PAUL SCHERRER INSTITUT PSI

Wir schaffen Wissen – heute für morgen

Workshop Research Integrity at PSI 2013

Data management

Tuesday June 4 2013, 13.30 – 17.00

Louis Tiefenauer, PSI

Program

	Dur.	End
Welcome by Thierry Straessle	5 min	13.40
Ethical issues in data management	40 min	14.20
Group discussions	40 min	15.00
Coffee break, informal discussions	20 min	15.20
Presentation of outcomes from 15.20	50 min	16.10
General discussion moderated by L. Tiefenauer	20 min	16.30

End 16.30

Data management: the last fifty years

Figure 2.1 Gazing back: a recent history of computational and data science[56]

1960 ··· 1970 ··· 1980 ··· 1990 ··· 2000

1960: Hypertext
Imagining connectivity in the world's knowledge
The concept of links between documents begin to be discussed as a paradigm for organizing textual material and knowledge.

1960: Full-Text Search
Finding text without an index
The first full-text searching of documents by computer is demonstrated.

1962: Roger Tomlinson
Computerizing geographic information
Roger Tomlinson initiates the Canada Geographic Information System, creating the first GIS system.

1963: ASCII Code
A standard number for every letter
ASCII Code defines a standard bit representation for every character in English.

1963: Science Citation Index
Mapping science by citations
Eugene Garfield publishes the first edition of the Science Citation Index, which indexes scientific literature through references in papers.

1963: Data Universal Numbering System (D-U-N-S)
A number for every business
Dun & Bradstreet begins to assign a unique number to every company.

1966: SBN Codes
A number for every book
British SBN codes are introduced, later generalized to ISBN in 1970.

You

196[illegible]: DIALOG
Retrieving information from anywhere
The DIALOG online information retrieval system becomes accessible from remote locations.

1968: MARC
Henriette Avram creates the MAchine-Readable Cataloging system at the Library of Congress, defining metatagging standards for books.

1970s: Relational Databases
Making relations between data computable
Relational databases and query languages allow huge amounts of data to be stored in a way that makes certain common kinds of queries efficient enough to be done as a routine part of business.

1970-1980s: Interactive Computing
Getting immediate results from computers
With the emergence of progressively cheaper computers, it becomes possible to do computations immediately, integrating them as part of the everyday process of working with knowledge.

1970—1980s: Expert Systems
Capturing expert knowledge as inference rules
Largely as an offshoot of AI, expert systems are an attempt to capture the knowledge of human experts in specialized domains, using logic-based inferential systems.

1973: Black-Scholes Formula
Bring mathematics to financial derivatives
Fischer Black and Myron Scholes give a mathematical method for valuing stock options.

1973: Lexis
Legal information goes online
Lexis provides full-text records of US court opinions in an online retrieval system.

1974: UPC Codes
Every product gets a number
The UPC standard for barcodes is launched.

1980s: Neural Networks
Handling knowledge by emulating the brain
With precursors in the 1940s, neural networks emerge in the 1980s as a concept for storing and manipulating various types of knowledge using connections reminiscent of nerve cells.

1982: GenBank
Collecting the codes of life
Walter Goad at Los Alamos founds GenBank to collect all genome sequences being found.

1983: DNS
The Domain Name System for hierarchical Internet addresses is created; in 1984, .com and other top-level domains (TLDs) are named.

1984: Cyc
Creating a computable database of common sense
Cyc is a long-running project to encode common sense facts in a computable form.

1988: Mathematica
Language for algorithmic computation
Mathematica is created to provide a uniform system for all forms of algorithmic computation by defining a symbolic language to represent arbitrary constructs and then assembling a huge web of consistent algorithms to operate on them.

1989: The Web
Collecting the world's information
The web grows to provide billions of pages of freely available information from all corners of civilization.

1990: IMDb
Indexing movies
The Internet Movie Database is launched.

1991: Gopher
Burrowing around the internet
Gopher provides a menu-based system for finding material on computers connected to the internet.

1991: Unicode
Representing every language
The Unicode standard assigns a numerical code to every glyph in every human language.

1991 arXiv.org established: open access e-print repository for journal articles from physics physics, mathematics, computer science, and related disciplines.

1993: Tim Berners-Lee
A catalog of the web
Tim Berners-Lee creates the Virtual Library, the first systematic catalog of the web.

1994: QR Codes
Quick Response (QR) scannable barcodes are created in Japan, encoding information for computer eyes to read.

1994: Yahoo!
Jerry Yang and David Filo create a hierarchical directory of the web.

1995: CDDB
Indexing music
Ti Kan indexes CDs with CDDB, which becomes Gracenote.

1996: The Internet Archive
Saving the history of the web
Brewster Kahle founds the Internet Archive to begin systematically capturing and storing the state of the web.

1997 Launch of SETI@ home Individuals can provide their computing resources to help in data analysis for the search for extra terrestrial intelligence.

1998: Google
An engine to search the web
Google and other search engines provide highly efficient capabilities to do textual searches across the whole content of the web.

2000: Sloan Digital Sky Survey
Mapping every object in the universe
The Sloan Digital Sky Survey spends nearly a decade automatically mapping every visible object in the astronomical universe.

2000: Web 2.0
Societally organized information
Social networking and other collective websites define a mechanism for collectively assembling information by and about people.

2001: Wikipedia
Self-organized encyclopedia
Volunteer contributors assemble millions of pages of encyclopedia material, providing textual descriptions of practically all areas of human knowledge.

2003: Human Genome Project
The complete code of a human
The Human Genome Project is declared complete in finding a reference DNA sequence for every human.

2004: Facebook
Capturing the social network
Facebook begins to capture social relations between people on a large scale.

2004: OpenStreetMap
Steve Coast initiates a project to create a crowdsourced street-level map of the world.

2004: the UK Government's Office of Public Sector Information began pilot scheme to use the Semantic Web to integrate and publish information from across the Public Sector.

2009: Wolfram|Alpha
www.wolframalpha.com
An engine for computational knowledge
Wolfram|Alpha is launched as a website that computes answers to natural-language queries based on a large collection of algorithms and curated data.

History:
Data are primary resources not only in science

Scientific data:
Data produced in Science

Metadata (connected to)

- personal data
- other data

Science as an open enterprise,
Open data for open science
The Royal Society, June 2012 p. 25

Motivation

“The practice of science: Open inquiry is at the heart of the scientific enterprise. Publication of scientific theories - and of the experimental and observational data on which they are based - permits others to identify errors, to support, reject or refine theories and to reuse data for further understanding and knowledge. Science’s powerful capacity for self-correction comes from this openness to scrutiny and challenge.”

Science as an open enterprise, Open data for open science The Royal Society, June 2012

Research Integrity at PSI, EMPA, eawag, WSL

Guidelines for Good Scientific Practice

ON FRONT PAGE

Honesty, openness, self-criticism and fairness are the basis for credibility and acceptance in science. Researchers at PSI are committed to these values and to the guidelines which derive from them.

Wahrhaftigkeit, Offenheit, Selbstkritik und Fairness sind die Grundlage für die Glaubwürdigkeit und Akzeptanz der Wissenschaft. Wir Forschende am PSI sind diesen Werten verpflichtet und halten uns an die daraus abgeleiteten Richtlinien.

Science data management

Scientific experiments: e.g. Gravitation

Experience

Idea

Observation

Proposal

Hypothesis

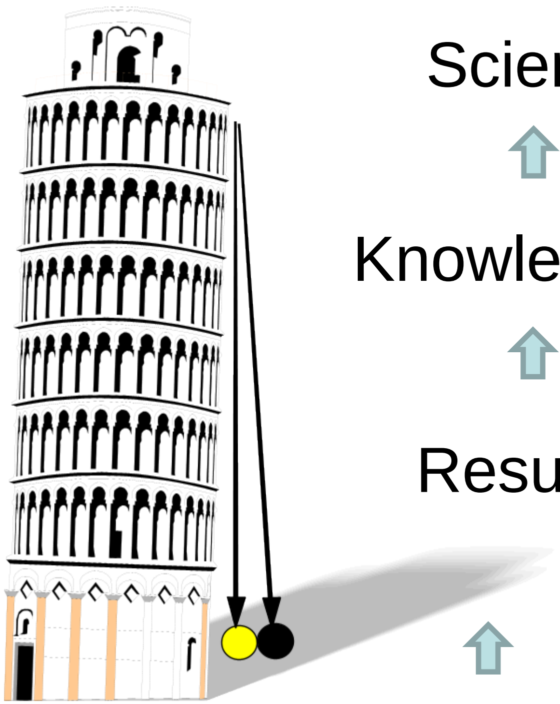
Project Appoval

Experiment design

Facilities

Measurements

Raw data



Data storage

Derived Data

Results

Knowledge

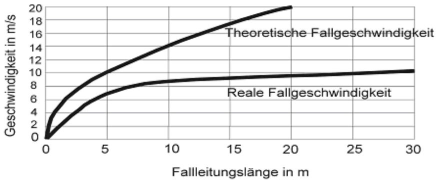
Science

Applications

Publication

$h = g/2\ t^2$

Data analysis



Data curation

Science data management

Raw data · Derived data · Information · Knowledge · Applications

Metadata

simulation

theory

Observation signal

numbers

properties

understanding

communication

speed matters e.g. in epidemies

coordination

generation · *analysis* · *description* · *publication*

Data should be:

accessible	*intelligible*	*comprehensible*	*usable*
zugänglich	verständlich	nachvollziehbar	*brauchbar*
↓	↓	↓	↓
Data storage / property	Data reduction	Data interpretation	Data format

Data life cycle and research integrity

Raw data

Storage

fabrication, falisification,theft
safety and security

Duration
Access
Ownership

responsibilities (PI) and others

Derived Data

Analysis

intelligibile, usable data
benefit and verifiability

Group discussion
Communication plan
Simulations, modelling
Interpretation

Metadata

Indexing
Communication
Indenfication sources

Curation

Readable data
Migration
Data (sets) access

privacy, fairness, usability
freedom of research
confidentiality

Results

Authorship
Visualization
Conclusions
Applications

Publication

fairness (plagiarism)
maximise benefit
avoid misinterpretation

TechTransfer

benefit (science, economies, poverty)
conflict of interest

Data management I

Acquisition

Detectors:

- Validation
- Deletion
- Processing

Maintenance

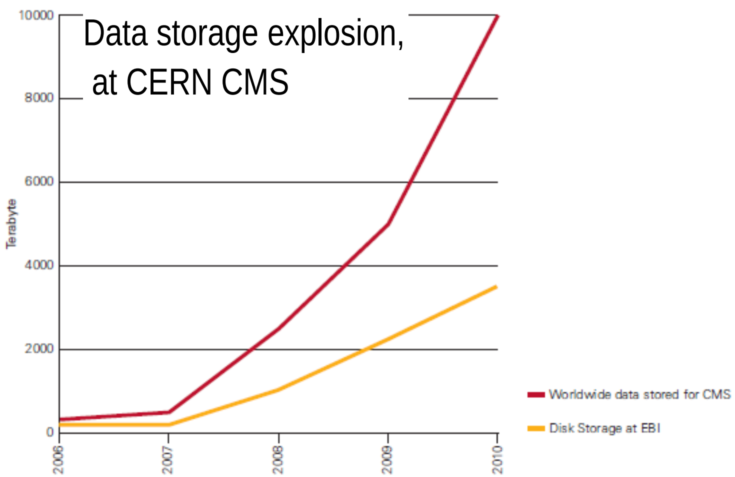
- Storage
- Curation
- Migration
- Safety (lost)
- Security (misuse)

Deposition of raw data

before publication

e.g. Bioscience-papers:

- DNA
- Proteins
- Microarrays (-omics)

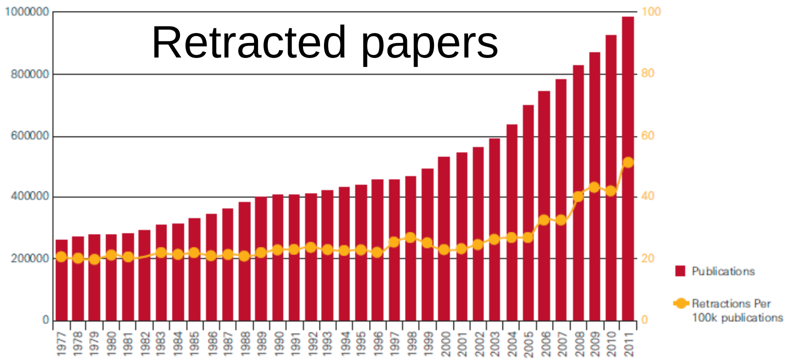


Policy depends
on the research field!

Scientific practice: Verification of results

Data management II

Figure 2.5 Number of publications (columns) and number of retractions (line) relating to clinical trials: 1977 - 2011[112]



Honest error, plagiarism 1:1

Personal data

- Clinical studies (side effects)
- Data banks (cancer, inheritary disease)
- Anonymization (how)
- Informed consent (test person's agreem.)
- Safe haven

Privacy (stigmatization, discrimination)

Public health

Restrictions

- Health Safety (DNA sequence infection)
- National security (terrorism)
- Ethical issues (dual use: avian flu paper)

Safety and security

TechTransfer

- Contract research
- Patent of process, product, apparatus
- Patent in force: licensing use, data free
- Public-private partnership

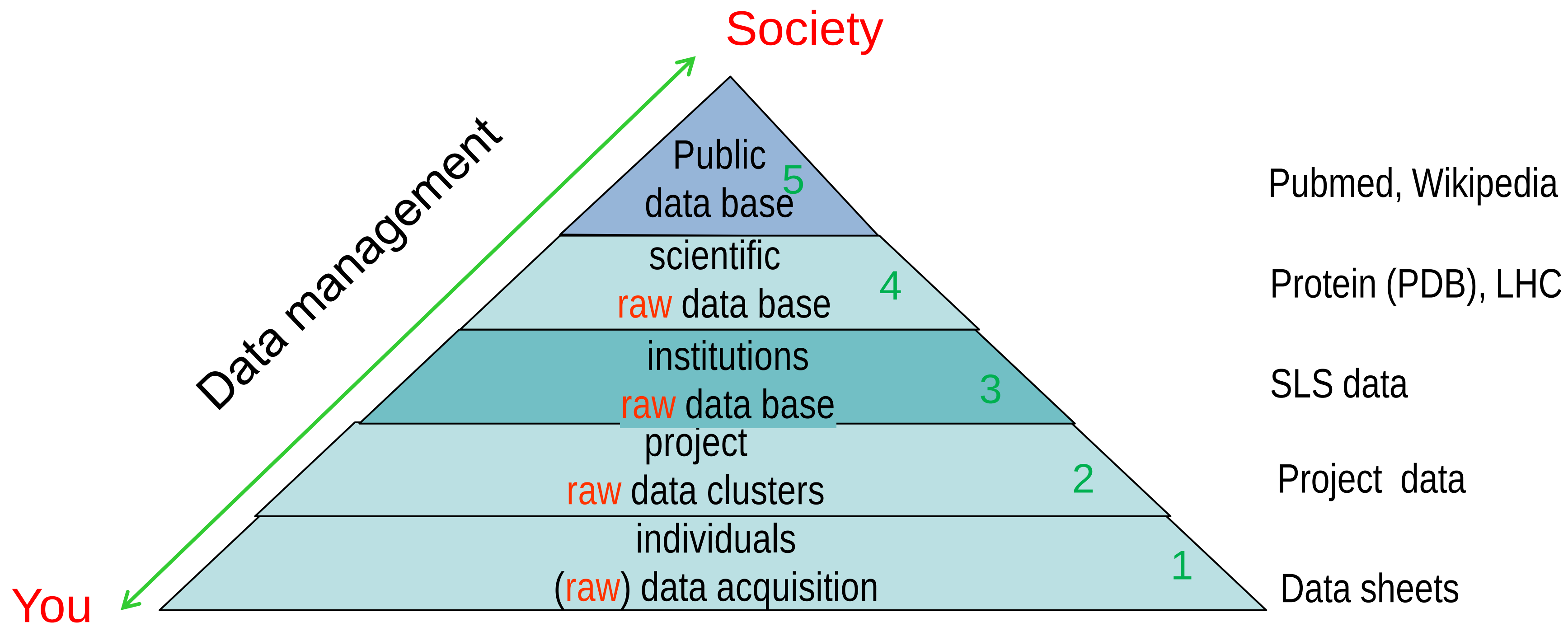
Conflict of interest

Independency

Freedom of research

Responsibilities

Data pyramide, **raw** data



Accord. Science as an open enterprise, Open data for open science The Royal Society, June 2012

Guidelines for good scientific practice (p.26 & 27)

- General aim: Foster credibility and acceptance of science, efficiency and quality
- Specific aims: verifiability (reproduction) (p.27), avoid misconduct , fairness (p. 28)

Duty of researchers: ***make use of your data!***

- publish upon completion of a project
- Transfer them into technologies to the benefit of society
- Conditions: freedom of research which is restricted by rules (legal and ethical)

Addressed points in PSI guidelines (code)

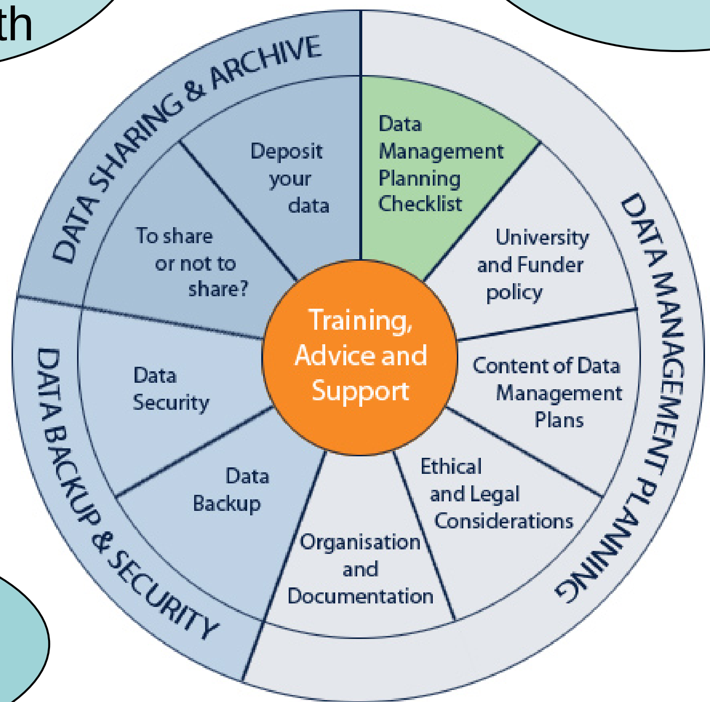
- Primary responsibility: PI
- Primary (raw) data (verification); processed (derived) data
- Storage (long-term), deletion, archiving
- Data cycle: analysis, publication
- Transfer for applications: technologies
- Analysis and interpretation: gray zone, self-crtiticism
- Communication: publish and share (scientific community, public)
- Rights and duties: sharing, ownership, access, proprietary

Data management policies

Regulations
what, when, where, how, who, whith

Advices
what, when, where, how

Training
help for researcher



Policy
Strategy & standards

Plan
who, what, when, how

Support
hard- & software procedures

Guidelines
what, why, how, who

PAUL SCHERRER INSTITUT

Data management policies

Points to be addressed (policy, plan, regulations)

- Responsibilities
- Application and use (data banks)
- quality (accessible, intelligible, usable)
- formats (for verification)
- storage (length) (cost: 1 Gb/5 years: 2 $)
- safety and security (dual use)
- curation and migration (costs: up to 10% of project cost)
- access & ownership (collaborations, proprietary)
- privacy metadata (researcher and users)
- sharing and communication rules
- Training and teaching

Supports

- Software data cycle: generation, analysis, curation, visualization
- Support for data curation: indexing, tracking,

Data management policies

Recommendations from Royal Society (2012)*

1. *Scientists:* create accessable, intelligible and usable data
2. *Institution*: data communication as a criterion for career promotion
3. *Ranking system*: institution output indicators (publications, data)
4. *Academies, learned societies*: promote open science
5. *Funding agencies*: require data management plan
6. *Scientific journal*: repository before publication, etc.
7. *Data in public interest*: industry and regulators agreements
8. *Governments*: support open science, also by skilled personnel
9. *Governance*: release privacy rules
10. *Good practices*: assure safety and security (openness & secrecy)

* Science as an open enterprise, Open data for open science The Royal Society, June 2012

Supports, tools, rules at PSI

- *Large-scale facilities*: acquistion, storage, access, sharing curation, metadata
- *Departments:* data storage, analysis, access, proprietary curation
- *AIT:* acquisition, format, storage, safety, migration, costs

List to be completed in the group discussions !

List of topics (I)

1. **Responsible actors**: experimentor, **PI!**, supervisors, leaders
2. **Data management plan**: education, responsibilities, communication
3. **Acquisition**: raw data, metadata, statistics, formats, ~~fabrication~~
4. **Treatment**: analysis, validation (grey zones), processing (~~falsification~~), conversion, statistical evaluation, reduction, presentation (tables, graphics, images)
5. **Utilization of results**: publications, authorship ~~(plagiarism)~~, tech-tansfer, spin-offs
6. **Storage and archiving**: IT facilities, costs, migration

ethical issues   Legal and financial issues

List of topics (II)

7. **Metadata**: associated metadata, data-catalogue (privacy, freedom of research)
8. **Ownership**: research data, patents, external users (scientific, proprietary), ~~theft~~, metadata,
9. **Disclosure practice**: ongoing project, auditing (conflict of interest), reviewing, collaborations (NDA)
10. **Access**: identified persons, passwords strategy, raw data access
11. **Deletion**: public data, storage
12. **Curation**: migration, backups, transformation (history)
13. **Data sharing**: open access, exchangeable formats

ethical issues

Legal and financial issues


ONE MORE
OF THESE
DATA-PIRACIES!

Group discussions

be back 15.20

1. Complete or adjust list of data management list

2. Discuss recommendations

3. Which point is most important for you?

4. Can you give specific recommendations or hints?

Summary*:

Open data to open science

1. Free exchange of data between researchers

2. Research institutions are primary actors (major influencing factors: reward and promotion system)

3. Additional indicators are needed to assess success

4. Promote open science policy by academies

5. Incentives given by funding agencies

* Science as an open enterprise, Open data for open science, The Royal Society, June 2012

Data management for science, recommendations

6. Improve free access to data (raw & processed) for readers

7. Publication of data (negative & null) of public interest

8. Politics and regulations should foster open science

9. Research data management practice (privacy, metadata, risk minimization)

10. Consider security (avoid lost of data) and safety (avoid damage to people) issues

Research Integrity Workshop: Topics

1 Publication / Authorship	**2011**
2 Research Misconduct FFP (Plagiarism)	**2012**
3 Data Management	**2013**
4 Collaborative Science, decided	2014

Future plans:
5 Mentorship
6 Conflicts of Interest / Commitments
7 Peer Review / Audits