# RTG Seminar – Research Data Management

Susanne Mocken, Dirk von Suchodoletz

Freiburg, 16/01/2019

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

RDMG

Research Data Management Group

# Outline

- Development of (field-specific) RDM strategies
  - Research data management
  - Scientific and institutional requirements
  - Planned technical systems
  - Organizational challenges
  - Internal and external cooperation
  - Managing the data influx (quality vs. quantity)
  - Governance
  - Field specific activities
  - Future steps

# Motivation for Research Data Management

# Digitization of Science

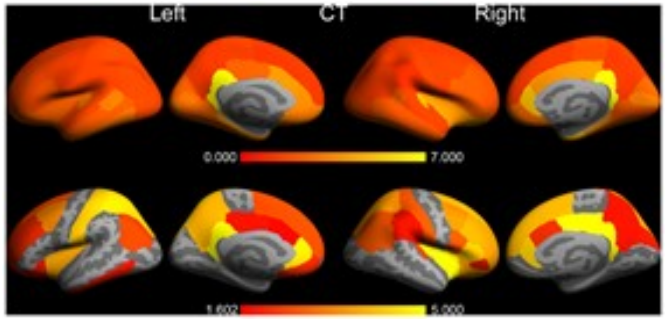
- Digital tools and digitized scientific workflows of rising significance (in most fields the relevant working environment)
  - Completely new possibilities (fast searching, retrieval, exchange, use in automated digital scientific workflows, …)
  - New forms of science: Big data, ...
  - New challenges: (longterm) access, distribution
- Physics – field of early adopters of computers for research in it's various areas

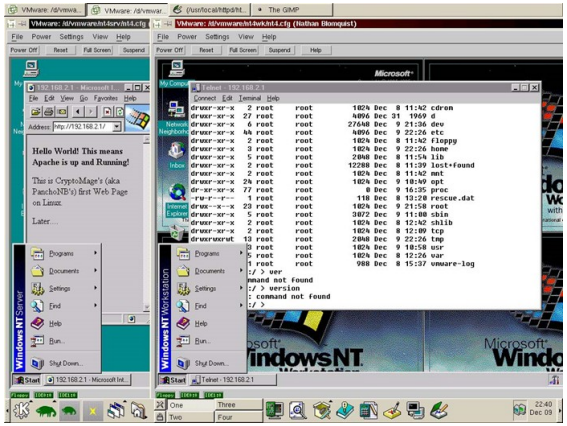
# Good Scientific Practice

- Concept relevant in all disciplines, developed over time
- Part of “Safeguarding Good Scientific Practice”
  - Lab books, reports
  - Scientific publishing
  - Various materials stored / kept in whatever repositories
- Concepts and tools for good scientific practice differ from domain to domain but share a common ground

# Research Data in Disciplines

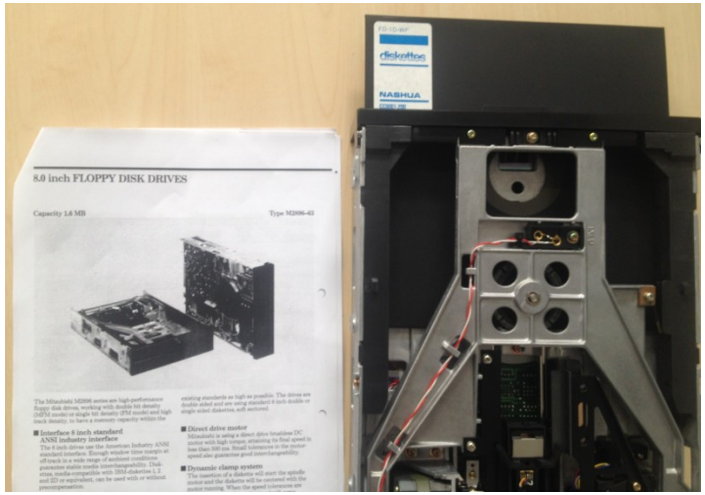
- Research data (like in the analog world) is rather discipline specific (like e.g. citations)
  - Often common, community specific standards exist
  - Discipline specific meta data
  - Complete technical environments (hardware, software stacks) become relevant beside the digital data alone (most data objects un-usable withou the technical context)



# Research Data Management

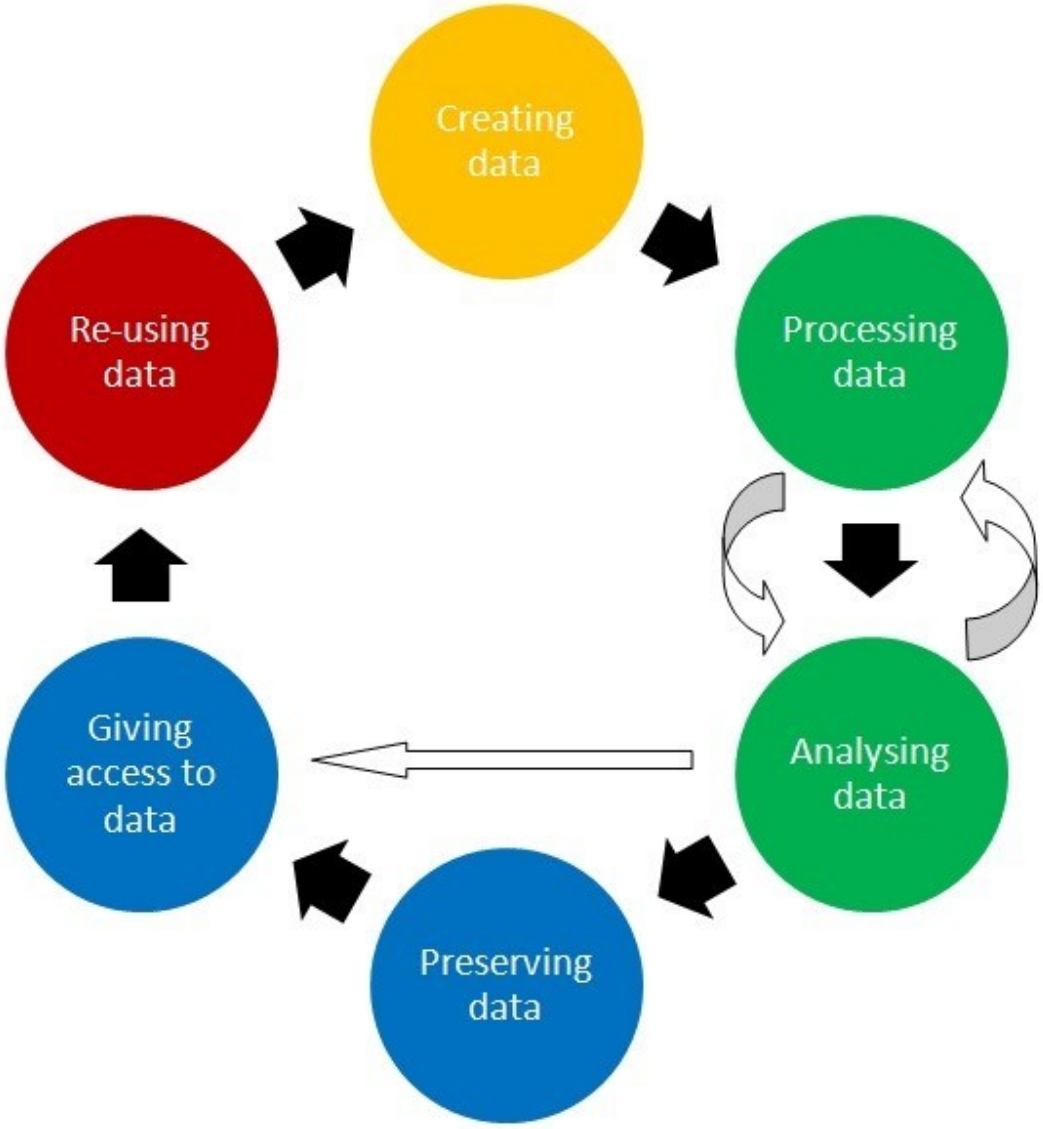
- Digital tools and workflows driven by fast technological change and scientific advance
  - Obsolescence of machines, data formats, tools a phenomenon present pretty much from the beginning
  - Lost data, results because of decay of storage media, unavailability of media readers, adapters, software, ...
- Good Scientific Practice to be established for the digital domain too → *Research Data Management*



# FAIR principles

- Requirements towards (digital) research data defined in 2105 by major institutions
  - Findable
  - Accessible
  - Interoperable
  - Reusable
- Idea to support knowledge discovery and innovation, support data and knowledge integration, promote sharing and reuse of data
- Discipline independent and allow for differences in disciplines

# Research Data Lifecycle




- Model for creation, processing, use and re-use of research data
- Re-use: Often old data could be input to new research (even in a completely different field)

# Institutional RDM

- Rising demand for Research Data Management because of digitization
  - Reuse of data
  - Ensuring good scientific practice
  - Creation, production of data involves significant infrastructure
  - End-to-end data management concepts
- Requirements set by funding institutions

→ RDM becomes new column in research organisations framework beside library, arts collections, zoology exhibition, ...

# Funding requirements

- Requirements will get more and more implemented on national level
  - Universities should have a data management policy
  - DFG started to require Data Management Plans for a while (as well as EU Horizon 2020 is requiring it already)
  - Same in the ongoing Excellency Strategy
  - In the future: New research projects might get formally rejected if no (proper) data management got described and implemented
  - Data management will become part of project evaluation in the future

# Funding requirements

- Increasingly formal requirements to get research funded
  - e.g. COMMISSION RECOMMENDATION (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information:

1. Member States should set and implement clear policies (as detailed in national action plans) for the dissemination of and open access to scientific publications resulting from publicly funded research. Those policies and action plans should provide for:
— concrete objectives and indicators to measure progress,
— implementation plans, including the allocation of responsibilities and appropriate licensing,
— associated financial planning.

Member States should ensure, in compliance with the EU acquis on copyright and related rights, that as a result of these policies or action plans: ...

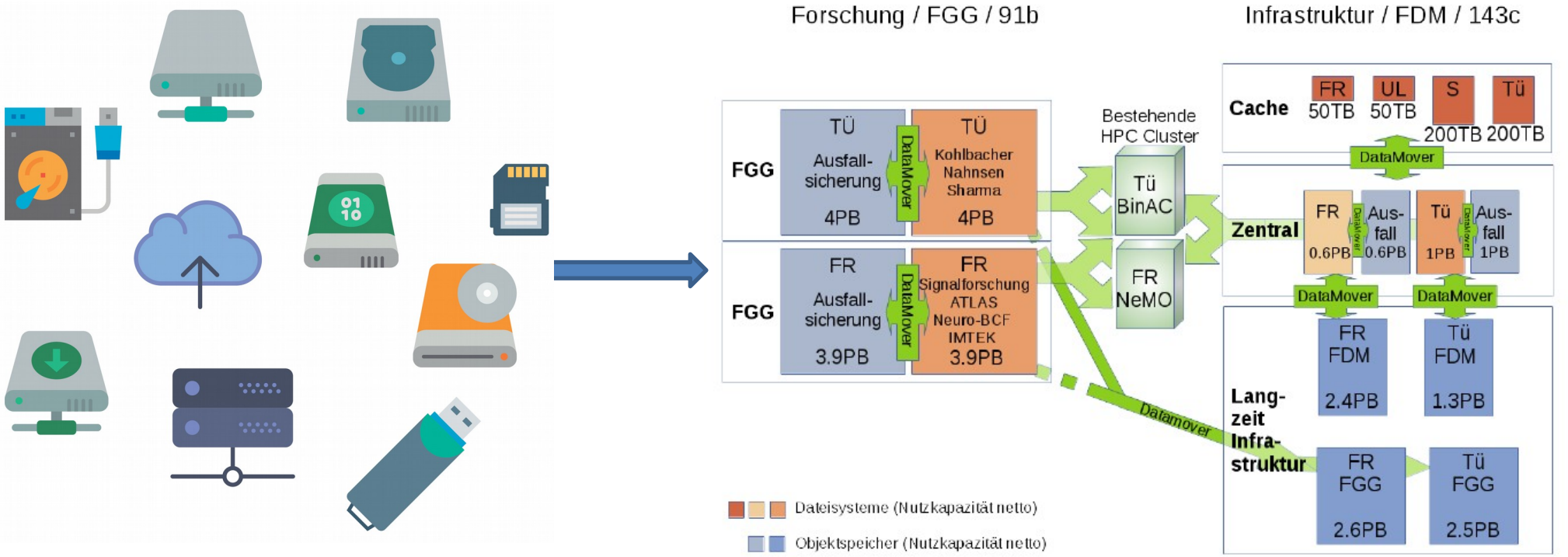
# Research Infrastructures

# Research Infrastructures

- Computer Center operates significant infrastructures
  - VMware ESX Virtualization
  - General data storage for homes & groups
  - bwCloud IaaS
  - NEMO HPC
  - TSM Tape Disaster Recovery
- Future: bwSFS
  - Storage for Science as system for the introduction of university level research data management system

# From NFS to RDM system

- Traditional file servers not enough to cope with modern requirements

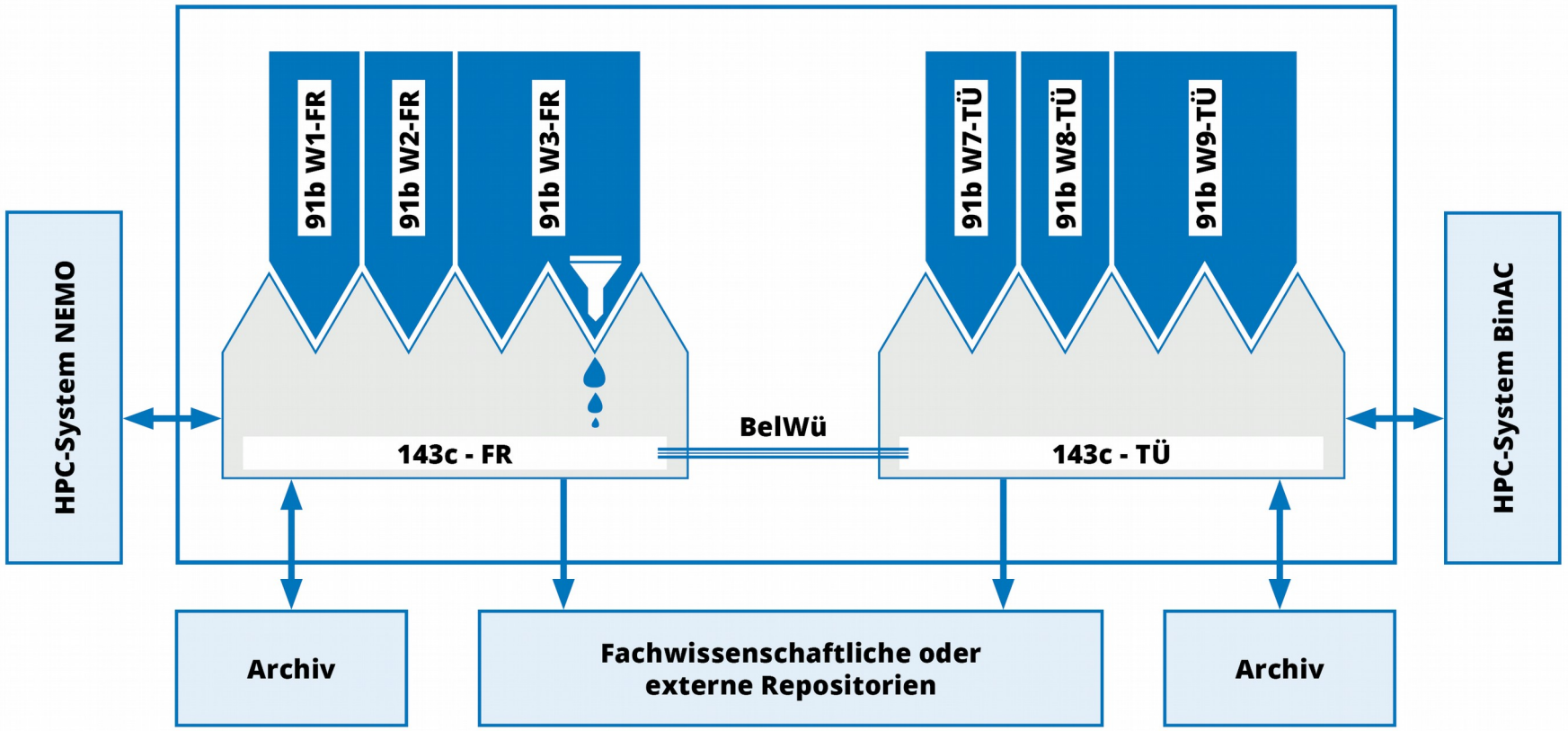


# Envisioned core RDM services

- Repository services like e.g. iRODS
  - Evaluated within the bwDATAbib project
  - Provide the linking of data to it’s meta data
- Object store or filesystem services
- Versioning services like e.g. GIT
- (Longterm) Archival services

# Initial Financing

- Structure of the DFG grant
  - General part covered by 143c



# Meta Data & Decision Making

# Technical and Scientific Meta Data

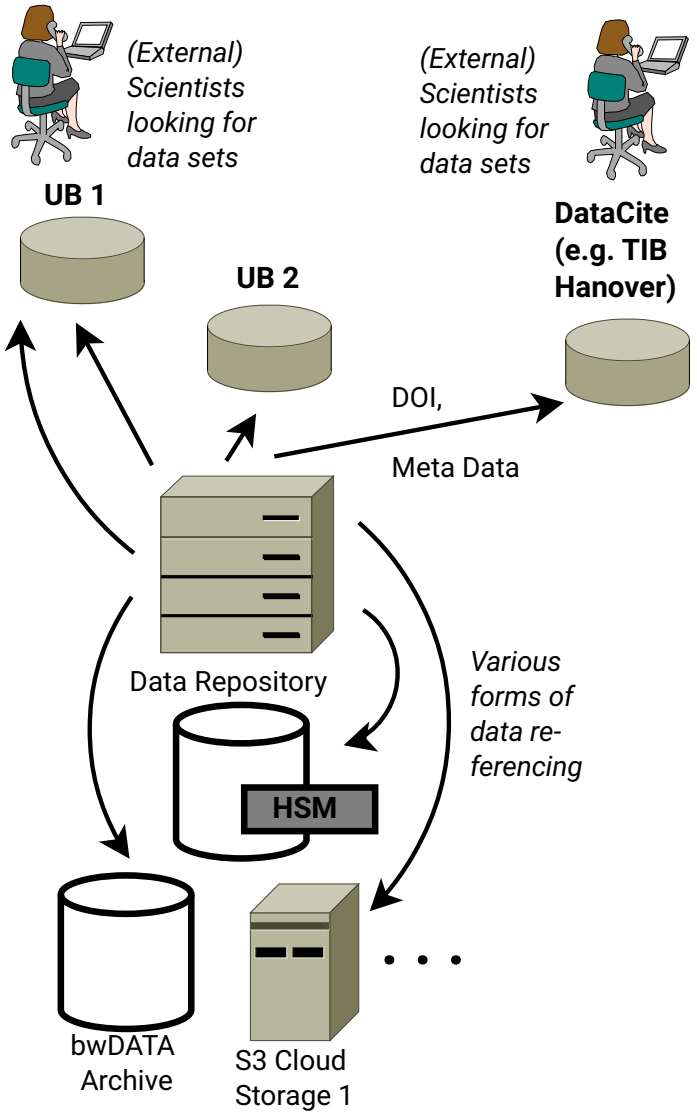
- Required for both automated and manual decision making
- Meta Data required for the Data Mover (some information only relevant for technical purposes)
- Meta Data required for re-use of data sets (defined by respective scientific community)
- Meta Data required to identify research projects (to link to e.g. grants in Research Information Systems)

# Definition of Services & Meta Data

- *Scientific communities* define their RDM / storage needs
  - Amount of data and expected time spans
  - Amount of data over the data life cycle
  - Access methods to the stored data
  - Required scientific Meta Data
  - Policies for access and re-use

# Data Cite and Object Access

- Libraries reference scientific material for long time
- Library systems provide access to various materials and can reference data sets already (with e.g. PUIs)
- Different forms and ways of referencing scientific data sets



# Data Cite and Object Access

- Data will become more and more part of scholarly communication and publication
- DataCite – using DOIs (Digital Object Identifiers or handles for less static objects)
- DOIs should be persistent and reference data set independent of its actual location within the SfS system
- Open Archives Initiative (OAI) Protocol for Metadata Harvesting
- Many libraries use OAI to exchange information between repositories

# Governance & Financing

# Governance

- Very diverse requirements by different groups
- *University library & computer* center can not decide on data lifetime, validity, worth, ...
  - Costs associated with amount of data (single data sets, storage capacity required by a project, faculty, ...)
  - Keeping costs under control requires  regular optimizations and cleanup
  - Cost models need to be created in the mid and long run

# Financing Model

- Core institutions of a university usually have their own budget
- Often research & service launch funds available from state or science funding institutions
- Not sustainable in the long run
  - Costs need to be covered by the university (and might get re-financed by instituts, projects, ...)

# Re-Financing

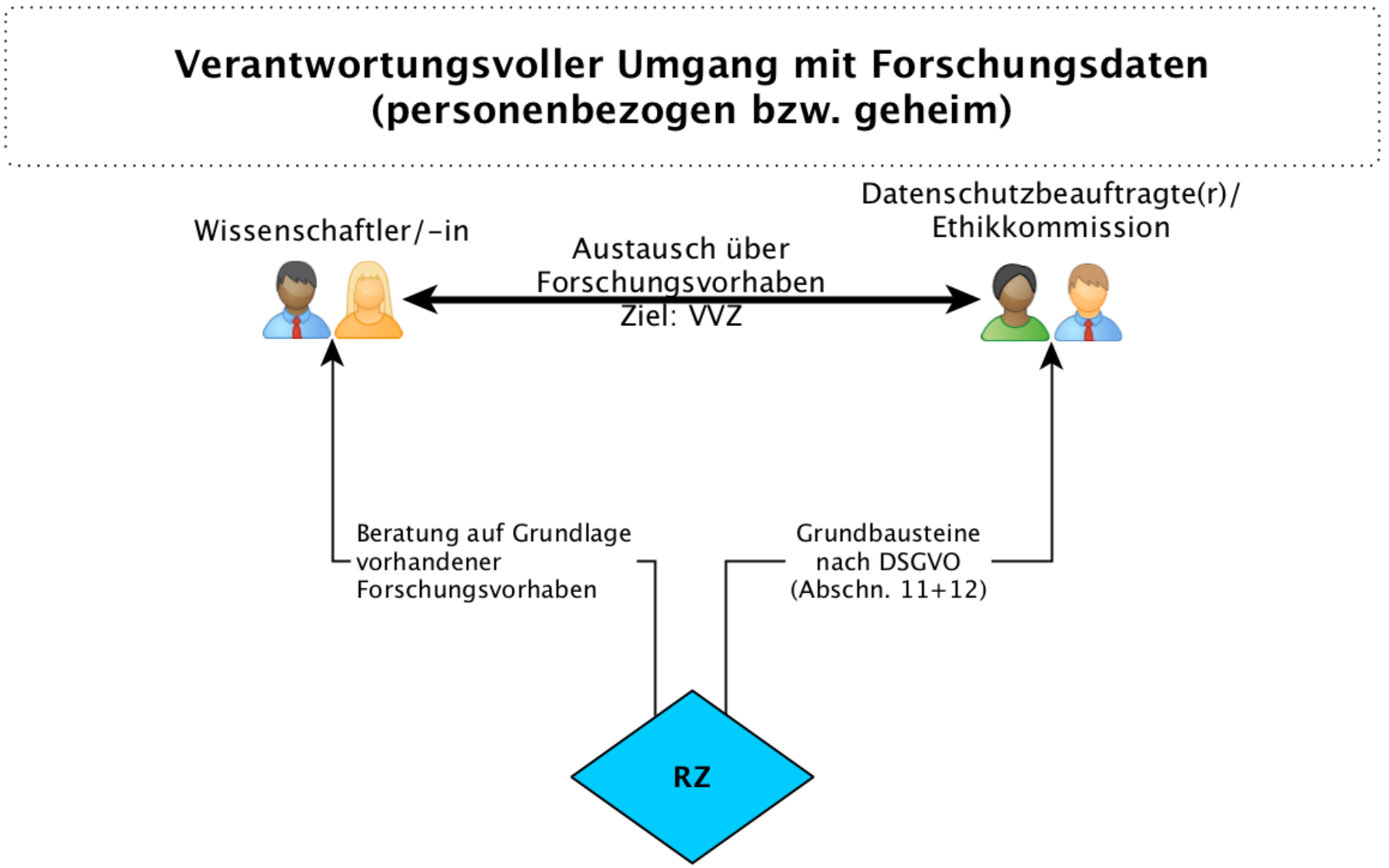
- Possible for researchers to get grant money for data management
  - Some funding institutions require open data and expect data to be published
- Challenges
  - Data ownership changing over time
  - Access policy might change over time (e.g. from restricted to open)
- Faculties need to get involved e.g. via academic self governing structures
  - Create committees to decide on RDM related issues

# Data Management Plans

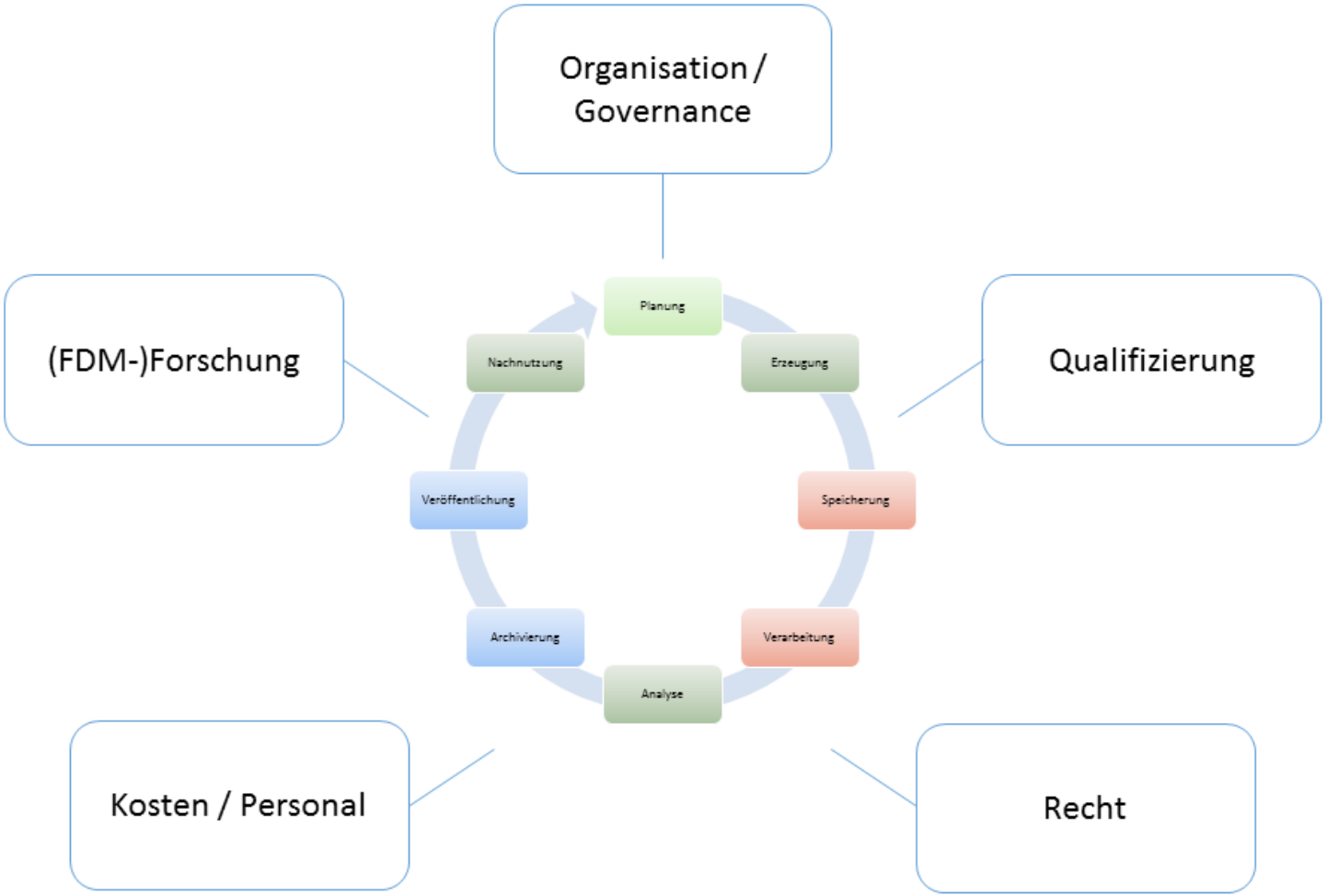
# Data Management Plan (DMP)

- Holistic planning of data management
  - Expected amounts of data (which sources)
  - Licenses, data protection, privacy, IPR, ...
  - Processing steps over the project life time
  - Relevance of primary data
  - Data planned to be thrown away
  - ...
- Planning of storage during project runtime
- Planning of archival storage and access methods after project end
- Cost estimation for grant application

# Next steps (researchers)



# Next steps (RDMG)



# Creation of the RDMG

- University of Freiburg quite late in the process but various groups were active already before
  - Research Data Management Group (RDMG) - Cooperation of the university library, the Freiburg Research Services and the computer center
- Computer center as core service provider for research infrastructures
  - Storage for various purposes
  - Tape storage for backup and disaster recovery
  - Facilities to host large scale IT equipment
  - Focus on technical expertise but less experience with data curation, annotation, …

RDMG
Research Data Management Group

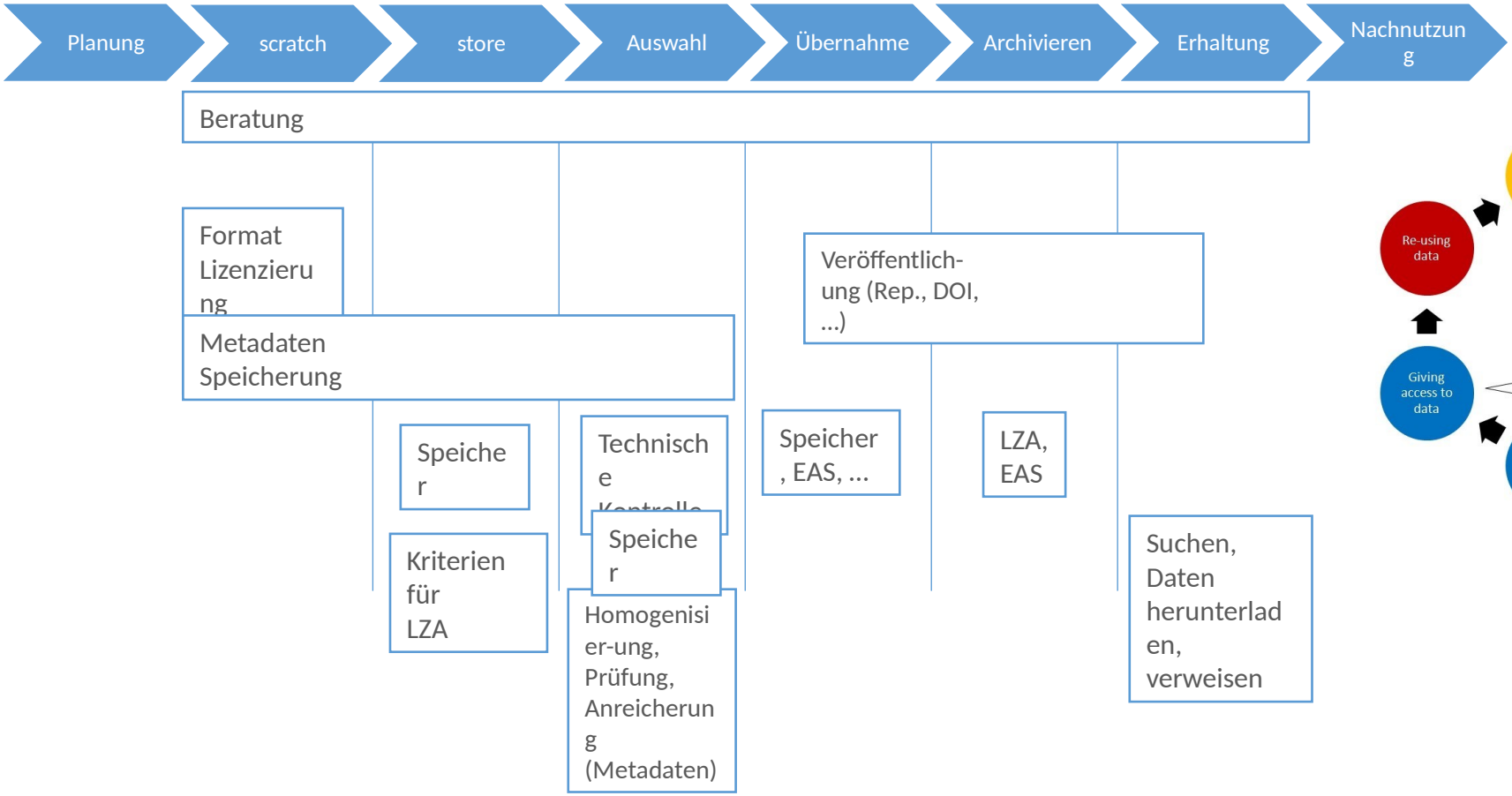
# Creation of an RDMG

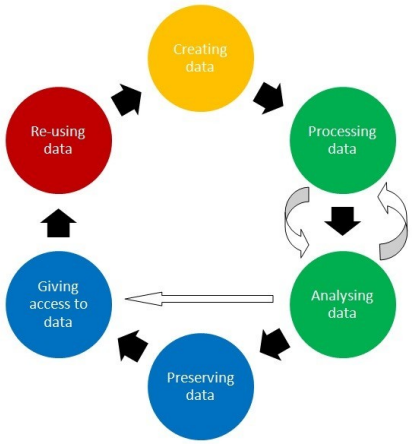
- Research data policy passed by the university / rectorate
  - Implementation will take some time
- Data related services distributed
  - Consulting on technical issues of RDM done by the *computer center*
  - Traditional research support services (on literature and other resources) → *university library*
  - *Science support center* department in central administration to advise on grants and applications
- End-to-end RDM requires holistic approach

# Creation of the RDMG

- Professorship in Communication Systems
  - Research in functional longterm access and archiving
  - Development of concepts and teaching
- Required
  - More advanced training for scientists, junior researchers and students
  - Make data management part of the general scientific practice

# Linearization of the Lifecycle



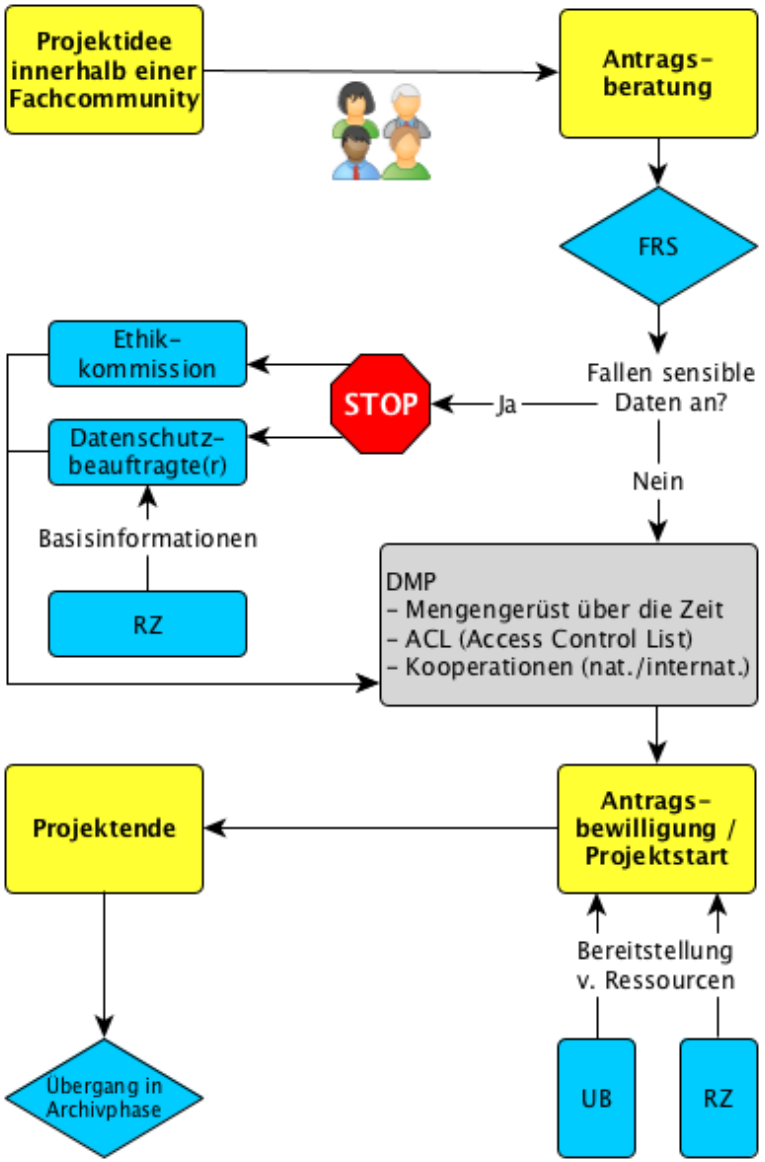

- Tasks attachable to phases in the lifecycle

# RDMG: Communication & material

- Development of a “RDM strategy for the university” by the eScience groups of the university library and the computer center (draft made available)
- OTRS queue for requests and consultation: *fdm@mail.uni-freiburg.de*
- Internal e-Learning platform for “Weiterbildung”, see https://wb-ilias.uni-freiburg.de/goto.php?target=frm_130918_10125&client_id=unifreiburgwb
- State-wide: https://forschungsdaten.info and the RDM WG (AK FDM)

# RDMG: Establishing Workflows

- Create support and advisory structures during grant application and project runtime
  - Involve FRS, ethics commission and data protection officer (if applicable)



# Licenses of research data

- Many funding institutions require the “openness” of data
  - Openness increasingly core concept for many scientific communities
  - Depending on the type of data and community
  - Requirement to use common infrastructures
- Option: Creative Commons with attributions!?
  - Part of the DMP statements
- University Library: Preexisting experience with the publication of PhD theses (publication contract)

# Support structures

- Personnel of the VO via (?)
  - Existing personnel within the VO partners and preexisting tasks (in various roles)
  - Personnel acquired via SFB Inf extensions
  - HPC storage / RDM person via the S5 project (official start 1st July)
  - Brought in 3rd party expertise (like e.g. from the state level: *forschungsdaten.info* or AK FDM)
- Will most probably not suffice in the long run

# Further steps

# Long-term Object Access

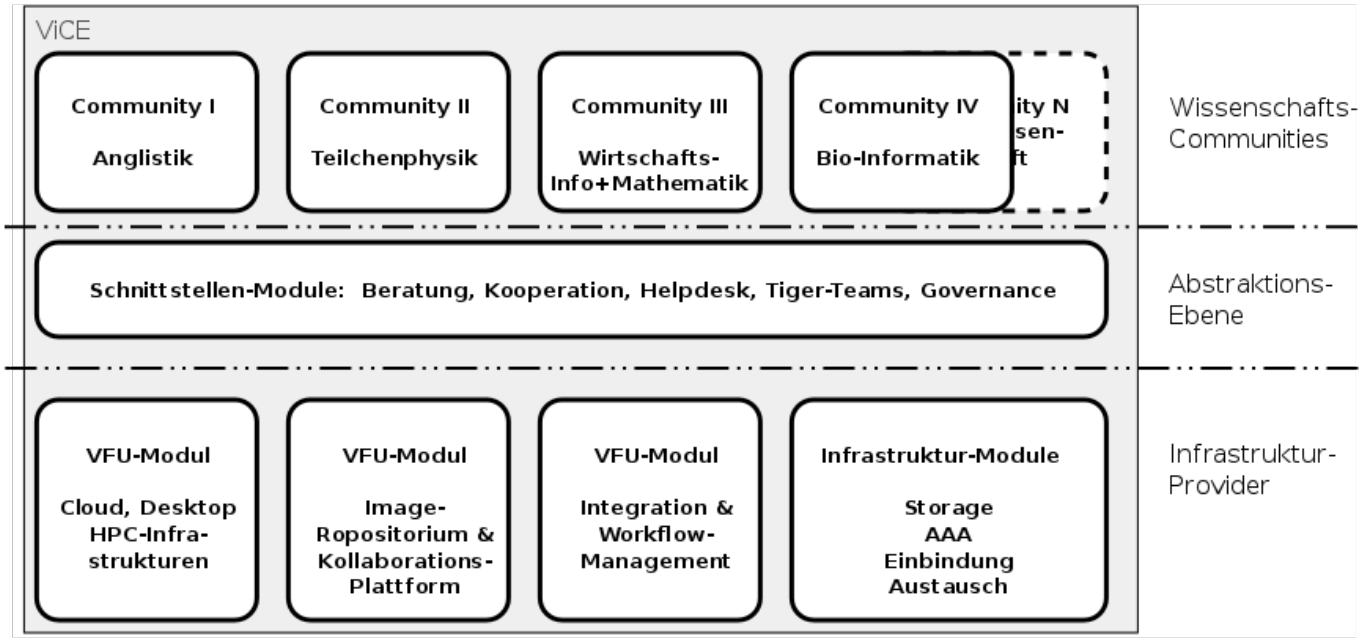
- Research data mostly useless without creation context
  - Archiving of methods and workflows
  - Virtual Research Environments can help here to abstract from concrete hardware / location
- eScience projects like CiTAR and ViCE work on citeable digital methods and workflows

ViCE
Virtual Open Science Collaboration Environment

CiTAR Citing & Archiving Research

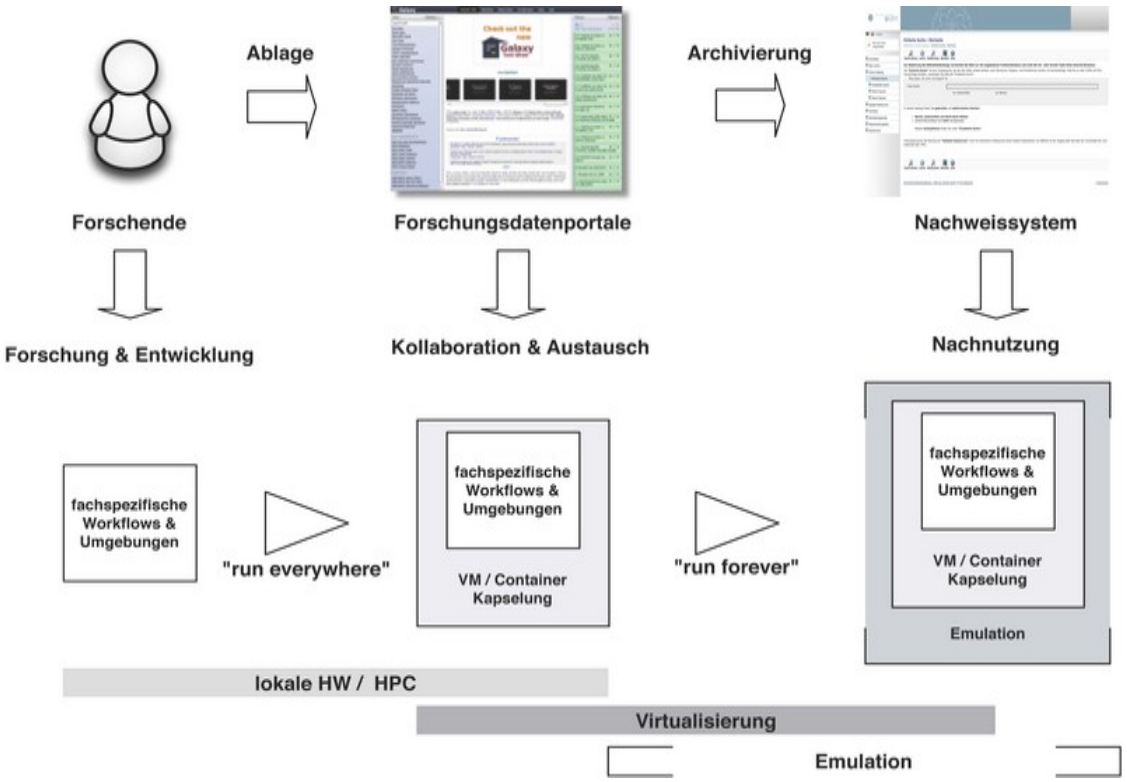
# ViCE project (2016 – 2018)

- Virtualized research environments
  - Loose the tight connection of concrete workflows from its underlying hardware
  - Usecases in particle physics for CMS and ATLAS workflows



# CiTAR (2015 – 2019)

- Cite-able research environments
  - Abstraction from underlying hardware precondition for longterm access to original environments




- Archive original research environments and relevant software
- Provide a framework to orchestrate and run past workflows
- Emulation-as-a-Service as a common base

# Field specific activities

- Data Preservation in High Energy Physics
  - Collaboration for Data Preservation and Long Term Analysis in High Energy Physics: https://hep-project-dphep-portal.web.cern.ch
  - Repository for publication-related High-Energy Physics data: https://www.hepdata.net
- Community active for quite a while (examples)
  - Paper (2009): https://arxiv.org/pdf/0912.0255.pdf
  - Paper (2010): “Data Preservation in High Energy Physics – why, how and when?”
  - Event (2014): https://indico.cern.ch/event/313634
  - ...

# Sum-up / Conclusion

- RDM becoming increasingly relevant
  - Future assets of a university
- Challenges
  - Various preexisting solutions need to get properly included, referenced, handled
  - (Internationally) Cooperating scientists vs. local institutions
  - Very diverse requirements
  - ...

!Thank you very much for your attention!