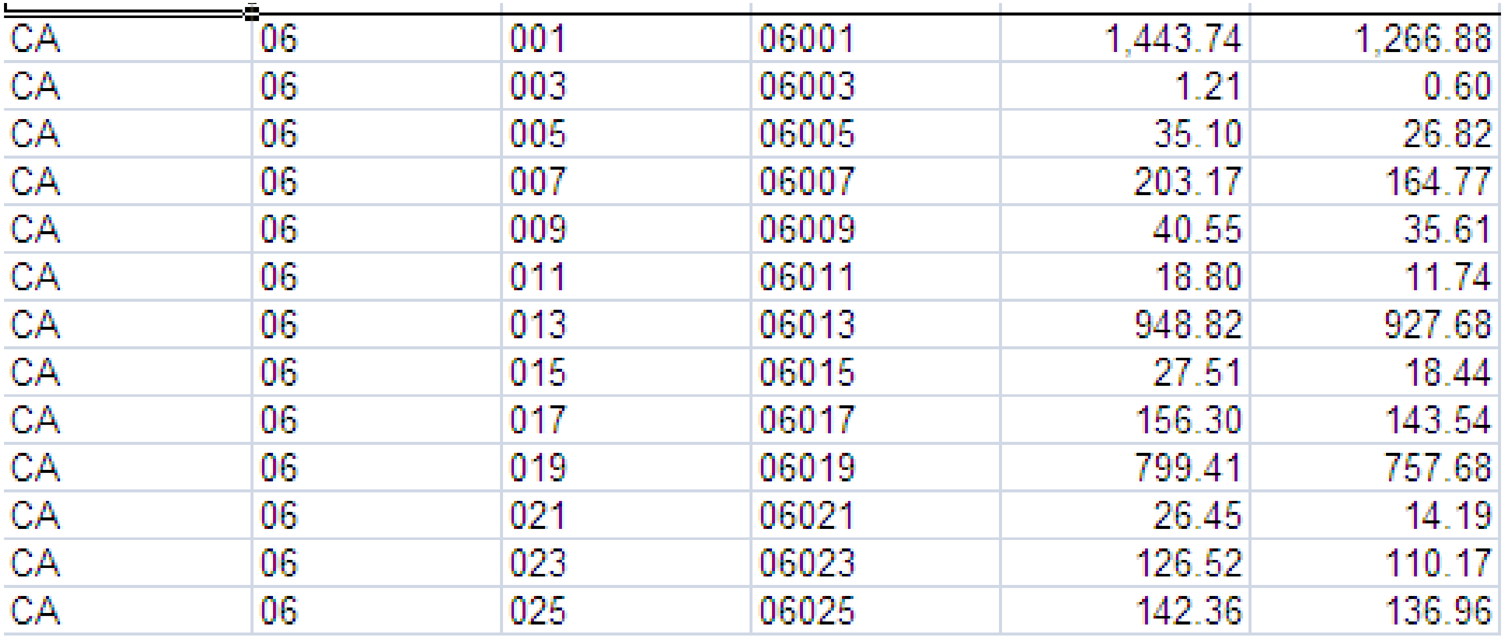
~~`Presentation_GGNB_20170906_V13final-finalv1.pptx`~~

~~`Presentation_GGNB_revised_v01a.pptx`~~

Best practice:

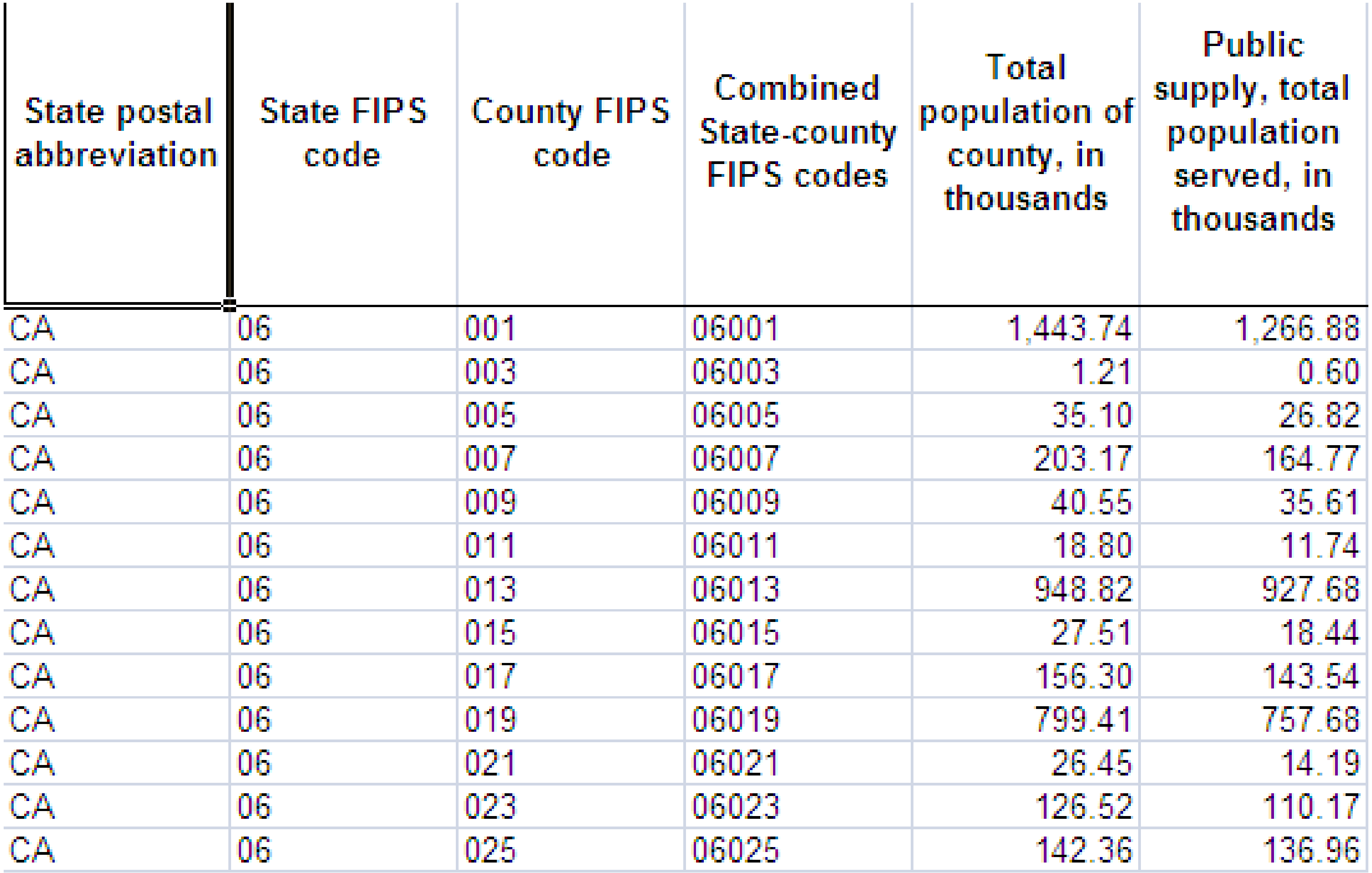
- Save a new version of a file with a **new name** before continuing work
- Use consecutive **version numbers** and eventually **author initials**
  - no „final“ or other unreliable descriptors in filenames
  - Rather **use folders** to mark/sort different purposes and avoid confusion
- If you collaborate on a document, **use “track changes”** if possible

Explain it



CA	06	001	06001	1,443.74	1,266.88
CA	06	003	06003	1.21	0.60
CA	06	005	06005	35.10	26.82
CA	06	007	06007	203.17	164.77
CA	06	009	06009	40.55	35.61
CA	06	011	06011	18.80	11.74
CA	06	013	06013	948.82	927.68
CA	06	015	06015	27.51	18.44
CA	06	017	06017	156.30	143.54
CA	06	019	06019	799.41	757.68
CA	06	021	06021	26.45	14.19
CA	06	023	06023	126.52	110.17
CA	06	025	06025	142.36	136.96

Explain it



State postal abbreviation	State FIPS code	County FIPS code	Combined State-county FIPS codes	Total population of county, in thousands	Public supply, total population served, in thousands
CA	06	001	06001	1,443.74	1,266.88
CA	06	003	06003	1.21	0.60
CA	06	005	06005	35.10	26.82
CA	06	007	06007	203.17	164.77
CA	06	009	06009	40.55	35.61
CA	06	011	06011	18.80	11.74
CA	06	013	06013	948.82	927.68
CA	06	015	06015	27.51	18.44
CA	06	017	06017	156.30	143.54
CA	06	019	06019	799.41	757.68
CA	06	021	06021	26.45	14.19
CA	06	023	06023	126.52	110.17
CA	06	025	06025	142.36	136.96

Image from: https://www.e-education.psu.edu/geog860/print/l2.html
Data courtesy of the U.S. Geological Survey.

Explain your data

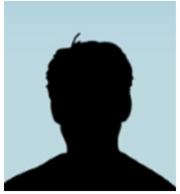
- Why?
- ➢ Make data ***FAIR***: Findable, Accessible, Interoperable, Reusable!
- ➢ Not only for others, but also mainly **for yourself**!
- How?
- ➢ Directly write down which **methods/materials** you used. Write down what fails and what was successfully analysed.
- ➢ Write down **time, place, persons involved** in creation of data.
- ➢ Include title, name of **primary and processed data**.
- ➢ **Add a text file** with this information to each data file/folder ***or:*** maintain and *update* an **overview spreadsheet**
- ➢ **Do not change/erase your original notes** but add more infos chronologically (with date of insertion).

What are metadata?

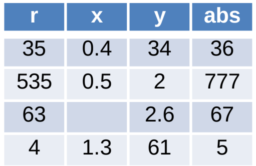
- Many definitions depending on the perspective
- Practical approach: metadata…
  - describe objects in a structured and standardised way
  - can help to select and identify resources
  - can describe how to use them correctly or how to reproduce them
  - can describe anything: literature, a painting, places, a dataset, …
  - can be digitally connected with objects (embedded) or added separately

What to include?

- **Who** created **what, how, when, where** and **why?**



Timo Gnadt
gnadt@sub.uni-goettingen.de



r	x	y	abs
35	0.4	34	36
535	0.5	2	777
63		2.6	67
4	1.3	61	5

Excel spreadsheet with test data for training purposes

Used random number generator to modify original field data

Aug 22 2017

At my office Windows PC

To be used in training workshop

- Include:
  - **Description** of the item
  - **Methodology**
  - **Units** of measurement
  - **References** to related data
  - **Definitions of** jargons, acronyms, code
  - **Technical information** about the file

CAN SOMEBODY ELSE UNDERSTAND YOUR DATA WITHOUT YOU?

“Metadata describe objects in a structured and standardised way…”

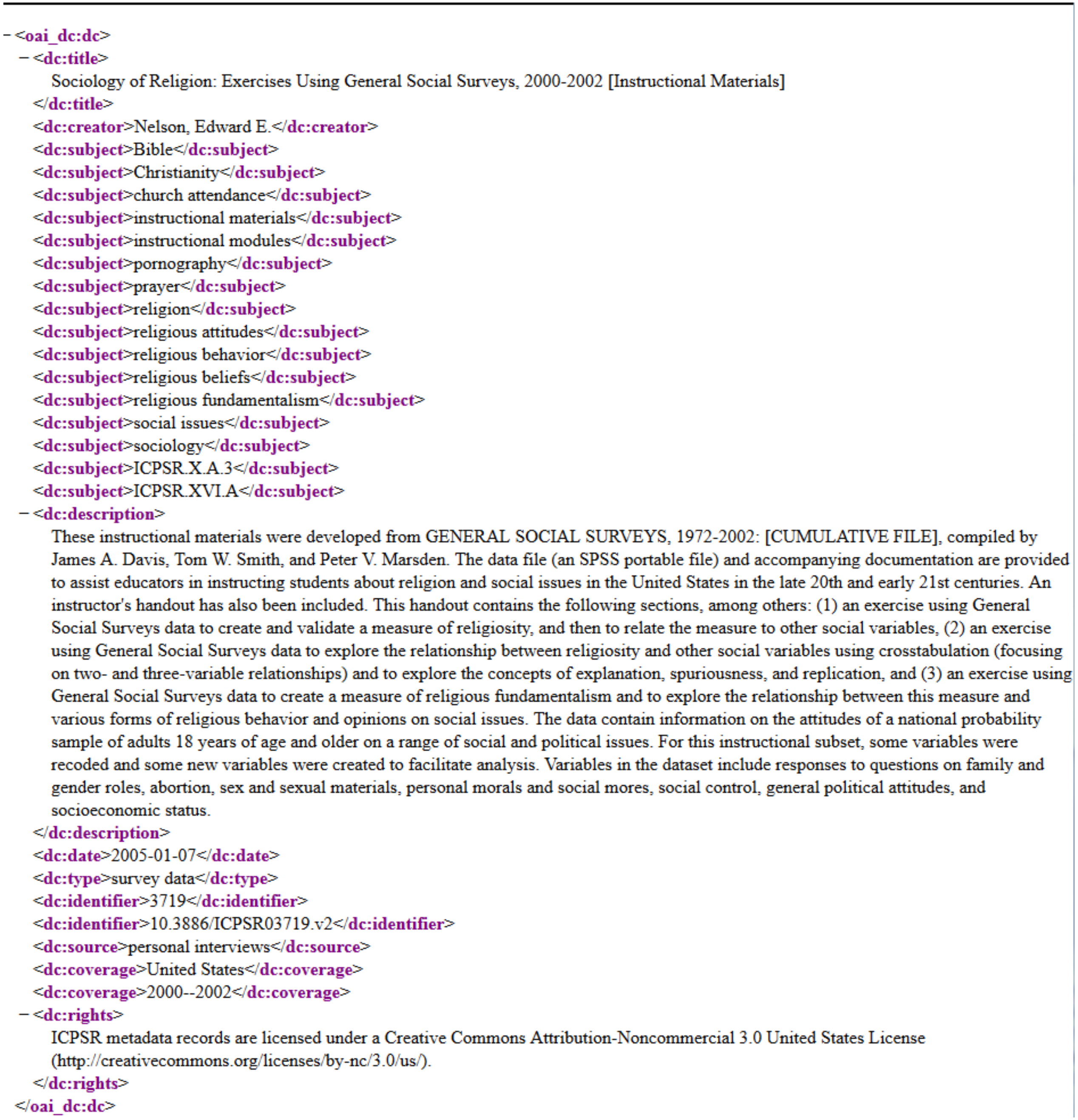
Many existing metadata standards, e.g.:

Dublin Core Metadata Element Set (15 optional elements)

ID:	identifier
Technical Data:	format, type, language
Content:	title, subject, coverage, description
Persons & Permissions:	creator, publisher, contributor, rights
Provenance:	source, relation
Life cycle:	date

Can be extended to 55 elements (DCMI Metadata Terms):

abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, created, dateAccepted, dateCopyrighted, dateSubmitted, educationLevel, extent, hasFormat, hasPart, hasVersion, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, license, mediator, medium, modified, provenance, references, replaces, requires, rightsHolder, spatial, tableOfContents, temporal, valid



```
- <oai_dc:dc>
  - <dc:title>
      Sociology of Religion: Exercises Using General Social Surveys, 2000-2002 [Instructional Materials]
    </dc:title>
    <dc:creator>Nelson, Edward E.</dc:creator>
    <dc:subject>Bible</dc:subject>
    <dc:subject>Christianity</dc:subject>
    <dc:subject>church attendance</dc:subject>
    <dc:subject>instructional materials</dc:subject>
    <dc:subject>instructional modules</dc:subject>
    <dc:subject>pornography</dc:subject>
    <dc:subject>prayer</dc:subject>
    <dc:subject>religion</dc:subject>
    <dc:subject>religious attitudes</dc:subject>
    <dc:subject>religious behavior</dc:subject>
    <dc:subject>religious beliefs</dc:subject>
    <dc:subject>religious fundamentalism</dc:subject>
    <dc:subject>social issues</dc:subject>
    <dc:subject>sociology</dc:subject>
    <dc:subject>ICPSR.X.A.3</dc:subject>
    <dc:subject>ICPSR.XVI.A</dc:subject>
  - <dc:description>
      These instructional materials were developed from GENERAL SOCIAL SURVEYS, 1972-2002: [CUMULATIVE FILE], compiled by
      James A. Davis, Tom W. Smith, and Peter V. Marsden. The data file (an SPSS portable file) and accompanying documentation are provided
      to assist educators in instructing students about religion and social issues in the United States in the late 20th and early 21st centuries. An
      instructor's handout has also been included. This handout contains the following sections, among others: (1) an exercise using General
      Social Surveys data to create and validate a measure of religiosity, and then to relate the measure to other social variables, (2) an exercise
      using General Social Surveys data to explore the relationship between religiosity and other social variables using crosstabulation (focusing
      on two- and three-variable relationships) and to explore the concepts of explanation, spuriousness, and replication, and (3) an exercise using
      General Social Surveys data to create a measure of religious fundamentalism and to explore the relationship between this measure and
      various forms of religious behavior and opinions on social issues. The data contain information on the attitudes of a national probability
      sample of adults 18 years of age and older on a range of social and political issues. For this instructional subset, some variables were
      recoded and some new variables were created to facilitate analysis. Variables in the dataset include responses to questions on family and
      gender roles, abortion, sex and sexual materials, personal morals and social mores, social control, general political attitudes, and
      socioeconomic status.
    </dc:description>
    <dc:date>2005-01-07</dc:date>
    <dc:type>survey data</dc:type>
    <dc:identifier>3719</dc:identifier>
    <dc:identifier>10.3886/ICPSR03719.v2</dc:identifier>
    <dc:source>personal interviews</dc:source>
    <dc:coverage>United States</dc:coverage>
    <dc:coverage>2000--2002</dc:coverage>
  - <dc:rights>
      ICPSR metadata records are licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License
      (http://creativecommons.org/licenses/by-nc/3.0/us/).
    </dc:rights>
  </oai_dc:dc>
```

Some metadata standards for neurosciences

- MIBBI - Minimum Information for Biological and Biomedical Investigations
  - set of guidelines for reporting data derived by relevant methods in biosciences. If followed, it ensures that the data can be easily verified, analysed and clearly interpreted by the wider scientific community.
- MINI - Minimum Information about a Neuroscience Investigation
  - minimum information required to report the use of electrophysiology in a neuroscience study, for submission to the CARMEN system
- ISA-Tab - Investigation/Study/Assay (ISA) tab-delimited (TAB) format
  - general purpose framework with which to collect and communicate complex metadata (i.e. sample characteristics, technologies used, type of measurements made) from 'omics-based' experiments employing a combination of technologies.
- Genome Metadata
  - consists of 61 different metadata fields (attributes), organized into seven categories: Organism Info, Isolate Info, Host Info, Sequence Info, Phenotype Info, Project Info, and Others.

Organization & Documentation: Best practice

- **Plan before you start**
  - Organize your folders & files
  - **D**efine, **D**iscuss and **D**ocument naming conventions
- **Explain your data**
  - Use standards if possible, do not re-invent
  - If standards are too complex or not complex enough then try to customize on the basis of them.
- **Discuss your approach** with your colleagues
- **Be specific and consistent**
  - Don’t alter the past, but document changes in your RDM practice

➢ *Somebody else should be able to* ***find and understand your research data without you –*** *ideally even years later*

Göttingen/ /eResearch Alliance

RDM @ GGNB
Data sharing and legal aspects

06.09.2017

Sander van der Wel: Young toekans sharing food, CC-BY-SA-2.0

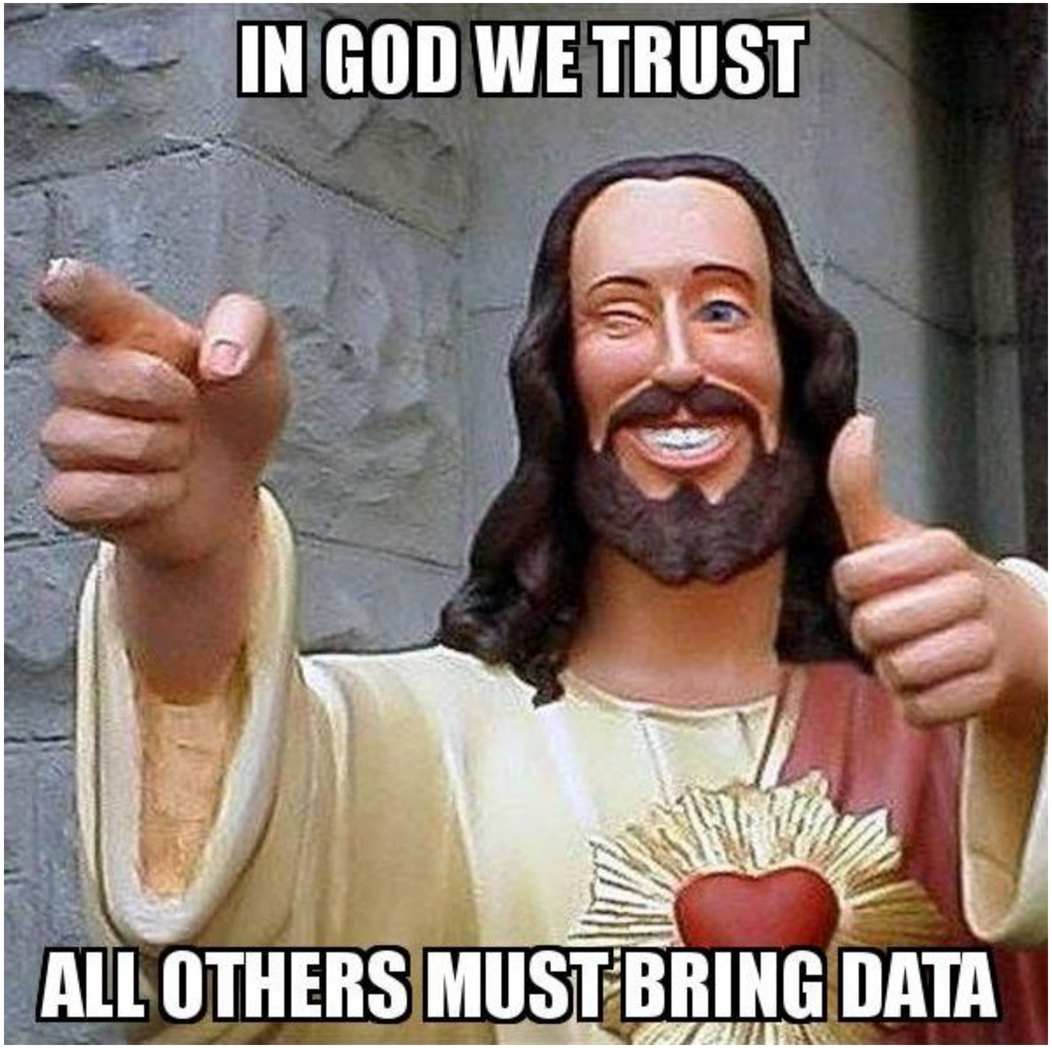
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

SUB

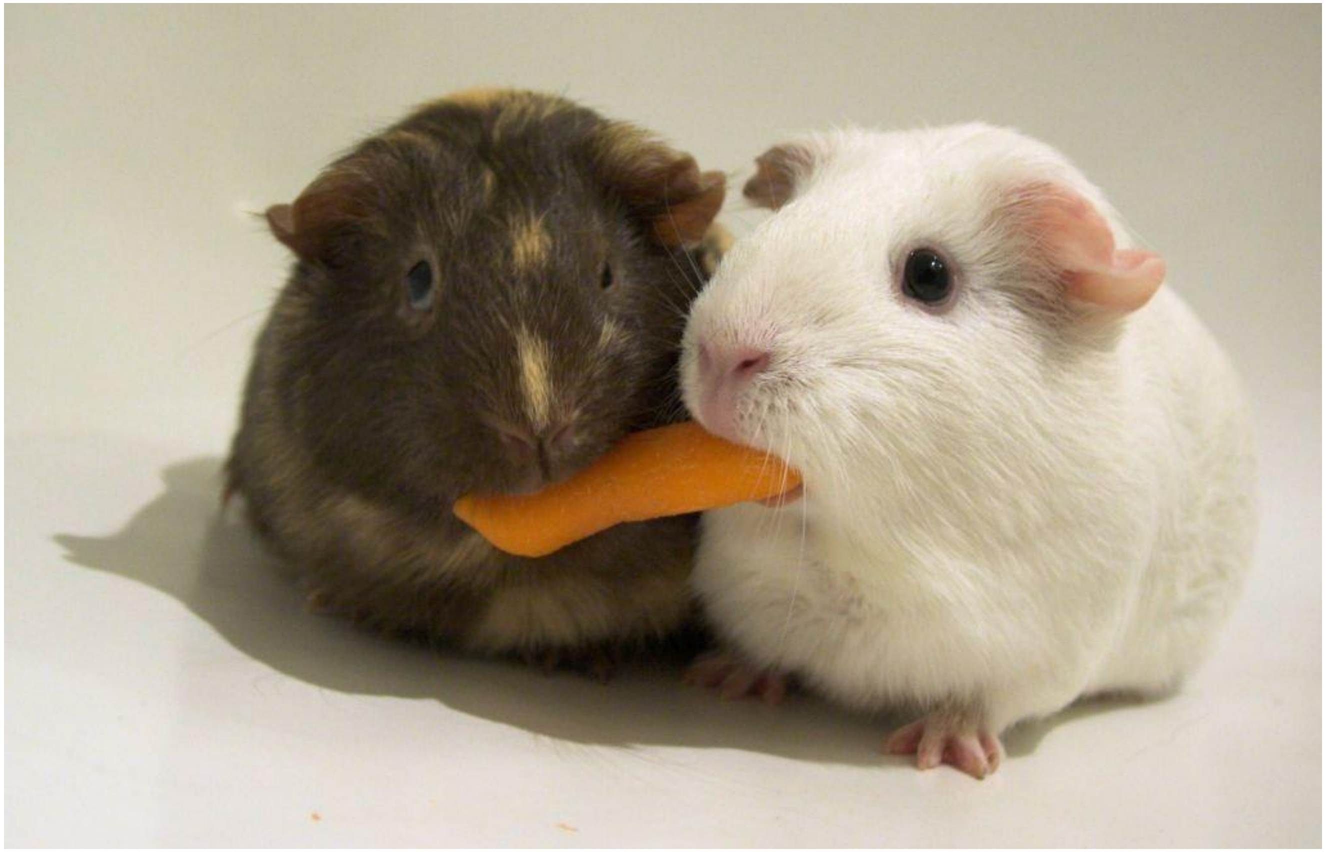
GWDG

Data sharing - motivation



Quote from: William E. Demming (1900-1993)

… but active, open, free sharing?



Source: Sharing by ryancr via flickr
CC-BY-NC 2.0

Why share?

Reputation

- Get credit for high quality research
- Increased understanding of your methods
- Allows work to be verified by others
- Recognition for contribution to research community

Funding

- Making data and/or publications available may be a requirement of your funding body
- It may make your funding proposal more attractive when sharing data is not essential

Why share?

Source: Richard Matthews, flickr: dart (2011) online at: https://commons.wikimedia.org/wiki/File:Darts_in_the_middle_of_a_dartboard.jpg?uselang=de CC-BY 2.0

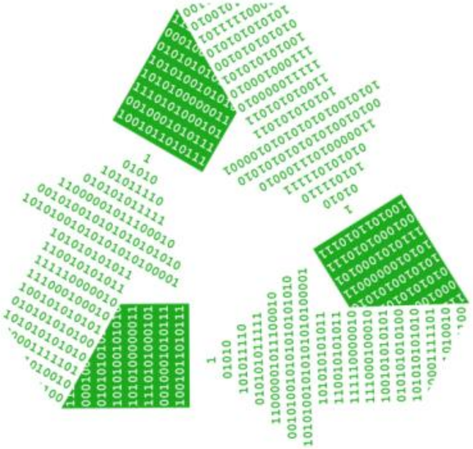
Impact

- Sharing makes your data:
  - easier to find
  - easier to access

- Open data/publications leads to increased citations

Reuse

- Starting point for a complementary study
- Test data for new software and algorithms
- Teaching purposes
- Contexts not currently envisioned
- Useful in completely different fields



Data sharing – concerns

Source: All he does is eat eat eat via flickr
CC-BY 2.0 Jannes Pockele

• Stockpiling for bad times	***Self-use***
• No one likes polishing	***No documentation***
• Dirt behind the scenes	***Work in progress***
• Atmosphere of fear	***Theft and misuse***
• Small fishes & unicorns	***Un-importance***

Value over time?

Embargo!

Do it for yourself!

„Working data set“

Trust law & science

Future is unpredictable

Data sharing - credits?

- Well documented research data

  helps your own (future) research
- Shared data may serve as

  facilitator for cooperation
- Increased accessibility and usability

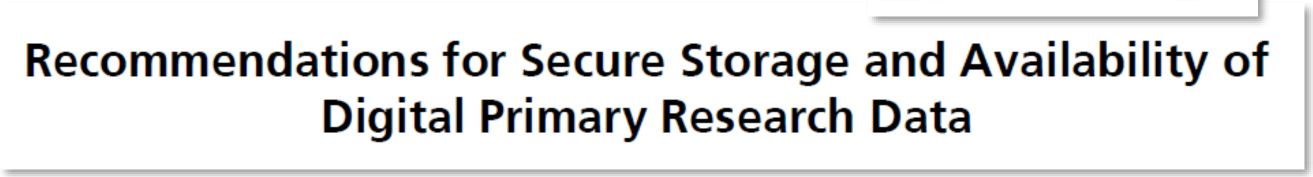
  enable reuse and citations
- Public and open access

  extend the range of your data and research

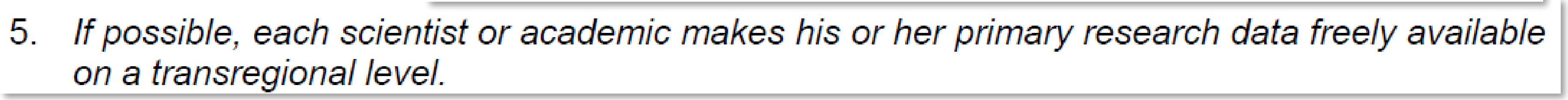
Responsibilities

- Funders

DFG

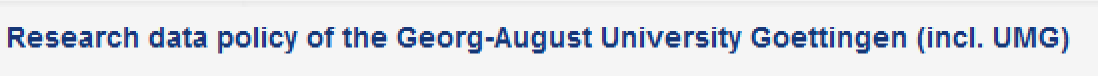


Recommendations for Secure Storage and Availability of Digital Primary Research Data

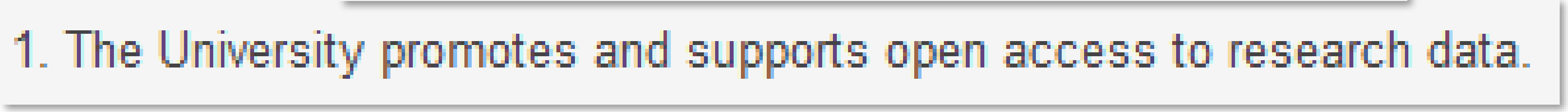


5. *If possible, each scientist or academic makes his or her primary research data freely available on a transregional level.*

- Institutions



Research data policy of the Georg-August University Goettingen (incl. UMG)



1. The University promotes and supports open access to research data.

- Public
  - Results from publicly funded research should be public. If this holds true for publications, why not for research data?
- Science
  - Evolving science

Data sharing – real barriers

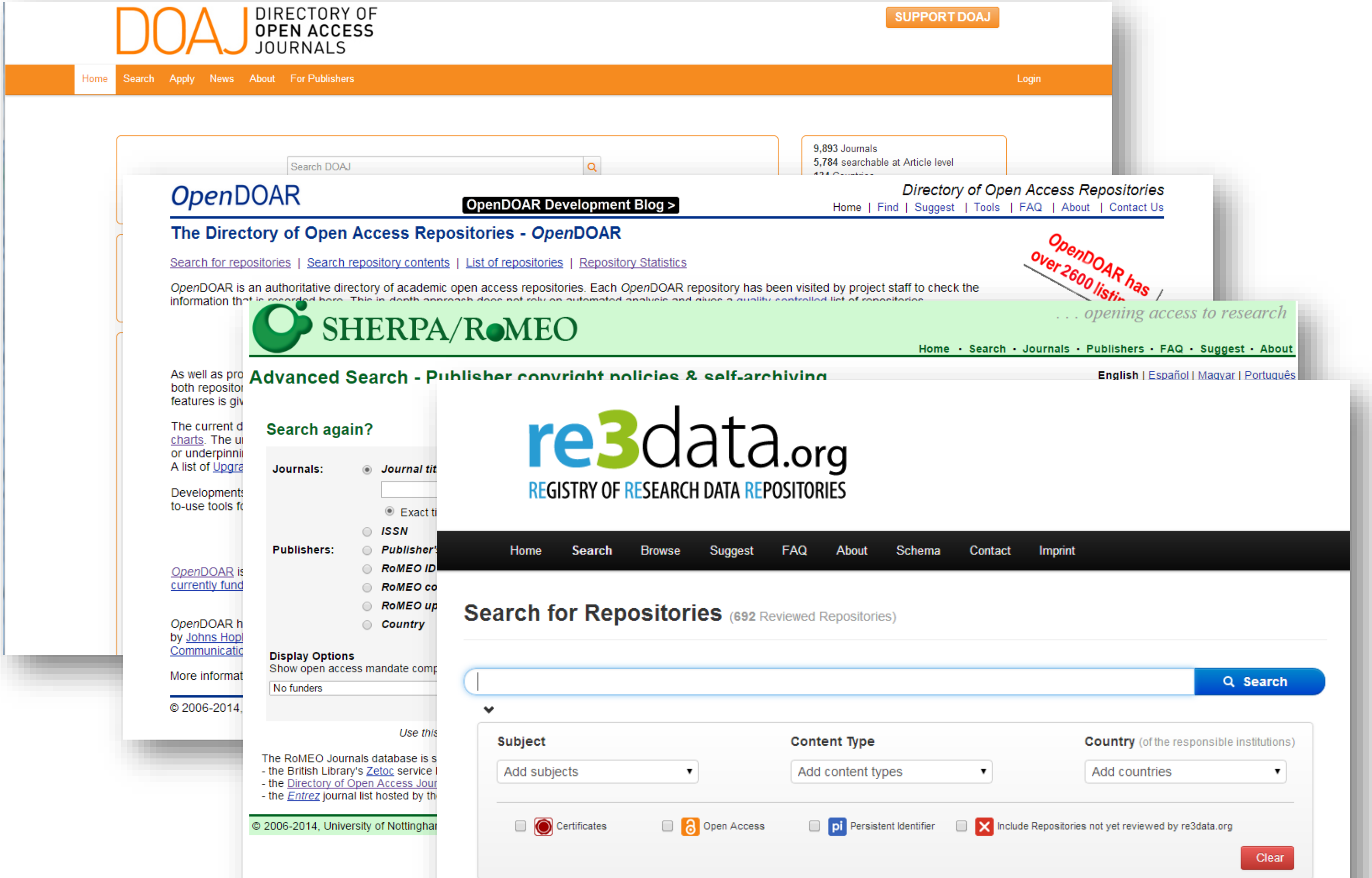
- Place
  - no sharing tradition
  - no repository
  - no expertise
- Funds
  - no money
- Rights
  - no carte blanche

Source: Simatai_Great_Wall by Arian Zwegers on Wikimedia Commons, CC BY SA 2.0

Modes of Sharing

Transfer Way	**Access Mode**	**Use Condition**
peer-to-peer	restricted	none
webspace	on demand	agreement
repository	embargo	licence
	open	

Finding OA journals and repositories



DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

SUPPORT DOAJ

Home | Search | Apply | News | About | For Publishers | Login

Search DOAJ

9,893 Journals
5,784 searchable at Article level

*Open*DOAR

OpenDOAR Development Blog >

Directory of Open Access Repositories

Home | Find | Suggest | Tools | FAQ | About | Contact Us

The Directory of Open Access Repositories - *Open*DOAR

Search for repositories | Search repository contents | List of repositories | Repository Statistics

OpenDOAR has over 2600 listin

*Open*DOAR is an authoritative directory of academic open access repositories. Each *Open*DOAR repository has been visited by project staff to check the information th

As well as pro
both repositor
features is giv

The current d
charts. The u
or underpinni
A list of Upgra

Developments
to-use tools fo

*Open*DOAR is
currently fund

*Open*DOAR h
by Johns Hop
Communicatio

More informat

© 2006-2014,

SHERPA/RoMEO

. . . opening access to research

Home • Search • Journals • Publishers • FAQ • Suggest • About

Advanced Search - Publisher copyright policies & self-archiving

English | Español | Magyar | Português

Search again?

Journals: ● *Journal tit*

● Exact ti

○ *ISSN*

Publishers: ○ *Publisher'*

○ *RoMEO ID*

○ *RoMEO co*

○ *RoMEO up*

○ *Country*

Display Options
Show open access mandate comp
No funders

Use this

The RoMEO Journals database is s
- the British Library's Zetoc service
- the Directory of Open Access Jour
- the *Entrez* journal list hosted by th

© 2006-2014, University of Nottinghar

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Home | Search | Browse | Suggest | FAQ | About | Schema | Contact | Imprint

Search for Repositories (692 Reviewed Repositories)

Search

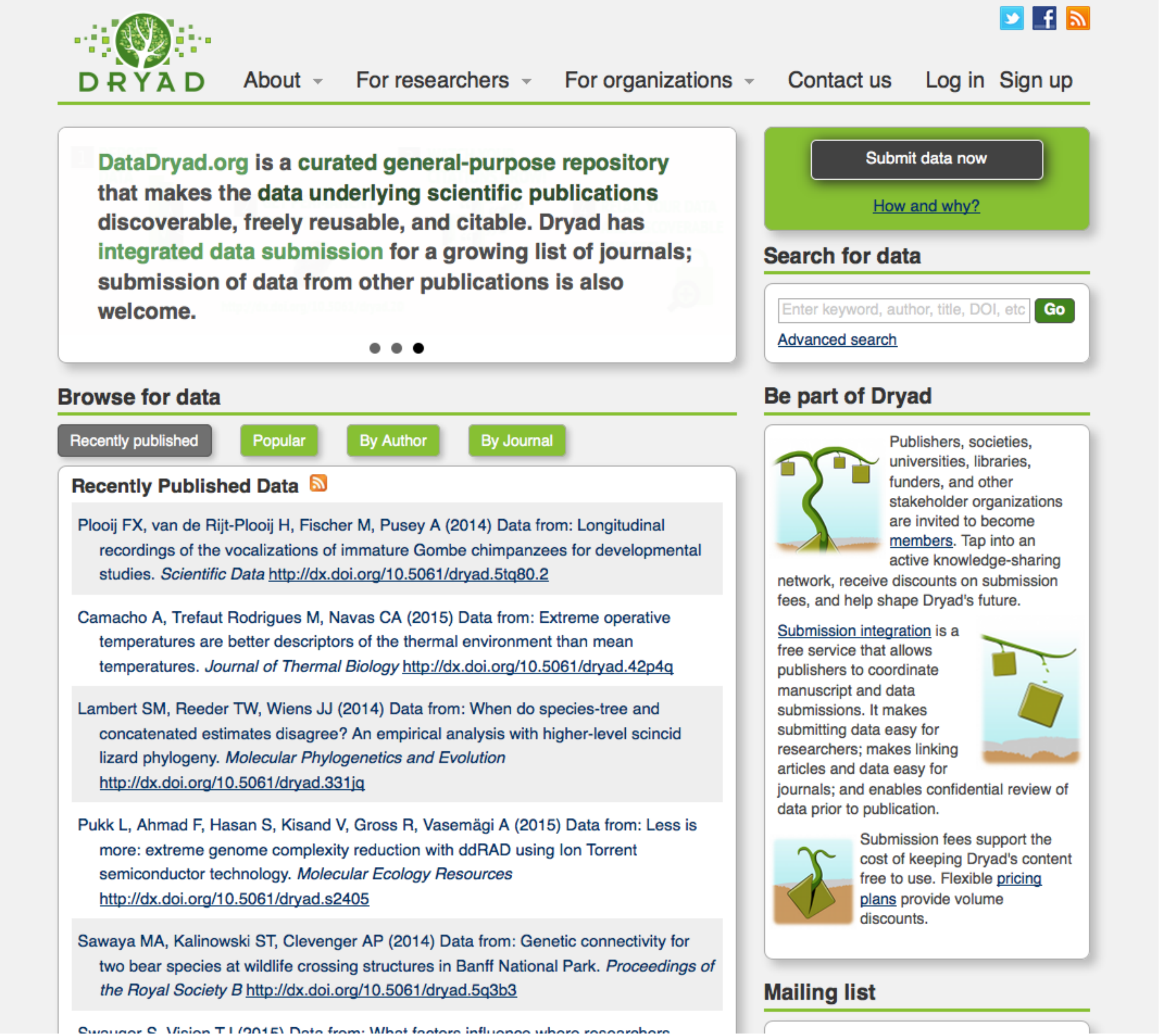
Subject: Add subjects ▾

Content Type: Add content types ▾

Country (of the responsible institutions): Add countries ▾

□ Certificates □ Open Access □ Persistent Identifier □ Include Repositories not yet reviewed by re3data.org

Clear



DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad has integrated data submission for a growing list of journals; submission of data from other publications is also welcome.

Search for data

Browse for data

Recently Published Data

Plooij FX, van de Rijt-Plooij H, Fischer M, Pusey A (2014) Data from: Longitudinal recordings of the vocalizations of immature Gombe chimpanzees for developmental studies. *Scientific Data* http://dx.doi.org/10.5061/dryad.5tq80.2

Camacho A, Trefaut Rodrigues M, Navas CA (2015) Data from: Extreme operative temperatures are better descriptors of the thermal environment than mean temperatures. *Journal of Thermal Biology* http://dx.doi.org/10.5061/dryad.42p4q

Lambert SM, Reeder TW, Wiens JJ (2014) Data from: When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. *Molecular Phylogenetics and Evolution* http://dx.doi.org/10.5061/dryad.331jq

Pukk L, Ahmad F, Hasan S, Kisand V, Gross R, Vasemägi A (2015) Data from: Less is more: extreme genome complexity reduction with ddRAD using Ion Torrent semiconductor technology. *Molecular Ecology Resources* http://dx.doi.org/10.5061/dryad.s2405

Sawaya MA, Kalinowski ST, Clevenger AP (2014) Data from: Genetic connectivity for two bear species at wildlife crossing structures in Banff National Park. *Proceedings of the Royal Society B* http://dx.doi.org/10.5061/dryad.5q3b3

Swauger S, Vision TJ (2015) Data from: What factors influence where researchers

Be part of Dryad

Publishers, societies, universities, libraries, funders, and other stakeholder organizations are invited to become members. Tap into an active knowledge-sharing network, receive discounts on submission fees, and help shape Dryad's future.

Submission integration is a free service that allows publishers to coordinate manuscript and data submissions. It makes submitting data easy for researchers; makes linking articles and data easy for journals; and enables confidential review of data prior to publication.

Submission fees support the cost of keeping Dryad's content free to use. Flexible pricing plans provide volume discounts.

Mailing list

Terms & legal concepts

- Intellectual Property (Geistiges Eigentum)
- Copyright (Urheberrecht)
- Copyright transfer (Nutzungsrecht)
- Fair Use / Fair Dealing (Schranken UrhG)
- Licence
- Copyleft
- Information privacy (Datenschutz)

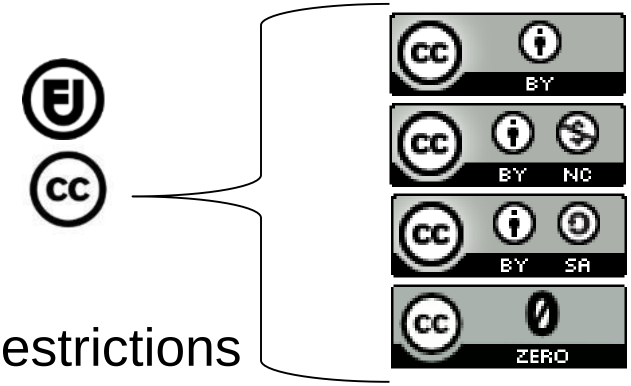
Intellectual property law

Touched rights

- Copyright
- Trade secret
- Patent
- Data privacy

Strategies

- Fair use
- Contracts and licences
- Clarifying terms of use
- Removing or limiting rights restrictions
- Anonymising your data

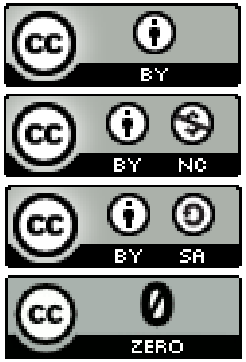


List of rights after: Carroll MW (2015) Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol 13(8): e1002235. doi:10.1371/journal.pbio.1002235

Data on Humans

- Confidential Data
  - are given in confidence
- Personal Data
  - identify a person
- Sensitive Data
  - can compromise a person:
    racial/ethnic origin; political opinions; religious/philosophical beliefs; or other beliefs of a similar nature; trade-union membership; physical/mental health/condition; sexual life

Licences



Proper licensing and attribution: TASL

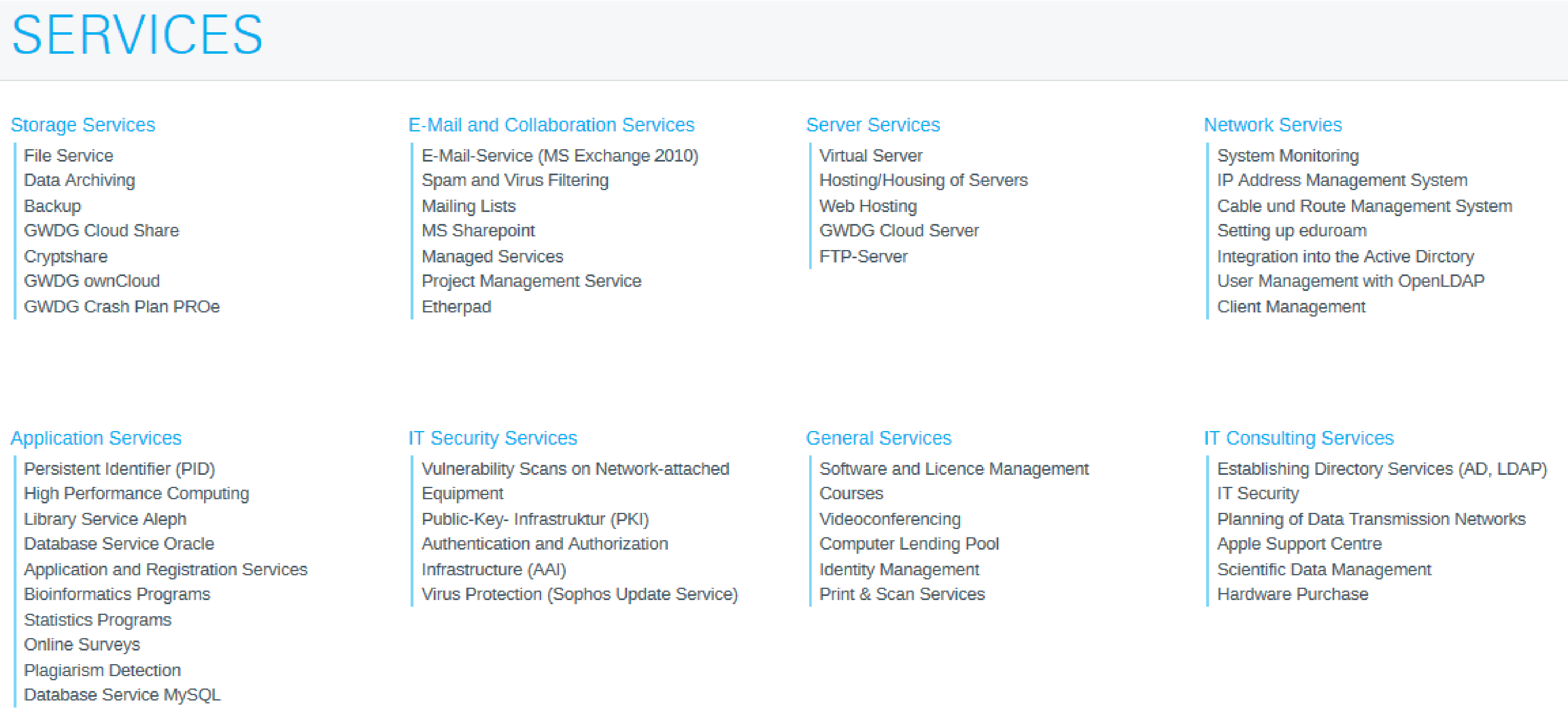
Title, Author, (Source), License (incl. Link)

e.g. “RDM Training for GGNB” by Timo Gnadt, CC-BY 4.0, http://creativecommons.org/licenses/by/4.0/

Some services on Campus

Name	Provided by	Purpose / Comments
Sharepoint	GWDG	Collaboration, Sharing of documents, lists, calendars, ...
Etherpad	GWDG	Collaborative notepad editing
Electronic lab notebook	UMG	(Re-)Organizable, searchable and Backupable research documentation
Biophysical Software	GWDG	analysis and sequencing software like MASCOT (proteome research), Delta2D (2D-Analysis of gel electrophoresis), GeneiousPro (sequential analysis) or for Next Generation Sequencing
Open Access Publication Fund	SUB	complete coverage for up to €2.000,- for publication in OA journal
Videoconferencing	GWDG via DFN	including option to join via phone call

GWDG services



SERVICES

Storage Services

- File Service
- Data Archiving
- Backup
- GWDG Cloud Share
- Cryptshare
- GWDG ownCloud
- GWDG Crash Plan PROe

E-Mail and Collaboration Services

- E-Mail-Service (MS Exchange 2010)
- Spam and Virus Filtering
- Mailing Lists
- MS Sharepoint
- Managed Services
- Project Management Service
- Etherpad

Server Services

- Virtual Server
- Hosting/Housing of Servers
- Web Hosting
- GWDG Cloud Server
- FTP-Server

Network Servies

- System Monitoring
- IP Address Management System
- Cable und Route Management System
- Setting up eduroam
- Integration into the Active Dirctory
- User Management with OpenLDAP
- Client Management

Application Services

- Persistent Identifier (PID)
- High Performance Computing
- Library Service Aleph
- Database Service Oracle
- Application and Registration Services
- Bioinformatics Programs
- Statistics Programs
- Online Surveys
- Plagiarism Detection
- Database Service MySQL

IT Security Services

- Vulnerability Scans on Network-attached Equipment
- Public-Key- Infrastruktur (PKI)
- Authentication and Authorization Infrastructure (AAI)
- Virus Protection (Sophos Update Service)

General Services

- Software and Licence Management
- Courses
- Videoconferencing
- Computer Lending Pool
- Identity Management
- Print & Scan Services

IT Consulting Services

- Establishing Directory Services (AD, LDAP)
- IT Security
- Planning of Data Transmission Networks
- Apple Support Centre
- Scientific Data Management
- Hardware Purchase

https://www.gwdg.de/services

Wrap up: Best Practices

- Plan your RDM before you start
- Discuss your approach
- Backup your data
- Explain your data
- Share your data

Thank you!

Questions?

CONTACT:

info@eresearch.uni-goettingen.de

www.eresearch.uni-goettingen.de