Introduction to

Research Data Management

Stefan Kühne, kuehne@uni-leipzig.de
University Computing Center, Research and Development
Introduction to
Research Data Management

Stefan Kühne, kuehne@uni-leipzig.de
University Computing Center, Research and Development
ABOUT ME

- Background in **Computer Science**
  - Diploma in Computer Science
  - Scientific assistant at Department Business Information Systems
  - PhD thesis: Transformation Space for model-driven service-oriented Integration
- Head of **Research and Development at UCC** since 2014
  - Research projects on big data, semantic data, linked open data
  - Scientific Computing
  - **Research Data Management**

[Kühn09]
WHAT IS SO SPECIAL ABOUT RESEARCH DATA MANAGEMENT?

It is about long-term preservation of data (10 years).
- Re-use of data in new contexts
- But it is expensive!

It is about data publishing.
- Open Access
- But maybe we want to use our data (again) exclusively in the future.

It is about reproducible research.
- Avoid scientific misbehaviors
- But it is time-consuming.
DIGITALISATION

- increased **amount of data**: digitisation of objects (books, maps, events, …), data streams, social networks, open data repositories, …
- world wide **interconnectedness** of resources (data resources, computing resources, services, …)
- **mobile computing**
- **machine learning** techniques

digitalisation ≠ digitisation
DOCUMENTS

BACH RESEARCH

Passacaglia BWV 582
(transcript from J.S. Bach’s brother)
STREAM DATA

ANOMALY DETECTION IN WIND POWER PLANTS
BIG DATA

THE COMMON CRAWL ARCHIVE, http://commoncrawl.org

- open repository of web crawl data
- web pages are available as raw data, meta data, text data)
  - 2.8 billion web pages (09/2018)
  - 242 TByte uncompressed content
- common crawl web graph represents the linking structure of the web as graph data
  - 886 m. nodes (07/2018)
  - 5.4 bn. edges (07/2018)
from DATA SCIENCE to RESEARCH DATA MANAGEMENT
DATA SCIENCE

- traditional research approaches
  - scientific findings based on observations
  - scientific findings based on model development and theory building
  - new options enabled by IT, e.g. HPC-based simulation of complex phenomena

- data-driven approaches
  - exploration and analysis of huge amounts of data from different sources
  - selection and limitation of data is affected by the analysis approach
  - generation of new hypothesis
  - stimulation of interdisciplinary collaborations
INTERDISCIPLINARY RESEARCH

Civilisation Diseases
- **Examples**: structured data of lifestyles, diseases, states of health
- **Sources**: questionnaires, examinations, images, audios, videos

Globalisation
- **Examples**: data of global streams (assets, financial capital, cultural goods), data resources to cultural artifacts (e.g. digital humanities)
- **Quellen**: questionnaires, archives, text corpora, images, audios, videos

Biodiversity
- **Examples**: data of zur diversity of species, genetic diversity, ecosystems, process data (such as nutrient cycles, evolutions, carbon dioxide cycle, nitrogen cycles, biochemical time series of proteins)
- **Sources**: repositories of genetic data, databases of occurrences of species, microarrays, images, audios, videos

research question: How to treat obesity in a global context?
research question: Do limited biodiversity forsters infectious diseases?
research question: What is the impact of local conflicts to biodiversity?
How can an integrated processing and analyzing of data from different sources be achieved?

- Trust in data security and safety
- Reliable infrastructures
- Sufficient information to interpretate and re-use of data
- Data workflows (transformation, storage, communication)
- Explicit and reproducible processes
- Independence of data producer and data user

REQUIREMENTS OF A SYSTEMATIC MANAGEMENT
VARIABILITIES IN REALISATION

Integrated processing and analyzing of data from different sources

- organizational governance structures
- infrastructure resources
- human resources, new roles and job profiles

Realisation is context-specific:

- interchange formats
- meta data
- data publishing services
- data curation services
- incentives
- ...
a closer look at
RESEARCH DATA
WHAT IS (RESEARCH) DATA?
MEGAMODEL

has long hair, has bushy tail, ...

concept

intension

represents

extension

element

element

represents

extension

Language

intension

Technical Space (XML, DB, OWL, CSV, …)

L0

L1

L2

Collie

Class

Object

Meta Class

has long hair, has bushy tail, ...

Lassie

instance-of

type

instance

instance-of

instance-of

instance-of

instance-of

instance-of

intension

extension

represents

represents

represents

represents
RESEARCH DATA
IS A REPRESENTATIVE FOR SOMETHING TO DESCRIBE
OR A FACT FROM REALITY TO EXPLAIN SOMETHING

- research process (inspired by critical rationalism)
  - generation of a research hypothesis
  - evaluation of the hypothesis in reality
  - for this purpose we use data as marks/traces of reality

- Data may be collected under certain research questions or even generated.
- Some data is used from a purely observational perspective, e.g., the use of process-generated data.
- Some data are "re-used" after being collected for analytical purposes other than scientific research (e.g., clinical data).
- Existing data collected under other research aspects are re-used in new contexts (secondary analysis).
Research data refers to all information (independent of form or presentation) needed to support or validate the development, results, observations or findings of a research project, including contextual information. Research data include all materials which are created in the course of academic work, including digitisation, records, source research, experiments, measurements, surveys and interviews. This includes software and code. Research data can take on several forms: during the lifespan of a research project, data can exist as gradations of raw data, processed data (including negative and inconclusive results), shared data, published data and Open Access published data, and with varying levels of access, including open data, restricted data and closed data. [Lear18a]
development and definition of RESEARCH DATA MANAGEMENT
POTENTIAL DISADVANTAGES OF PUBLISHING RESEARCH DATA

- Time and money required for preservation and publishing data

- Disadvantages in scientific competition (research data may be evaluated by others)

- Reputation-damaging detection of methodical inadequacies of the data analysis

- Disadvantages in the economic use of research data

- Legal problems due to the publication of data (violation of data protection or property rights, e.g. in the case of protected measuring methods)

- Misinterpretation by third parties
SYSTEMATIC HANDLING OF RESEARCH DATA

- **local storage within institutes** without specifications, guidelines

  vs.

- **data-intensive sciences (e.g. climate research, meteorology)**
  - continuous increase in the amount of data to be stored and processed
  - required new information systems for long-term preservation (mechanism and services for storage, backup, registration, deposit and retention) to support current and future access
  - development was driven by economic considerations
  - handling of the dynamic research data lifecycle (long-term interpretability)
INITIATIVES, RECOMMENDATIONS, GUIDELINES

- 1998/2013 German Research Foundation (DFG)
  - Safeguarding Good Scientific Practice
- 2009 German Research Foundation
  - Recommendations for secure storage and provision of digital research primary data
- 2010 Alliance of Science Organizations in Germany
  - Principles for handling research data
- 2014 German Rectors' Conference (HRK)
  - Management of research data - a key strategic challenge for university management
- 2015 German Research Foundation
  - Guidelines for handling research data

Recommendation 7 – Protection and Storage of Primary Data: Primary data as the basis for publications should be kept on durable and secure media in the institution where they were created for ten years.
INITIATIVES, RECOMMENDATIONS, GUIDELINES

- 1998/2013 German Research Foundation (DFG)
  - Safeguarding Good Scientific Practice
- 2009 German Research Foundation
  - Recommendations for secure storage and provision of digital research primary data
- 2010 Alliance of Science Organizations in Germany
  - Principles for handling research data
- 2014 German Rectors' Conference (HRK)
  - Management of research data - a key strategic challenge for university management
- 2015 German Research Foundation
  - Guidelines for handling research data

1. Subject-specific definition of primary data, e.g. level of aggregation
2. Subject-specific organisational concept
3. Standards
4. Storage of data is related to researcher
5. Open Access
6. Meta-data
7. Data quality
INTIATIVES, RECOMMENDATIONS, GUIDELINES

- 1998/2013 German Research Foundation (DFG)
  - Safeguarding Good Scientific Practice
- 2009 German Research Foundation
  - Recommendations for secure storage and provision of digital research primary data
- 2010 Alliance of Science Organizations in Germany
  - Principles for handling research data
- 2014 German Rectors' Conference (HRK)
  - Management of research data - a key strategic challenge for university management
- 2015 German Research Foundation
  - Guidelines for handling research data

Concretization of recommendations in the context of DFG funding programs

+ Research Data Management Plan
+ Provision/Publishing of RD
+ Long-term archiving
INTITIATES, RECOMMENDATIONS, GUIDELINES

- 1998/2013 German Research Foundation (DFG)
  - Safeguarding Good Scientific Practice
- 2009 German Research Foundation
  - Recommendations for secure storage and provision of digital research primary data
- 2010 Alliance of Science Organizations in Germany
  - Principles for handling research data
- 2014 German Rectors' Conference (HRK)
  - Management of research data - a key strategic challenge for university management
- 2015 German Research Foundation
  - Guidelines for handling research data
INITIATIVES, RECOMMENDATIONS, GUIDELINES

- 1998/2013 German Research Foundation (DFG)
  - Safeguarding Good Scientific Practice
- 2009 German Research Foundation
  - Recommendations for secure storage and provision of digital research primary data
- 2010 Alliance of Science Organizations in Germany
  - Principles for handling research data
- 2014 German Rectors' Conference (HRK)
  - Management of research data - a key strategic challenge for university management
- 2015 German Research Foundation
  - Guidelines for handling research data

Establishing RDM at Universities

- Development guidelines
- Establishing collaborations across organizations
- Strengthening the information competence with training programs
- Extension of institutional infrastructures
ADVANTAGES OF RESEARCH DATA MANAGEMENT

- supports the scientific progress
- getting access to financial resources (e.g. grants, funding)
- direct scientific reputation gain by publishing high quality research data
- increased visibility of the own research work
- recognition in the scientific community (fulfillment of social norms)
- opening up new opportunities for cooperation and publication (e.g. co-authoring with data users)
- scientific suggestions through collaborative work with the same data
“Research data management concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. Research Data Management is part of the research process, and aims to make the research process as efficient as possible, and meet expectations and requirements of the university, research funders, and legislation.” [Leic18]

- describes the process including all methods and procedures of handling RD
- covers the complete lifecycle: includes planning, collecting/generating, processing, preservation and documenting of research data
- ensures controlled access, reuse, reproducibility and quality assurance of RD
- enables reproducibility and the reusability of scientific results
- requires the establishment of governance structures
- creates a requirement for interdisciplinary work and future project applications
KEY PERFORMANCE INDICATORS

PREPARATION OF RDM
- Institutional policy
- Steering committee dedicated to RDM
- RDM Services
- Dedicated staff to RDM
- Job profiles dedicated to RDM
- Training sessions on RDM

IMPLEMENTATION OF RDM
- Monitoring activities
- Review of the policy
- Engagement of disciplines within the research institution
- data sets stored
- data sets published
- persistent identifiers for published RD

[ Lear18b ]
the starting point

RESEARCH DATA POLICY
RESEARCH DATA POLICY

- defines the principles of handling research data in a specific context
- context: interdisciplinary vs. disciplinary, institutional, journal
- within an institution a research data policy represents a starting point for RDM

- components of a research data policy
  - preample
  - definitions (researchers, research data, research data management, research data lifecycle)
  - jurisdiction
  - intellectual property rights
  - principles of handling of research data
  - responsibilities, rights, duties for researchers and for the research institution
  - validity
A RESEARCH DATA POLICY DEFINES RESPONSIBILITIES

- different roles
  - IT service provider
  - scientific library
  - research support
  - legal services
  - researcher
- storage systems for long term preservation
- legal compliance, legal and ethical issues
- handling of sensitive data
- creation of data management plans
- quality assurance
CONCLUSION
OUTLOOK
CONCLUSION

- benefits and risks of Research Data Management for scientists
  - offers potential for excellent, reproducible and sustainable research
  - defines requirements for scientific work
  - increases effort

- benefits and risks of Research Data Management for research institutes
  - offers potential for increased visibility
  - represents a requirement for interdisciplinary research (proposals in the future)
  - requires new information services, infrastructure resources and consulting services
OUTLOOK

- **Publishing of Research Data**
  - Stephan Wünsche
  - Thomas Riechert

- **Research Data Lifecycle**
  - Marc Schönwiesner
  - Toralf Kirsten

- **Data Repositories, Research Infrastructures**
  - Gerhard Heyer

- **Data Scientists**
  - Chris Armbruster

- **Other aspects**
  - Panel discussion

- **Data Management Plan**
  - Barbara Weiner
THANK YOU FOR YOUR ATTENTION!

Dr. Stefan Kühne
Universitätsrechenzentrum

Augustusplatz 10, 04109 Leipzig
T +49 341 97-33303  F +49 341 97-33399

kuehne@uni-leipzig.de
http://www.urz.uni-leipzig.de
LITERATURE


[Leic18] University of Leicester: What is Research Data Management. https://www2.le.ac.uk/services/research-data/rdm/what-is-rdm