Enabling a data management system to support the good laboratory practice

Master Thesis Final Report – Miriam Ney (09.06.2011)

Deutsches Zentrum für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

Freie Universität Berlin

Overview

- Description of Task
- Phase 1: Requirements Analysis
  - Good Laboratory Practice
  - Scientific Workflow
  - Laboratory Notebook
  - DataFinder
- Phase 2: Implementation
  - DataFinder Configuration
  - Origin of Data – Provenance Integration
  - Evidential Archiving
  - Signing Data
- Phase 3: Evaluation
- Conclusion

Description of Task:
Enabling a data management system to support the good laboratory practice

- Data management system: DataFinder
  - Open Source Project developed by DLR
  - Data management and workflow management
  - Meta data handling

- Good laboratory practice (GLP)
  - Scientific conduct of experiments
  - Regulatories from DFG, OECD, Universities, …
  - Part of GLP: laboratory notebook

Phase 1: Requirements analysis
What is the good laboratory practice?

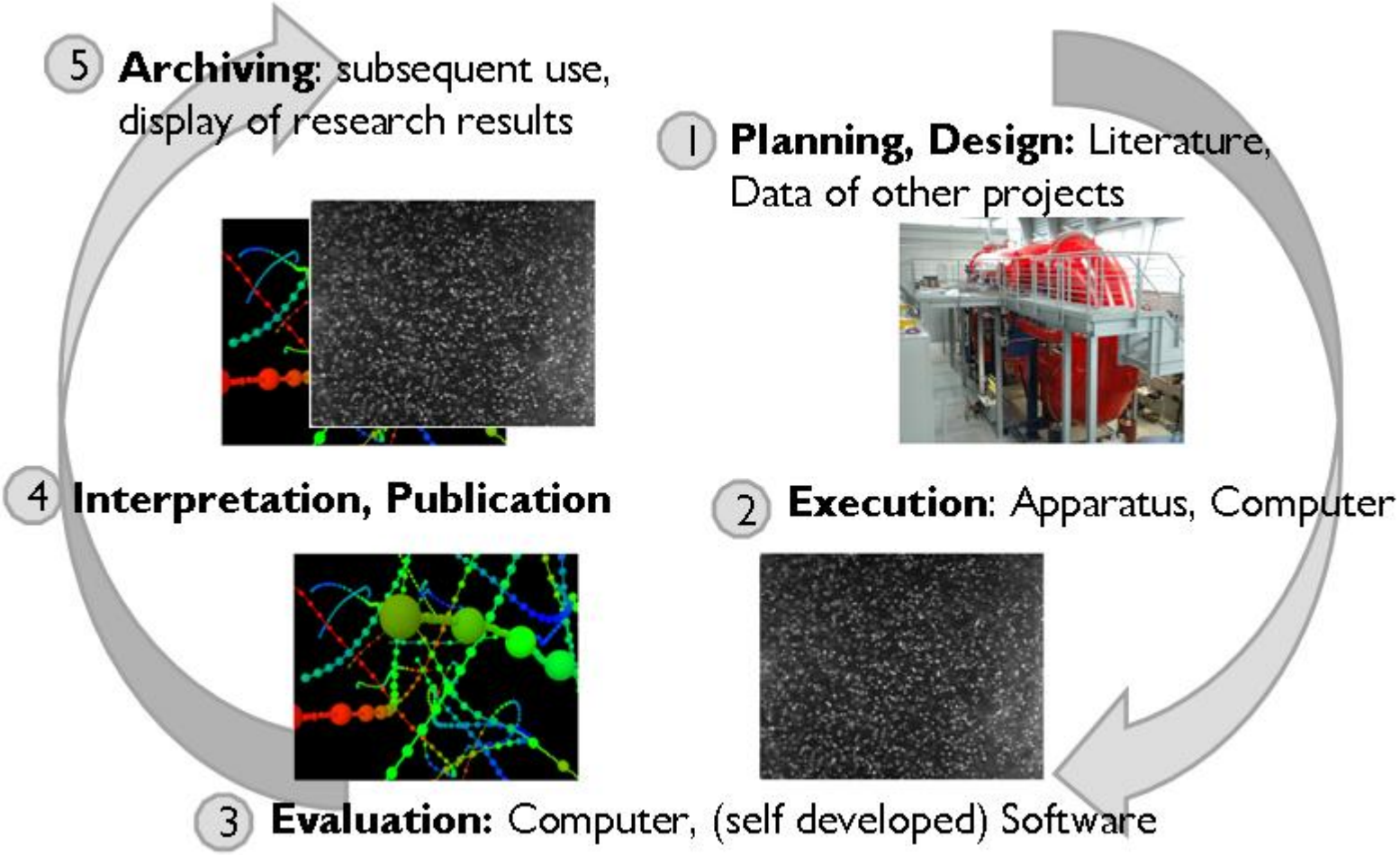
The principles of Good Laboratory Practice (GLP) have been developed to ***promote the quality and validity*** *of test data used for determining the safety of chemicals and chemicals products.*

OECD Principles on Good Laboratory Practice
(as revised in 1997)

[The recommendations] are designed to provide ***a framework for the deliberations and measures*** *which each institution will have* ***to conduct*** *for itself according to its constitution and its mission*

Deutsche Forschungsgemeinschaft:
Sicherung guter wissenschaftlicher Praxis (Safeguarding good scientific practice) 1998 (p.50).

Phase 1: Requirements analysis
What does a scientific workflow look like?



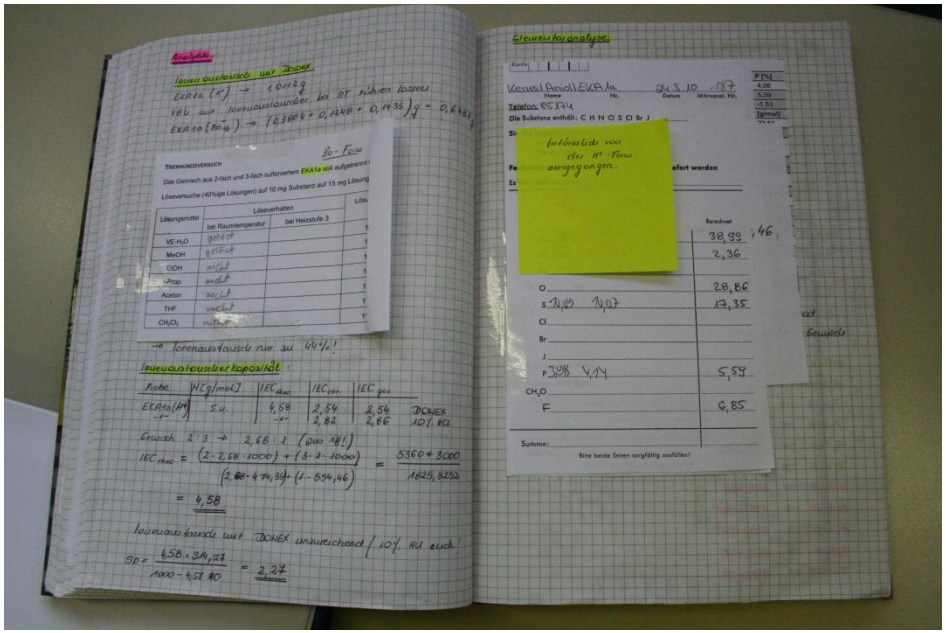
Picture adapted from: www.belab-forschung.de

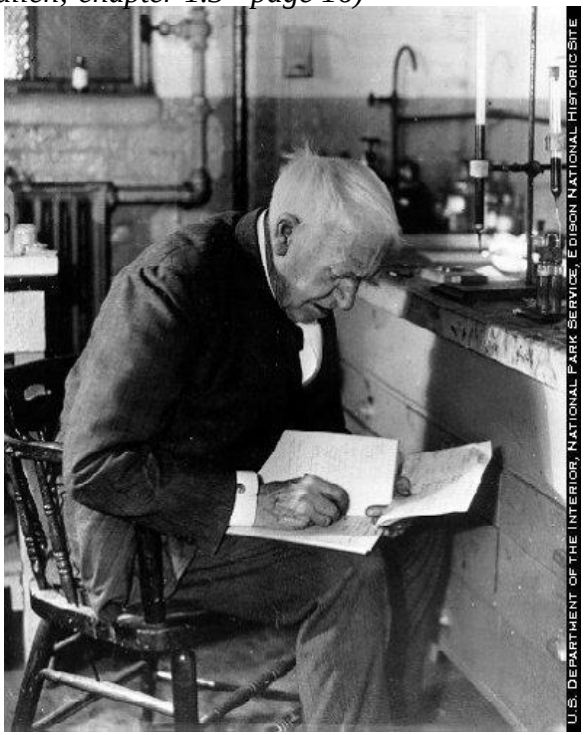
Phase 1: Requirements analysis: What is a laboratory notebook?

"Das Laborbuch ist ein Tagebuch des experimentierenden Naturwissenschaftlers"

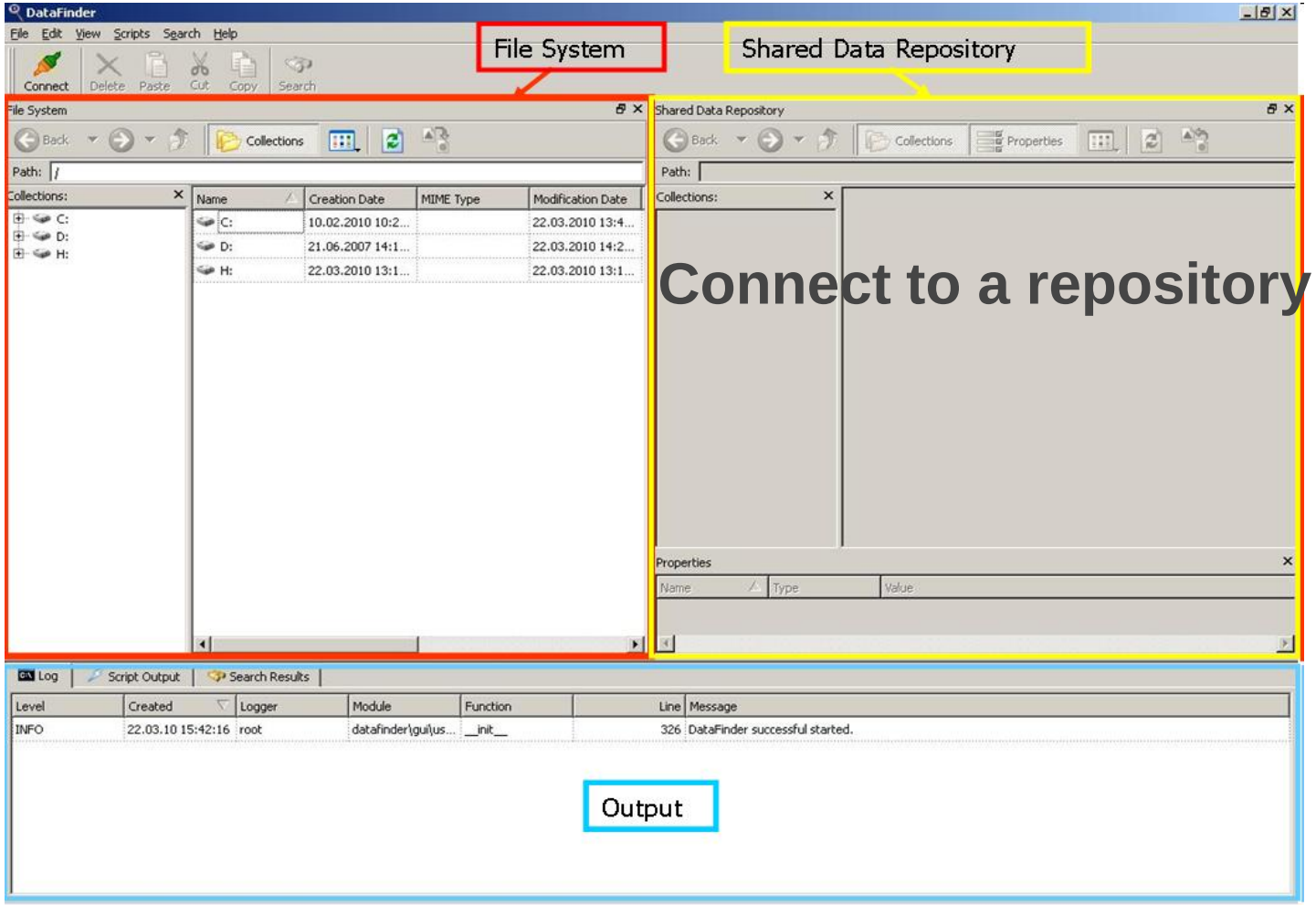
(The laboratory notebook is the diary of the experimenting scientist)

(Schreiben und Publizieren in den Naturwissenschaften
Von Hans F. Ebel,Claus Bliefert,Walter Greulich; chapter 1.3 - page 16)



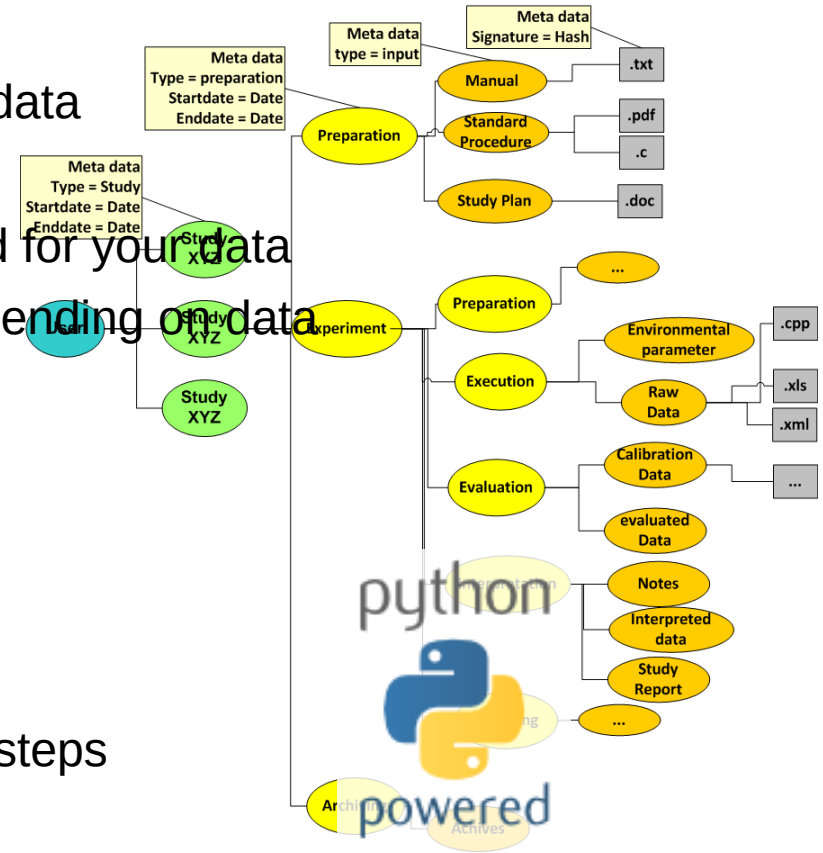


Phase 1: Background
What is the DataFinder?



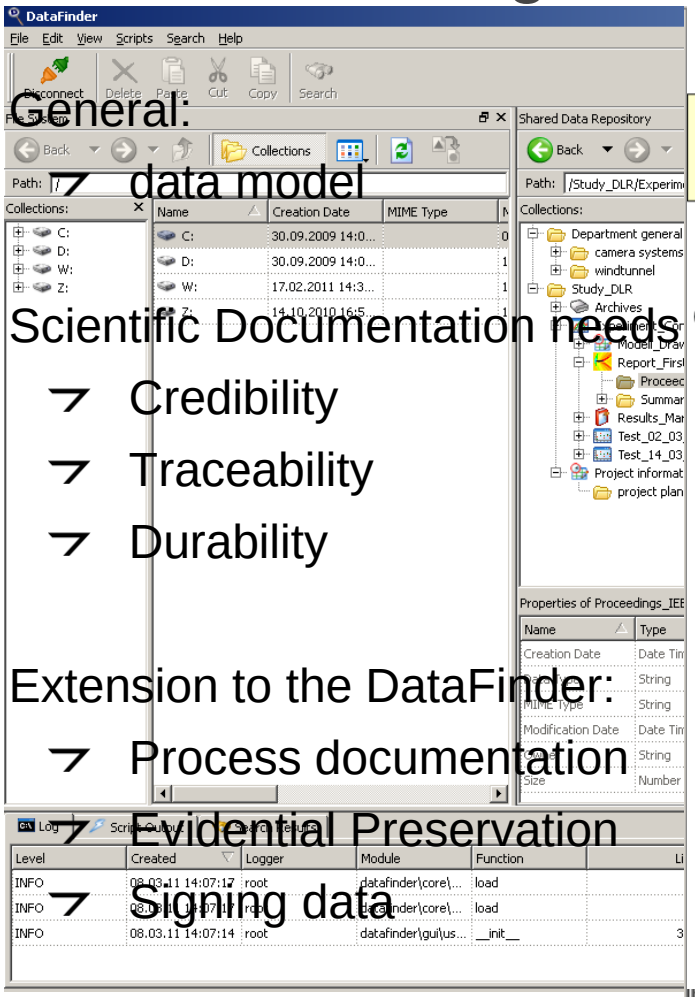
Phase 1: Background
What is special about the DataFinder?

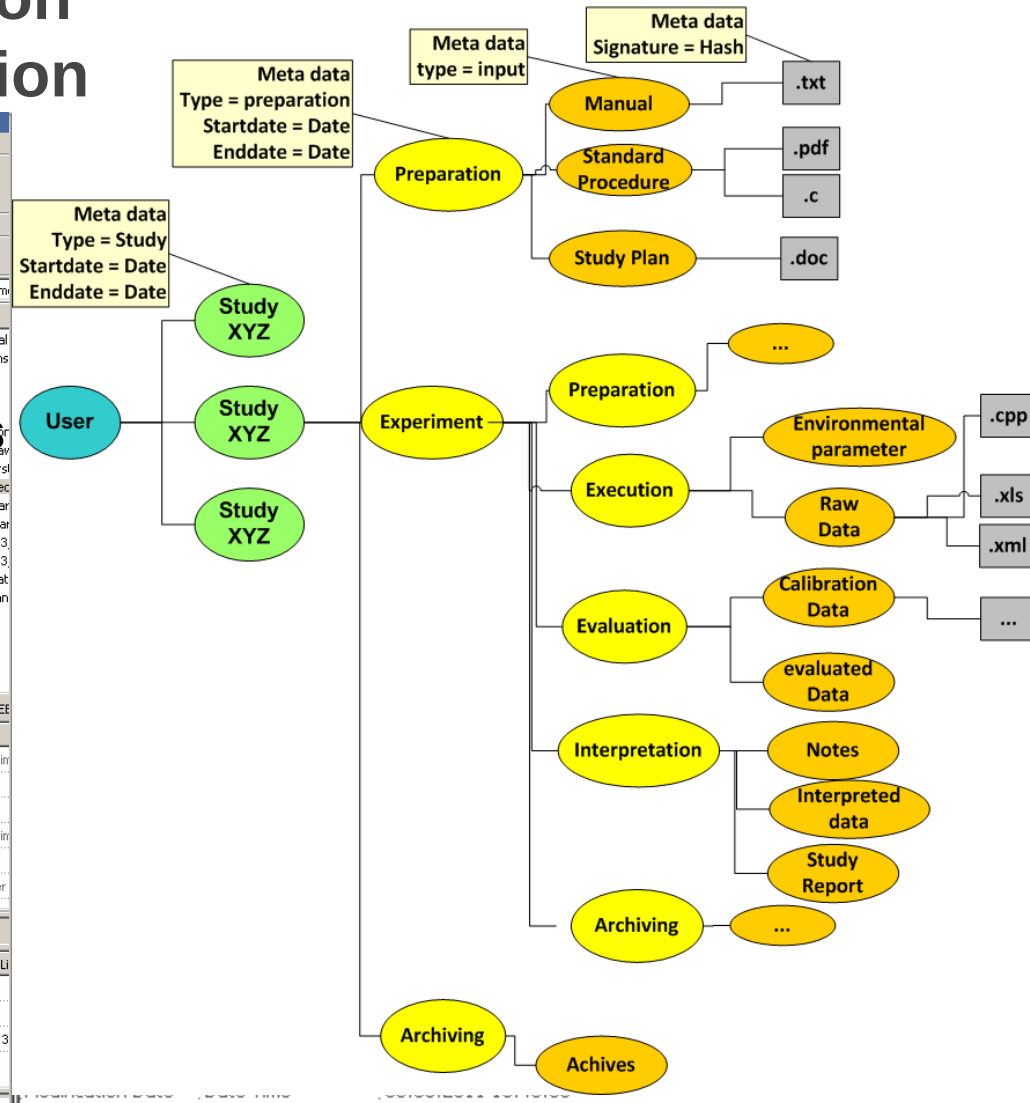
- Structuring of data in a standardized way through a data model
  - Restricting the user to a layout
  - Forcing the user to enter meta data
- Using heterogenous storage backend for your data
  - Best fitting storage solution depending on data
  - Existing solutions can be kept
  - Offline storages possible
- Extension by scripts
  - Adjusting to your environment
  - Automation of data processing steps



Phase 2: Implementation DataFinder Configuration

- General:
  - data model
- Scientific Documentation needs:
  - Credibility
  - Traceability
  - Durability
- Extension to the DataFinder:
  - Process documentation
  - Evidential Preservation
  - Signing data





Phase 2: Implementation
Feature 1: Process documentation – Provenance Integration

Provenance (lat. provenire = to come from): origin of data, source

Motivation:

- Credibility
- Chain of events
- Integrity

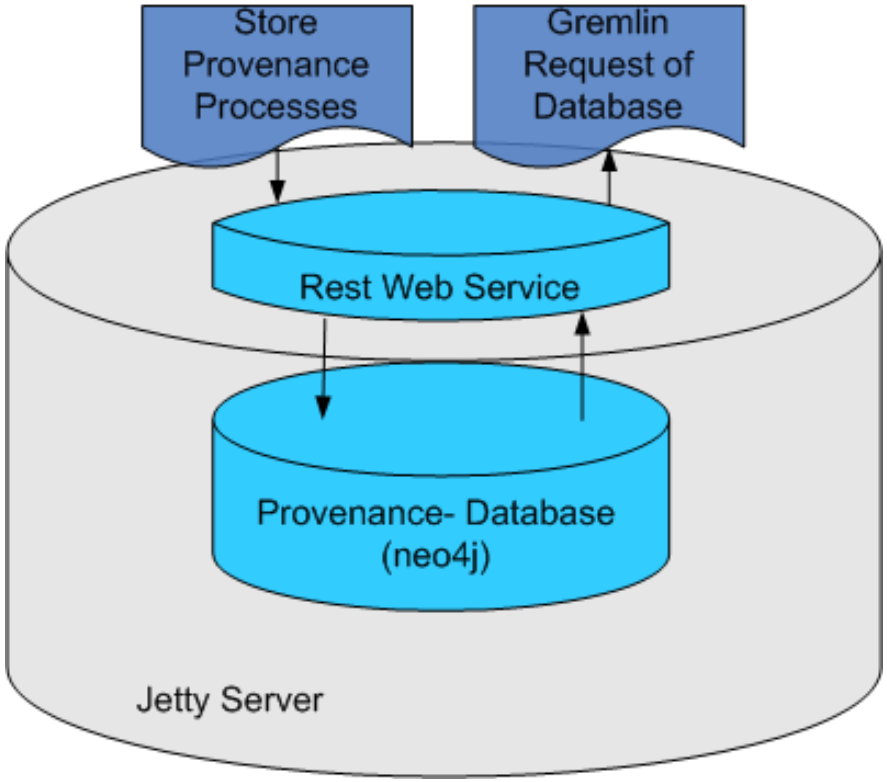
Tasks:

- Developing provenance model for the „Good Laboratory Practice“
  - PrIMe and OPM
- Enhance provenance storing system „noblivious“
  - Graph database with REST interfaces
- Integration into DataFinder
  - Developing Python script extension

Feature 1: Origin of Data – Provenance Integration
Developing the OPM model

Feature 1: Origin of Data – Provenance Integration Provenance Storing System

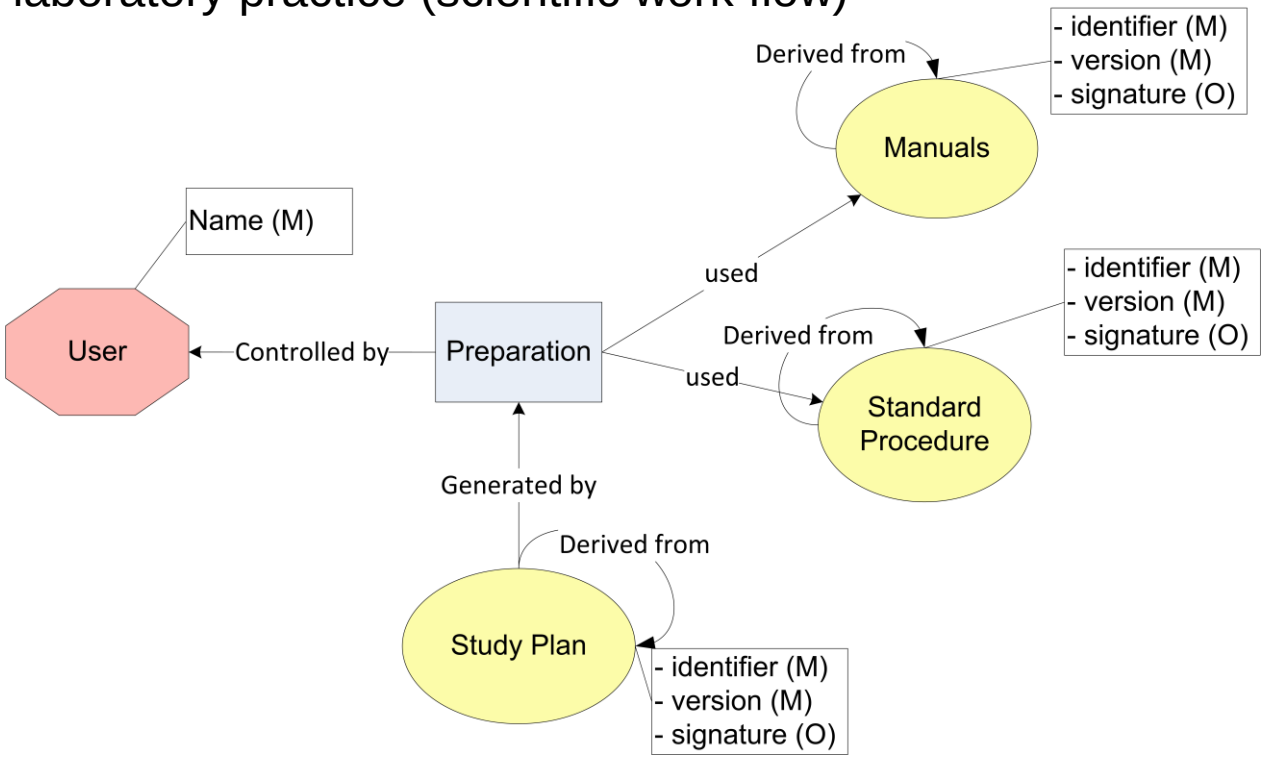
- Java Implementation
- Server: Jetty
- Graph Database: neo4j
- Interfaces
  - Storing Provenance (REST)
  - Extracting Provenance (REST)
  - Extracting Provenance (Servlet)
- Open Source (Apache License Version 2.0)



Feature 1: Origin of Data – Provenance Integration DataFinder Extension: Process documentation

Goal:

- Documentation of the data items added: according to the good laboratory practice (scientific work flow)

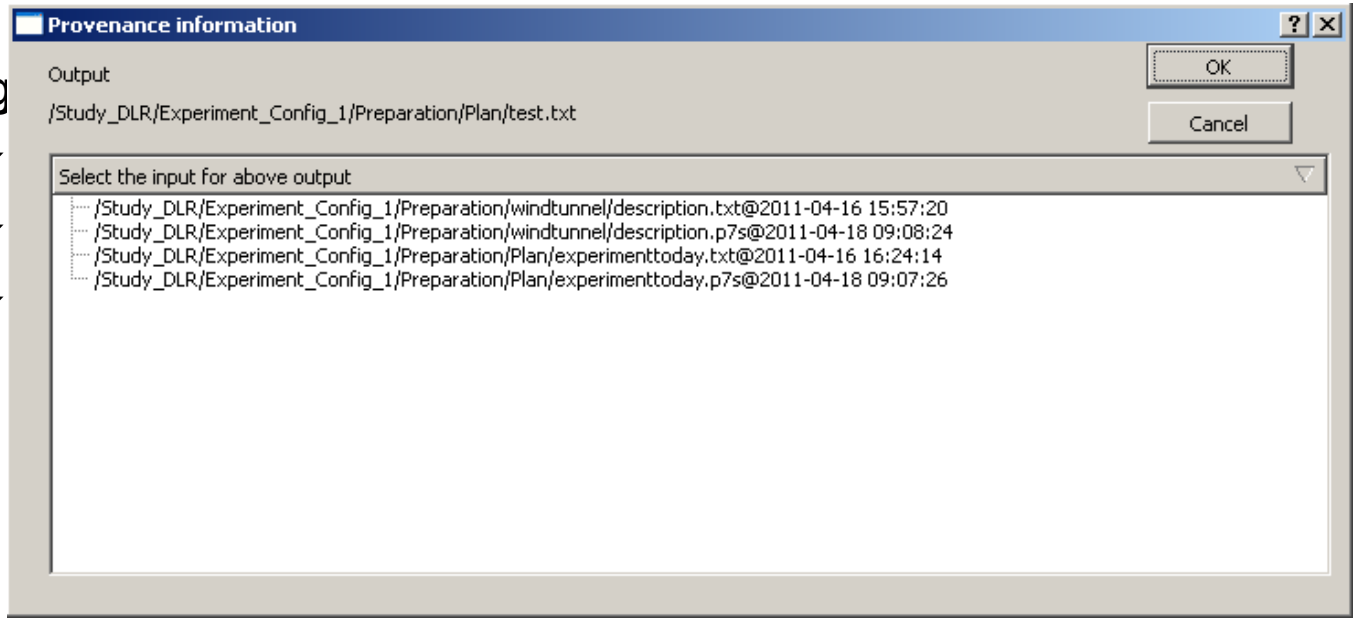


Feature 1: Origin of Data – Provenance Integration DataFinder Extension: Process documentation

Realization:

- User imports File (output) to DataFinder repository
- Active Import Listener reacts and dialog for input pops up
- Information on input and output is sent to provenance store

Chang



Phase 2: Implementation
Feature 2: Evidential Preservation

Motivation:

„Recommendation 7: Primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin.“

Deutsche Forschungsgemeinschaft:

Sicherung guter wissenschaftlicher Praxis (Safeguarding good scientific practice) 1998 (p.55).

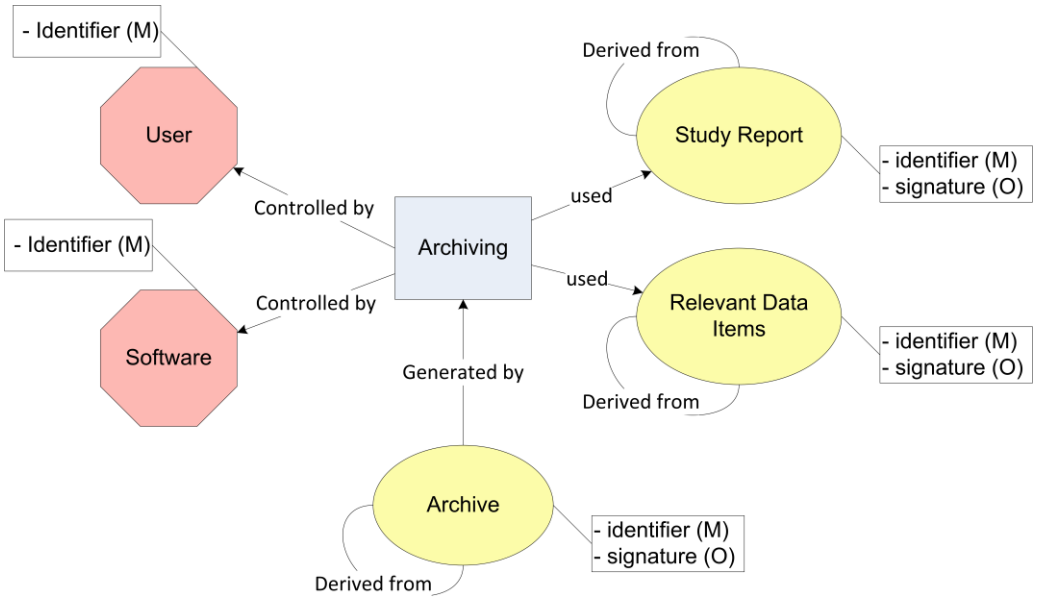
Tasks:

- Evaluation of important archiving characteristics
  - Extraction of relevant data – using provenance
- Integration of a preservation service
  - Connecting to DataFinder

Feature 2: Evidential Preservation
DataFinder Extension: Use of Provenance Information

Goal:

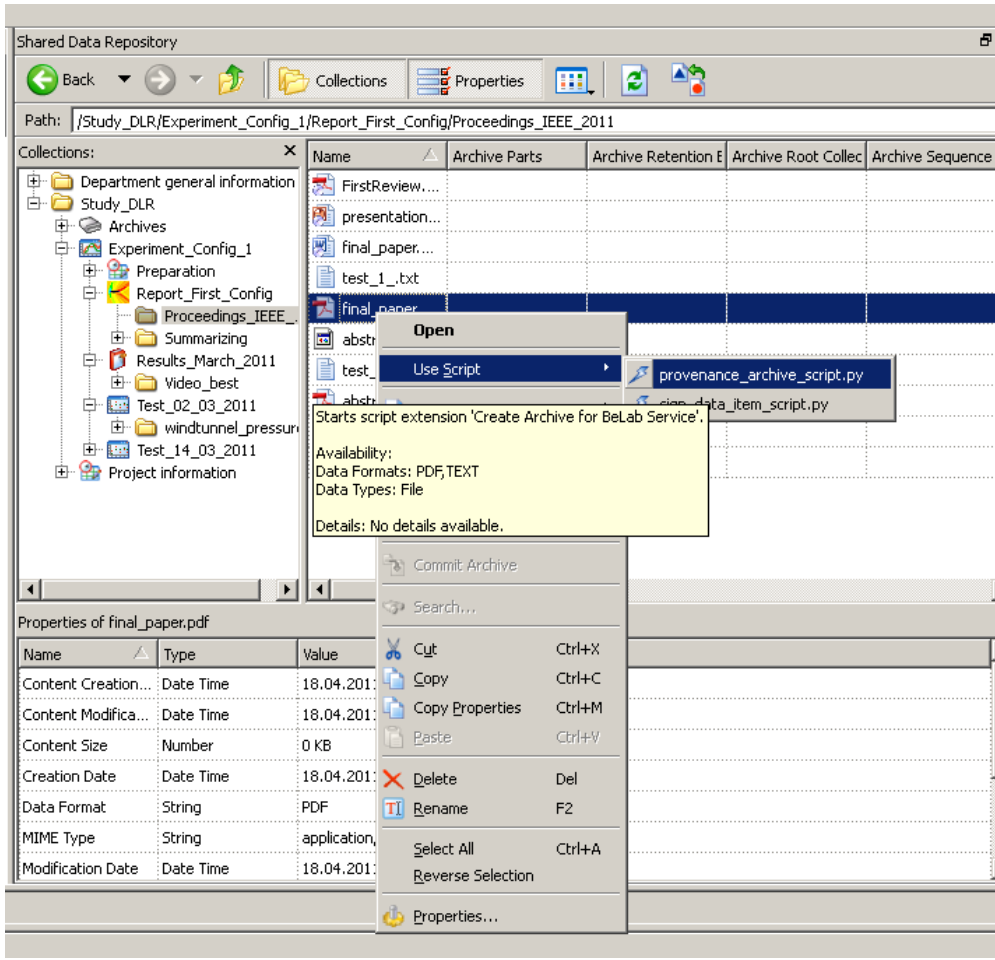
- Extraction of data relevant for the preservation process



Feature 2: Evidential Preservation
DataFinder Extension: Use of Provenance Information

Realization:

- User chooses a study report
- User activates script
- Script extracts from the provenance system relevant input items
- Relevant data items are added to an archive
- Archive is stored in the DataFinder



Feature 2: Evidential Preservation
Integration of a preservation service

Beweissicheres Laborbuch Project (BeLab):

DFG Project from:

- Physikalisch Technische Bundesanstalt Braunschweig
- Karlsruher Institute of Technology
- Universität Kassel

Goal of the project:

Development of a service, which

- characterizes the preservation time of an item
- characterizes the legal trustworthiness of an item
- stores the archive securely

Feature 2: Evidential Preservation DataFinder Extension: Integration of BeLab

Goal:

- Use the service to get a classification
  - on time span
  - in legal issues
- Increase trustworthiness of DataFinder items

Realization:

- User chooses an archive and activates script
- Script sends the archive to BeLab service via WS-Secure
- The service processes the archive
- Service returns preservation information, which is stored

Phase 2: Implementation
Feature 3: Signing digitally

Motivation:

- Authenticity in general
- Attesting authentication

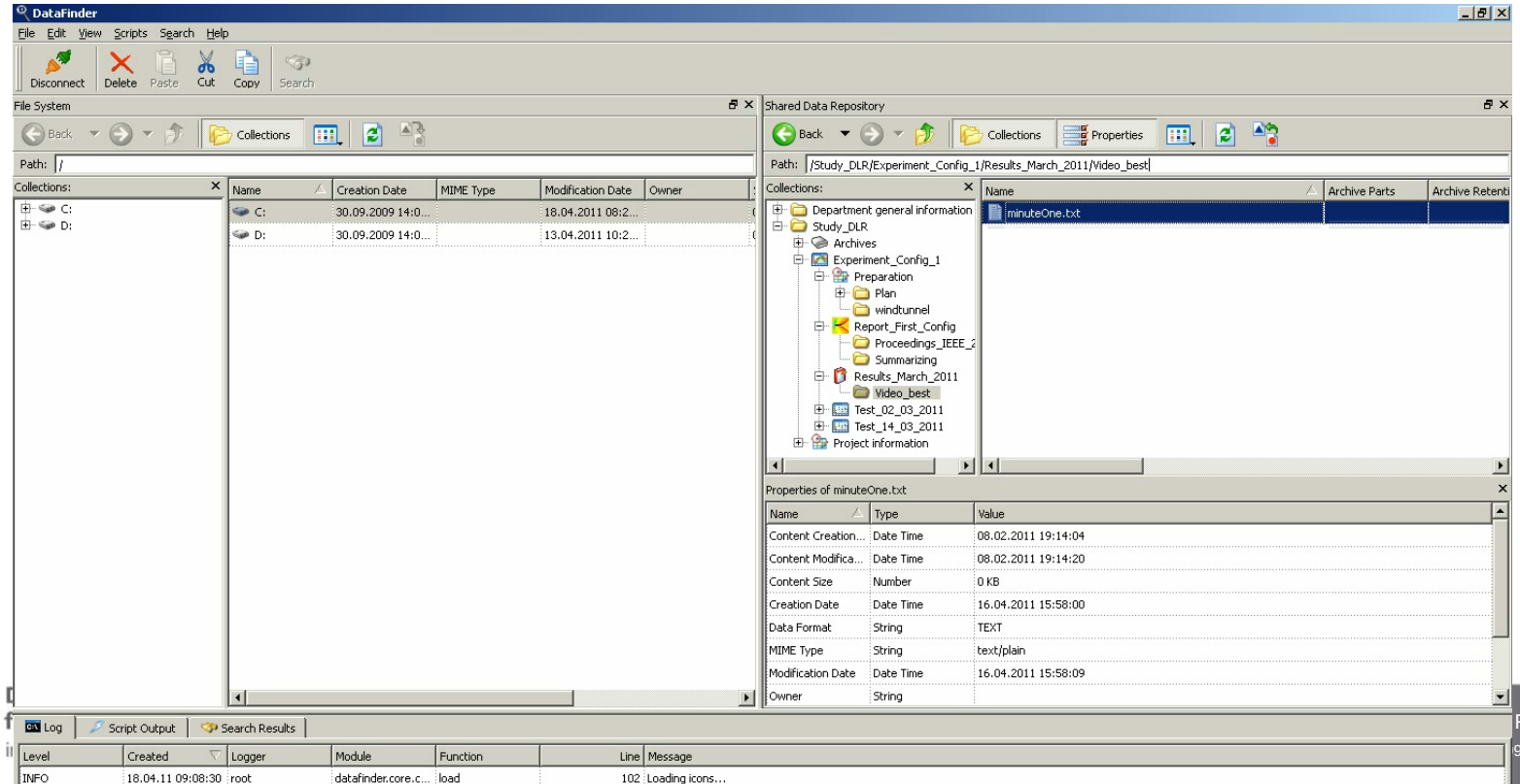
Tasks:

- Concept:
  - Data: as Meta Data item
  - Meta Data: extraction to XML and then store as XML Signature
- Integration into DataFinder
  - Simple Concept: Signature of the data as separate file

Feature 3: Signing digitally
DataFinder Extension: Integration of signatures

Realization:

- User chooses a file and executes script
- A signature file is generated (PKCS #7)
- Signature file is stored in the DataFinder



Phase 3: Evaluation
Development approaches

Choice for each feature:

- Feature 1: Provenance integration in DataFinder:
  - Prototyping
- Feature 2: Evidential Preservation
  - Test Driven Development
- Feature 3: Signing Data
  - For Concept: Pair Programming
  - In Thesis: Test Driven Development

Phase 3: Evaluation
Design of Extensions

Each feature is evaluated on three characteristics

- **Usability**:
  - Effectiveness, efficiency and satisfaction
  - *Example: automatic execution of task (Process Documentation)*
- **Adaptability**
  - Change in environment
  - *Example: no provenance system is needed (Preservation)*
- **Adjustments**
  - Improvements
  - *Example: Choice of signature algorithms,…(Signing)*

Phase 3: Evaluation
DataFinder as Electronic Laboratory Notebook

Implementations

- Documentation of a chain of events
- Credibility
- Durability/ Preservation

Adjustments

- Automatic Report generation
- Mobile version
- Graphical representation and integration of the provenance information

Conclusion
Impact and Future

Impact of my work

- Basic use case of DataFinder
- Basic scientific documentation workflow modelled with provenance
- Design of a general provenance storage system

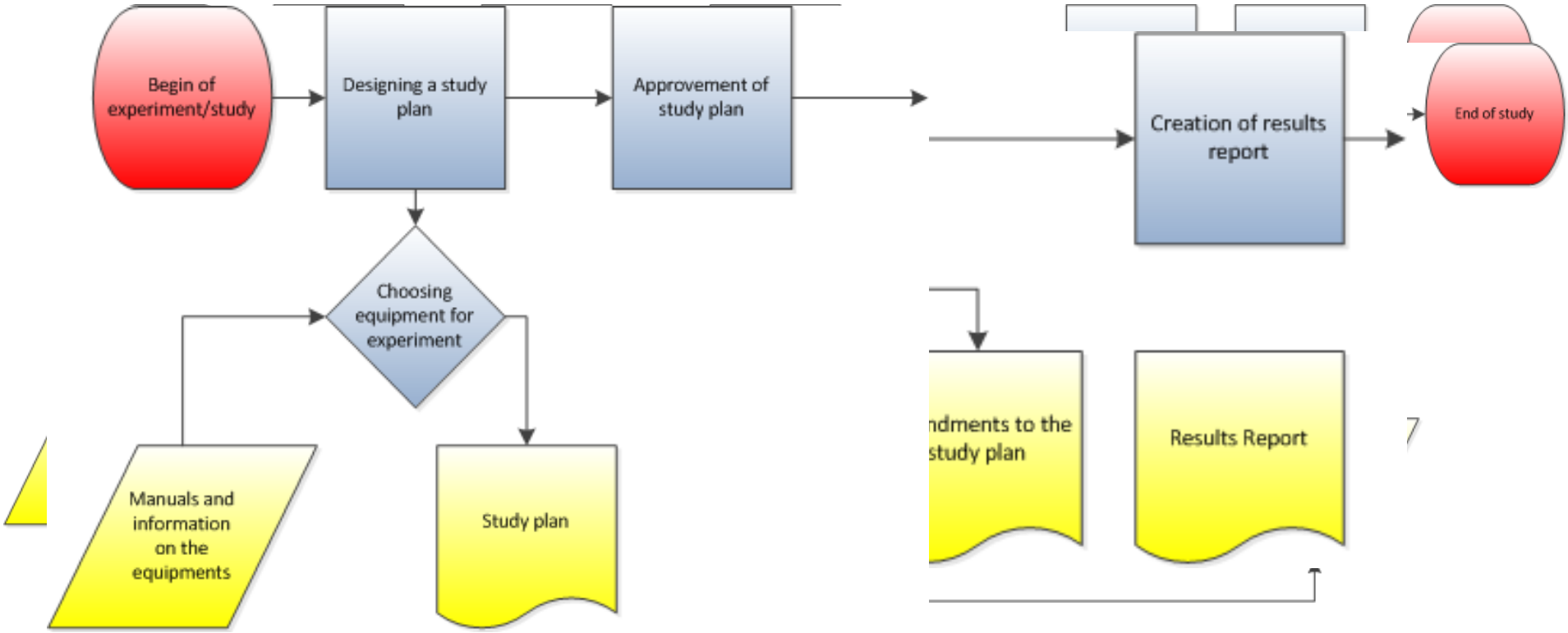
Future Work

- Integration of signing data into core
- Marketing for the provenance system and laboratory notebook application
- Publication of provenance system (source code and design)

Questions?

Kontakt:
Miriam Ney
DLR Simulations- und Softwaretechnik, Berlin
Email: Miriam.Ney@dlr.de

Phase 1: Requirements analysis
What does a scientific workflow look like? (Detail)



Feature 1: Origin of Data – Provenance Integration Rest Interface Storing Provenance information

- localhost:9999/rest/provenance?
  - **@param** process String having the information about the process and its general type e.g. *"process"*
  - **@param** input String symbolizing all inputs e.g. "*InputType~input1@29;InputType~input2@30;InputType~input2@40"*
  - **@param** ouput String symbolizing the output e.g. *"OutputType~output21@29"*
  - **@param** actor String symbolizing the actor e.g. "*ActorType~actor1;ActorType~actor2"*

- OutputType/ InputType = Type defined in the model
- ~ @ ; → Delimiter
- Input1 = identifier of the object
- 29 : Version of the object