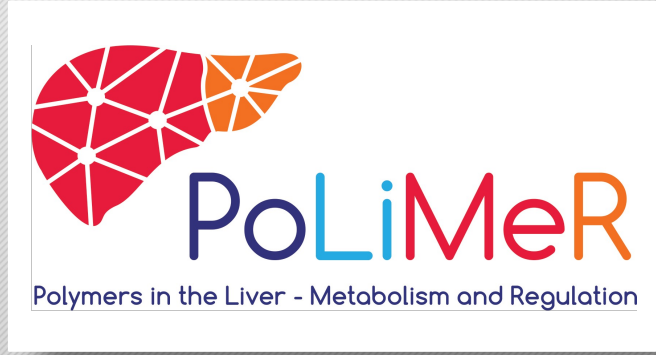
Information Retrieval and Data&Model Management for

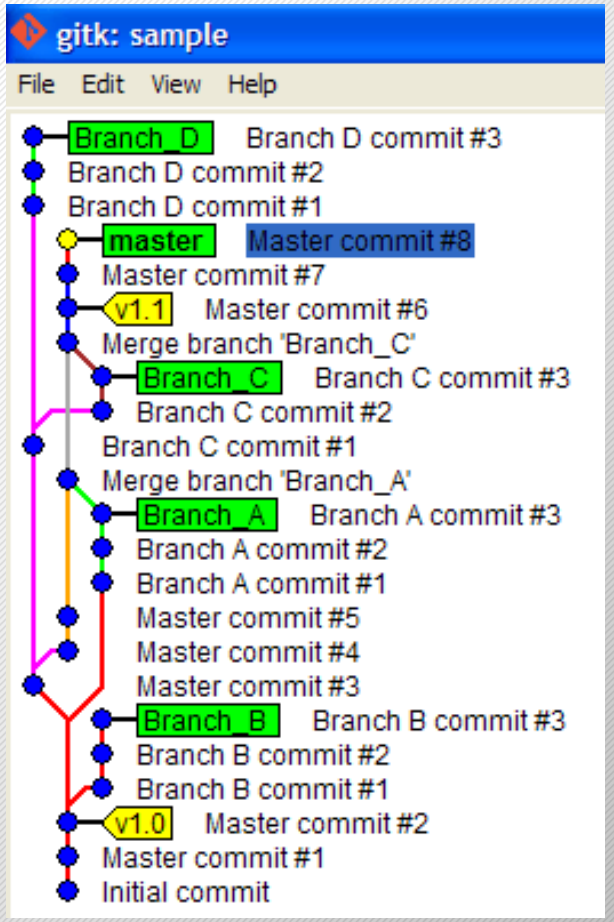


Wolfgang Müller + Olga Krebs, HITS

GIT

Versioning for Software

- Base idea
  - Code consists of files and lines
  - Track changes by noting
    - which lines changed
    - Which files added/removed
- Important concepts
  - Branching
  - Merging
- Use for: Code
- Don‘t use for: Experimental files

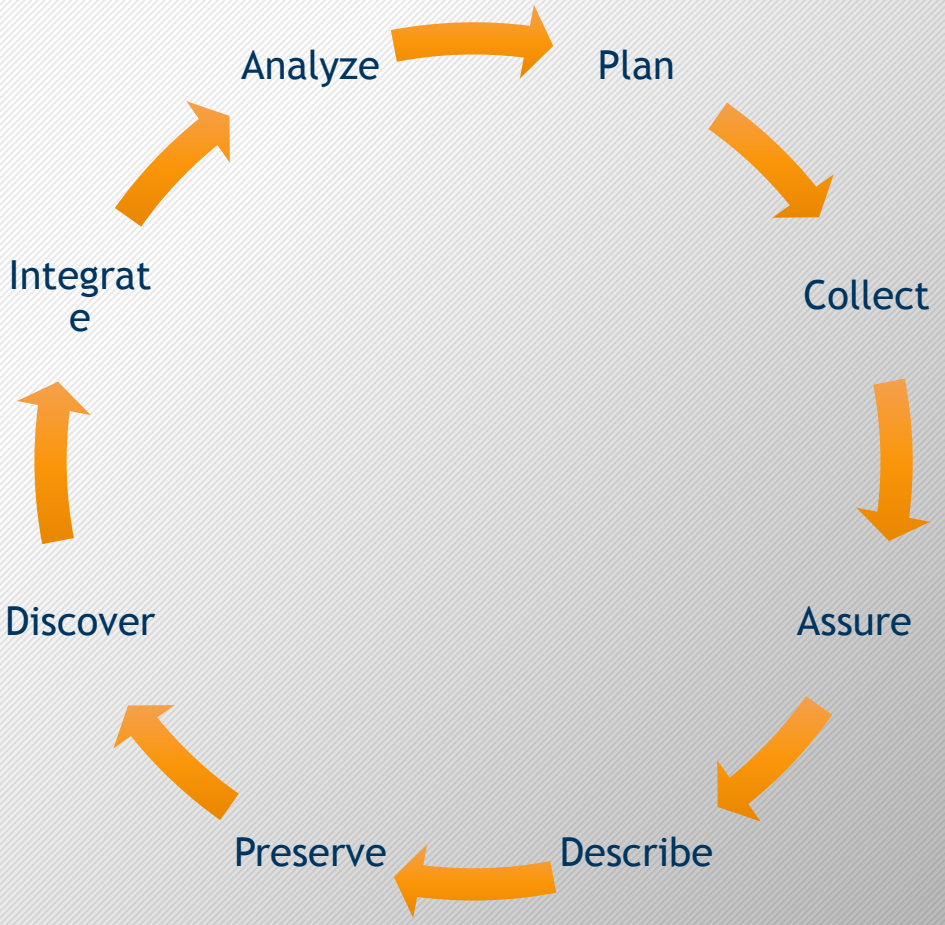


Way forward

- Please agree with Oliver Ebenhöh
- What to put in there
- How to use
  - Policies for branches
  - Setup of repository (directory structure)

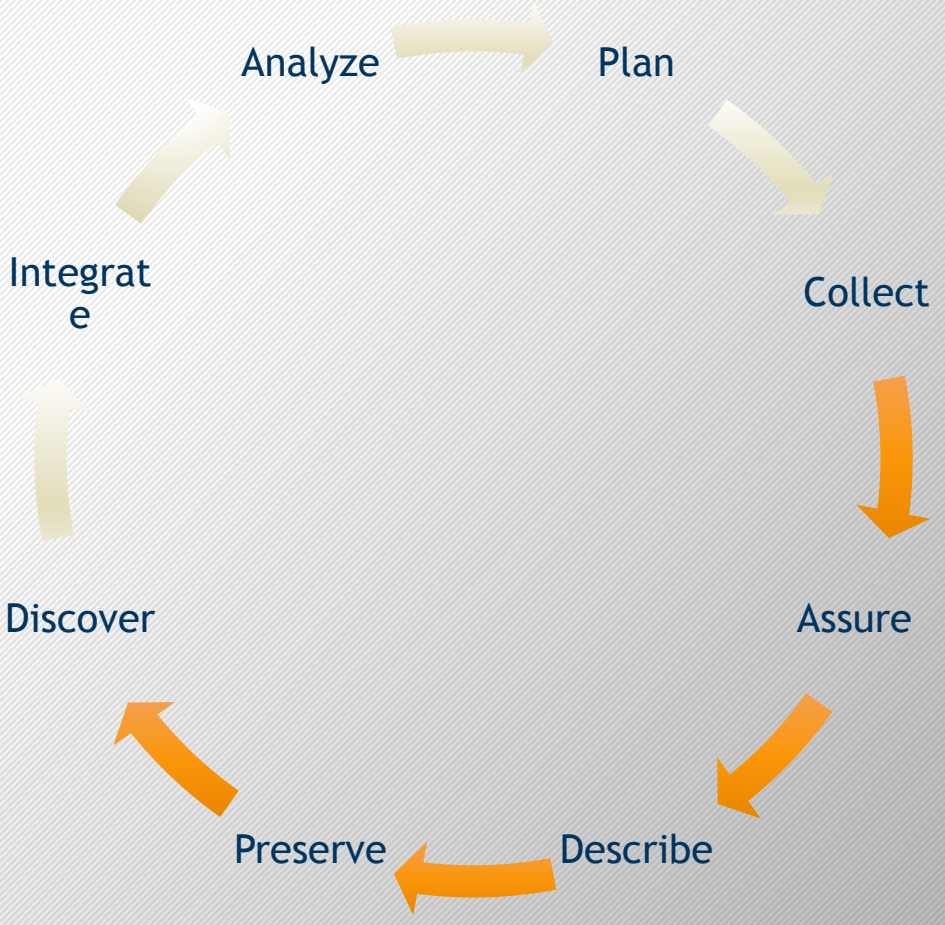
Data life cycle

- This cycle consists of…



Data life cycle

- …parts people do…
- …and parts that are less prioritary…



Reproducibility Rampancy

Drive for open, reproducible, Long-term maintained „FAIR“ data

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2*], Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

1. For Every Result, Keep Track of How It Was Produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

Findable (Citable)
Accessible (Trackable)
Interoperable (…for machines)
Reusable (Reproducible)

Record All
Automate All
Contain All
Access All

JOIN IN THE DISCUSSION - LEAVE YOUR COMMENTS BELOW

FAIR DATA PRINCIPLES

Preamble

One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows. Here, we describe **FAIR** - a set of guiding principles to make data **Findable, Accessible, Interoperable, and Re-usable**. The FAIR principles have now been **published**.

TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.

TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

Enable identification, registration, search

Link to metadata

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent id

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

Machine retrievable,
machine reusable
AOpenAP
Metadata outlive data

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

FAIR and accessible description

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

Use metadata standards

Some analysis

- 4+4+3+4 rules
- 4+**2**+3+4 rules contain „**metadata**“
- 4 about **description languages** and **standards**
- 3 about **access protocols**
- 3 about **identifiers** and **references**
- 1 about **licensing**
- 1 about **lifetime**

FAIR

- It‘s FAIR principles, and not FAIR standards, because
- Standards are too strongly binding
- Politics:
  - First make people love the abbreviation
  - Make them get on the FAIR boat
  - Then get them to know what they need to stay on the boat

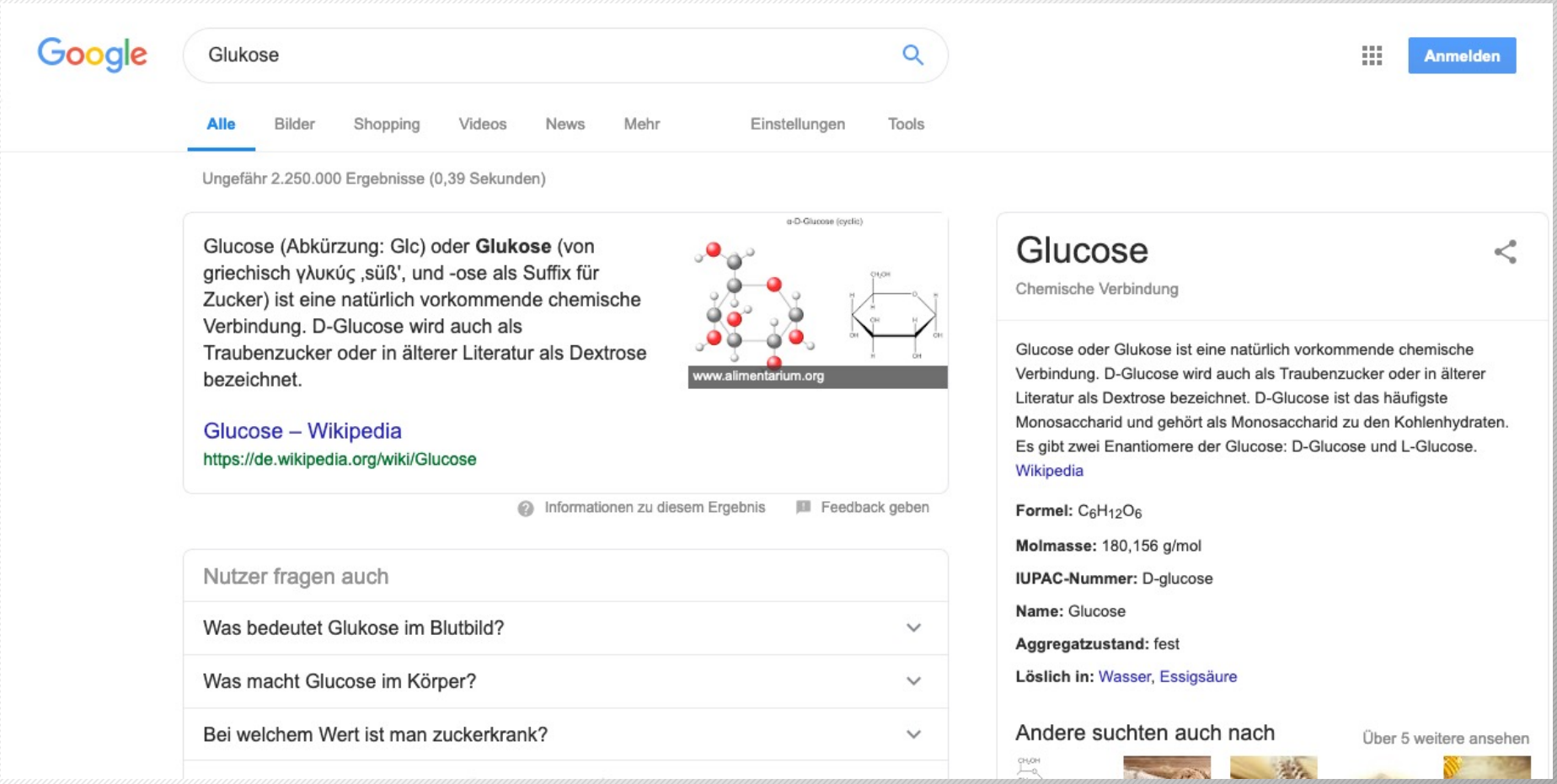
Barend Mons, Leiden

The ESR7: Information Retrieval and SEO

Wolfgang Müller

What is information retrieval?

Find results **fulfilling an information need** of a user, giving an **imprecise** query formulation

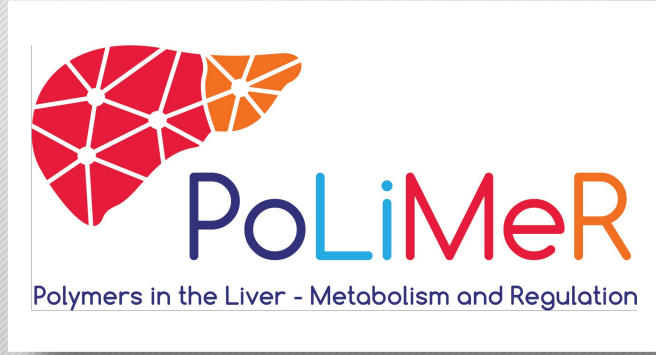


What do we try to achieve?

- Web
  - We don‘t have control over the data
  - We know much about the structure&statistics of data
  - Use that
- This project:
  - Small, very diverse data set
  - Full control about the data we store
- Idea: See this as a **search enginge optimisation** problem
  - Tailor metadata for subsequent retrieval
  - Simplify generation of metadata for the user
  - Tailor retrieval methods to metadata

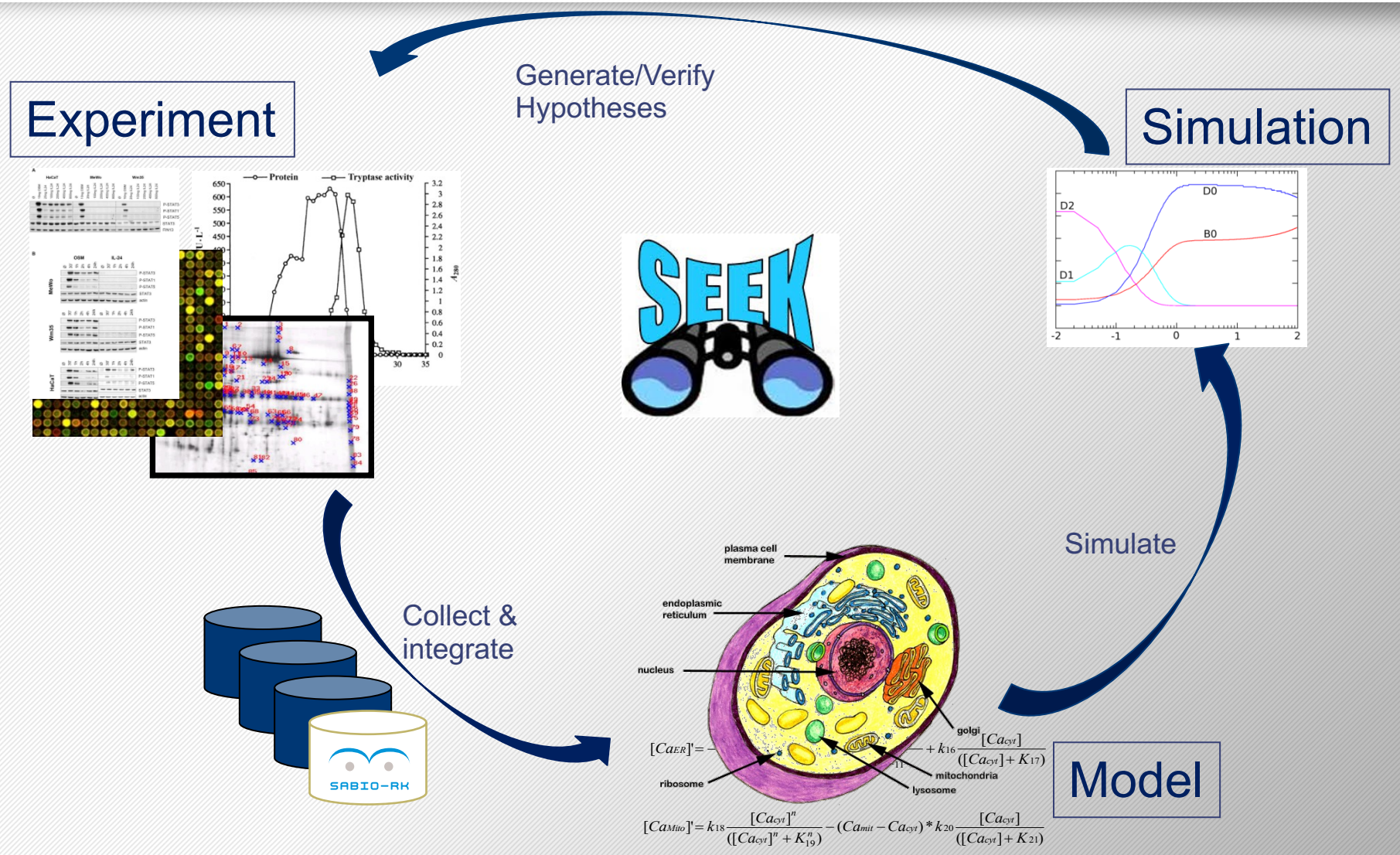
For now, let‘s go back to Systems Biology

Data management for systems biology

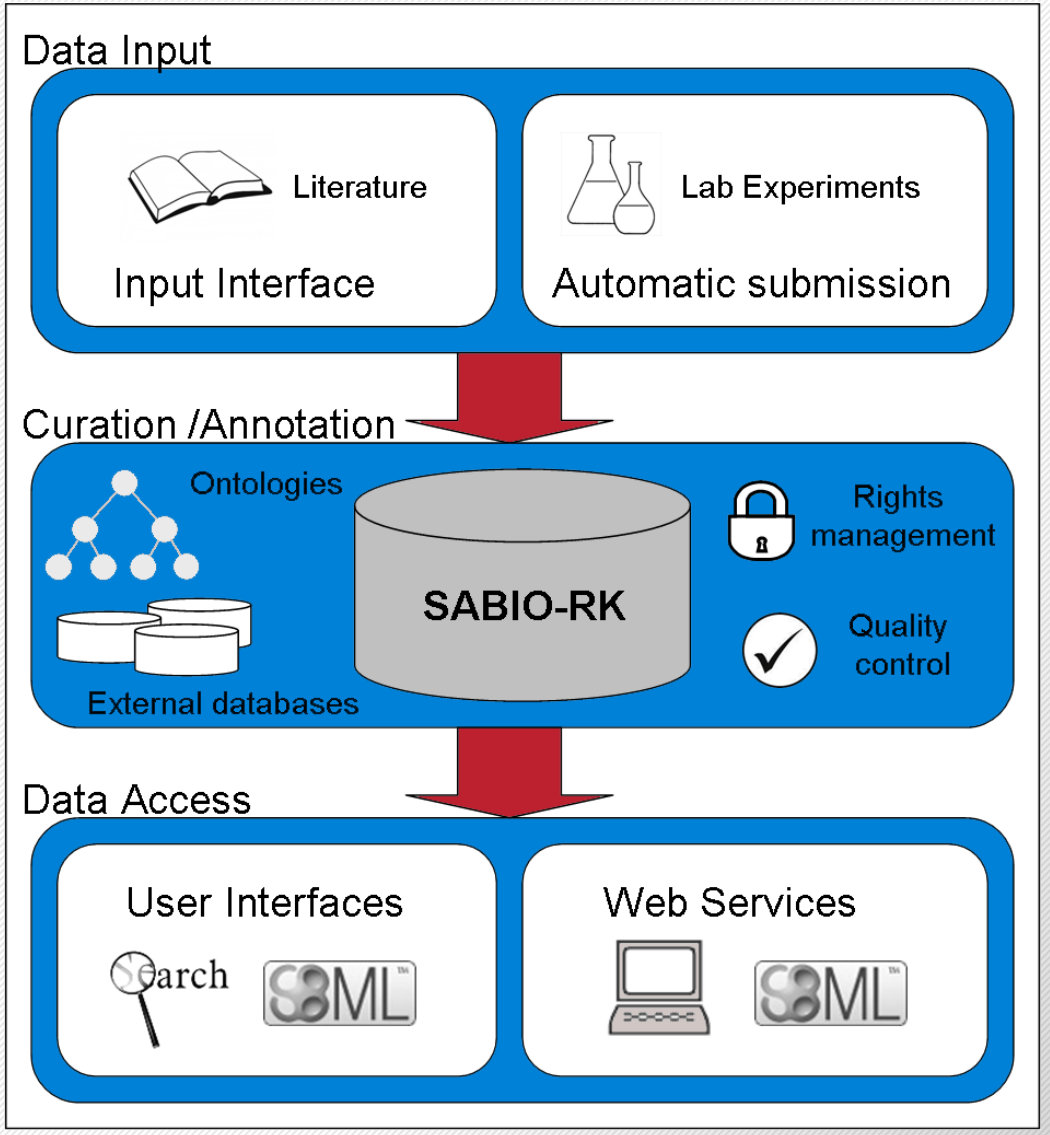


for de.NBI NBI-SysBio:
Wolfgang Müller

Systems biology



SABIO-RK: Reaction Kinetics



SABIO-RK

http://sabiork.h-its.org

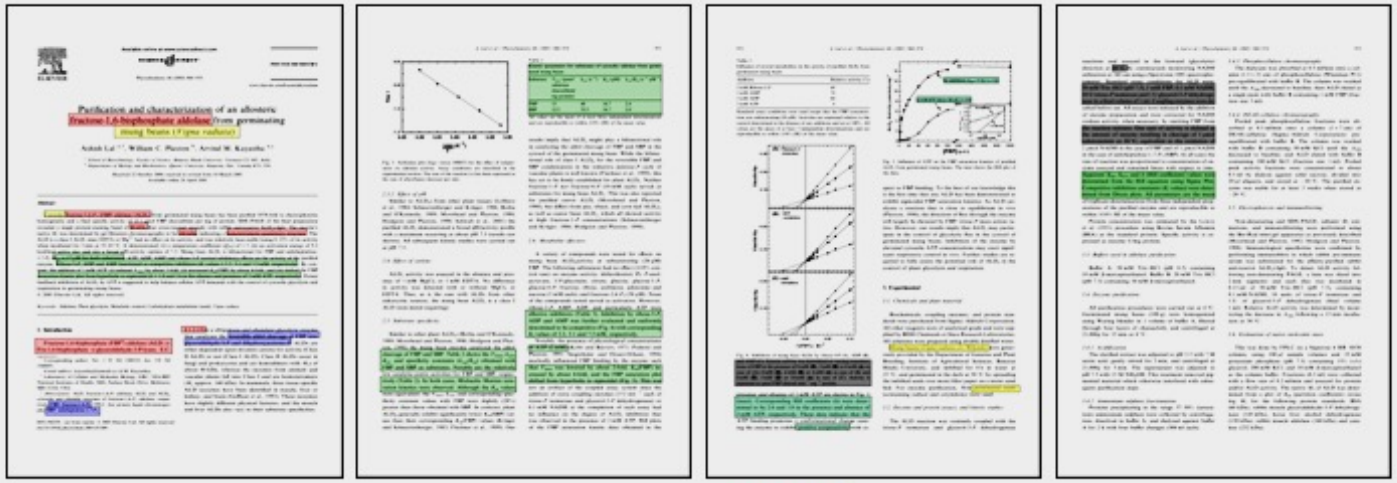
Biochemical reaction kinetics

Data

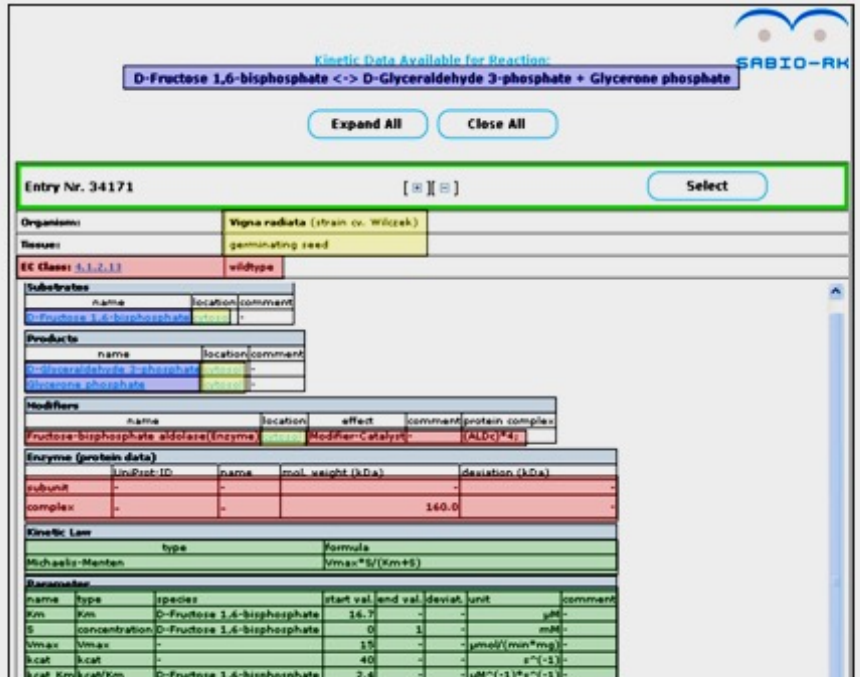
- Integrated
- Structured
- Unified
- Interlinked
- Annotated

Curation: Add value to data

A



B



Advantage:

- Clean
- Integrated
- Consistent
- Interlinked/annotated

→ **High quality data**

- Protein- bzw. Enzymdaten
- Reaktionen und chemische Verbindungen
- kinetische Daten
- experimentelle Bedingungen
- biologische Quelle (Organismus, Gewebe, Zelltyp)

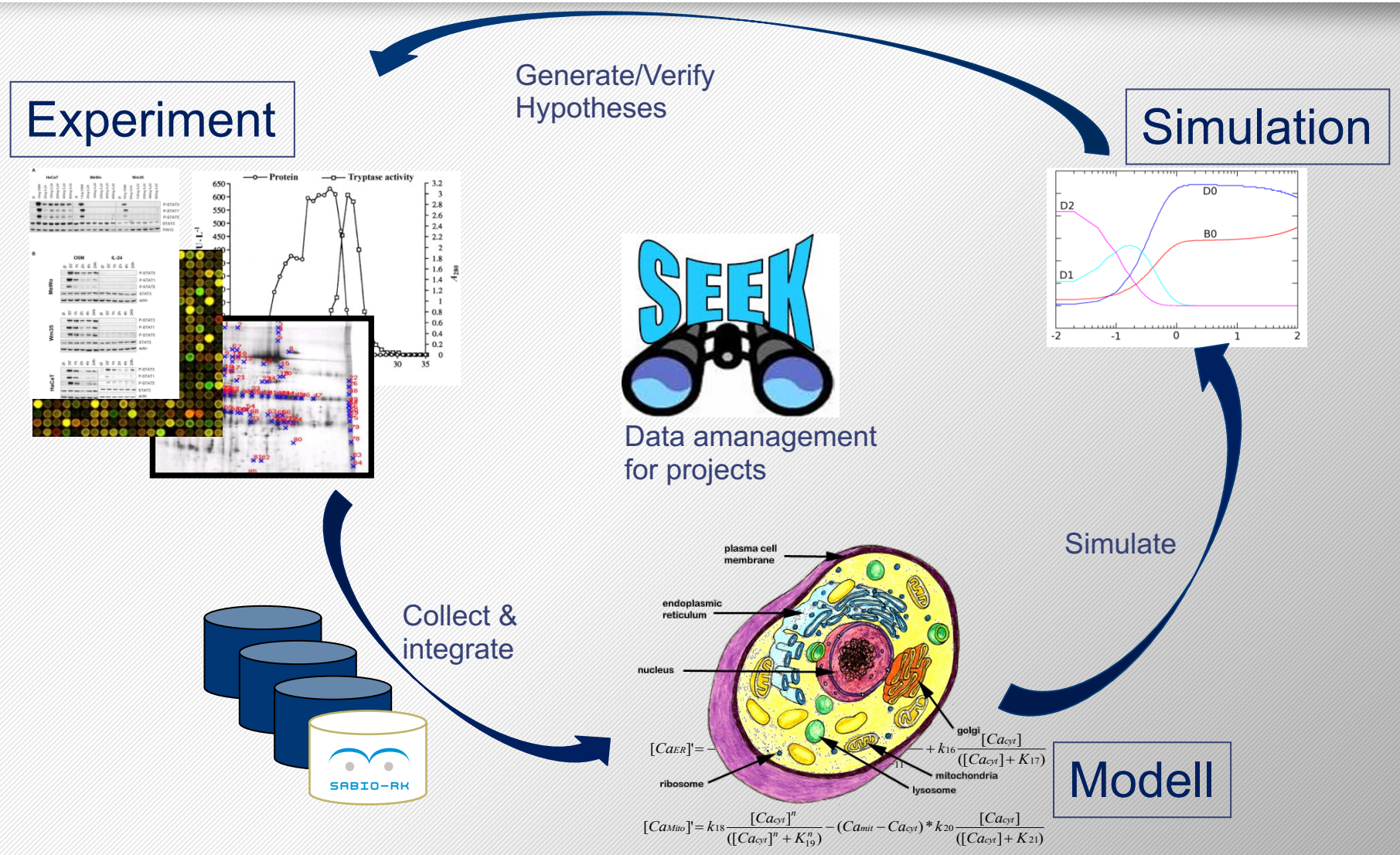
Disadvantage

- One paper takes 0.5 to 1 workday
- Reliable automatic curation impossible in our setting
- Limits number of papers you can curate per year
- Even multiple people cannot catch up with everything

➔**Parallelisation** is the solution

➔**Let the user contribute data quality**

Systems biology



Data & Model management and self curation

PoLiMeR

Polymers in the Liver - Metabolism and Regulation

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

- **D**ata
- **O**perations
- **M**odels

ISBE Infrastructure for Systems Biology Europe

ERASys APP ERA-Net for Applied Systems Biology

SystemsX.ch The Swiss Initiative in Systems Biology

Purpose of project data management

- Organisation
- Communication
- Dissemination

of data associated with a project

Helps you find your way

Reuse later

Helps others find out

Enable team to reuse

Tell more and take credit

Reuse with new partners

Why does not everyone do it?

- Hard to know how much you need
- Afraid of sharing too much
- Takes time
- …

- …and other personal factors

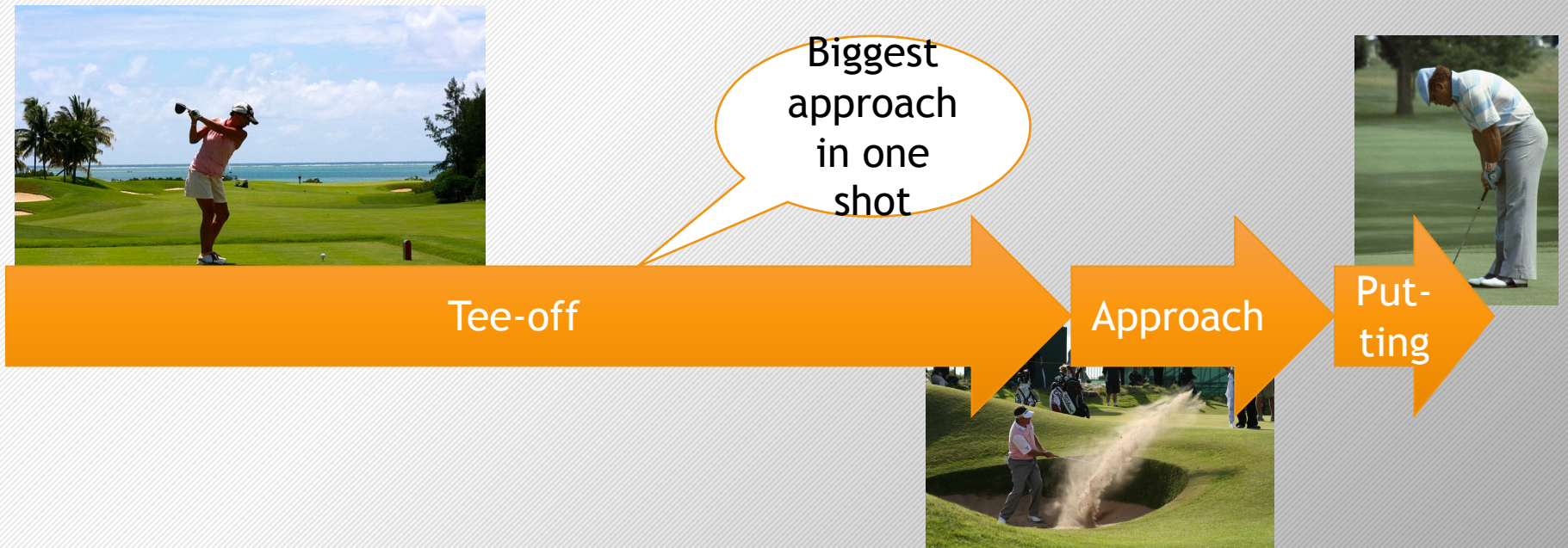
LiMeR

etabolism and Regulation

80-20 rule

Voltaire: „The best is the enemy of the good“

80-20 rule: Often you can get **80%** of the benefits using **20%** of the effort.

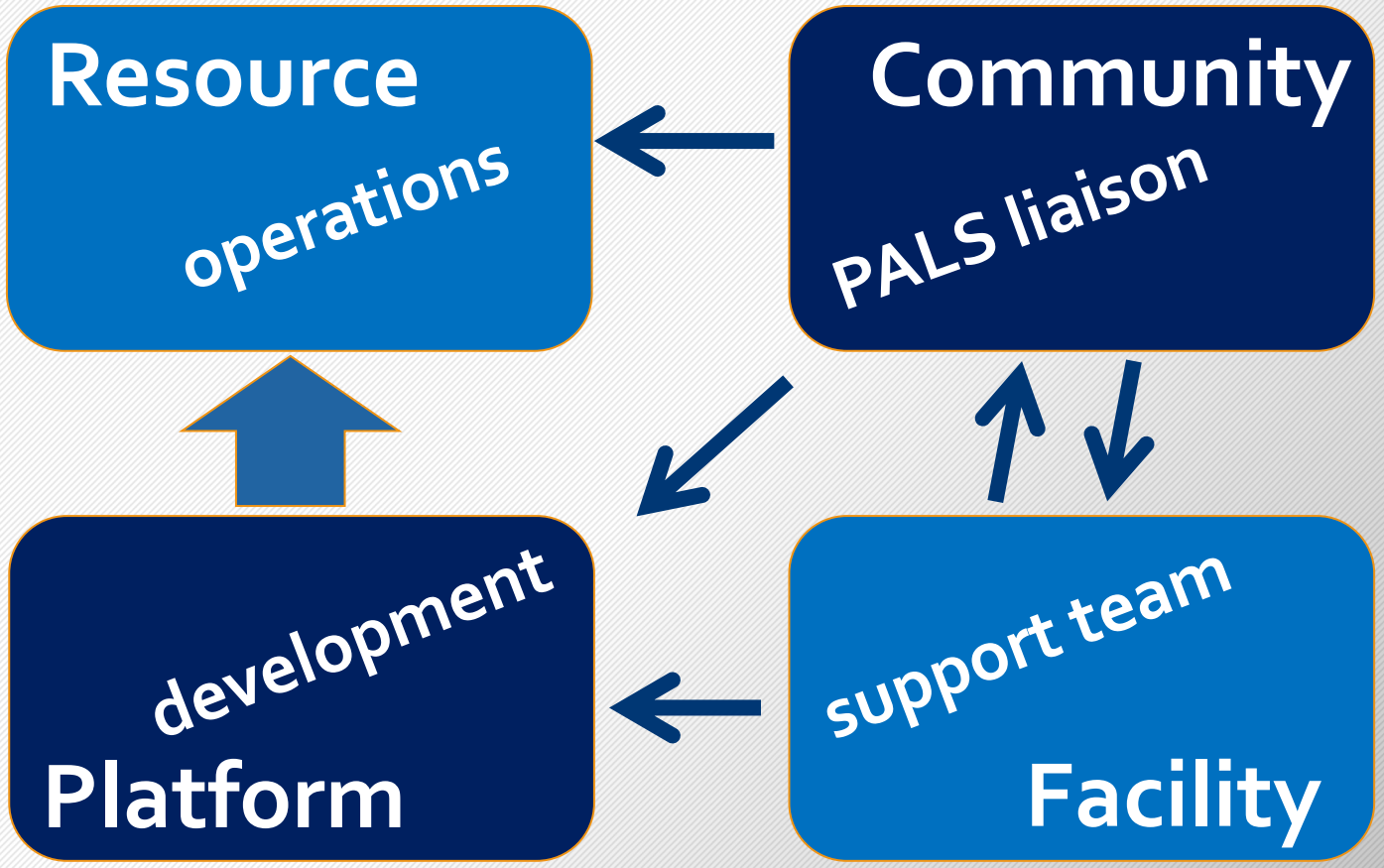


Challenge of data sharing

- Most data never gets shared
  - Wrong experimental method
  - Hidden parameter discovered
  - Faulty experiment
- How to prepare data in this situation?
  - Don‘t want to waste time
  - Want to be prepared if we share
- We propose useful way forward

What to share?

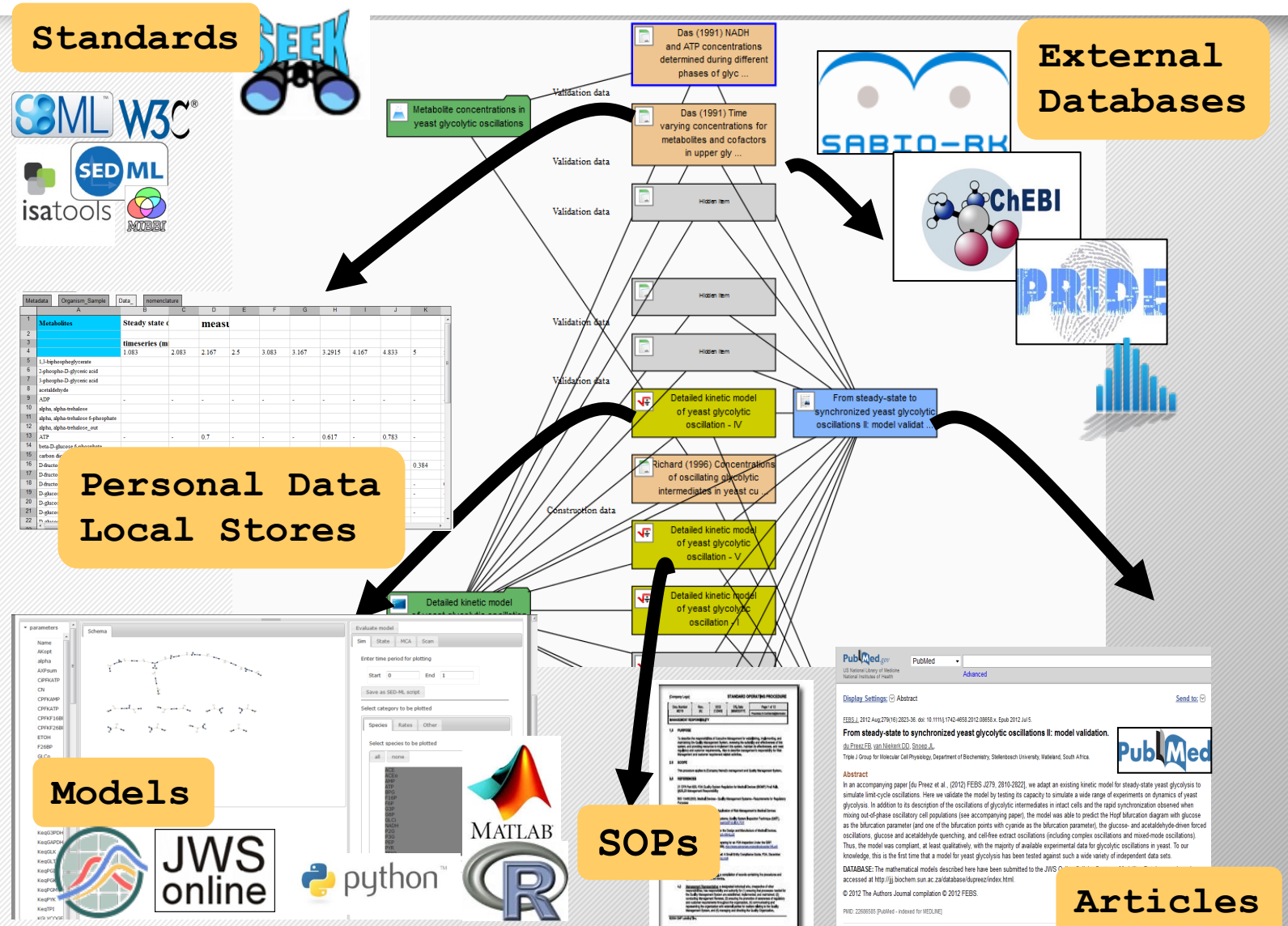
- Raw data (sometimes)
- Condensed, interpreted data
- Metadata: Data about the data
  - **Conditions & Procedures** of the measurements
  - **Information** about the samples
    - What was sampled?
    - How was it prepared?
    - How was it treated after sampling?


Resource
operations
Community
PALS liaison
development
Platform
support team
Facility

WP1

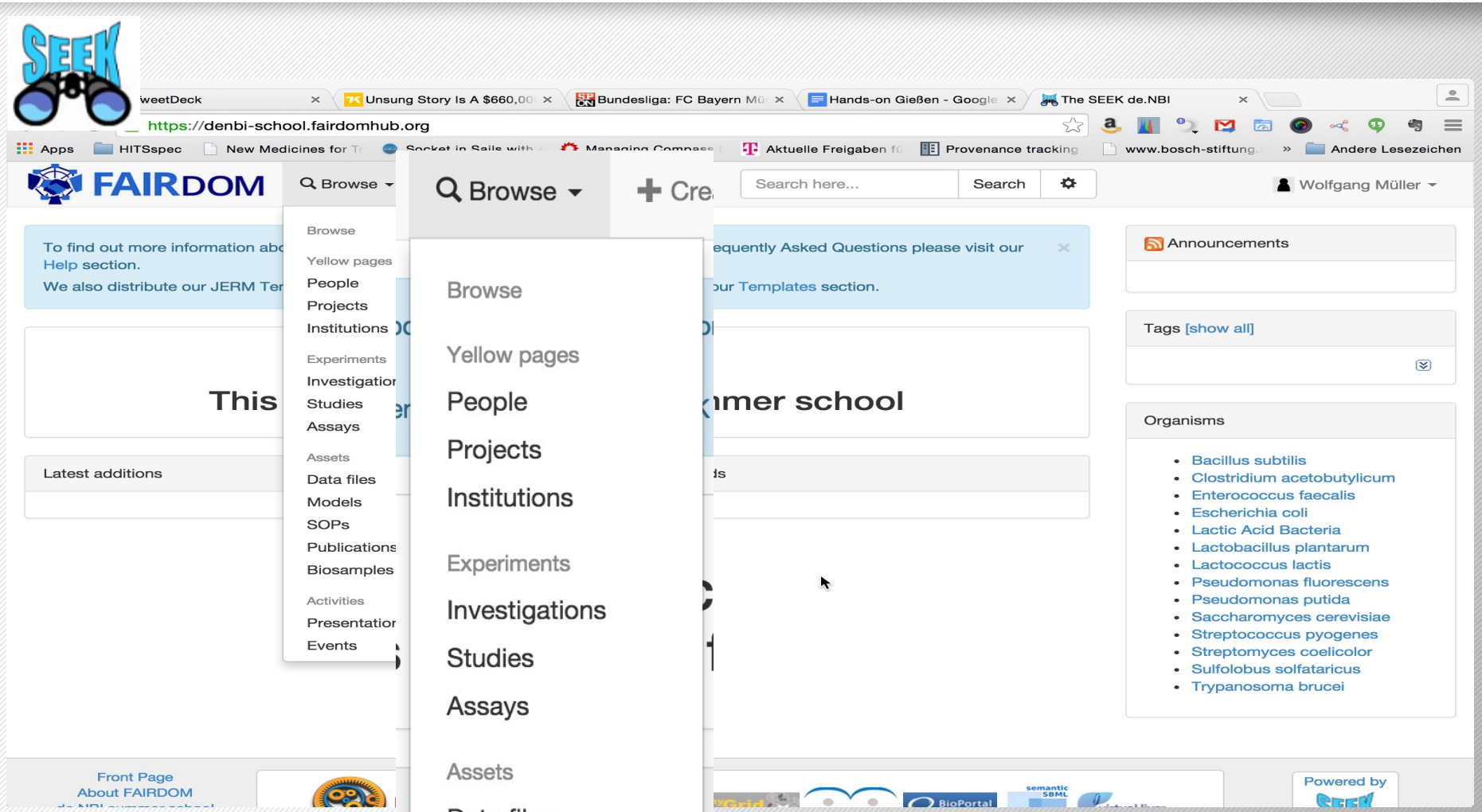
WP2

gSEEK as integration point

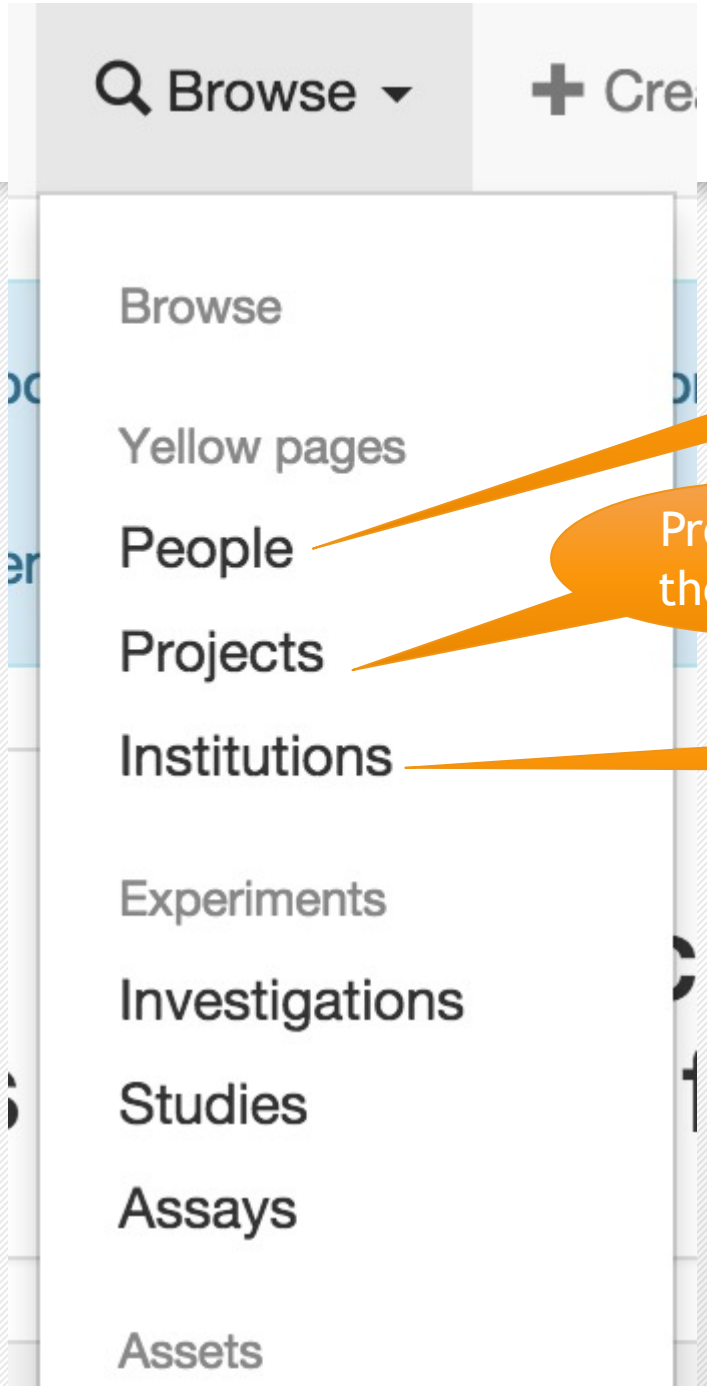


Organisation – Collaboration - Dissimination

https://www.fairdomhub.org



Science is about people and credits

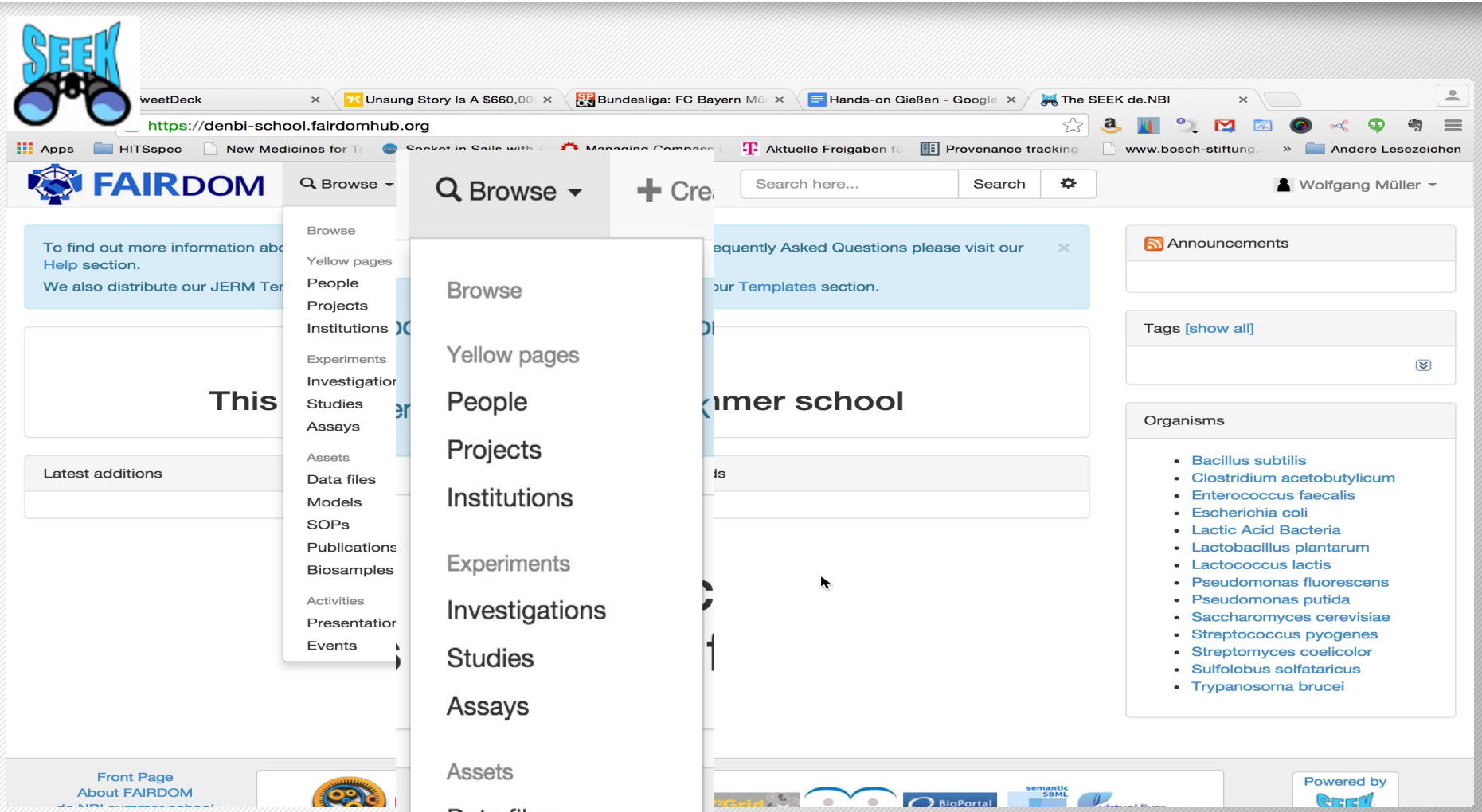


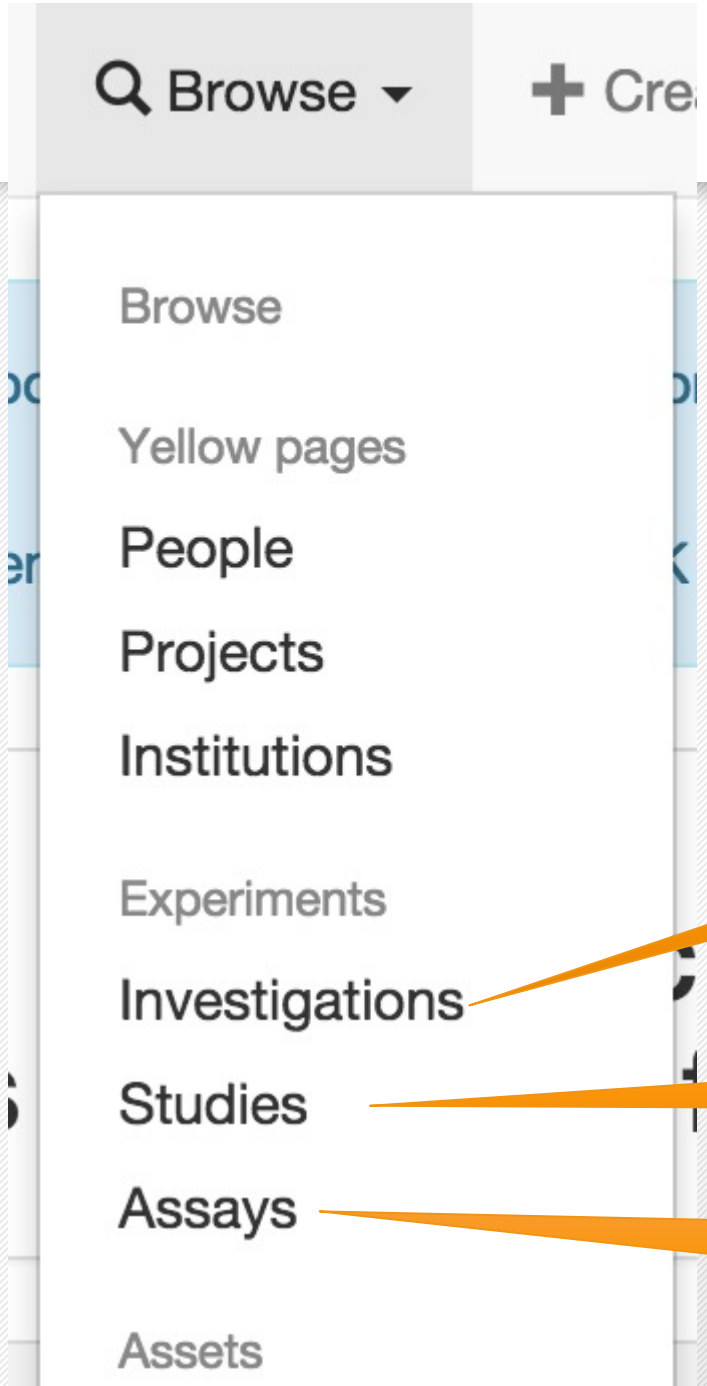
People in the system

Projects in the system

Participating institutions

https://www.fairdomhub.org



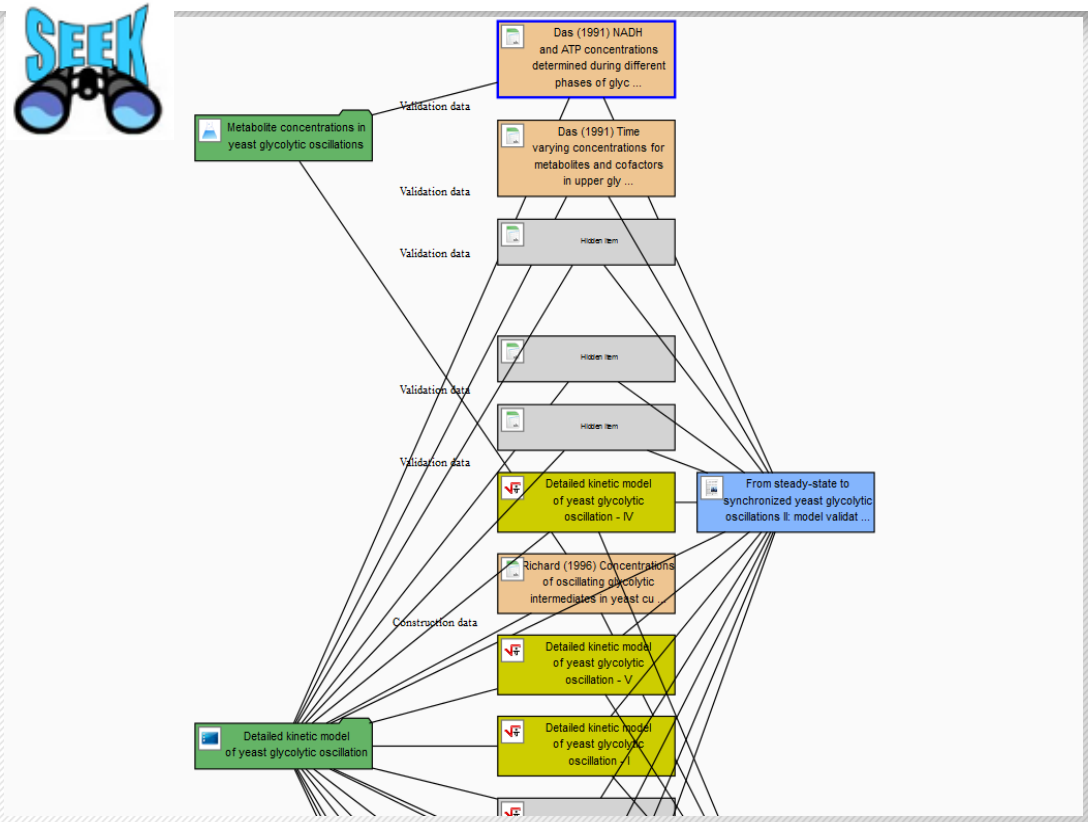


A few per project

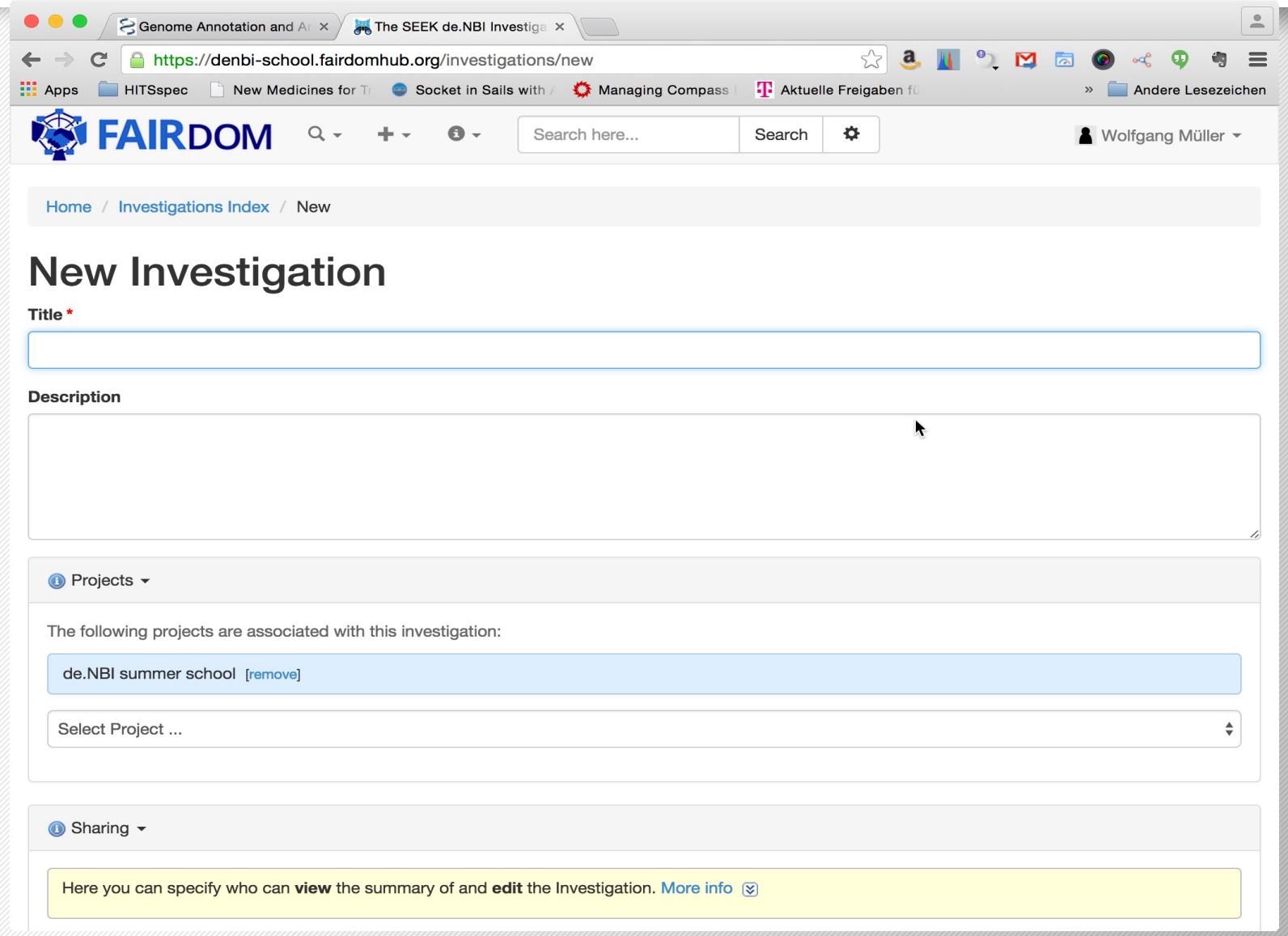
Roughly one paper

„smallest meaningful group of measurements“

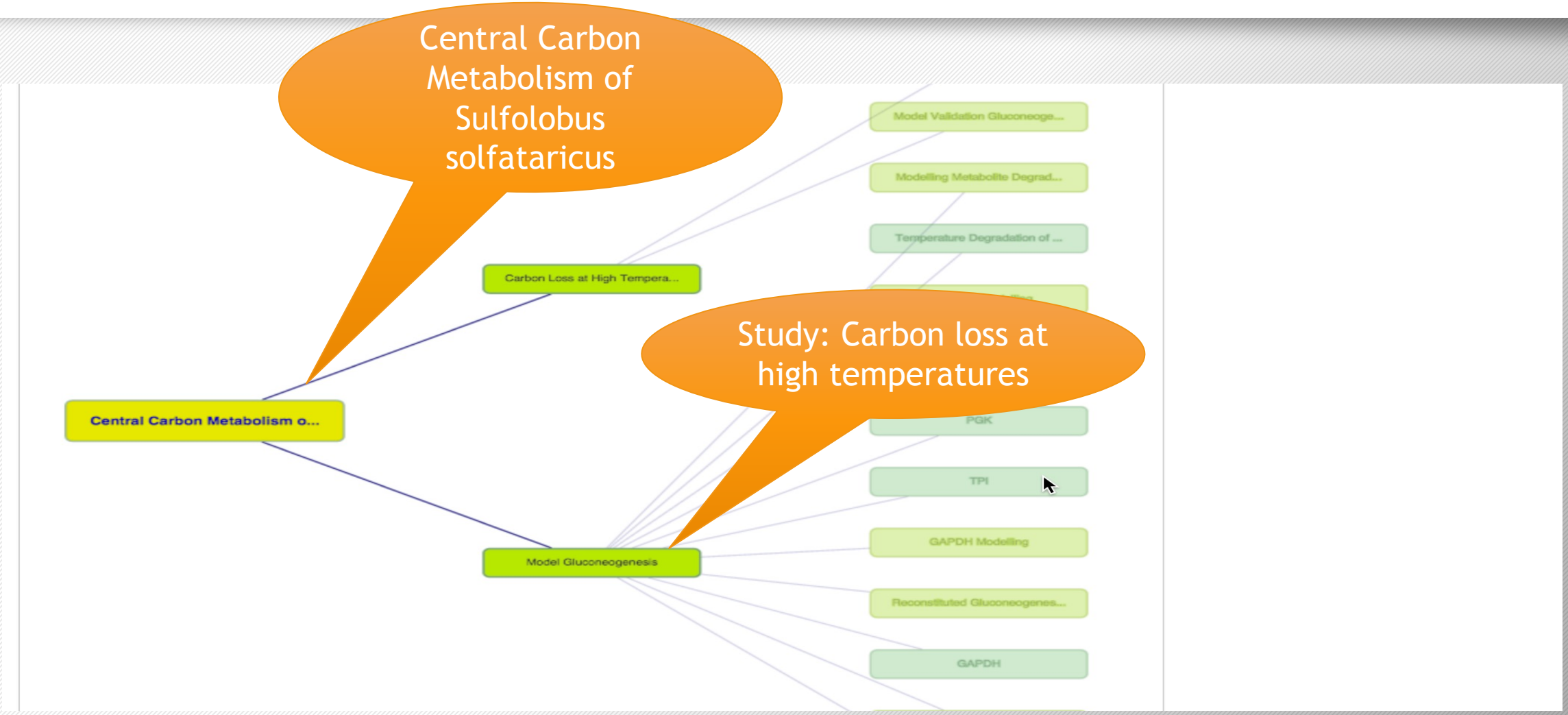
Core: The ISA structure



New investigation

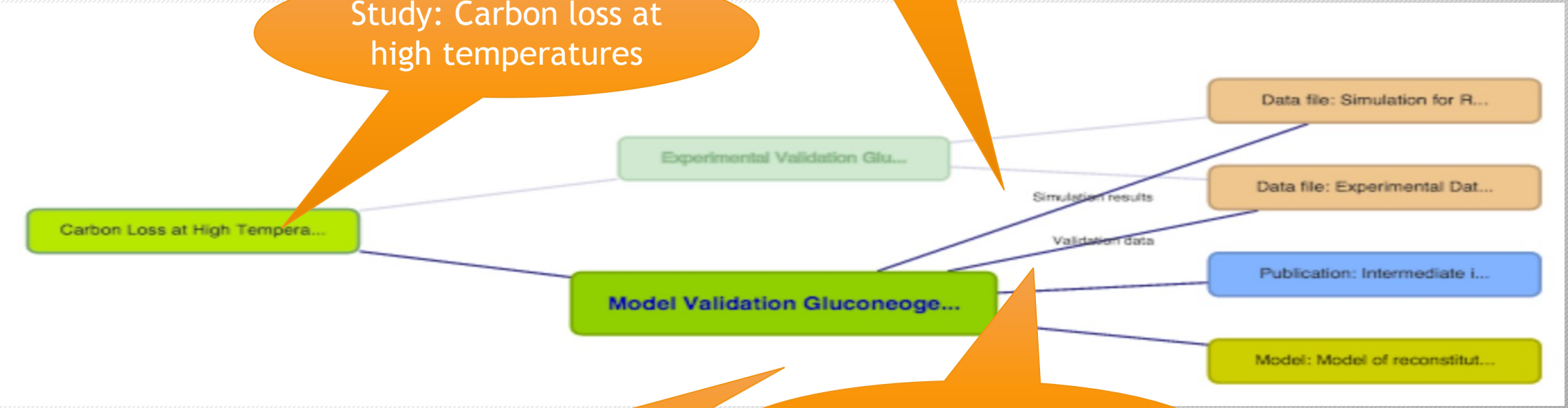


An example: Investigation, Study, Assay



Simulation results

Study: Carbon loss at high temperatures



Model Validation Gluconeogenesis in S. solfataricus

Validation data

Model

Publicatio n

Takeaway

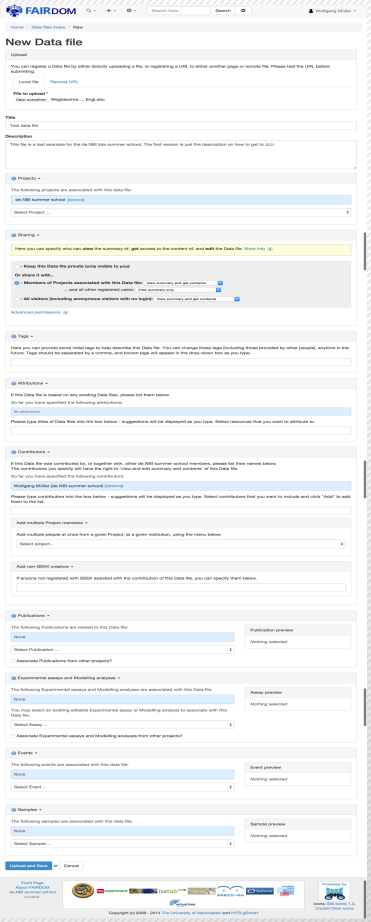
- Structure tells you a lot about what happens
- Structure a good first step

SEEK and data files

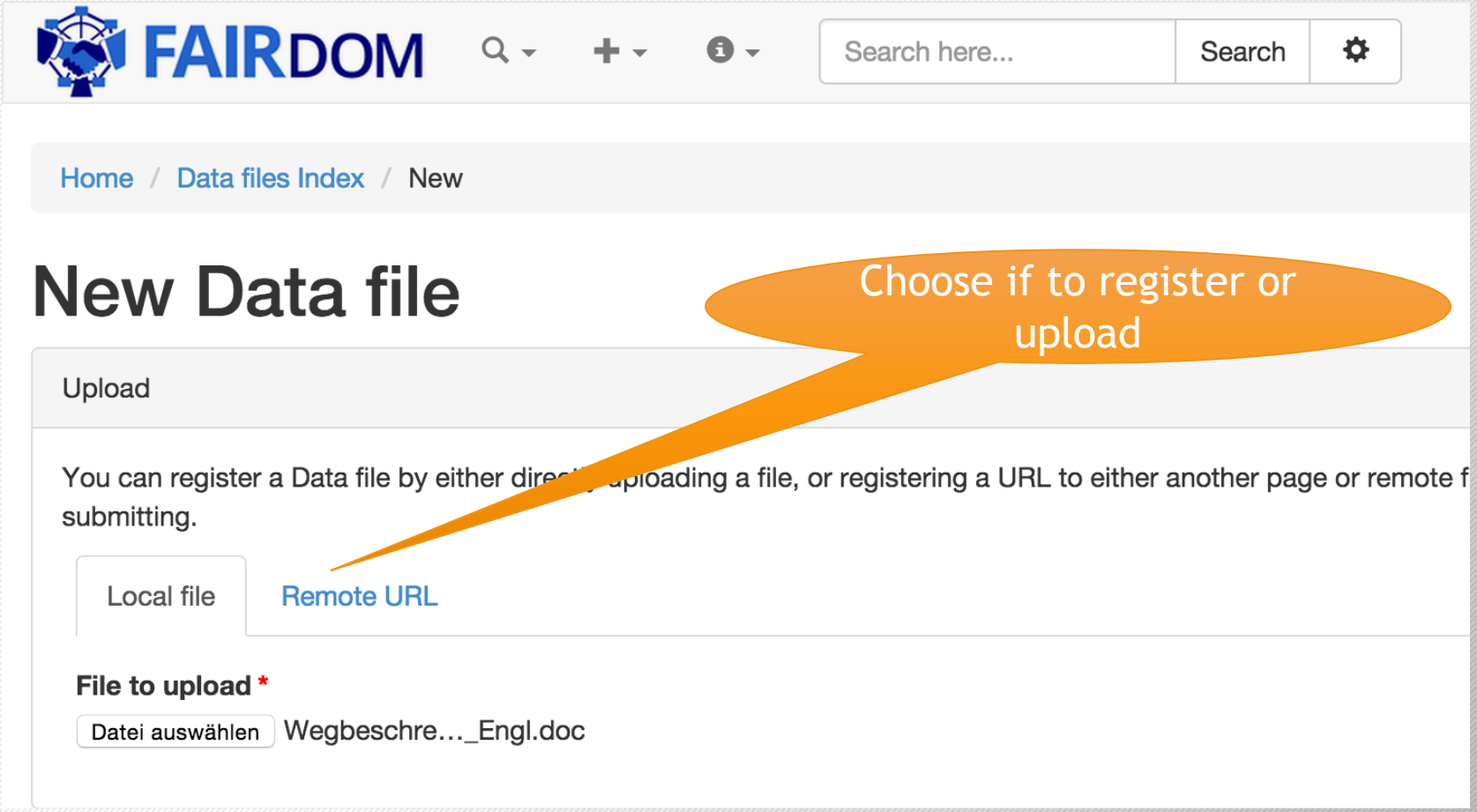
PoLiMeR

Polymers in the Liver - Metabolism and Regulation

Plenty of opportunity for configuration



Basic description; Register vs. Upload



Fa

Factors Studied

No factors studied for this Data file.

Edit factors studied (latest version)

Version History

Version 1 Created 24th Sep 2015 at 04:03 by Wolfgang Müller

No revision comments

Factors studied

Related Items

People (1) Projects (1)

Versioning!

People (1)

Wolfgang Müller

Projects: de.NBI summer school
Institutions: HITS
Email: wolfgang.mueller@h-its.org
Web page: *Not specified*

Disciplines: *Not specified*
Roles: *Not specified*
Expertise: *Not specified*
Tools: *Not specified*

Summarising

- Functionality to fill the ISA structure
- Local or remote files + metadata
- Elaborate possibilities for sharing with right people
- Add more elaborate metadata later

Data files and Standards

PoLiMeR

Polymers in the Liver - Metabolism and Regulation

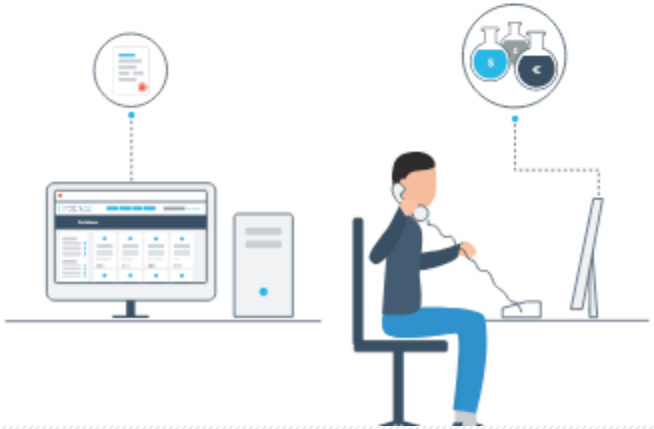
FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*.

HOW CAN WE HELP?

We guide consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited.



Journal editors & publishers

Create and maintain an interrelated list of citable standards, databases and repositories to recommend to your authors, users or their community, and revise this recommendation over time...
[read more]

FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

Standards

Contribute by adding a standard | Any problems? Please tell us!

The standards in FAIRsharing are manually curated from a variety of sources, including BioPortal, MIBBI and the Equator Network.

Search Standards | Search | Search | Reset | Advanced

1314 Standards registered!

View as Table | View as Grid

Sort by

Name

Recommended Records

Showing records **1 - 50** of **1314**.

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 »

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database
	ABA Adult	ABA	Standard	None	Brain	Mus musculus	NeuroMorpho.O

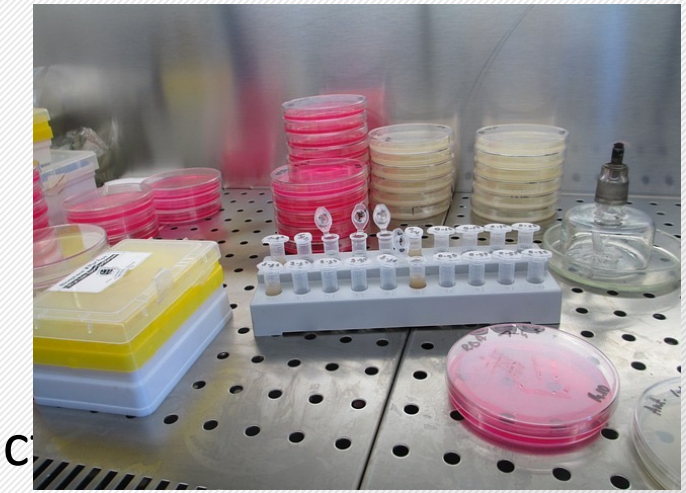
Levels of detail

- Action **guidelines** (e.g. SOP)
- Structure **guidelines** (e.g. F1000 data preparation guidelines)
- Semantics **guidelines** (metadata + content, e.g. some MIBBIs)
- **File format standards** (e.g. ISA-TAB, SBML)
- **Ontologies** + **vocabularies** (e.g. ChEBI)

Standardisation scales

Increased usability for others

- Self
- Group
- Collaborative project
- Field scale



SBML Systems Biology Markup Language

Self-standardisation

- **Store** same things in same structure
  - Test question: „Does Excel cell (e.g.) A2 have the same meaning in all files about the same experiment type“?
- **Name** same things the same way
  - Test question: „Does ‚gl‘ mean exactly the same in all occurences“?
- **Identify** uniquely things that you **reference**.
- Use **open source file formats** wherever possible

Benefit:
Automatic adaptation of your data much easier

Recognizing old data much simpler

Modify Excels reproducibly

RightField

OPEN Refine

Open for Innovation KNIME

jupyter

Standardisation within group or project

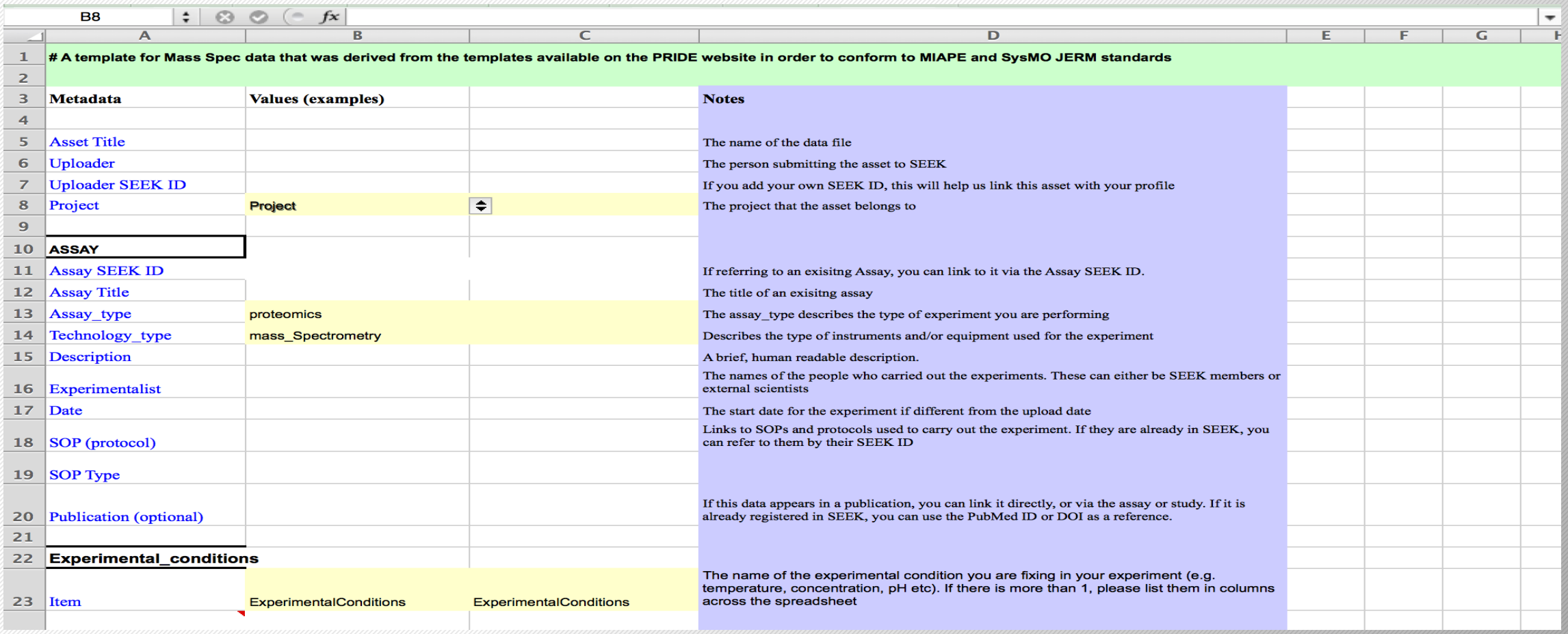
Same as before, but **in addition**:

- Needs agreeing on **how** to do things the same way
- Needs looking into **standards for your domain**
  - Inspiration how to proceed
  - Clear insight into migration paths

e.g. F1000 data preparation guidelines

- Give each column a **descriptive heading**
- Use a **single header row**
- Ensure you have **used the first cell**, i.e. A1
- Include **Title & Legend** for each spreadsheet
- Save each data file with a **telling name**
- Submit each **table** as a separate **file**
- Submit each **work sheet** as a separate **file**

JERM templates



	A	B	C	D
1	# A template for Mass Spec data that was derived from the templates available on the PRIDE website in order to conform to MIAPE and SysMO JERM standards			
2				
3	**Metadata**	**Values (examples)**		**Notes**
4				
5	Asset Title			The name of the data file
6	Uploader			The person submitting the asset to SEEK
7	Uploader SEEK ID			If you add your own SEEK ID, this will help us link this asset with your profile
8	Project	Project		The project that the asset belongs to
9				
10	**ASSAY**			
11	Assay SEEK ID			If referring to an exisitng Assay, you can link to it via the Assay SEEK ID.
12	Assay Title			The title of an exisitng assay
13	Assay_type	proteomics		The assay_type describes the type of experiment you are performing
14	Technology_type	mass_Spectrometry		Describes the type of instruments and/or equipment used for the experiment
15	Description			A brief, human readable description.
16	Experimentalist			The names of the people who carried out the experiments. These can either be SEEK members or external scientists
17	Date			The start date for the experiment if different from the upload date
18	SOP (protocol)			Links to SOPs and protocols used to carry out the experiment. If they are already in SEEK, you can refer to them by their SEEK ID
19	SOP Type			
20	Publication (optional)			If this data appears in a publication, you can link it directly, or via the assay or study. If it is already registered in SEEK, you can use the PubMed ID or DOI as a reference.
21				
22	**Experimental_conditions**			
23	Item	ExperimentalConditions	ExperimentalConditions	The name of the experimental condition you are fixing in your experiment (e.g. temperature, concentration, pH etc). If there is more than 1, please list them in columns across the spreadsheet

Systems Biology Markup Language

- XML-Based format
  - Levels and Versions
  - Packages
- Model of relations within SBML files as UML
- Library implementations
- MIRIAM guidelines for proper annotation of SBML files
- MIRIAM resources, MIRIAM resolver
  for providing identifiers and links
- ...

Identify uniquely
(e.g. McCurry et al. preprint)

1. If you create identifiers, **do not DIY** (Do Identifiers by Yourself)
2. Help identifiers **travel well**: don't let them leave home without a Prefix and a Namespace
3. Make Local Resource Identifiers **rugged to realworld** use
4. Make the full URI **simple and durable**
5. Carefully consider whether to embed meaning
6. Make the full URI and CURIE **clear and easy to find**
7. Implement a **version management** policy
8. Manage complex **lifecycles** without deletion
9. **Document** the identifiers you issue and use
10. **Reference** responsibly and rely on full URIs

Not covered so far...

- Viewing functionality
- SEEK support for metabolic models using
  JWS online

- Data **citation**
  (give FAIR data a long-term identifier that can be cited)
- Data **publication**
  (publish existing data to a longer-term/more specialized repository)

End
(of this bit)