Management of Scientific Data

ChemBioSys

Matthias Bruhns, Tobias Giesemann, Lucas Schneider, Tabitha Uphoff

29.04.2019
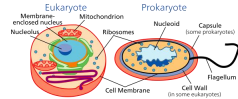
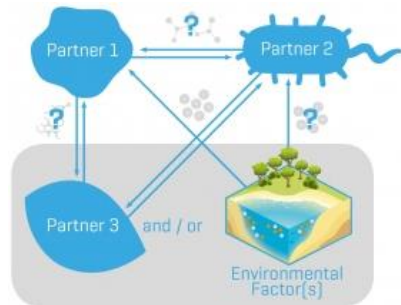Friedrich-Schiller-Universität Jena

Project Background

ChemBioSys -Goals

Goals of ChemBioSys:

1. Discovery of chemical mediators and their targets
2. Understanding the mechanisms that generate community structures
3. Directed modulation of complex Biosystems

Understanding Complex Communities

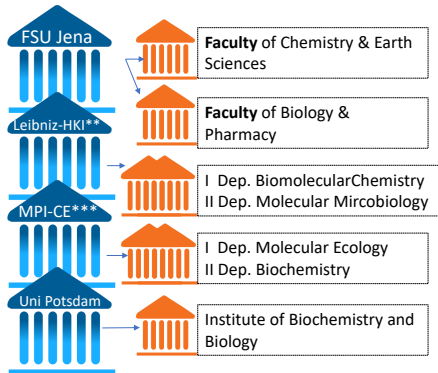http://www.chembiosys.de/en/research/

ChemBioSys - Project

- Second DFG funding period (2018 -2022)
- Large Project: Currently ca. 70 active scientists
- Phd- and Masters' Theses, Students' Projects

ChemBioSys – Organisational Structure

*not listed **Hans-Knöll-Insitute ***Max-Planck-Institute for Chemical Ecology, RG: Research Group

ChemBioSys – Organisational Structure

FSU Jena	**Faculty** of Chemistry & Earth Sciences
	Faculty of Biology & Pharmacy
Leibniz-HKI**	I Dep. BiomolecularChemistry II Dep. Molecular Mircobiology
MPI-CE***	I Dep. Molecular Ecology II Dep. Biochemistry
Uni Potsdam	Institute of Biochemistry and Biology

*not listed **Hans-Knöll-Insitute ***Max-Planck-Institute for Chemical Ecology, RG: Research Group

ChemBioSys – Organisational Structure

FSU Jena

Faculty of Chemistry & Earth Sciences → 3 Institutes*

Faculty of Biology & Pharmacy → 4 Institutes*

Chair for Natural Product Chemistry

Center for Molecular Biochemistry

Leibniz-HKI**

I Dep. BiomolecularChemistry
II Dep. Molecular Mircobiology

MPI-CE***

I Dep. Molecular Ecology
II Dep. Biochemistry

RG: Chemical Biology of Microbe-Host interactions

Uni Potsdam

Institute of Biochemistry and Biology

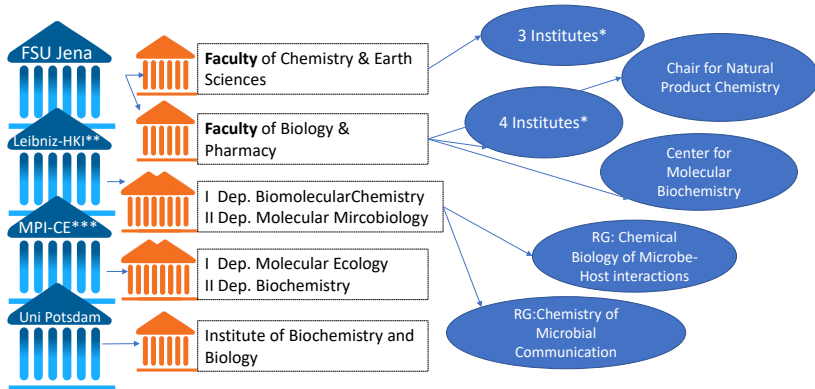RG:Chemistry of Microbial Communication

*not listed **Hans-Knöll-Insitute ***Max-Planck-Institute for Chemical Ecology, RG: Research Group
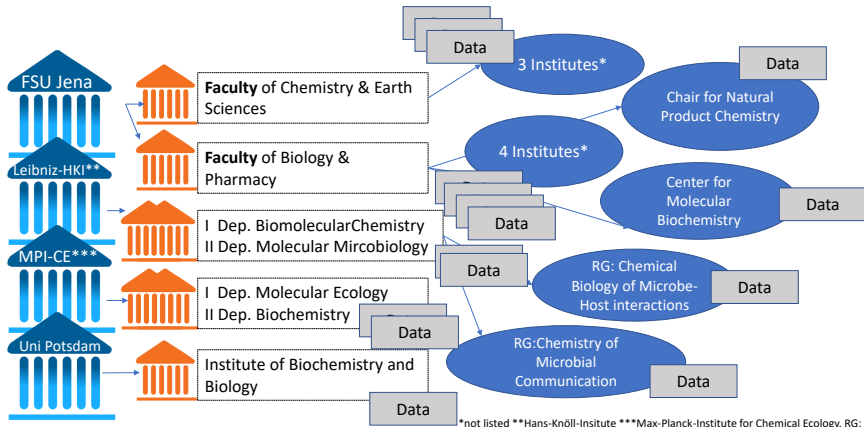
ChemBioSys – Organisational Structure

ChemBioSys – Data Overview

Mass-Specrometry LC-MS, GC-MS	Nuclear Magnetic Resonance Spectroscopy (NMR)	Gene Sequencing	Image Data	Metabolism Data

Mass-Specrometry LC-MS, GC-MS

Molecule Mass Data

LC-MS and GC-MS produce similar data

Nuclear Magnetic Resonance Spectroscopy (NMR)

Molecule / protein detection & analysis

Structure, dynamics, reaction state, chemical environ-ment of molecules

Gene Sequencing

Mostly DNA/RNA-Strings A,U,T,G,C

Image Data

Images (jpeg, png, svg, …)

Usages:
- Petri dishes
- Plants
- Bacteria
- etc.

Metabolism Data

…

Status of data management

Status of data management

- The previous data management team neglected their tasks
- Data management has to be implemented from scratch
- Problem: " We always did it this way!" - mentality
- Some labs use online repositories for their data management
- Disparate formats of meta data storage (`.txt`, `.csv`, `.docx`)
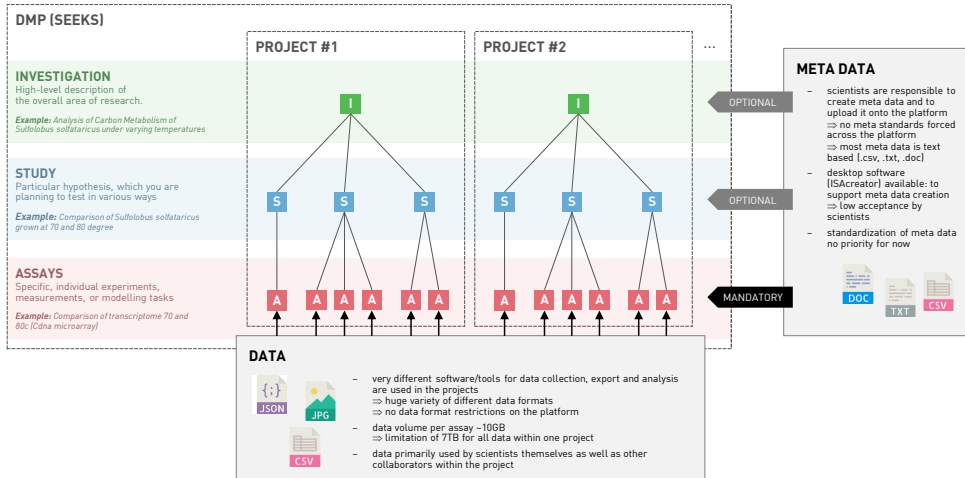- Objective of BioChemSys: show the advantages of a well implemented DMS

Challenges & Trends

- No data management mindset
  - People often don't see the benefit in propper data management
  - Leads to varying formats of datasets
- Space limitation for storing Data
- Limited time and money for data management
- Data access
  - Most scientists keep their data private

datasets/meta data storage

DMP (SEEKS)

PROJECT #1

PROJECT #2

...

INVESTIGATION
High-level description of
the overall area of research.

Example: Analysis of Carbon Metabolism of
Sulfolobus solfataricus under varying temperatures

STUDY
Particular hypothesis, which you are
planning to test in various ways

Example: Comparison of Sulfolobus solfataricus
grown at 70 and 80 degree

ASSAYS
Specific, individual experiments,
measurements, or modelling tasks

Example: Comparison of transcriptome 70 and
80c (Cdna microarray)

OPTIONAL

OPTIONAL

MANDATORY

META DATA
- scientists are responsible to
  create meta data and to
  upload it onto the platform
  ⇒ no meta standards forced
  across the platform
  ⇒ most meta data is text
  based (.csv, .txt, .doc)
- desktop software
  (ISAcreator) available: to
  support meta data creation
  ⇒ low acceptance by
  scientists
- standardization of meta data
  no priority for now

DOC TXT CSV

DATA
- very different software/tools for data collection, export and analysis
  are used in the projects
  ⇒ huge variety of different data formats
  ⇒ no data format restrictions on the platform
- data volume per assay ~10GB
  ⇒ limitation of 7TB for all data within one project
- data primarily used by scientists themselves as well as other
  collaborators within the project

JSON JPG

CSV

Questions?