Göttingen/eResearch Alliance

Research Data Management

or: How to love your data

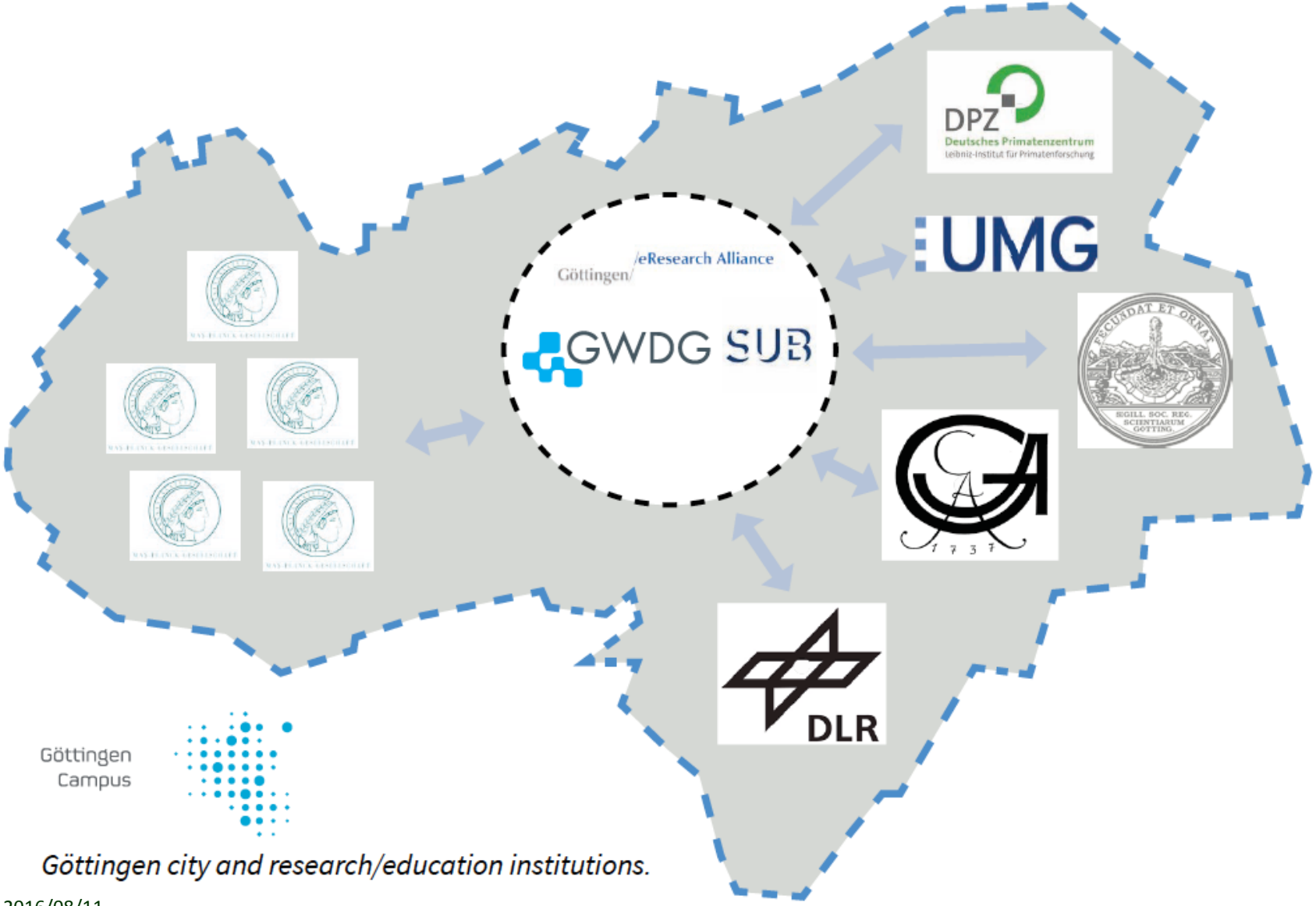
Jessika Rücknagel
ruecknagel@sub.uni-goettingen.de
Timo Gnadt
gnadt@sub.uni-goettingen.de

11/08/2016, Göttingen

Overview

- Göttingen eResearch Alliance
- Research *!* Data *…* Management *?*
  - Backup
  - File Organization
  - Services on Campus

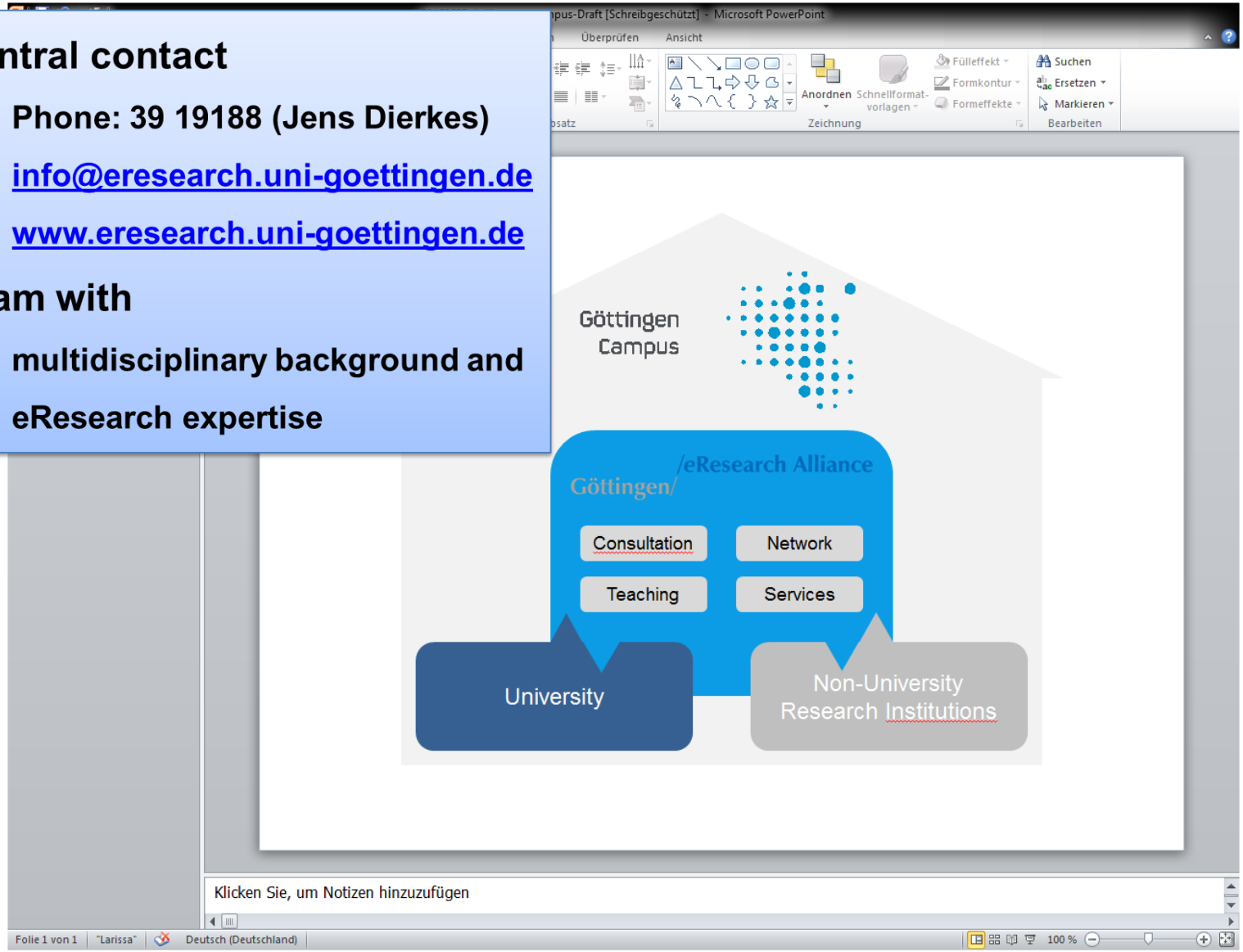
Göttingen eResearch Alliance



Göttingen city and research/education institutions.

eRA Services

- **Central contact**
  - **Phone: 39 19188 (Jens Dierkes)**
  - **info@eresearch.uni-goettingen.de**
  - **www.eresearch.uni-goettingen.de**
- **Team with**
  - **multidisciplinary background and**
  - **eResearch expertise**

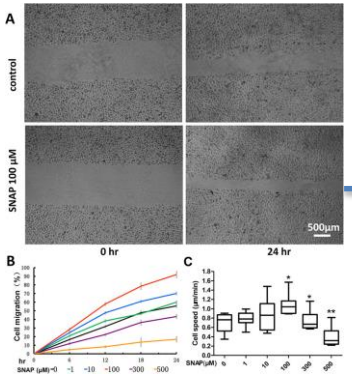


What is Research Data?

Any information you use in your research:

Statistics, interviews, simulations, measurement data from experiments, observational data from instruments, field sample analysis results, text with semantic annotations, 3D scans...

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



Source: Zhan, R. et al. Nitric oxide promotes epidermal stem cell migration via cGMP-Rho GTPase signalling. Sci. Rep. 6, 30687; doi: 10.1038/srep30687 (2016)

research object → research data → result / publication

Research data – a valuable investment

Source: European Space Agency: Rosetta and Philae at comet, on flickr.

Rosetta & Philae

Duration:

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

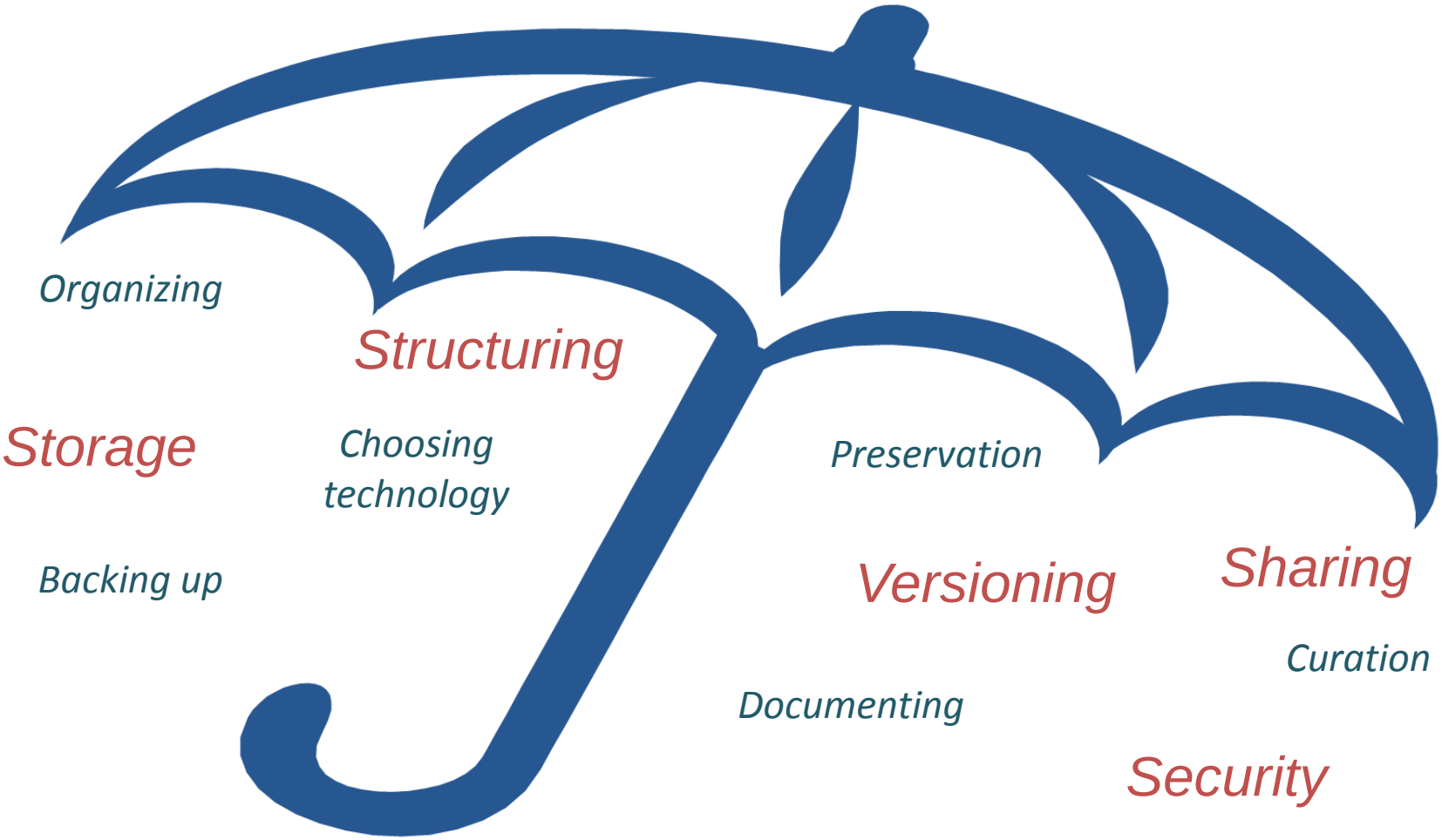
Costs:

- over € 1.000.000.000

Outcome:

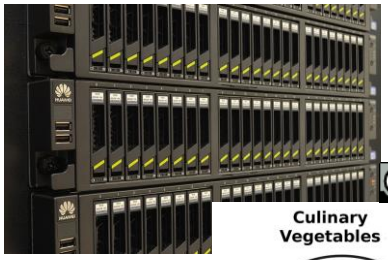
- some cool photos
- lots of data
- *a radically new theory on the origin of the universe?*

What is research data management?

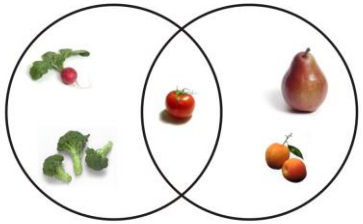


What is Research Data Management?

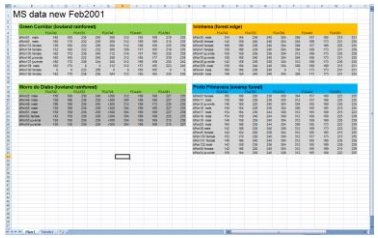
- Backup and Storage

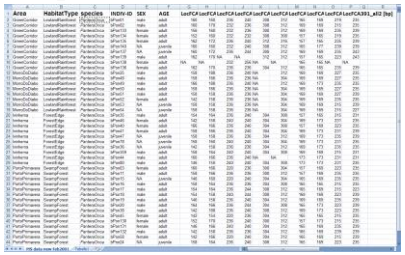


- Metadata and Documentation



- Data Quality





- File Names, Identifier and Versions

doi:10.10.1038/nphys1170

Thesis_final_v13b_revised.docx

- Ethics, Rights and Licenses

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?

Source: RSCD Protein Data Bank; http://www.rcsb.org/pdb/

Why Research Data Management?

NMR Ensemble

View in 3D JSmol or PV (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop Kiosk Viewer

Protein Symmetry: C1 (View in 3D)

Protein Stoichiometry: A

Macromolecule Content

- Unique protein chains: 1
- Unique nucleic acid chains: 1

Display Files Download Files

2N3O

Structure of PTB RRM1(41-163) bound to an RNA stemloop containing a structured loop derived from viral internal ribosomal entry site RNA

DOI: 10.2210/pdb2n3o/pdb **BMRB**: 25652

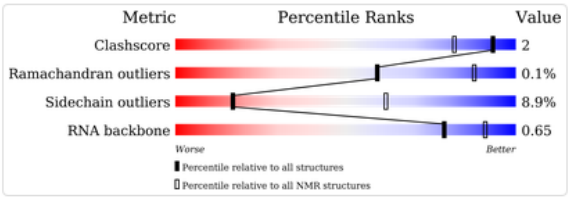
Classification: RNA BINDING PROTEIN / RNA
Deposited: 2015-06-08 **Released**: 2016-08-10
Deposition author(s): Maris, C., Jayne, S.F., Damberger, F.F., Ravindranathan, S., Allain, F.H.-T.
Organism: Homo sapiens | synthetic construct
Expression System: Escherichia coli

Structural Biology Knowledgebase: 2N3O SBKB.org

Experimental Data Snapshot

Method: SOLUTION NMR
Conformers Calculated: 250
Conformers Submitted: 20
Selection Criteria: Structures with the Least Restraint Violations

wwPDB Validation Full Report



Literature Download Primary Citation

C-terminal helix folding upon pyrimidine-rich hairpin binding to PTB RRM1. Implications for PTB function in Encephalomyocarditis virus IRES activity.

Maris, C., Jayne, S.F., Damberger, F.F., Ravindranathan, S., Allain, F.H.-T.

To Be Published

Source: RSCD Protein Data Bank; http://www.rcsb.org/pdb/

Why Research Data Management?

ScienceDirect Journals Books Sign in Help

Download PDF Export Search ScienceDirect Advanced search

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.

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

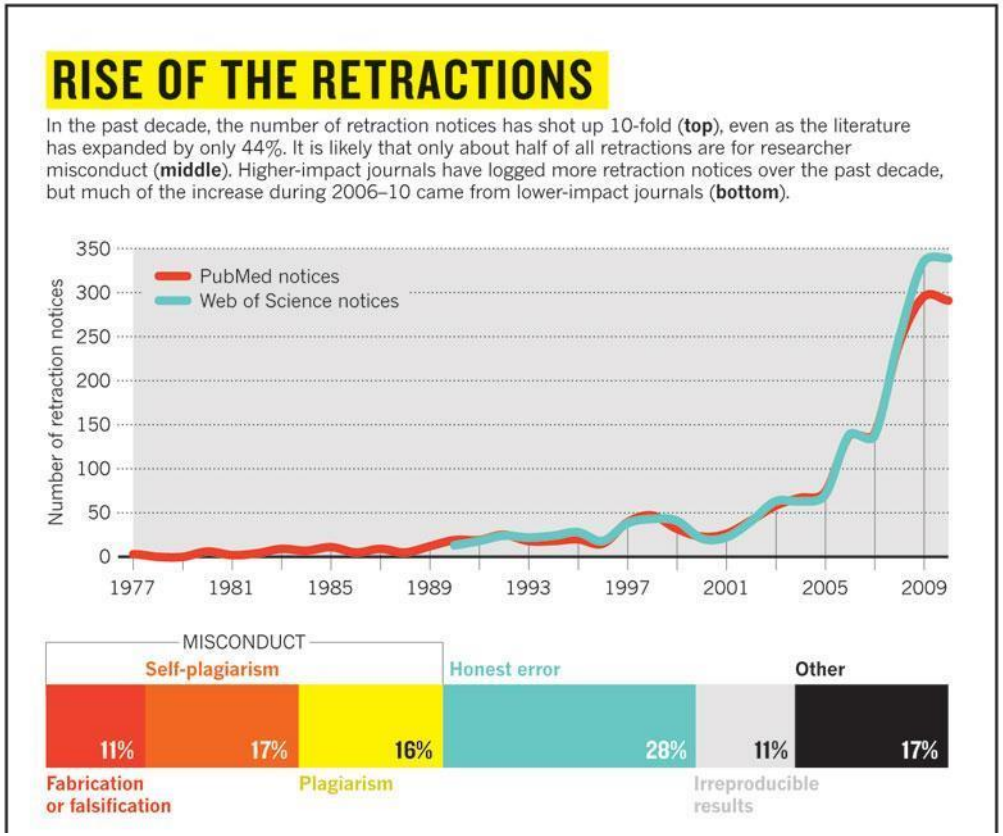
View more articles »

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.***

Source: Science Direct; http://www.sciencedirect.com/science/article/pii/S1355030614000537

Why Research Data Management?



Source: Science Publishing: The trouble with retractions; http://www.nature.com/news/2011/111005/full/478026a.html

Why Research Data Management?

1. Improve your research

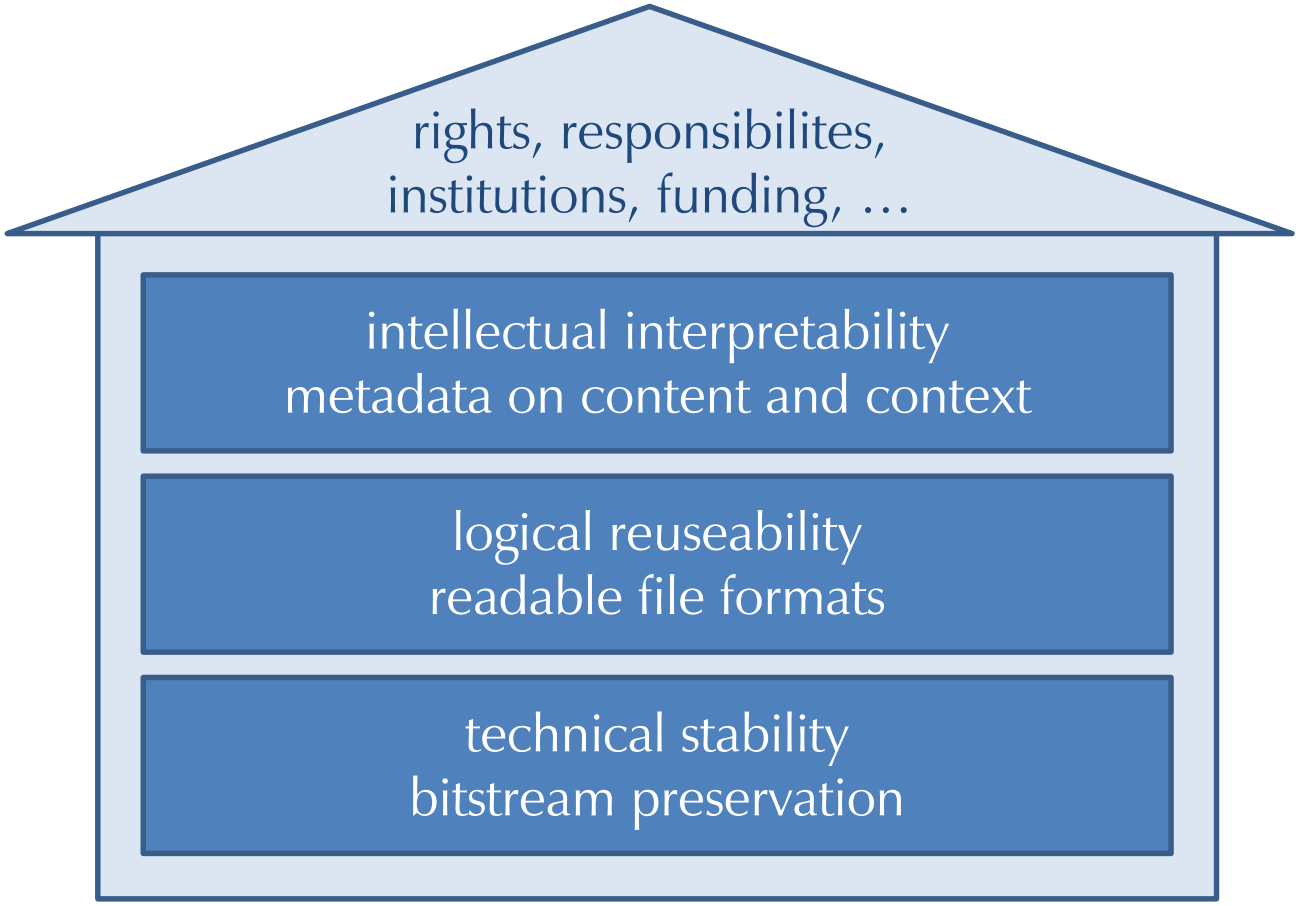
2. Good Scientific Practice

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

Levels of data preservation

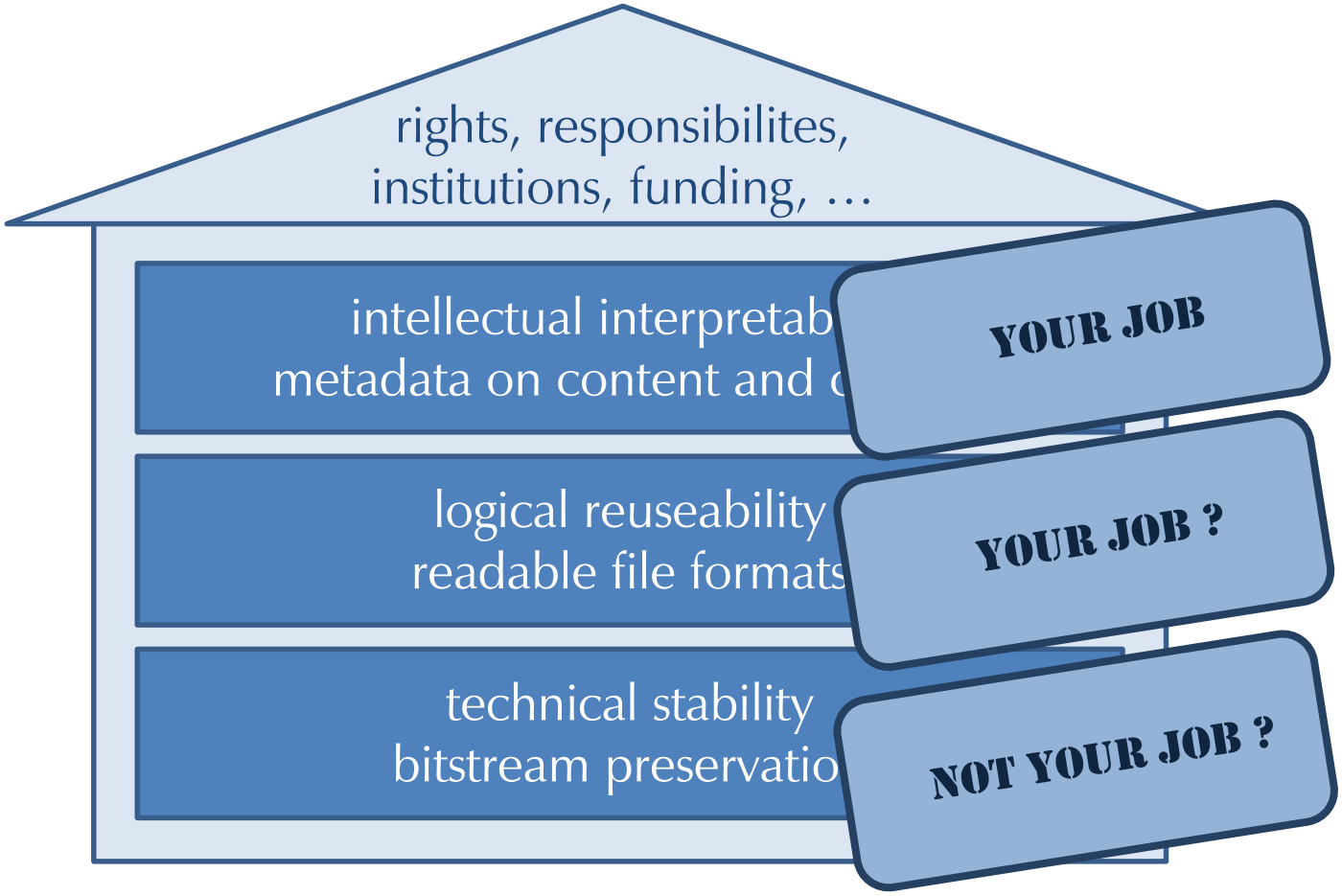


Data preservation motivation

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

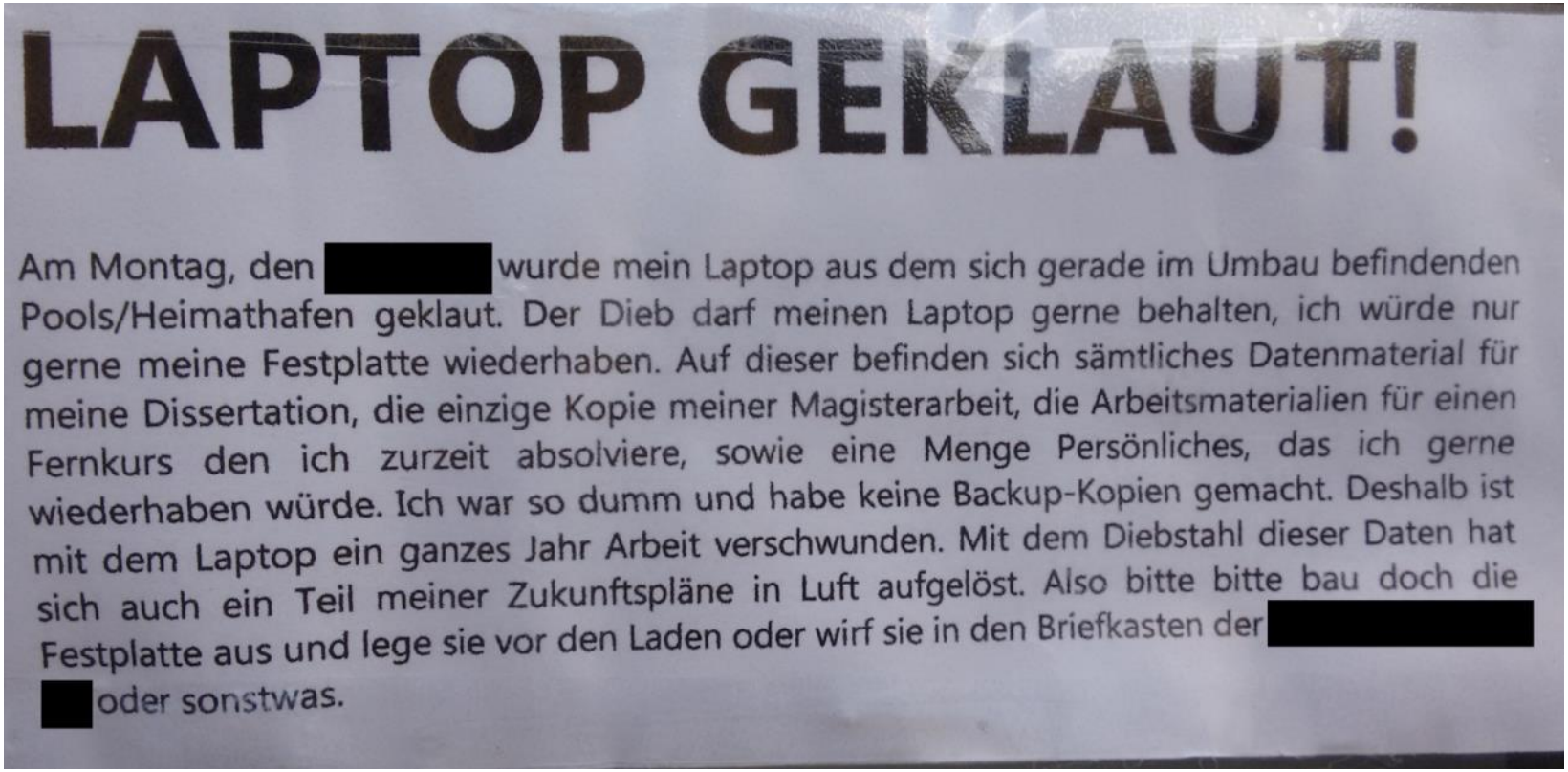


The deeper meaning of Research Data Management

Source: cmhughes on pgfplots, CC-BY 2.5

Why Backup?

Notice at bus stop Jüdenstrasse/Göttingen

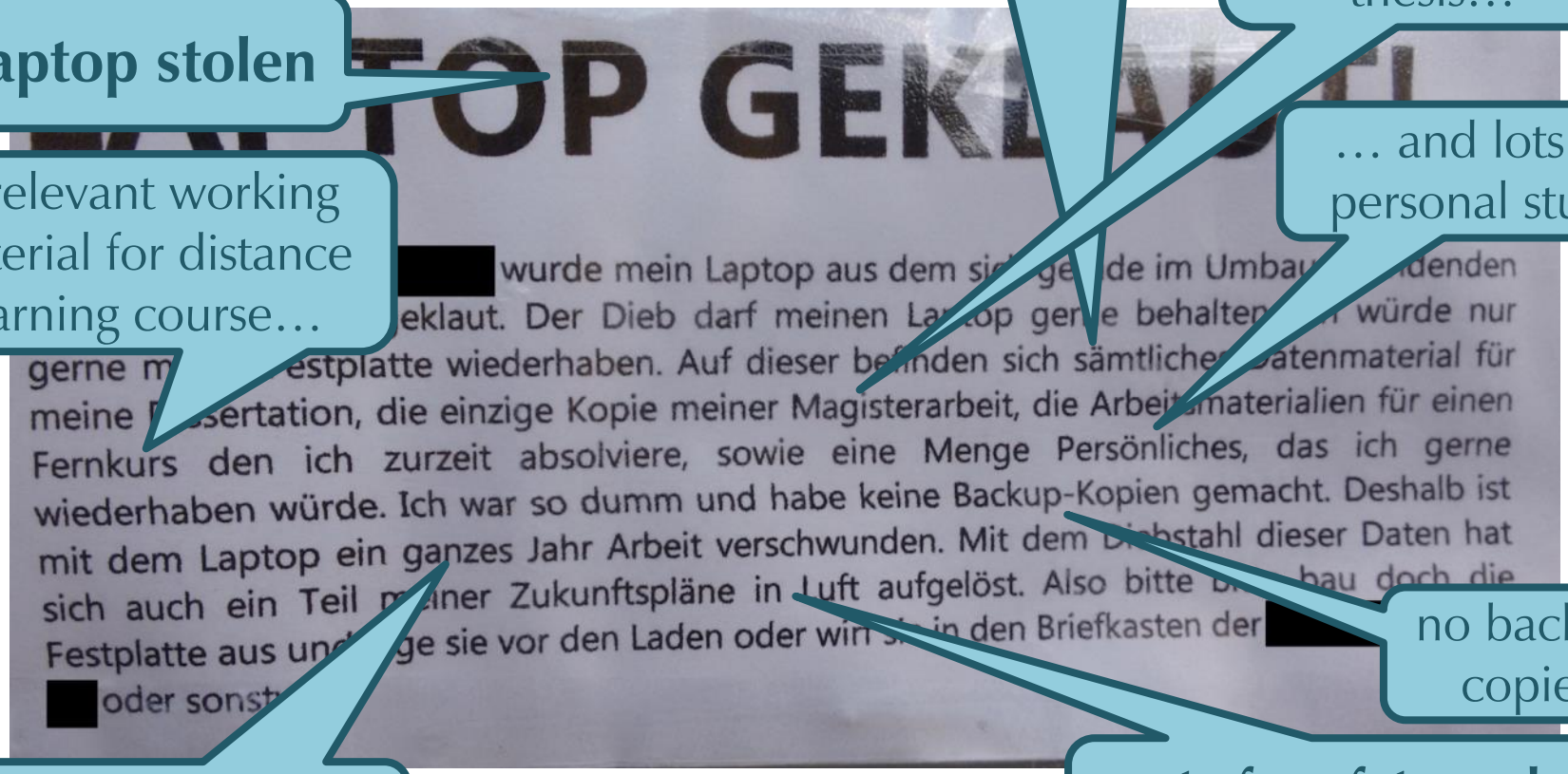


LAPTOP GEKLAUT!

Am Montag, den [redacted] wurde mein Laptop aus dem sich gerade im Umbau befindenden Pools/Heimathafen geklaut. Der Dieb darf meinen Laptop gerne behalten, ich würde nur gerne meine Festplatte wiederhaben. Auf dieser befinden sich sämtliches Datenmaterial für meine Dissertation, 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. Deshalb ist mit dem Laptop ein ganzes Jahr Arbeit verschwunden. Mit dem Diebstahl dieser Daten hat sich auch ein Teil meiner Zukunftspläne in Luft aufgelöst. Also bitte bitte bau doch die Festplatte aus und lege sie vor den Laden oder wirf sie in den Briefkasten der [redacted] [redacted] oder sonstwas.

Why Backup?

Notice at bus stop Jüdenstrasse/Gött[illegible]n



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?

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

Why Backup?

...because:

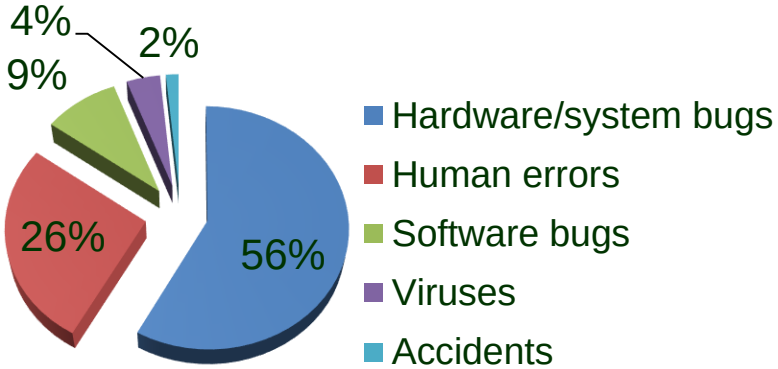
- Don't wait until data loss happens to your best friend. It might happen to you first!
- NOBODY is safe from data loss. But EVERYBODY can minimize the risk at a relatively low prize and effort.
- Once it's become a habit, you will hardly notice the required effort.

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

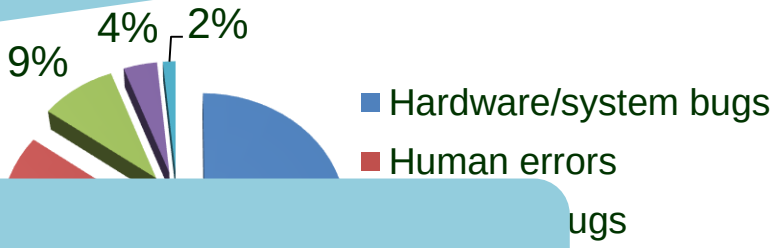
Sources of data loss

- Mal
- Soft
  - Program errors / bugs
  - Features
    (e.g.: Dropbox overwriting on synchronization)
- Hardware failures
  - Bad design /
  - A
  - D
  - Liquids (water, coffee, coke)
  - L
- Hu
  - Accidental deletion
  - Missing knowledge

How much of your work can you afford to loose?
- **an accidentally deleted file?**
- **a complete hard drive?**

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

Backup: Types, Methods & Media

How fast should data be recovered?
How much data can you afford to lose?

Backup Types:

- manually vs automated

Backup Methods:

- full vs. incremental – differential

Backup Media:

- USB HDD: fast, cheap, *but:* not shock resistant
- USB Sticks: cheap, small (also in storage), *but:* not very reliable
- USB SSD: mostly very resilient, *but:* expensive, often not recoverable
- NAS: safer, more features, *but:* 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 it 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)
- Do not use backup drives for sharing files

BACKUP: NOT IN BACKPACK

- 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 principles

- Create multiple backups
- Expect human errors (keep older versions)
- Do not use backup drives for sharing files
- Store backups physically separate from your PC / laptop
- Check your backups regularly
- Practice the worst case and make a full recovery dry-run
- Discuss the topic with friends to learn their best-practices
- Include your mobile devices in your planning

3-2-1

- **3 copies**
- **2 different media**
- **1 remote**

BACKUP: NOT IN BACKPACK

ONCE / MONTH

JUST DO IT. REGULARLY.

ONCE / YEAR

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 2-3 years
  - Allow some overlap time

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

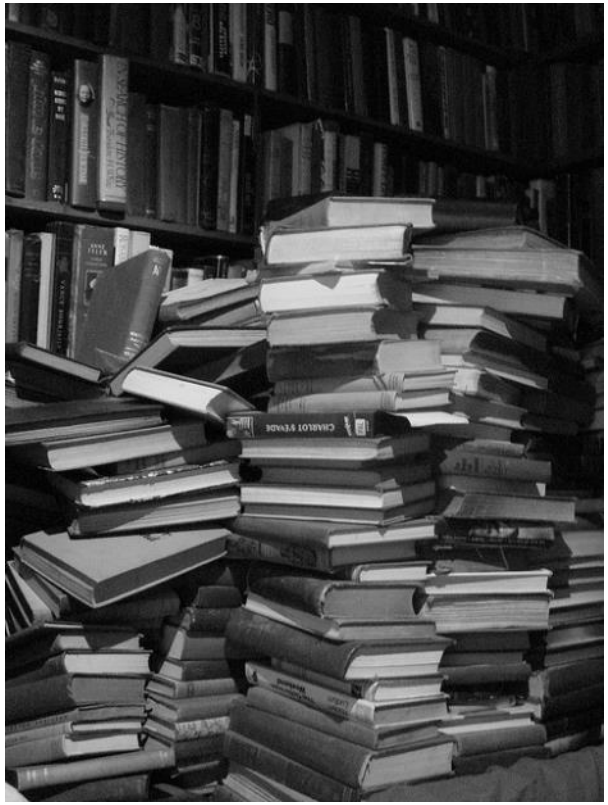
Operating system	Free Third Party Backup SW
Windows	Personal Backup, PureSync, Paragon Backup&Recovery, Robocopy, …
Mac OS	Carbon Copy Cloner, SuperDuper, …
Ubuntu	Rsync, Back in Time

as of 2016/08/11

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 a	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 to MPG
Hierarchical Storage Management (HSM)	No	No	For archival of data from closed project

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

USE WHAT WORKS FOR YOU

AND STICK TO IT !

… especially when working in a team!

Self-speaking file name:

`Presentation_MPIBPC_20160811_V42.pptx`

vs. short file name:

~~`MPIBPC_final.pptx`~~

Avoid special characters

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

Original file name:

`PICT7639.jpg`

Custom file name:

`20151103_experiment01_proband03_001.jpg`

Versioning

```
Presentation_MPIBPC_20160811_V13.pptx
Presentation_MPIBPC_20160811_V13final.pptx
Presentation_MPIBPC_20160811_V13new-final.pptx
Presentation_MPIBPC_20160811_V13final-finalv1.pptx
Presentation_MPIBPC_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
- **Archive obsolete versions** to avoid confusion
- If you collaborate on a document, **use “track changes”** if possible

Folder structure

Use (sub)folders to organise your files, e.g.:

- Literature (primary literature)
- Publications (your own articles)
- Thesis (files relevant for your PhD-Thesis)
- Emails (archived important e-mails, as PDF)
- Projects (material from other projects/side-projects)
- Pictures (images, graphs, illustrations, logos, …)
- Experiments (e.g. experiment or survey designs)
- Data („raw“ datasets, separated from processed data)

(How) Do you organise your e-mail inbox?

No Folder structure

Alternative: use tagging / metadata to describe your files

- Content type (literature, publication, experiment design, data,…)
- Project context (researchers/SPs involved,
- Topic
- Time (and place) of recording, creation, acquisition
- Related material

\+ *Any other information you or others might need to quickly find a specific file*

Best practice (suggestion):

- Use a maximum of two levels of folders
- Put other relevant information in the file name
- Use tagging/metadata to the extent you feel comfortable with
  - and to the extent your OS supports it

Explain your data

- Why?
  - Make data understandable, verifiable, findable and reusable!

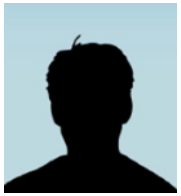
- How?
  - Directly write down which methods/materials you used. Write down what fails and what was successfully analysed.
  - Write down time, place, persons involved in creation of data.
  - Include title, name of primary and processed data.
  - Do not change/erase your original notes but add more 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 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

July 26 2016

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

Some Metadata standards for BPC (?)

OME-XML - Open Microscopy Environment XML

- vendor-neutral file format for biological image data, with an emphasis on metadata supporting light microscopy
- can be used as data file format or for encoding metadata within TIFF file
- maintained by the Open Microscopy Environment Consortium

Cell ML

- standard for encoding mathematical models
- particularly for models based on biophysical mechanisms
- substantial number of models available in CellML Model Repository

HDF 5

- open, free, versatile data model to represent complex data objects and wide variety of metadata
- completely portable file format with no limit on number or size of data objects in the collection

Organizing: 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**

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

Other 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

Questions?

Answers?

Your Feedback?