Omics Data Management

Dieter Beule

Translational Bioinformatics & CUBI, BIH@Charite, Berlin
Oktober 2024

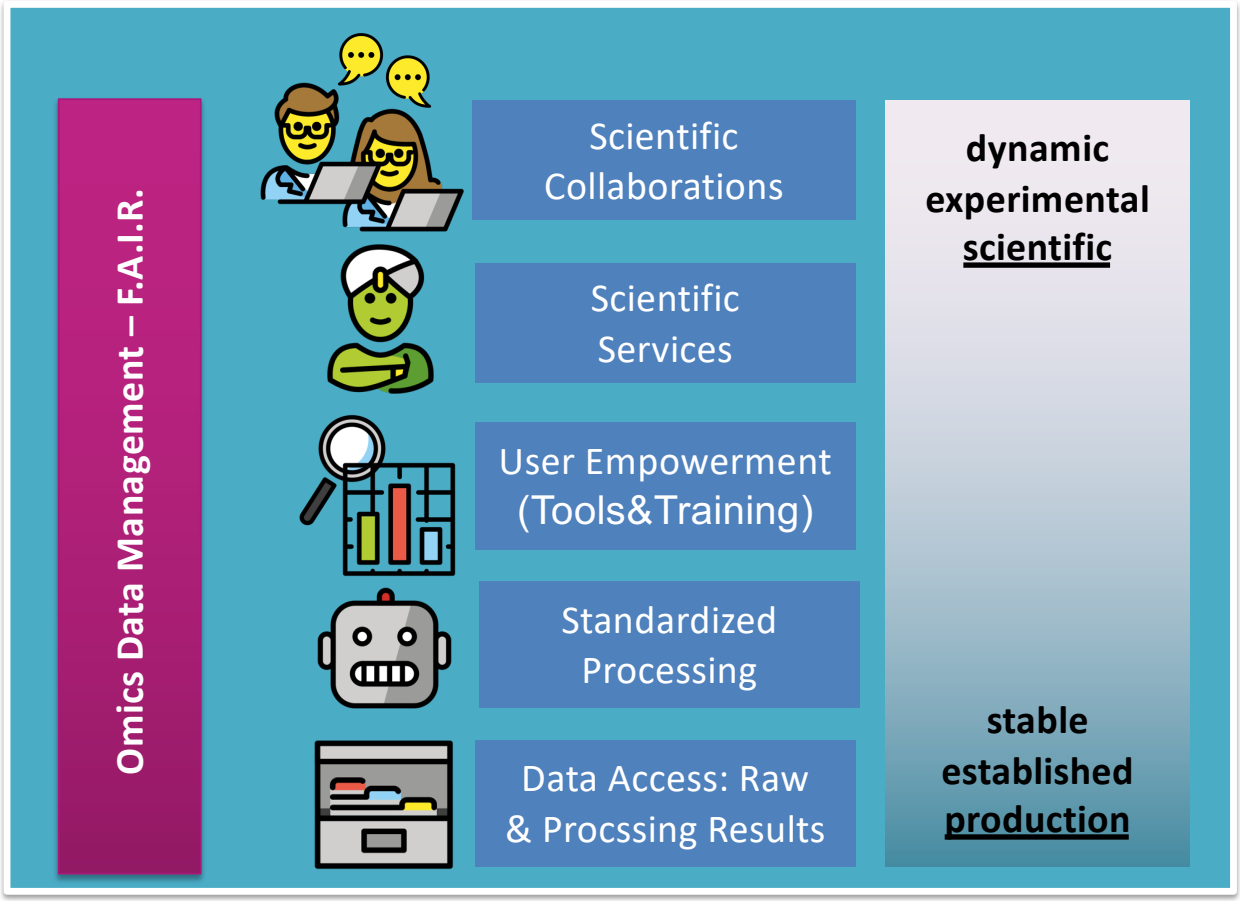
CUBI
BIH Core Unit Bioinformatics

BIH Berlin Institute of Health @Charité

Agenda

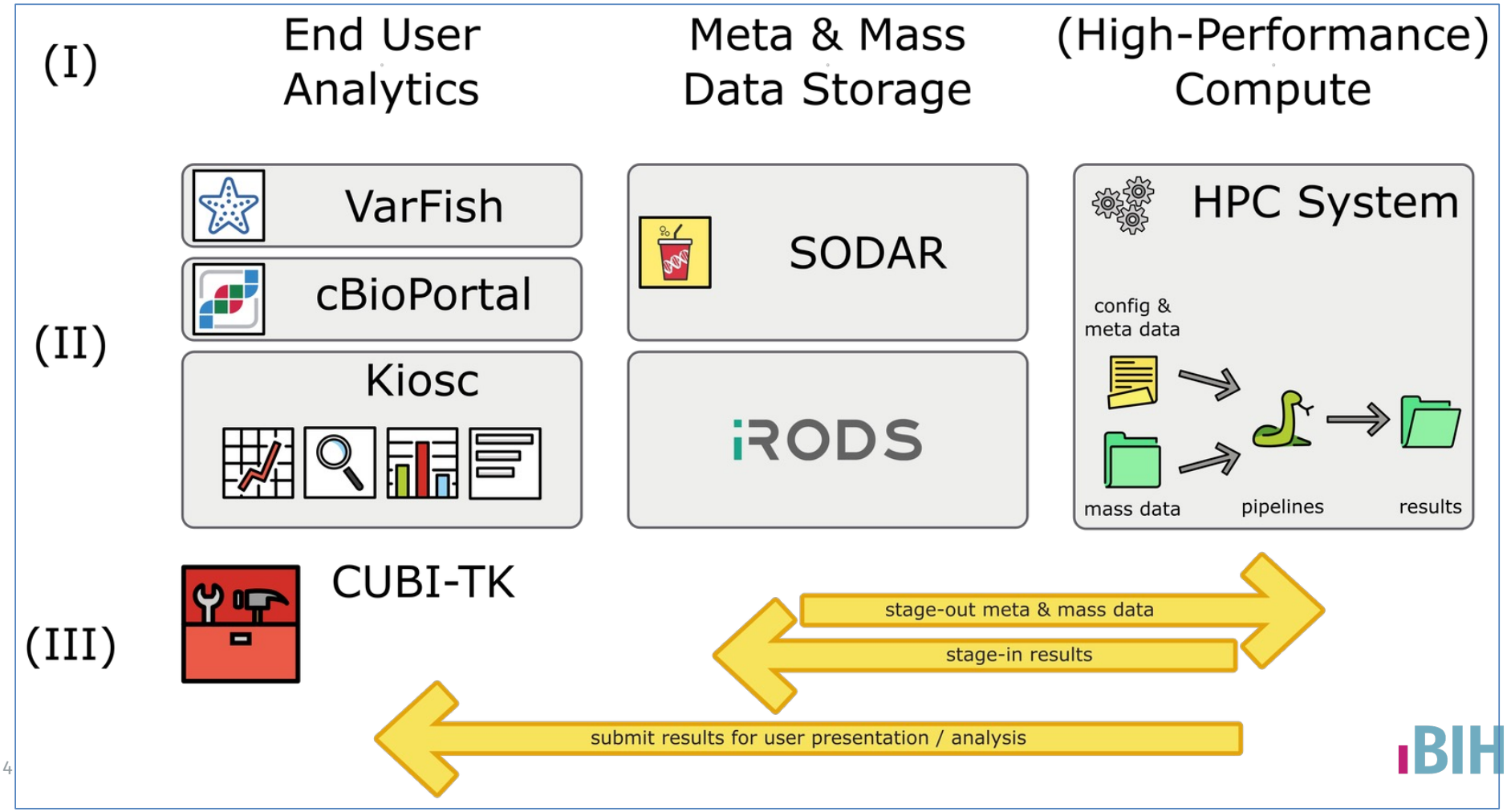
12:30	*Welcome & registration*	
13:00	**Opening remarks**	Dieter Beule
13:10	SODAR introduction and status	Mikko Nieminen
13:50	How-to: Sample sheets & ISA-tabs	Thomas Sell
14:20	*Soda break*	
14:30	User stories	
	Olufemi Bolaji [Sawitzkis Lab]	
	Victor Sikora [Sander & Gaebler Lab]	
15:00	Feedback and tutorial session	Mikko & Thomas
15:50	Closing remarks	Dieter
16:00	*Get together*	

Translation?



BIH Berlin Institute of Health @Charité

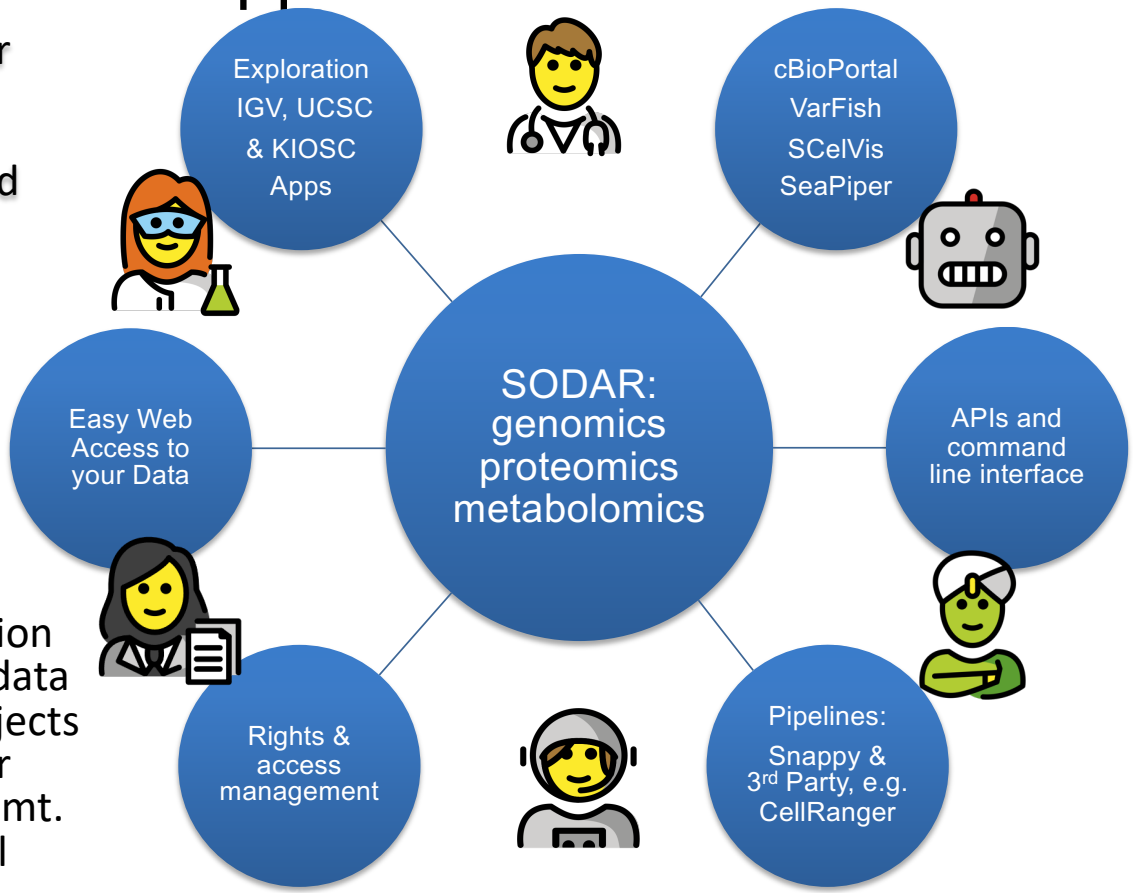
Data Processing & User Empowerment Environment



Berlin Institute of Health @Charité

Translational Research with Omics Data Ecosystem Approach – The SODARverse

System for Omics Data Access and Retrieval



In Production
- 400+ TB data
- 250+ projects
- 300+ user
- rights mgmt.
- audit trail

- Scalable
- Flexible
  - Use Cases, Data Types
- Open
  - Best Practice Integration
  - Workflows
  - Visualization
  - User Empowerment
- Reproducible
- Affordable & No Lock-In
  - Open Source is key
- Secure und F.A.I.R.

Icons: OpenMoji.org

BIH Berlin Institute of Health @Charité

Scientific Data Management Categories

- Four categories (Machina HK, Wild DJ. J Lab Autom. 2013)
  - laboratory information management systems (LIMS)
  - electronic laboratory notebooks (ELNs)
  - scientific data management systems (SDMSs)
  - and a chromatography data system that we generalize as an instrument-specific data system (IDS)

- Nieminen et al. GigaScience 2023 added
  - data repository systems (DRSs)
  - data warehouse management frameworks (DMFs).

Data Management Categories

LIMSs - laboratory information management systems

focus on storing **information** around **laboratory workflows**. This includes tracking of consumables, samples, instruments, and tests. They deal with daily tasks of laboratories such as billing and instrument calibration. They are often specific to certain domain areas.

ELNs - electronic laboratory notebooks

focus on allowing **humans to record** their laboratory **work**. They replace paper notebooks and capture experiments and their results, mostly in free-form text, pictures, tables, and so on. They play a key role in fulfilling regulatory requirements.

Data Management Categories

SDMSs - scientific data management systems

provide scientific **content management** functionality for scientific data and documentation. They allow for the management of **metadata** and potentially **mass data**. Their core functionality does not include data analysis, user-centric data collection, or laboratory workflow tracking. Such features may be potentially supported by **plugins or extensions**. Many such systems offer integration with surrounding systems

IDSs - instrument-specific data system

provide **data capturing**, storage, and analysis functionality in **instrument-specific** domains. Two examples are the CASAVA pipeline and the BaseSpace cloud-based service, both from Illumina. Such software often ships with the instruments themselves.

Data Management Categories

DRSs - data repository systems

provide **shared access** to data with appropriate documentation and metadata. Examples are FAIRdom Seek, Dataverse, and Yoda. There also specialized DRSs focusing on particular use cases, such as **dbGAP, MetaboLights, and GEO**, that allow for managing public or controlled public access to large research data collections.

DMFs - data warehouse management frameworks

allow for the **rapid development** of database and data **warehouse applications**. They often provide preexisting components to build on ready-made functionality and extension by implementing custom components. Such enable creating domain-specific databases and structured data capturing. Examples include Molgenis and Zendro.

People

PhD-Students

- Erika Zuljan
- Vinzenz May
- Aliki Grammatikaki
- Fatheme Habibolahi
- Cedric Moris
- Mireille Tchouto
- Tzu-Ting Wei

Bachelor&Master-Students

- Dzmitry Hramyka
- Pham Gia Cuong

Grants & Projects

- Miha Milek (CRC1588)
- Shuba Alampalli (TR241&2841)
- **Thomas Sell** (MSTARS)
- N.N. (GHGA)
- Divyaratan Popli (NCT)
- Daniel Wendisch (Locotac)
- Manuela Benary (PPK-C)
- Nicolai v Künkelen (CADS)

Research

- Benedikt Obermayer
- **Mikko Nieminen**
- Manuel Holtgrewe
- Mathias Kuhring

CUBI & HPC-Service

- January Weiner
- Eric Blanc
- Andranik Ivanov
- Oliver Stolpe
- Till Hartmann

BIH Berlin Institute of Health @Charité

Thank You!

www.cubi.bihealth.org

Berlin Institute of Health @Charité

Wrap Up & Outlook

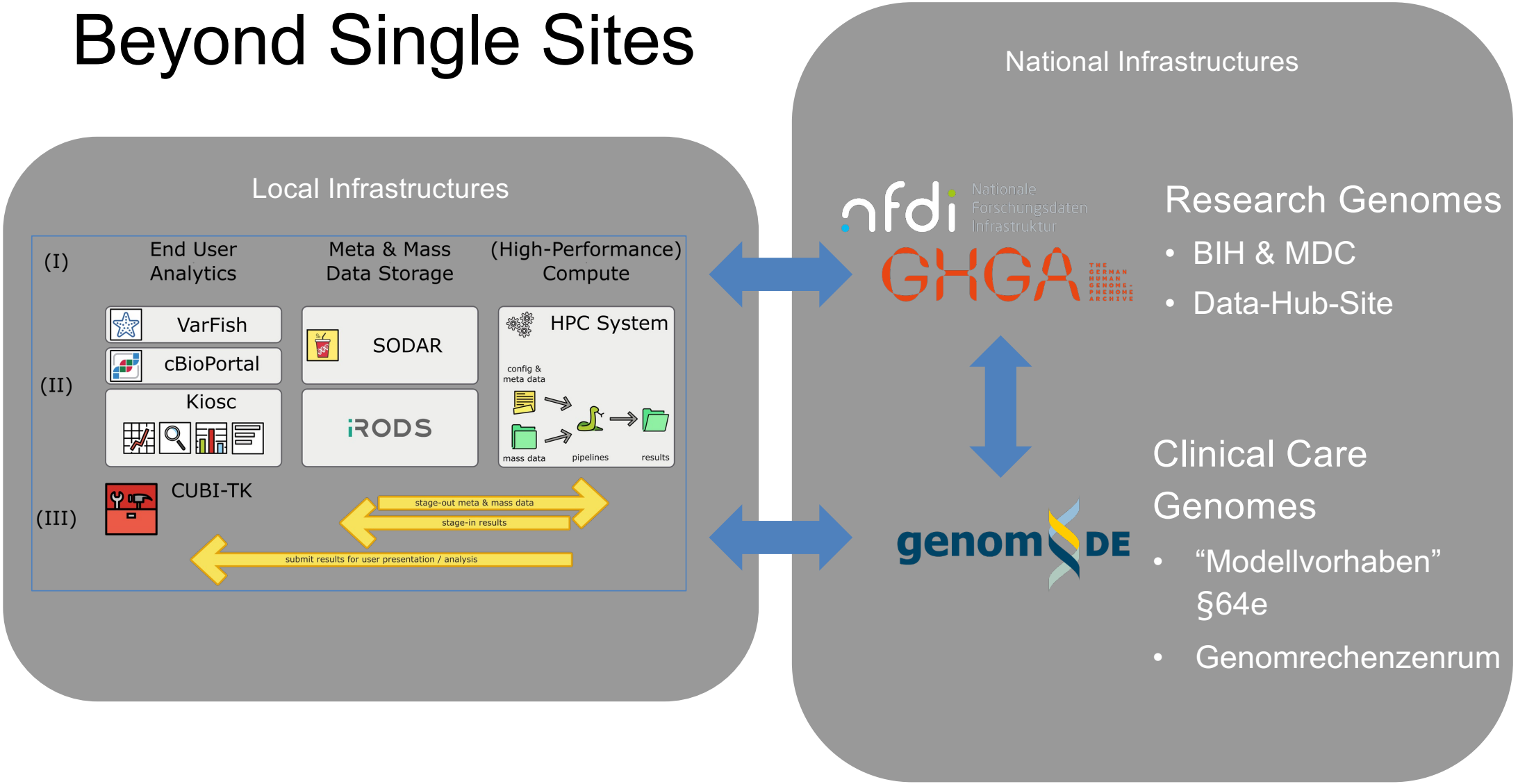
Dieter Beule

Translational Bioinformatics & CUBI, BIH@Charite, Berlin
Oktober 2024

CUBI
BIH Core Unit Bioinformatics

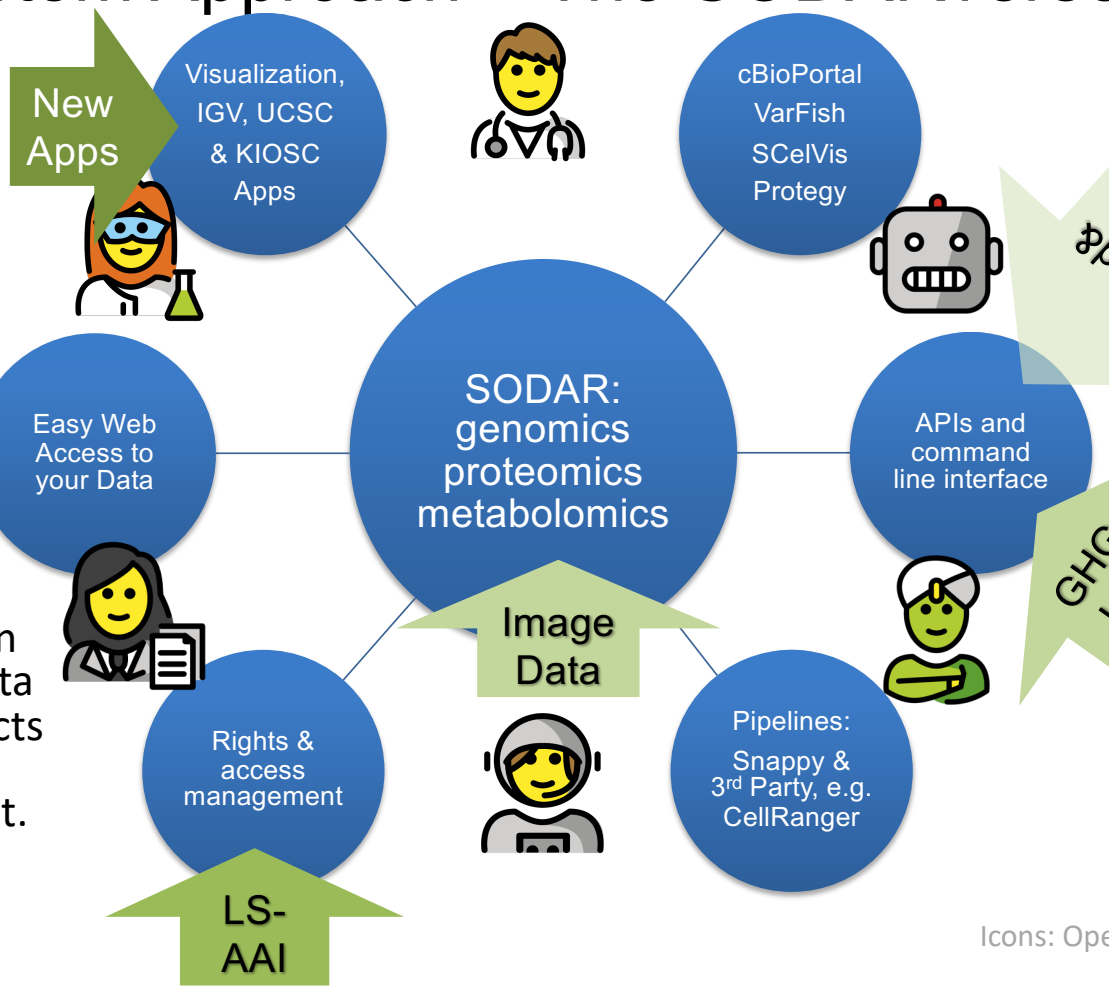
BIH Berlin Institute of Health @Charité

Beyond Single Sites



ROADMAP: Translational Research with Omics Data Ecosystem Approach – The SODARverse

System for Omics Data Access and Retrieval



In Production
- 800+ TB data
- 650+ projects
- 600+ user
- rights mgmt.
- audit trail

- Scalable
- Flexible
  - Use Cases, Data Types
- Open
  - Best Practice Integration
  - Workflows
  - Visualization
  - User Empowerment
- Reproducible
- Affordable & No Lock-In
  - Open Source is key
- Secure und F.A.I.R.

Icons: OpenMoji.org

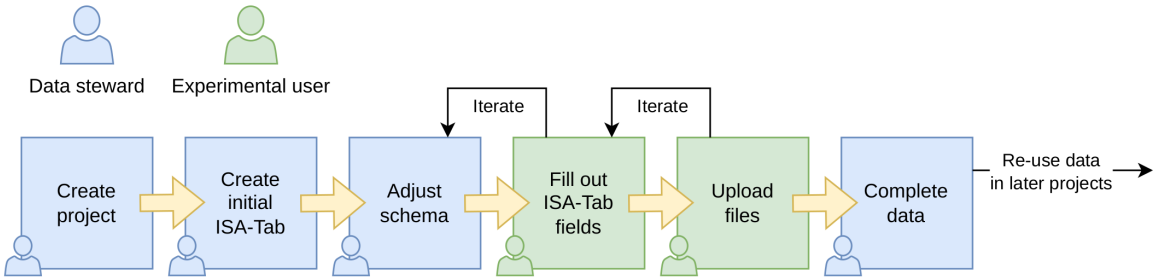
BIH Berlin Institute of Health @Charité

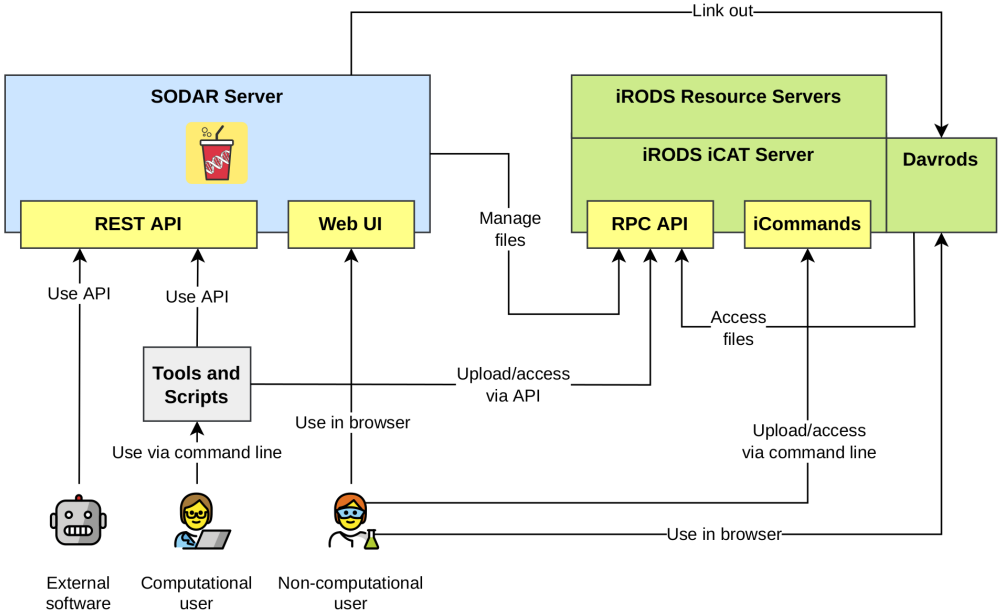
Summary

Paper

Nieminen M, **Stolpe O**, **Kuhring M**, **Weiner J**, **Pett P**, **Beule D**, **Holtgrewe M** (2023). “SODAR: managing multiomics study data and metadata.” *Gigascience*, *12*, giad052. doi:10.1093/gigascience/giad052

Video Tutorial:
https://www.youtube.com/watch?v=LQ8foUpjnqs
User Manual:
https://sodar-server.readthedocs.io/en/latest/
Demo Site:
https://sodar-demo.cubi.bihealth.org/
Source Code:
https://github.com/bihealth/sodar-server





Highlights

- ISAtab Elixir metadata standard
- MIT license, Installable via Docker
- Application cloud support
- More to come …

BIH Berlin Institute of Health @Charité

Thank You!

www.cubi.bihealth.org

Berlin Institute of Health @Charité