

# Information Retrieval and Data&Model Management for

Wolfgang Müller + Olga Krebs, HITS



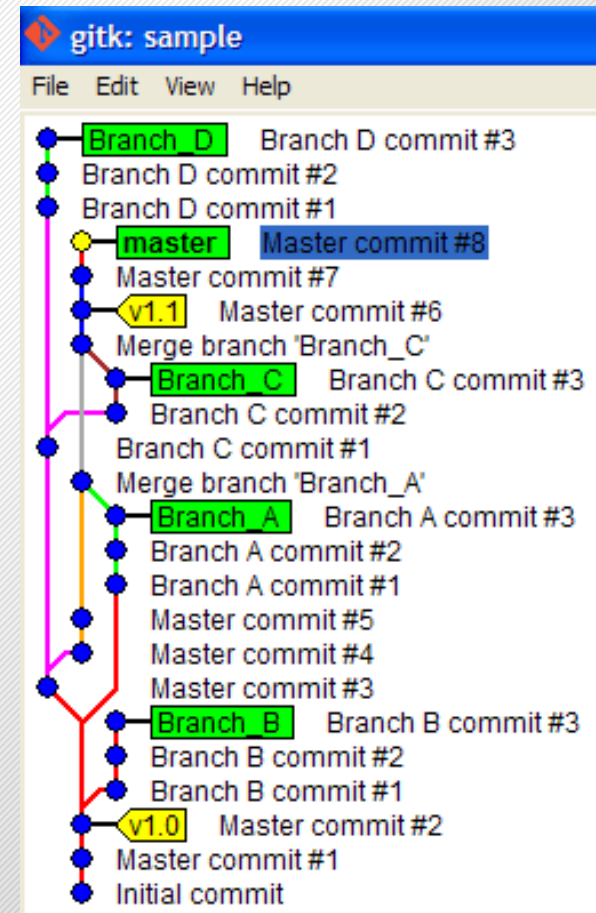
PoLiMeR

Polymers in the Liver - Metabolism and Regulation



## 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



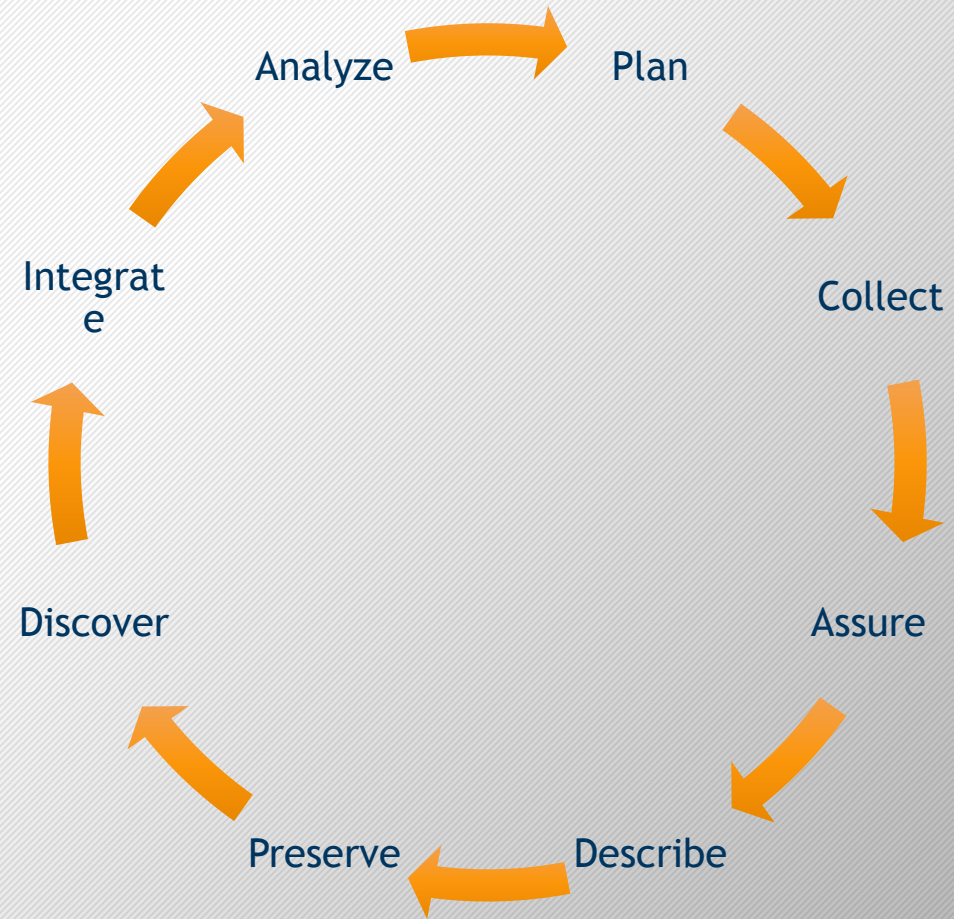


- Please agree with Oliver Ebenhöh
- What to put in there
- How to use
  - Policies for branches
  - Setup of repository (directory structure)

# Data life cycle



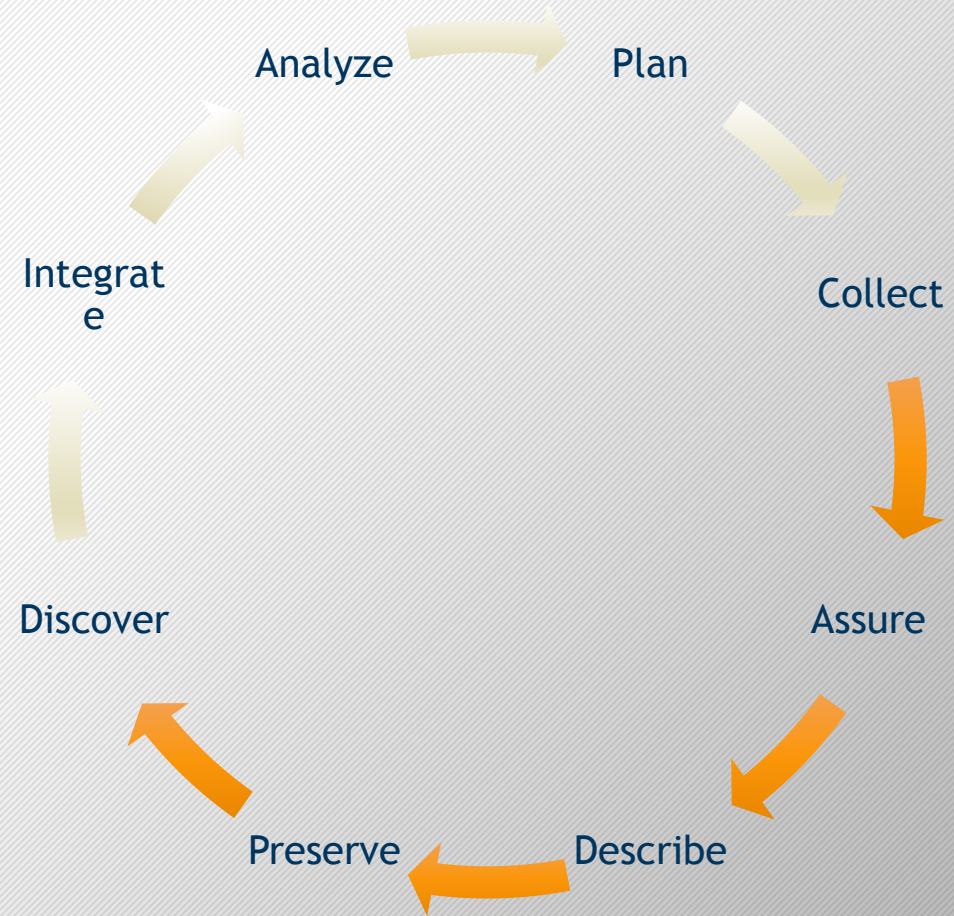
- This cycle consists of...



# Data life cycle



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



**The Economist**  
 OCTOBER 19TH-25TH 2013  
 Economist.com

Washington's lawyer surplus  
 How to do a nuclear deal with Iran  
 Investment tips from Nobel economists  
 Junk bonds are back  
 The meaning of Sachin Tendulkar

**nature** International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Specials & supplements archive > Challenges in irreproducible research

# Reproducibility and Rampancy



# HOW SCIENCE IS BEING DRIVEN FOR OPEN, WORKING, REPRODUCIBLE, AND MAINTAINABLE, FAIR DATA

**REPRODUCIBILITY**

**Journal of Open Access Bionomics**

Nature Biotechnology  
 accessibility of

these computer codes available at a level of detail regarding greater than the analogous non-empirical descriptions printed in natural language.

As for reproducibility in many computer code is no longer available in any concrete form. Even software systems often used for analysis typically do not keep track of the different results together. Addressing this problem will require the behavior of the software to be more amenable. Neither is likely to happen de facto, and many will discard the hours spent learning non-open source software can be their own, who may not view reproducibility in computing as a high priority.

For reproducibility in computing, contributions will need to be made in different directions. Journals can play a role in this effort by the scientific journal *Bioinformatics*, for which

**POLICYFORUM**

As the use of computation in research grows, new tools are needed to expand recording, reporting, and reproduction of methods and data.

**OPEN ACCESS** Freely available online



**Essay**

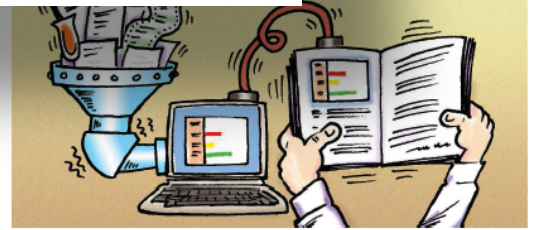
## How to Make More Published Research True

**John P. A. Ioannidis**<sup>1,2,3,4\*</sup>

**1** Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, United States of America, **2** Department of Medicine, Stanford Prevention Research Center, Stanford, California, United States of America, **3** Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, United States of America, **4** Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, California, United States of America

Experimental scientists in many fields of experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension.

Over the past ~35 years, computational science has posed challenges to this traditional paradigm—from the publication of the four-color theorem in mathematics (1), in which the proof was partially performed by a computer program, to results depending on computer simulation in chemistry, materials science, astrophysics, geophysics, and climate modeling. In these settings, the scientists are often sophisticated, skilled, and innovative programmers who develop large, robust software packages.



More recently, scientists who are not themselves computational experts are conducting data analysis with a wide range of modular software tools and packages. Users may often combine these tools in unusual or novel ways. In biology, scientists are now routinely able

between two types of acute leukemia, based on large-scale gene expression profiles obtained from DNA microarrays (3). This paper generated hundreds of requests from scientists interested in replicating and extending the results. The method involved a complex pipeline of steps, including (i) preprocessing of the

language that can produce all of the text, figures, code, algorithms, and settings used for the computational research (16). Although these approaches may accomplish the goal, they are not practical for many nonprogramming experimental scientists using other groups' or commercial software tools today.

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

SUBJECT CATEGORIES  
 » Research data  
 » Publication characteristics



## Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve<sup>1,2\*</sup>, Anton Nekrutenko<sup>3</sup>, James Taylor<sup>4</sup>, Eivind Hovig<sup>1,5,6</sup>

### 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



33  
WORK GROUPS

2,408  
ACTIVE MEMBERS

1,313  
MEMBER POSTS

Sign In

- OR -

Join Now!

JOIN AND BE A PART OF THE COMMUNITY DEFINING THE FUTURE OF SCHOLARSHIP



English

Search



ABOUT

COMMUNITY

GROUPS

RESOURCES

NEWS + BLOGS

EVENTS

PUBLICATIONS

MEDIA

DONATE

THE FAIR DATA PRINCIPLES



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 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.

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.

### 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.

FAIR and accessible  
description



### 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**



- 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



PoLiMeR

Polymers in the Liver - Metabolism and Regulation

# What is information retrieval?



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

The screenshot shows a Google search for "Glukose". The search bar contains "Glukose" and the search button is visible. Below the search bar, there are navigation tabs for "Alle", "Bilder", "Shopping", "Videos", "News", "Mehr", "Einstellungen", and "Tools". The search results show approximately 2.250.000 results in 0,39 seconds. The main result is a knowledge panel for "Glucose" (Chemische Verbindung). The panel includes a description: "Glucose (Abkürzung: Glc) oder **Glukose** (von griechisch γλυκύς ‚süß‘, und -ose als Suffix für Zucker) ist eine natürlich vorkommende chemische Verbindung. D-Glucose wird auch als Traubenzucker oder in älterer Literatur als Dextrose bezeichnet." It also features a ball-and-stick model and a Haworth projection of α-D-Glucose (cyclic). Below the description is a link to the German Wikipedia page for "Glucose" with the URL <https://de.wikipedia.org/wiki/Glucose>. To the right of the knowledge panel, there is a sidebar with the title "Glucose" and the subtitle "Chemische Verbindung". It provides more details: "Glucose oder Glukose ist eine natürlich vorkommende chemische Verbindung. D-Glucose wird auch als Traubenzucker oder in älterer Literatur als Dextrose bezeichnet. D-Glucose ist das häufigste Monosaccharid und gehört als Monosaccharid zu den Kohlenhydraten. Es gibt zwei Enantiomere der Glucose: D-Glucose und L-Glucose." It also lists the formula  $C_6H_{12}O_6$ , molar mass 180,156 g/mol, IUPAC name D-glucose, and name Glucose. The aggregate state is listed as "fest" and it is soluble in "Wasser, Essigsäure". At the bottom of the sidebar, there is a section "Andere suchten auch nach" with a link to "Über 5 weitere ansehen". Below the knowledge panel, there is a section "Nutzer fragen auch" with three questions: "Was bedeutet Glukose im Blutbild?", "Was macht Glucose im Körper?", and "Bei welchem Wert ist man zuckerkrank?".



# 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 engine 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



PoLiMeR

Polymers in the Liver - Metabolism and Regulation

# Data management for systems biology

for de.NBI NBI-SysBio:  
Wolfgang Müller

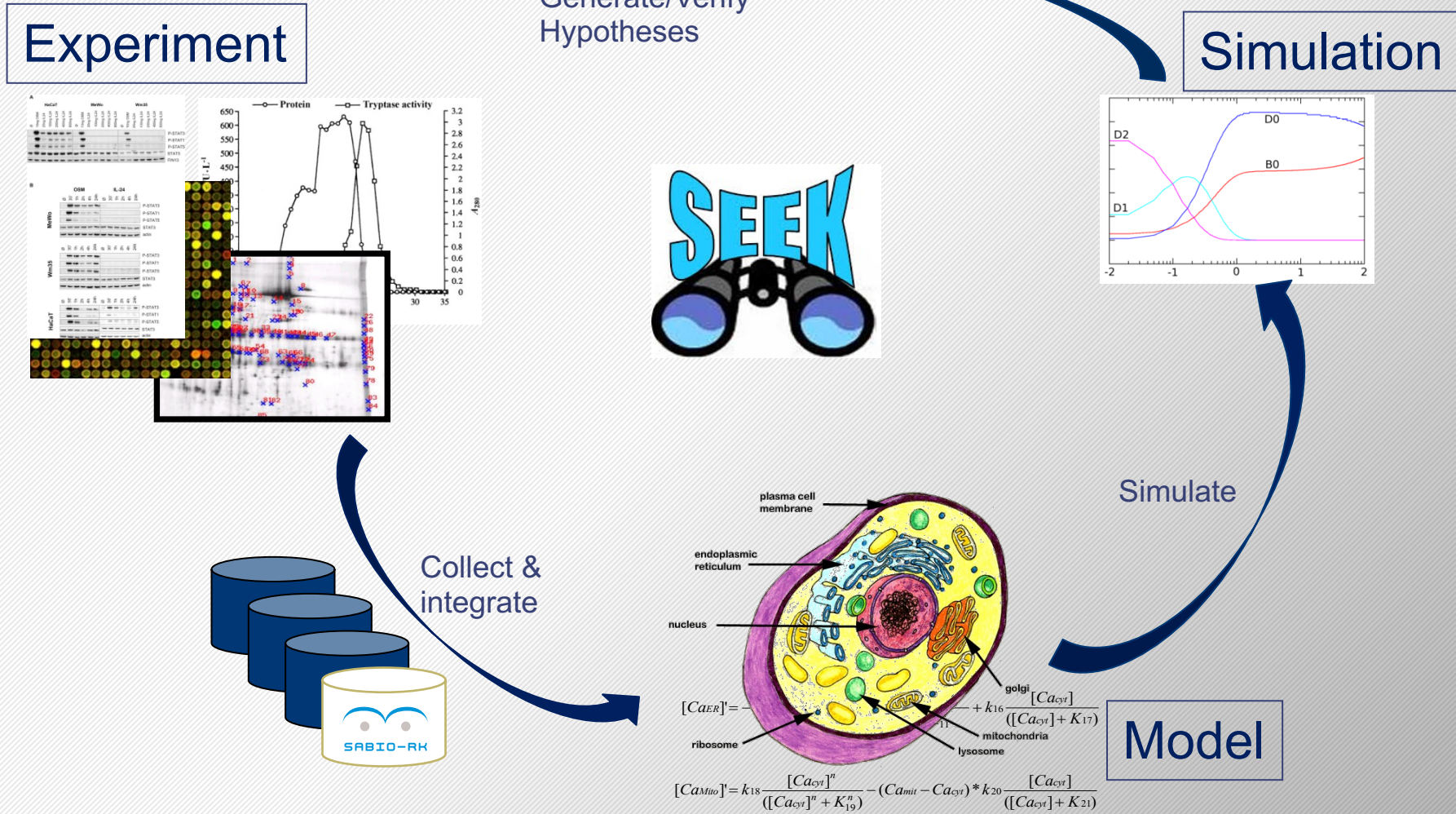


PoLiMeR

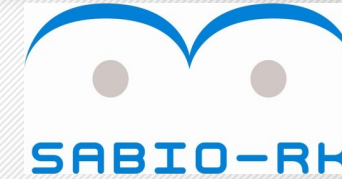
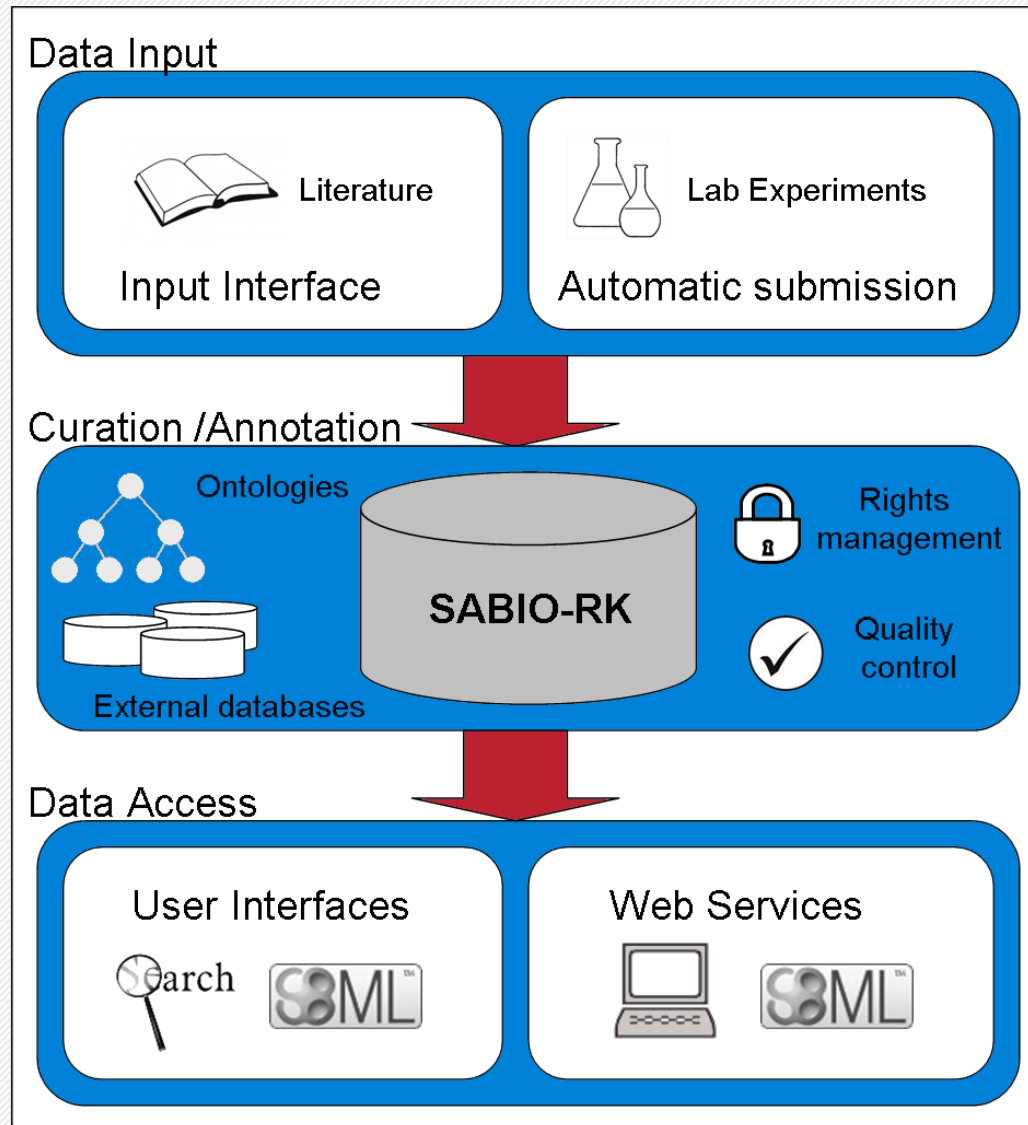
Polymers in the Liver - Metabolism and Regulation



# Systems biology



# SABIO-RK: Reaction Kinetics



<http://sabiork.h-its.org>

**Biochemical reaction kinetics**

## Data

- Integrated
- Structured
- Unified
- Interlinked
- Annotated

# Curation: Add value to data



A

Four panels of scientific publications showing original data. The first panel is a text-based abstract with green highlights. The second panel shows a graph with a linear fit and green highlights. The third panel shows multiple graphs with green highlights. The fourth panel is a text-based abstract with green highlights.

B

Kinetic Data Available for Reaction:  
**D-Fructose 1,6-bisphosphate <-> D-Glyceraldehyde 3-phosphate + Glycerone phosphate**

Expand All Close All

Entry Nr. 34171 [ ] [ ] Select

Organism: **Vigna radiata (strain cv. Wilczek)**

Issue: **germinating seed**

EC Class: **5.1.2.13** **wildtype**

Substrates:

name	location	comment
D-Fructose 1,6-bisphosphate	cytosol	

Products:

name	location	comment
D-Glyceraldehyde 3-phosphate	cytosol	
Glycerone phosphate	cytosol	

Modifiers:

name	location	effect	comment	protein complex
Fructose-bisphosphate aldolase(Enzyme)	cytosol	Modifier-Catalytic		ALDO4

Enzyme (protein data):

subunit	linProt-ID	name	mol. weight (kDa)	deviation (kDa)
complex	-	-	140.0	-

Kinetic Law:

type	formula
Michaelis-Menten	$v_{max} * S / (K_m + S)$

Parameters:

name	type	species	start val.	end val.	deviat.	unit	comment
Km	Km	D-Fructose 1,6-bisphosphate	14.7	-	-	µM	
S	concentration	D-Fructose 1,6-bisphosphate	0	1	-	µM	
Vmax	Vmax	D-Fructose 1,6-bisphosphate	17	-	-	µmol/(min*mg)	
kcat	kcat	-	40	-	-	s <sup>-1</sup> (-)	
kcat	Km/kcat/Km	D-Fructose 1,6-bisphosphate	2.4	-	-	µM <sup>2</sup> (-1)*s <sup>-1</sup> (-)	

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

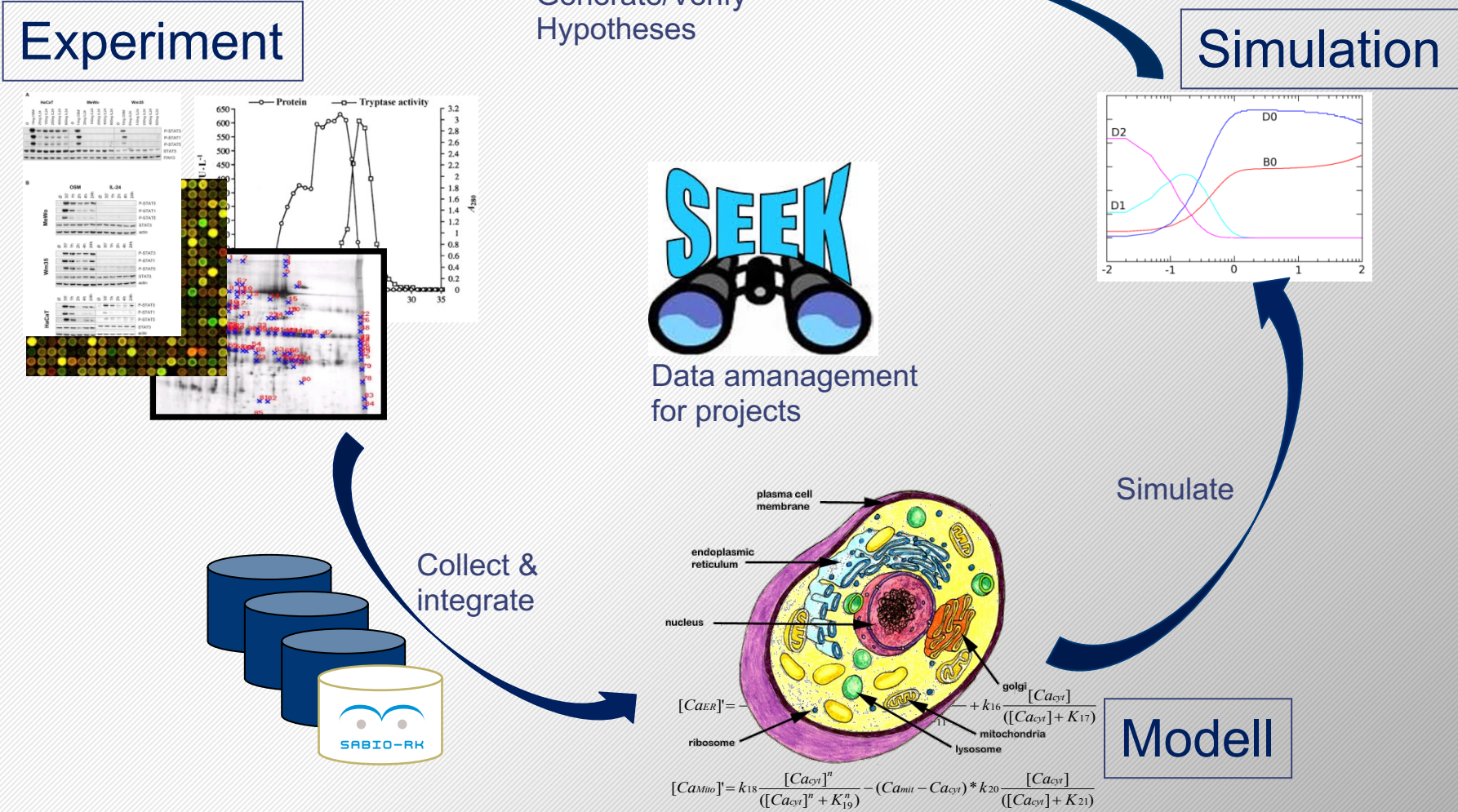


- One paper takes 0.5 to 1 workday
- Reliable automatic curation impossible in our setting
- Limits number of papers you can curate per year
- Even multiple people cannot catch up with everything

→ **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



- Findable
  - Accessible
  - Interoperable
  - Reusable
- 
- Data
  - Operations
  - Models



# Purpose of project data management



- Organisation

Helps you  
find your  
way

Reuse later

- Communication

Helps others  
find out

Enable team  
to reuse

- Dissemination

Tell more  
and take  
credit

Reuse with  
new  
partners

of data associated with a project

# Why does not everyone do it?



- Hard to know how much you need
- Afraid of sharing too much
- Takes time
- ...
  
- ...and other personal factors



# 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?

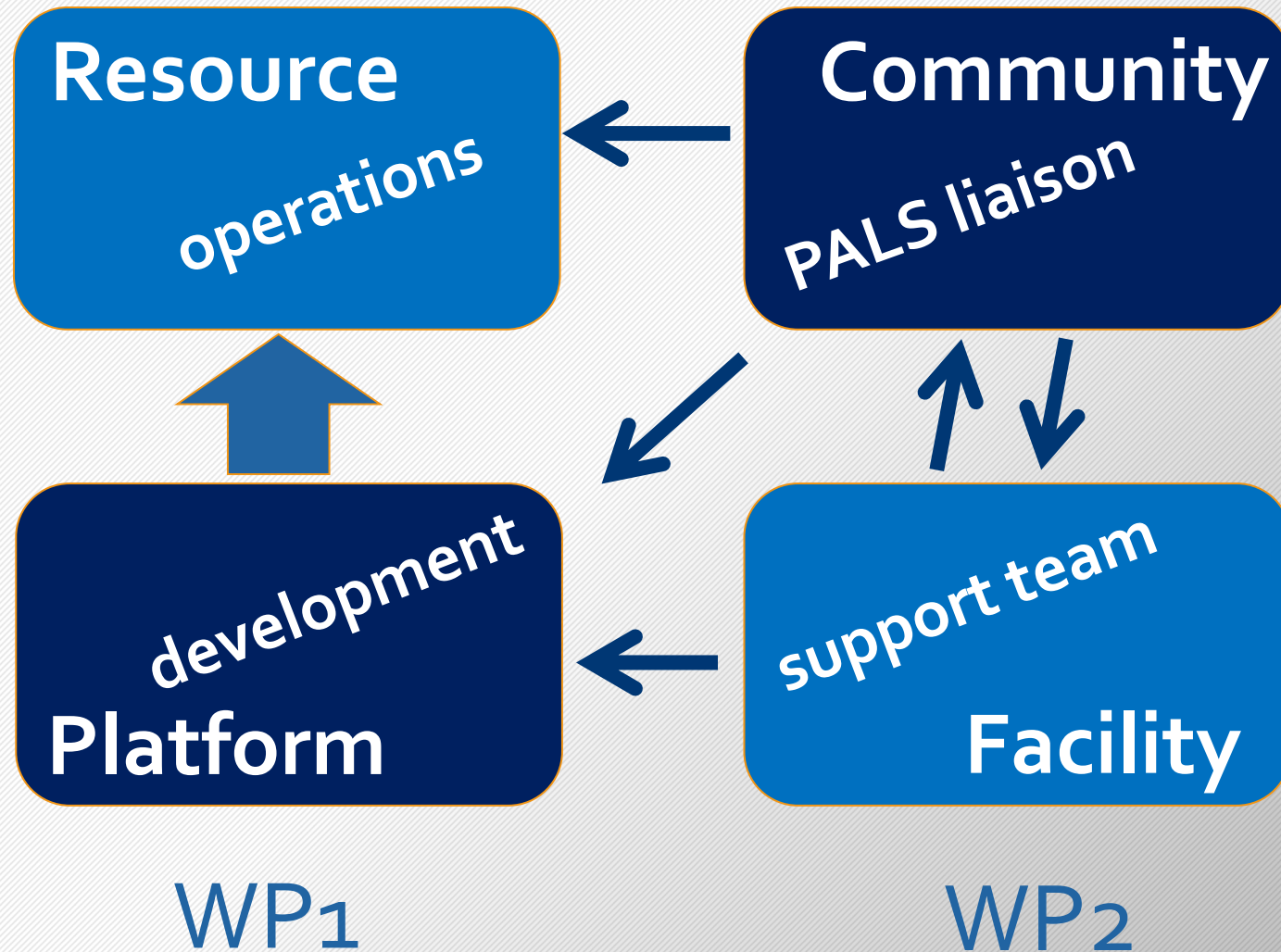


- Raw data (sometimes)
- Condensed, interpreted data
- Metadata: Data about the data
  - **Conditions & Procedures** of the measurements
  - **Information** about the samples
    - What was sampled?
    - How was it prepared?
    - How was it treated after sampling?





**ERASys APP**  
ERA-Net for Applied Systems Biology





# Organisation – Collaboration - Dissimination



https://www.fairdomhub.org



The screenshot displays the FAIRDOME website interface. At the top, there is a navigation bar with the FAIRDOME logo, a search bar, and a user profile for Wolfgang Müller. A search menu is open, listing various categories: Browse, Yellow pages, People, Projects, Institutions, Experiments, Investigations, Studies, Assays, Assets, Data files, Models, SOPs, Publications, and Events. The main content area features a 'Summer school' announcement, a 'Frequently Asked Questions' section, and a list of organisms including Bacillus subtilis, Clostridium acetobutylicum, Enterococcus faecalis, Escherichia coli, Lactic Acid Bacteria, Lactobacillus plantarum, Lactococcus lactis, Pseudomonas fluorescens, Pseudomonas putida, Saccharomyces cerevisiae, Streptococcus pyogenes, Streptomyces coelicolor, Sulfolobus solfataricus, and Trypanosoma brucei. The footer includes logos for YGCid, BioPortal, semantic SBML, and the FAIRDOME logo.



# Science is about people and credits

Q Browse ▾

+ Create

Browse

Yellow pages

People

Projects

Institutions

Experiments

Investigations

Studies

Assays

Assets

People in the system

Projects in the system

Participating institutions

https://www.fairdomhub.org



The screenshot shows the FAIRDOM website interface. At the top, there is a navigation bar with the FAIRDOM logo, a search bar, and a user profile for Wolfgang Müller. Below the navigation bar, there are several content sections:

- Left sidebar:** Contains a search bar, a "Browse" dropdown menu, and a "Latest additions" section.
- Main content area:** Features a large "Browse" dropdown menu with the following options: Browse, Yellow pages, People, Projects, Institutions, Experiments, Investigations, Studies, Assays, Assets, Data files, Models, SOPs, Publications, and Events.
- Right sidebar:** Includes an "Announcements" section, a "Tags [show all]" section, and an "Organisms" section listing various species such as *Bacillus subtilis*, *Clostridium acetobutylicum*, *Enterococcus faecalis*, *Escherichia coli*, *Lactic Acid Bacteria*, *Lactobacillus plantarum*, *Lactococcus lactis*, *Pseudomonas fluorescens*, *Pseudomonas putida*, *Saccharomyces cerevisiae*, *Streptococcus pyogenes*, *Streptomyces coelicolor*, *Sulfolobus solfataricus*, and *Trypanosoma brucei*.

At the bottom of the page, there are logos for various partners and a "Powered by SEEK" logo.



Q Browse ▾

+ Cre

Browse

Yellow pages

People

Projects

Institutions

Experiments

Investigations

Studies

Assays

Assets

A few per  
project

Roughly one  
paper

„smallest meaningful  
group of  
measurements“





# New investigation



Genome Annotation and Ar x The SEEK de.NBI Investiga x

https://denbi-school.fairdomhub.org/investigations/new

Apps HITSspec New Medicines for T Socket in Sails with Managing Compass Aktuelle Freigaben fu » Andere Lesezeichen

FAIRDOM Search here... Search Wolfgang Müller

Home / Investigations Index / New

## New Investigation

**Title \***

**Description**

**Projects**

The following projects are associated with this investigation:

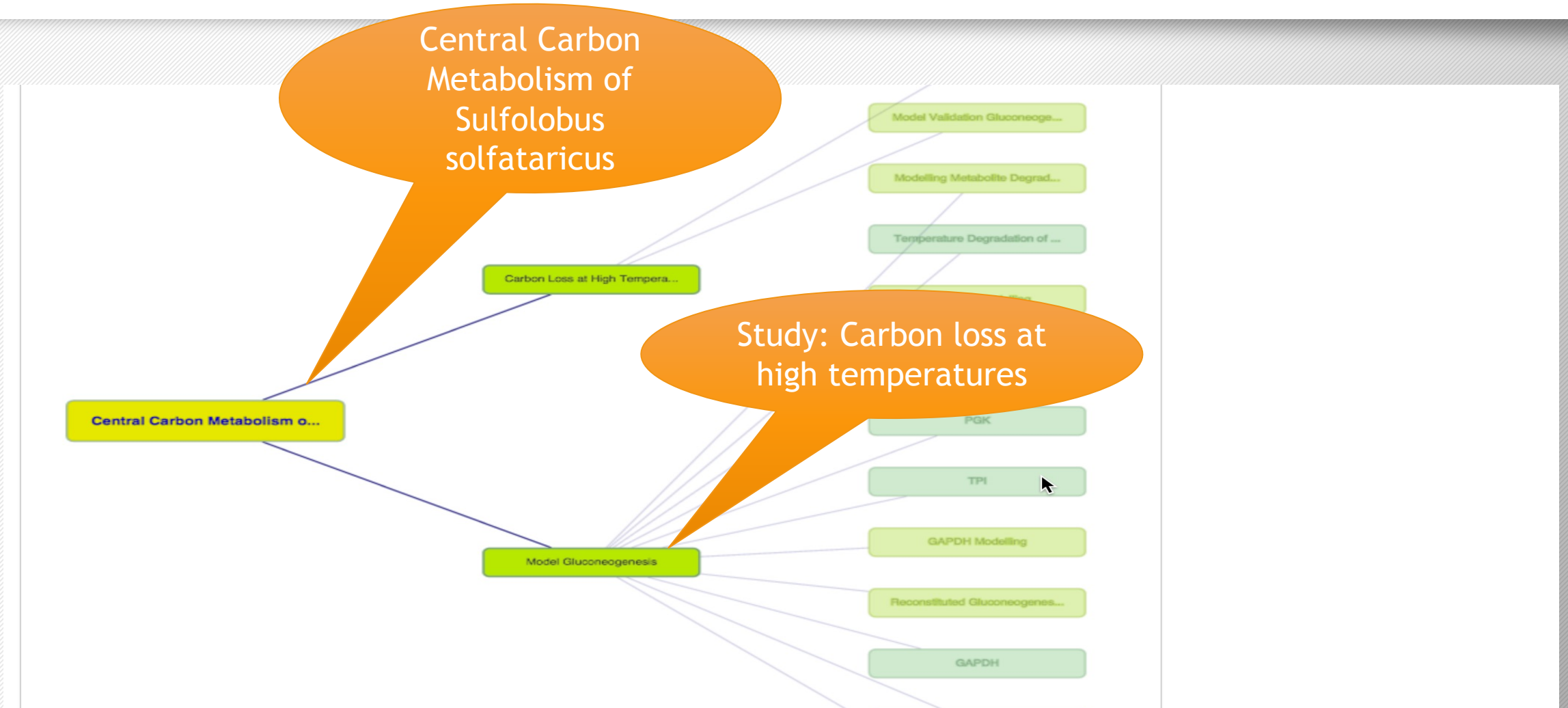
de.NBI summer school [\[remove\]](#)

Select Project ...

**Sharing**

Here you can specify who can **view** the summary of and **edit** the Investigation. [More info](#)

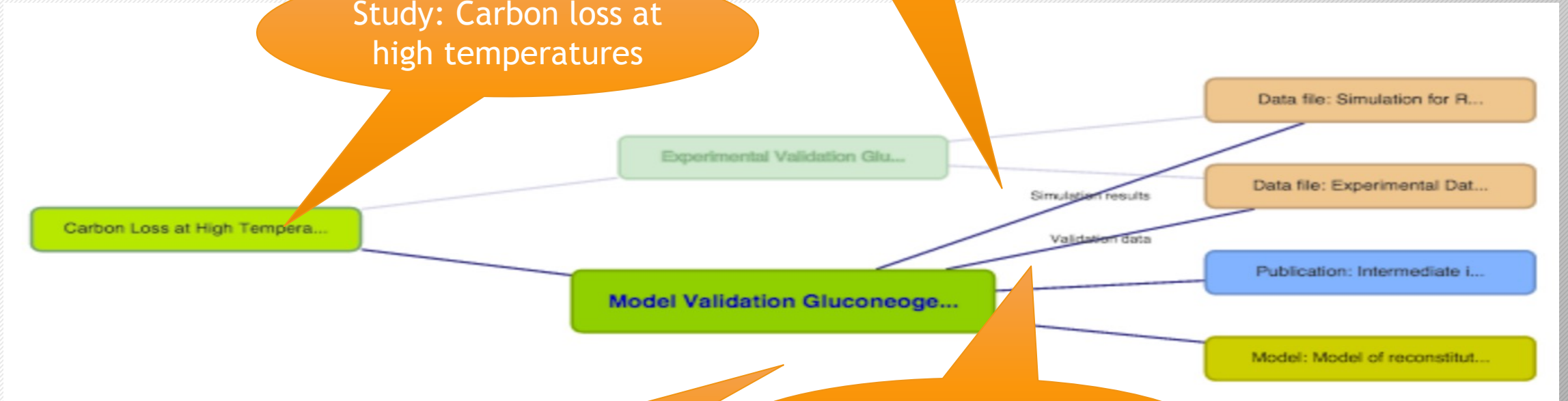
# An example: Investigation, Study, Assay





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



**FAIRDOM** | Search | My Account | My Account

## New Data file

**Overview**

You can upload a Data File by either directly uploading a file, or supplying a URL, to either another page or outside the Platform. Please read the URL before submitting.

Upload File:

File to upload?

**Title**

Text area for title

**Description**

This file is a best practice for the FAIR Data standard. This text content is just the description on how to get to the data.

**Projects**

The following projects are associated with this data file:

**Sharing**

How can you best specify who can view the summary of, get access to the content of, and edit the Data File? (Show this...)

Public: This Data File is publicly available to all.

Private: This Data File is private (only visible to you).

Members of Project: This Data File is only visible to members of the following project(s):

All: This Data File is publicly available to all.

**Tags**

How can you describe better your data to help describe the Data File? You can change these tags (including those provided by other projects, such as in the FAIR Data Tag) and add new tags to describe your data, and these tags will appear in the Data File's metadata.

**Attributions**

If this Data File is based on any existing Data File, please list them below:

Please type Data File ID for each attribution. Suggestions will be displayed as you type. Select attributions that you want to attribute to.

**Contributors**

If this Data File is based on any existing Data File, please list them below:

The contributors are people who have contributed to the creation of this Data File.

Please type Data File ID for each contributor. Suggestions will be displayed as you type. Select contributors that you want to include and click "Add" to add them to the list.

**Publications**

The following Publications are related to this Data File:

**Equipment**

The following Equipment details and Metadata are associated with this Data File:

**Events**

The following events are associated with this Data File:



**Samples**

The following samples are associated with this Data File:

© 2014 FAIRDOM. All rights reserved. FAIRDOM is a trademark of FAIRDOM. All other trademarks are the property of their respective owners.

# Basic description; Register vs. Upload



 **FAIRDOM**   

[Home](#) / [Data files Index](#) / [New](#)

## New Data file

Upload

You can register a Data file by either directly uploading a file, or registering a URL to either another page or remote file by submitting.

Local file  [Remote URL](#)

**File to upload \***


Wegbeschre...\_Engl.doc

*Note: An orange callout bubble points to the 'Remote URL' radio button with the text 'Choose if to register or upload'.*



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](mailto: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

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...

[\[read more\]](#)



# 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 Search Reset Advanced

1314 Standards registered!

Showing records 1 - 50 of 1314.

View as Table View as Grid

Sort by  
Name

Recommended Records

«	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
	<a href="#">ABA Adult</a>	<a href="#">ABA</a>	Standard	None	<a href="#">Brain</a>	<a href="#">Mus musculus</a>	<a href="#">NeuroMorpho.o</a>

# 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)

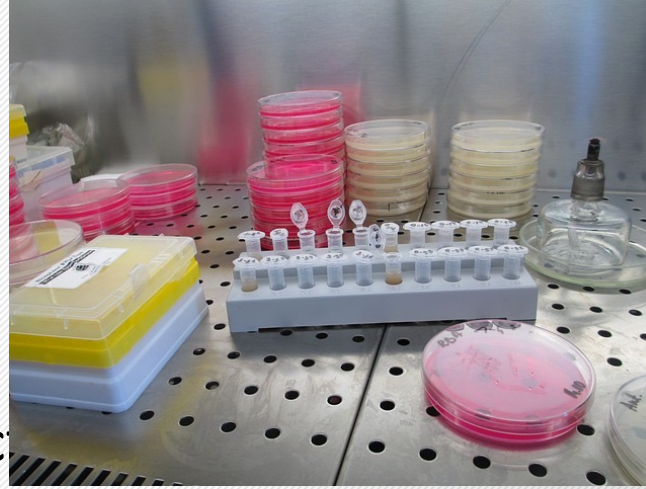
# Standardisation scales



Increased usability for others



- Self
- Group
- Collaborative project
- Field scale



# 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 occurrences“?
- **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**



**Refine** <sup>OPEN</sup> 





# 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



B8								
	A	B	C	D	E	F	G	H
1	<b># 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</b>							
2								
3	<b>Metadata</b>	<b>Values (examples)</b>		<b>Notes</b>				
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	<b>ASSAY</b>							
11	Assay SEEK ID			If referring to an existing Assay, you can link to it via the Assay SEEK ID.				
12	Assay Title			The title of an existing 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	<b>Experimental_conditions</b>							
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)

The



End  
(of this bit)