PaNOSC – WP2 Workshop
Data management plans

HZB Helmholtz Zentrum Berlin

POLICIES AND DMPS IN PANOSC AND EXPANDS TASKS

PaNOSC

- Task 2.2: Updated PaNOSC Data Policy framework
- Task 2.4: Create Guidelines
- Task 2.5: Implement DMP template

ExPaNDS

- Task 2.1: Alignment of policies and practices of EOSC relevant national and/or thematic initiatives for EOSC standards
- Task 2.2: Data Management Planning
- Task 2.3: The mainstreaming of standards for data management

HZB Helmholtz Zentrum Berlin

DMP IN ExPaNDS

- DMP on instrument level

- Leverage costs for experiments
  - **automatically populating metadata information based on the proposal and instrument information**

- Approach to active DMPs,
  - **integrating the DMP information into data lifecycle and metadata collections, and for policy enforcement and reporting**

HZB Helmholtz Zentrum Berlin

WHAT TO EXPECT FROM DMPs?

A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed.(Wikipedia)

- Helps thinking about how to organise the data
- Implementation of policies, decisions/agreements, workflows, rules
- Information source for stakeholders

HZB Helmholtz Zentrum Berlin

WHAT TO EXPECT FROM DMPs?

A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed.(Wikipedia)

- Helps thinking about how to organise the data
- Implementation of policies, decisions/agreements, workflows, rules
- Information source for stakeholders

A DMP needs to be:
Created, implemented, executed, exposed, validated

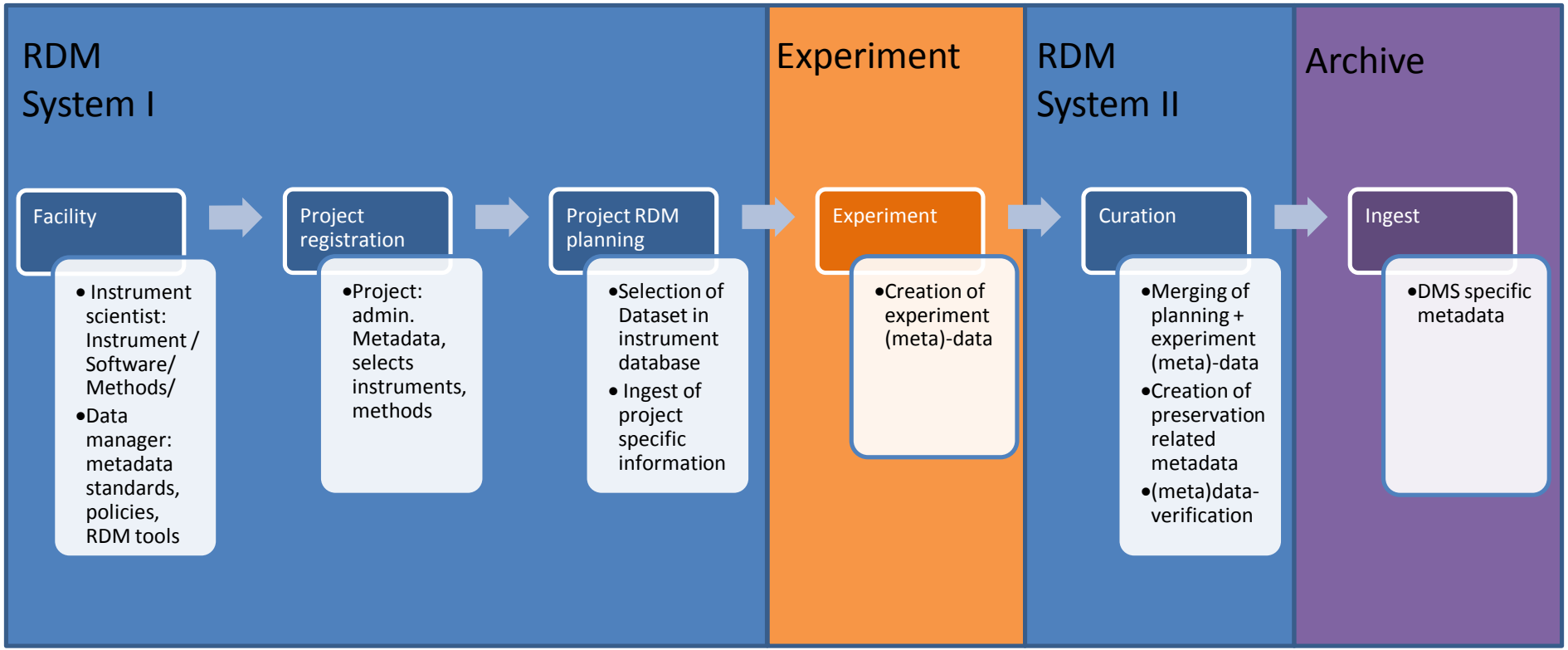
HZB Helmholtz Zentrum Berlin

DMP THEMES (DCC & UC3 - DMPRoadmap)

- Data description
- Data Format
- Data Volume
- Data Collection
- Metadata & Documentation
- Ethics & Privacy
- IPR
- Storage & Security
- Data Sharing
- Data Repository
- Preservation
- Roles & Responsibilities
- Costs
- Related Policies

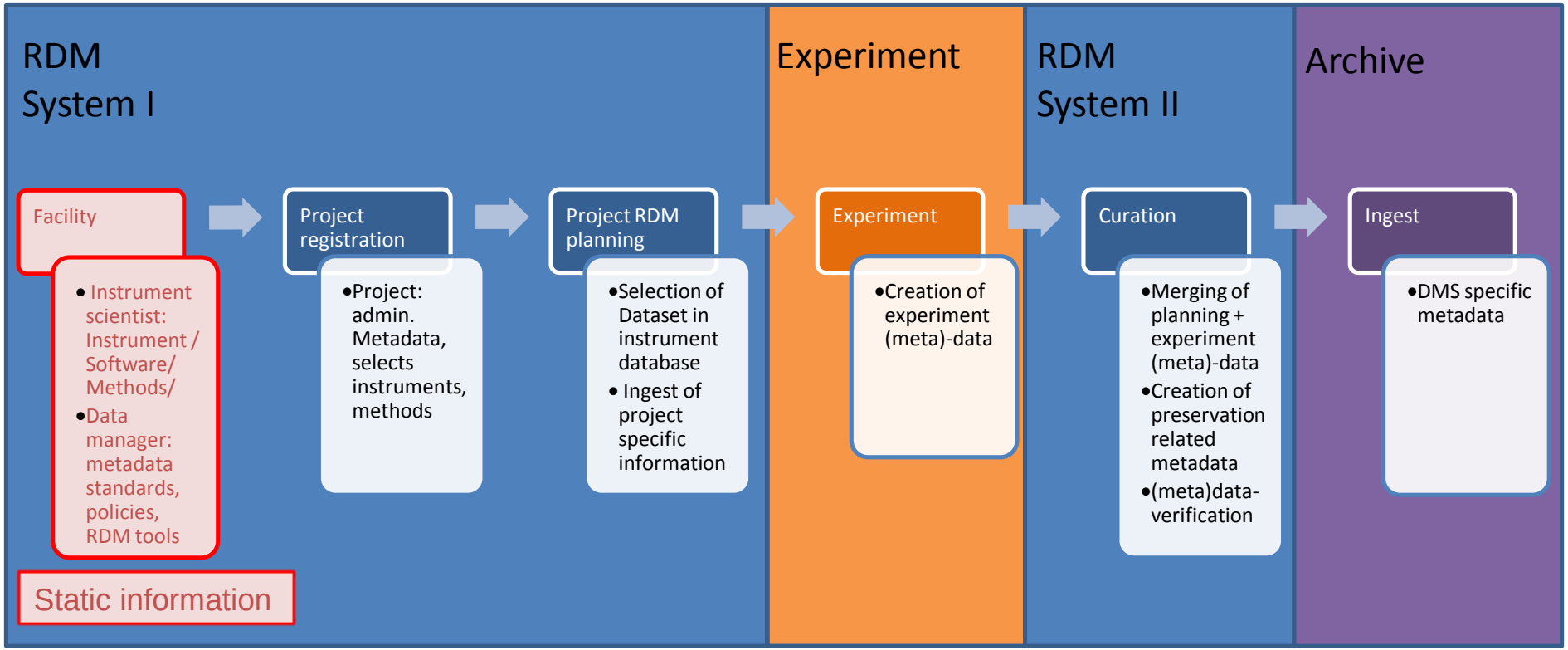
HZB Helmholtz Zentrum Berlin

RDM WORKFLOW AND KNOWLEDGE SOURCES



HZB Helmholtz Zentrum Berlin

RDM WORKFLOW AND KNOWLEDGE SOURCES



HZB Helmholtz Zentrum Berlin

WHAT ARE POLICIES?

- A policy is a deliberate system of principles to guide decisions and achieve rational outcomes. A policy is a statement of intent, and is implemented as a procedure or protocol. (Wikipedia)

- Policies themselves are often natural language documents that are not implementable on their own. A procedure needs to be followed that results in implementable processes that enforce the policy with each workflow corresponding to a particular policy statement. (SHAMAN, 2011)

HZB Helmholtz Zentrum Berlin

ROLES IN POLICY CREATION AND IMPLEMENTATION

Policy creators

- Policy and law makers
  International, national, regional, institutional, facility, department

Requirement definitions

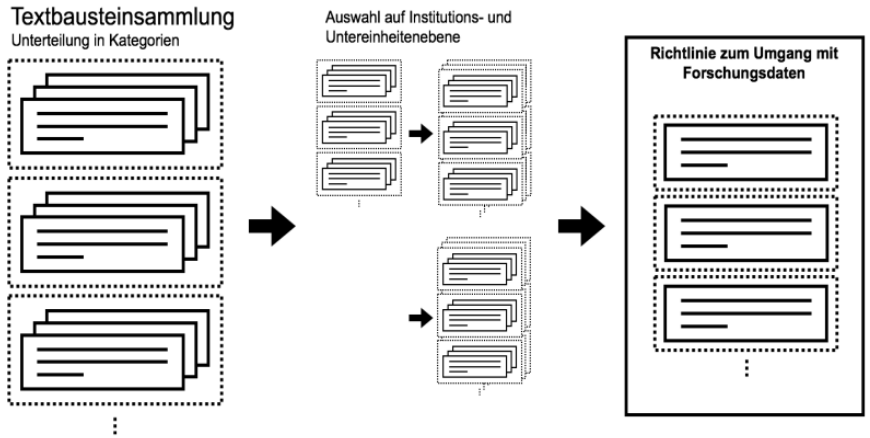
- Data users
  Scientists, reviewers, publishers
- Scientific community
- Data management experts
- Legal experts

Policy implementers

- Data producers
  Instrument scientists, researchers
- Data management
  Repository, catalogue, format, metadata, preservation

DATA POLICY KIT (CARL ALBRECHTS UNIVERSITY - KIEL)

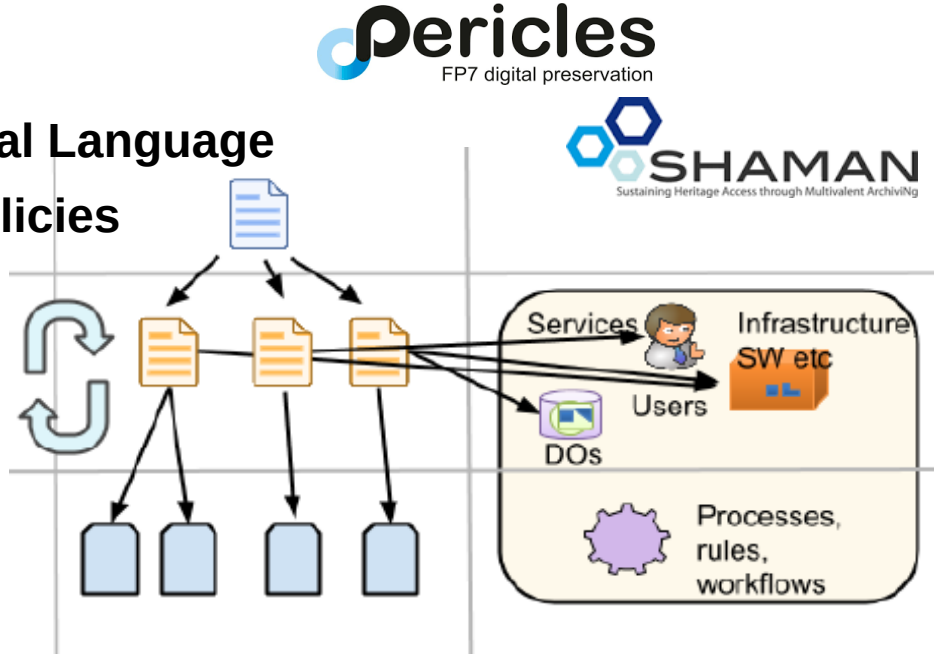
- There is a limited set of propositions all over the different policy documents
- The propositions might have different parameters
- The text blocks of the propositions can be used to create a policy document



HZB Helmholtz Zentrum Berlin

POLICY TO PROCESS DERIVATION

- three Levels of Abstraction
  - **Natural Language Policies**
  - **Intent and Constraints in Natural Language**
  - **Concrete Implementation of Policies**
- References Ecosystem Model

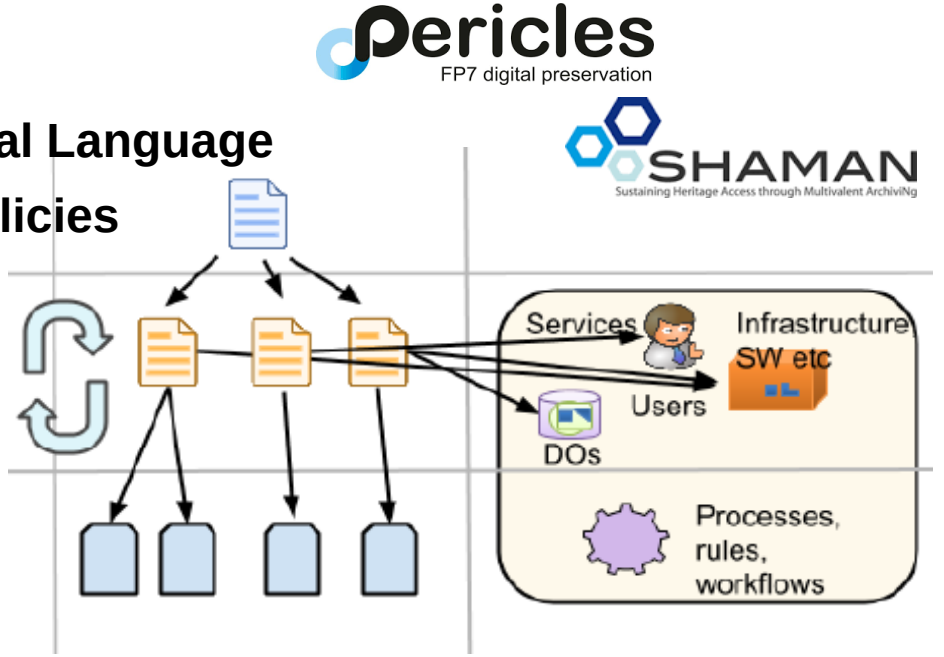


HZB Helmholtz Zentrum Berlin

POLICY TO PROCESS DERIVATION

- three Levels of Abstraction
  - **Natural Language Policies**
  - **Intent and Constraints in Natural Language**
  - **Concrete Implementation of Policies**
- References Ecosystem Model

A DMP (when executed) will implement these rules, processes and workflows.



PLACES OF POLICY APPLICATION

- Creation/Processing
- Storage
- Ingest
- Archival
- Dissemination
- Use

They are resulting e.g. in formats, file structures, metadata, storage, access rights, access protocols …

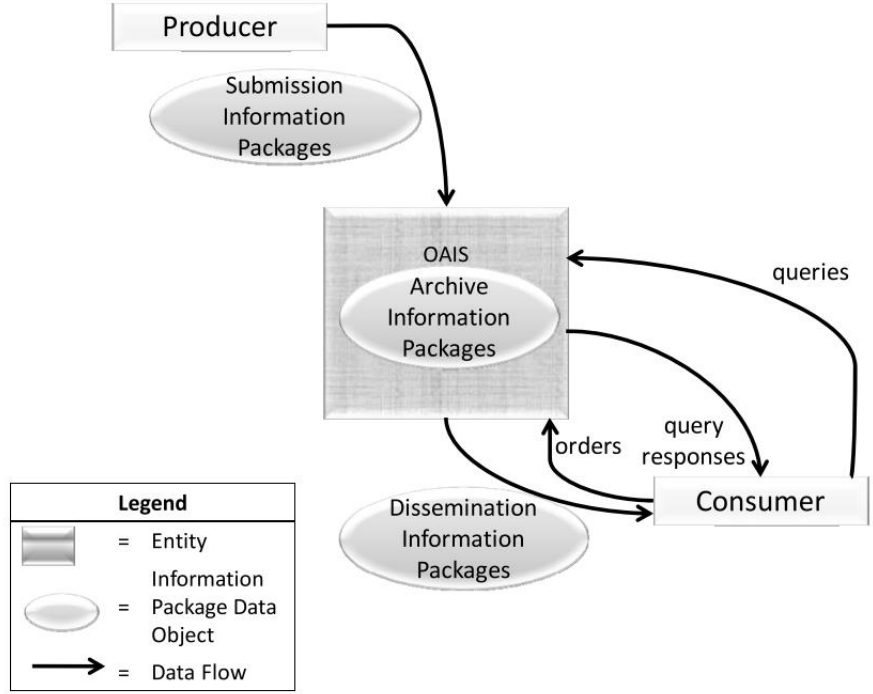


Figure 2-4: OAIS Archive External Data

HZB Helmholtz Zentrum Berlin

POLICY ACTIVITIES ON DATASET

RDA - WG PRACTICAL POLICIES

- Contextual metadata extraction
- Data access control
- Data backup
- Data format control
- Data retention
- Disposition
- Integrity (including replication)
- Notification
- Restricted searching
- Storage cost reports
- Use agreements

Data format control policy template

Format Requirements	**Constraint**	**State attributes for Constraint**
	On ingestion of file	
	Periodic check	Time interval between checks
	For specific format type	File-format_type
	For collection	Collection_name
	Operations	**State Attributes for Operation**
	Set file format	File_ID
		File_format_type
	Get file format	File_ID
		File_format_type
	Check file format	File_ID
		File_format_type
	Convert file format	File_ID
		File_format_type
		Desired_file_format_type
	Verify file format	Collection_name
		File_name
		File_format_type
		Desired_file_format_type

HZB Helmholtz Zentrum Berlin

STEPS FROM POLICY TO DMP

General policy by **funder/facility/department**:

➢ data has to be FAIR

Just one part: data has to be **interoperable**

Instrument scientist is responsible for implementation:

- Needs to ask **users and scientific community** about requirements for interoperability
- Gets help from **IT and data management** for implementation

➢ Policy results in workflow (verify or convert format)

POLICIES AND DMPs

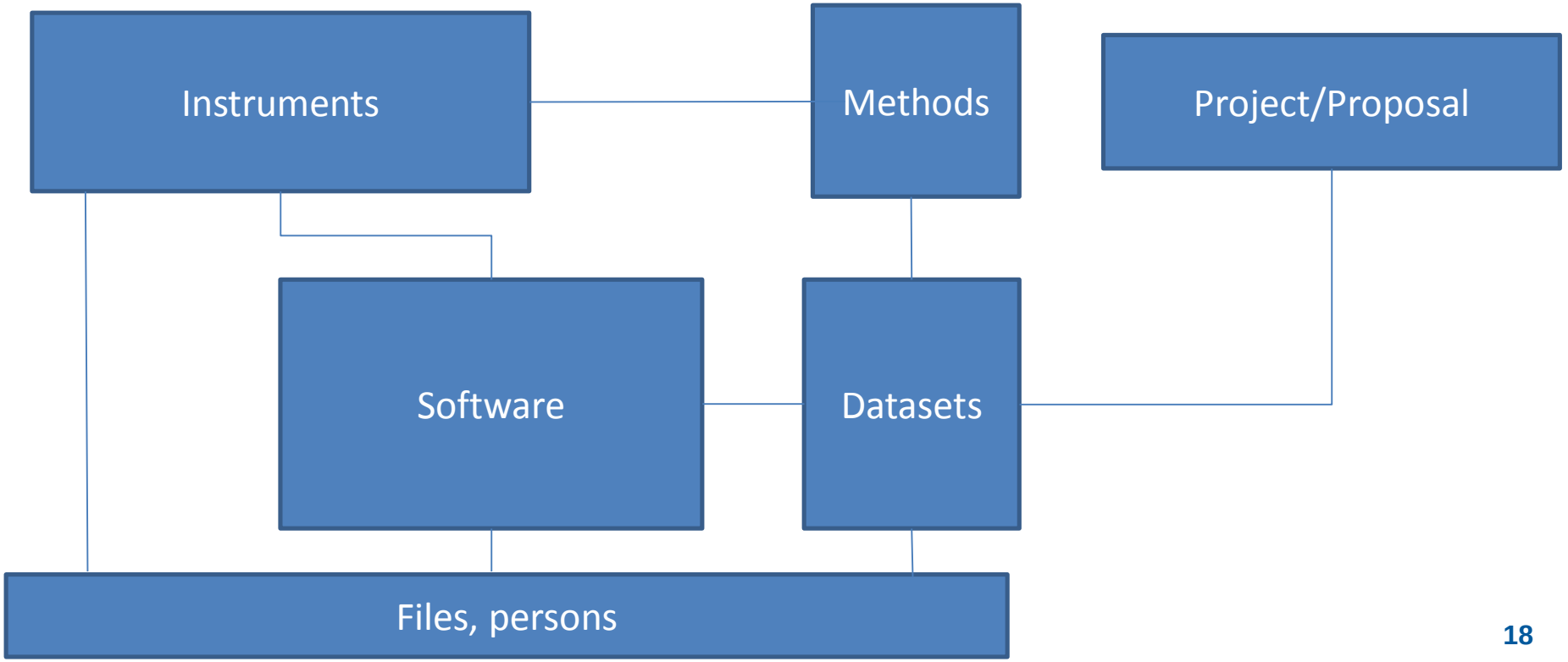
- **Policy implementation requires data management planning**

- **Static DMP answers at instrument**

- Is this dataset interoperable?
- What format has your dataset?
- What metadata do you use?
- How is your dataset structured?

HZB Helmholtz Zentrum Berlin

DATA SCHEMA INSTRUMENT DATABASE



JSON EXPORT FROM THE INSTRUMENT DATABASE

- Project with dataset
- A dataset consists of one or many filecolletions


```
"project": {
    "name": "Electronic Ground States of Fe2+ and Co2+",
    "description": "To determine the ground state of metall ions.",
    "fundRef": "DFG:123123213",
    "members": [
    ],
    "start_date": "2019-08-02",
    "end_date": "2020-08-02",
    "disciplines": [
    ],
    "jurisdictions": [
    ],
    "datasets": [
        {
            "id": "rdminfopool:datenmgt.basisit.de/dataset/?",
            "name": "UE52_Experiment_LISE_LabBook",
            "instrument": {
                "name": "UE52 Nanocluster Trap",
                "id": "rdminfopool:datenmgt.basisit.de/instrument/11",
                "alternate_id": [
                    "doi:1000/10",
                    "igama:1848"
                ],
                "type": "Beamline fixed endstation"
            },
            "policies": [
            ],
            "methods": [
            ],
            "metadataschema": [
            ],
            "filecollections": [
```

HZB Helmholtz Zentrum Berlin

JSON EXPORT FROM THE INSTRUMENT DATABASE

- A filecollection


```
"filecollections": [
    {
        "id": "rdminfopool:datenmgt.basisit.de/dataset/filecollection/8",
        "name": "UE52 NanoclusterTrap LISE Collection",
        "policies": [
        ],
        "protection": [
        ],
        "members": [
        ],
        "instrument": {
        },
        "hardware": {
        },
        "reading_software": [
        ],
        "software": {
        },
        "files": [
        ]
    },
```

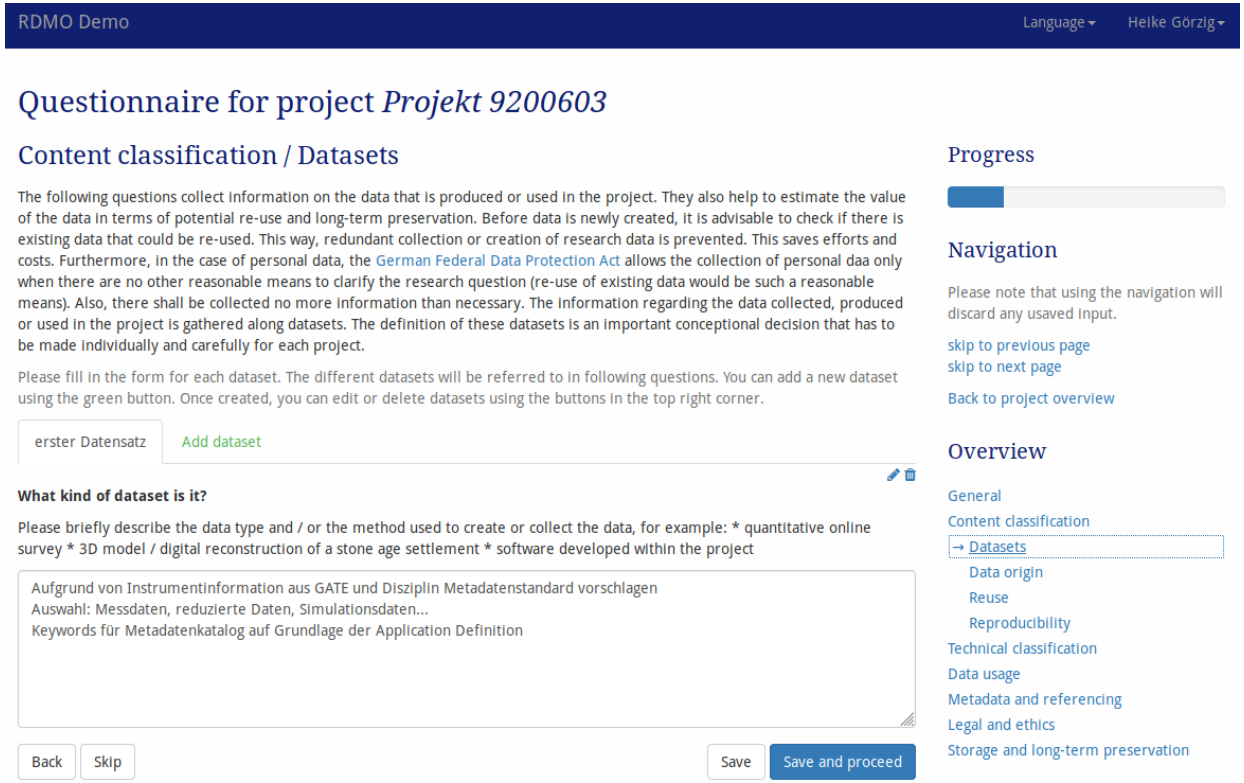
HZB Helmholtz Zentrum Berlin

POLICY MAPPING TO EXECUTABLE WORKFLOW

```
{
    "operation": "Get mime type",
    "origin": [
        "Instrument1 data policy",
        "Project1 DMP"
    ],
    "constraint": {
        "type": "event",
        "value": "onCreation"
    },
    "parameters": [
        "$file_path_in"
    ],
    "categories": [
        "format",
        "extract"
    ],
    "description": "Retrieving the mime type."
},
```

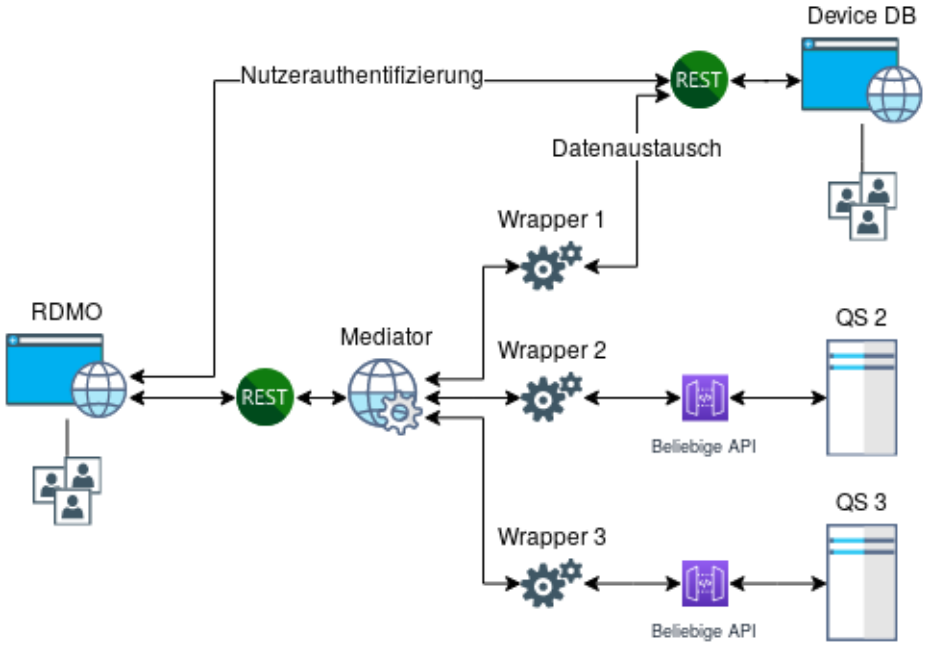
```
0:
    id:          "c2147c30.a82ad"
    type:        "subflow"
    name:        "GetMimeType"
    info:        ""
    category:    ""
    in:          []
    out:         []
1:
    id:          "94c3d5d7.b5c488"
    type:        "exec"
    z:           "c2147c30.a82ad"
    command:     "sh '/home/heike/fits-1.4.0/fits.sh'  -xc  -i "
    addpay:      true
    append:      ""
    useSpawn:    "false"
    timer:       ""
    oldrc:       false
    name:        "Datei analyse"
    x:           380
    y:           140
  wires:
    0:
      0:         "70bdbf8.2085dc"
      1:         []
      2:         []
2:
    id:          "e49514ca.fc11c8"
    type:        "watch"
    z:           "c2147c30.a82ad"
    name:        "Home"
    files:       "/home/heike/test-folder/files"
    recursive:   ""
    x:           50
    y:           140
  wires:
    0:
      0:         "bdd04557.3cab18"
```

TOOLS - RDMO

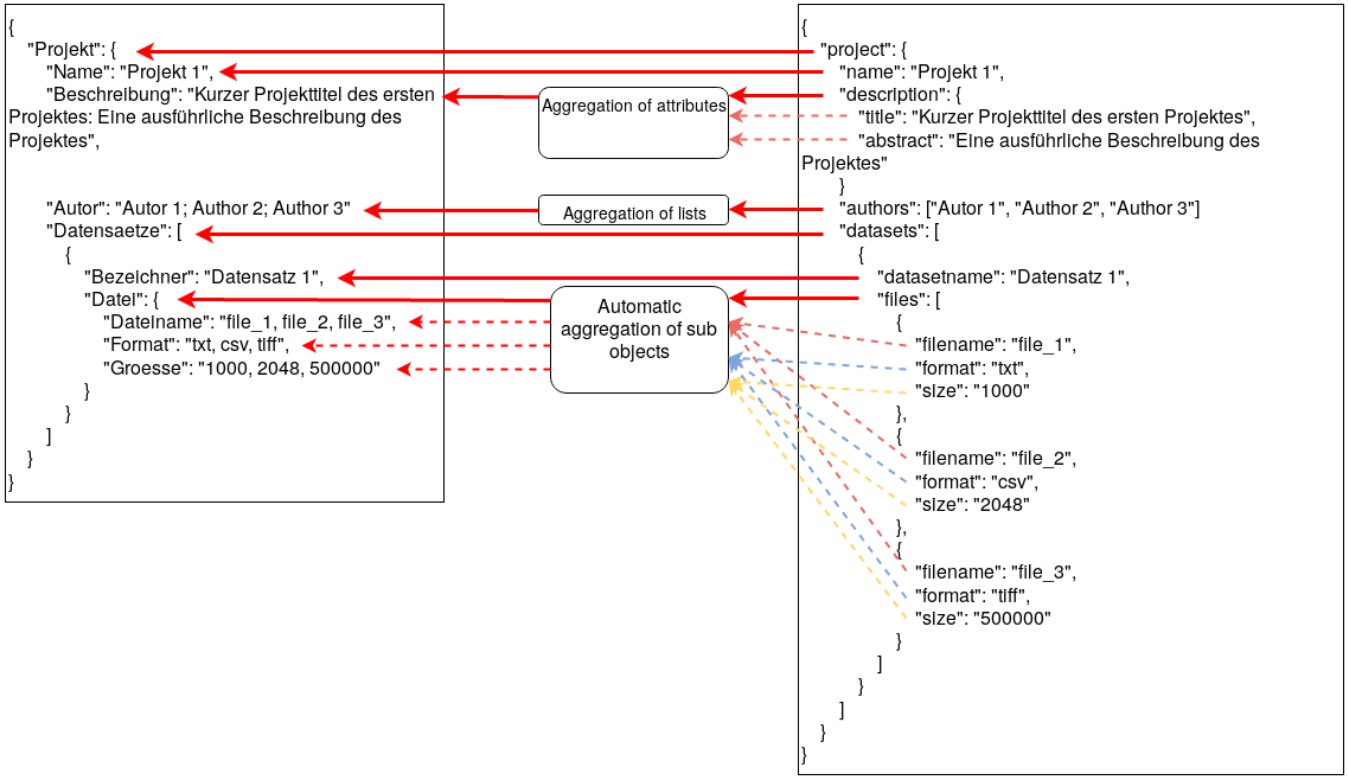

RDMO Demo
Language
Heike Görzig
Questionnaire for project *Projekt 9200603*
Content classification / Datasets
The following questions collect information on the data that is produced or used in the project. They also help to estimate the value of the data in terms of potential re-use and long-term preservation. Before data is newly created, it is advisable to check if there is existing data that could be re-used. This way, redundant collection or creation of research data is prevented. This saves efforts and costs. Furthermore, in the case of personal data, the German Federal Data Protection Act allows the collection of personal daa only when there are no other reasonable means to clarify the research question (re-use of existing data would be such a reasonable means). Also, there shall be collected no more information than necessary. The information regarding the data collected, produced or used in the project is gathered along datasets. The definition of these datasets is an important conceptional decision that has to be made individually and carefully for each project.
Please fill in the form for each dataset. The different datasets will be referred to in following questions. You can add a new dataset using the green button. Once created, you can edit or delete datasets using the buttons in the top right corner.
erster Datensatz
Add dataset
What kind of dataset is it?
Please briefly describe the data type and / or the method used to create or collect the data, for example: * quantitative online survey * 3D model / digital reconstruction of a stone age settlement * software developed within the project
Aufgrund von Instrumentinformation aus GATE und Disziplin Metadatenstandard vorschlagen
Auswahl: Messdaten, reduzierte Daten, Simulationsdaten...
Keywords für Metadatenkatalog auf Grundlage der Application Definition
Back
Skip
Save
Save and proceed
Progress
Navigation
Please note that using the navigation will discard any usaved input.
skip to previous page
skip to next page
Back to project overview
Overview
General
Content classification
→ Datasets
Data origin
Reuse
Reproducibility
Technical classification
Data usage
Metadata and referencing
Legal and ethics
Storage and long-term preservation

HZB Helmholtz Zentrum Berlin

CONNECTING SOURCES TO RDMO



MAPPING RDMO INSTRUMENT DATABASE



POSSIBLE NEXT STEPS

- Define where and on which level agreements and decisions are required
  - **Goals: FAIR, PID Graph, reproducebility, sample integration, calibration ….**
- Relate DMP questions and Policy blocks (propositions) to
  - **Roles**
  - **Infrastructure**
  - **Activities/Workflows**
- Determine constancy and reusability of
  - **Information on DMPlanning**
  - **Workflows**
  - **Tools**

THANK YOU !

HZB Helmholtz Zentrum Berlin

DISTRIBUTED KNOWLEDGE IN DMPs

- DMP questions vs. knowledge source to answer question

	project research team	project/ institutional administration	device/ instrument/ software responsible	institutional data manager	disciplinary community
Number of naming	62	8	52	44	4
Only column where named	16	4	18	22	0
Preparing project final decision		4	23	26	1