

KIT Data Manager: Demo Nanoscopy

Ajinkya Prabhune

Thomas Jejkal, Volker Hartmann, Jürgen Hesser, Margund Bach, Eberhard Schmitt

Institute for Data Processing and Electronics (IPE)



Motivation & Goal

Motivation:

- Extending the infrastructure and services for handling large data sets
→ biological and medical diagnosis are challenging

Goal:

- Efficient solution to manage large data sets
 - Research, design and build a nanoscopy open data reference archive
 - Seamless integration with KIT Data Manager, Large Scale Data Facility (LSDF) and High Performance Computing (HPC)

Nanoscopy

■ Scientific perspective

- Investigation on “*aggressive B-cell lymphomas*”
- Microscopy technique – Spectral Precision Distance Microscopy (SPDM)
- Microscope installations
 - University Mainz
 - University Heidelberg

■ Technical perspective

- Data produced in range of 150-200 TB
- Data is produced in multiple formats
- Associated metadata is unstructured
- Analysis workflow

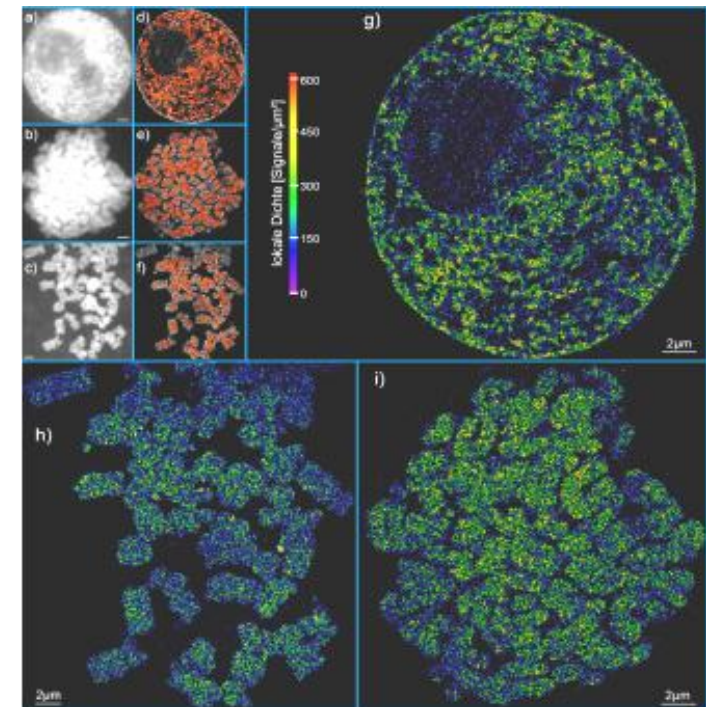


Figure 1: Histone H2B distribution in HeLa cell nuclei and (pro-)metaphase chromosomes

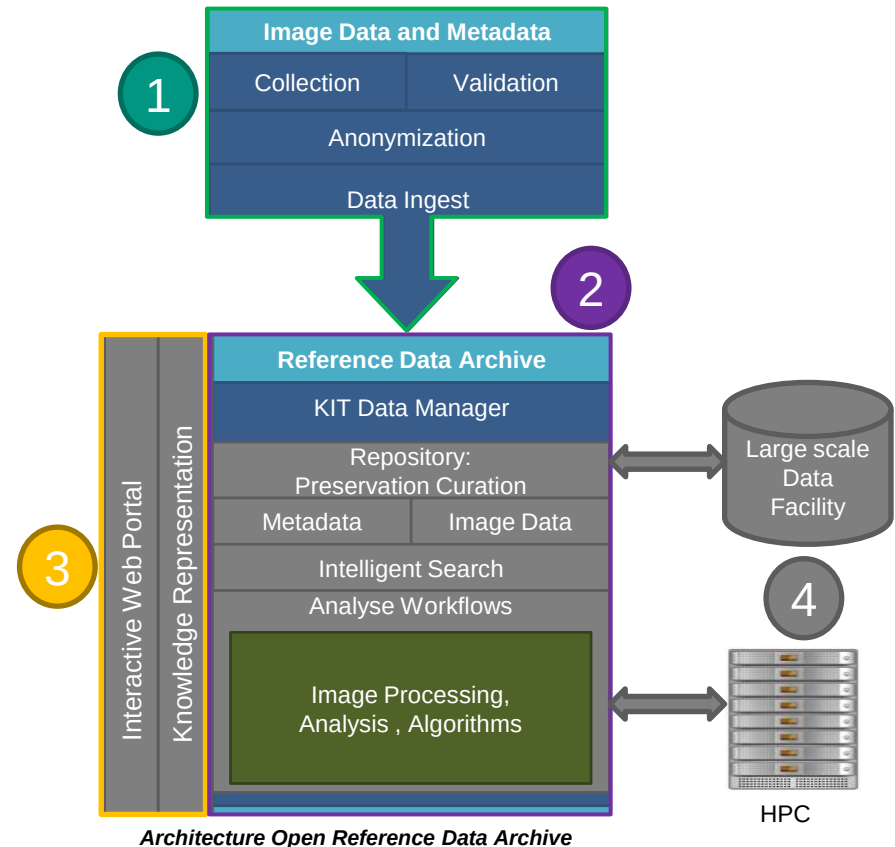
Requirements

- Easy to use Command Line Interface (CLI)
- Access provisioning between Heidelberg (Uni) and Karlsruhe (KIT)
- Features
 - Basic metadata management
 - Dataset registration
 - Data ingest
 - Data download
 - Data representation

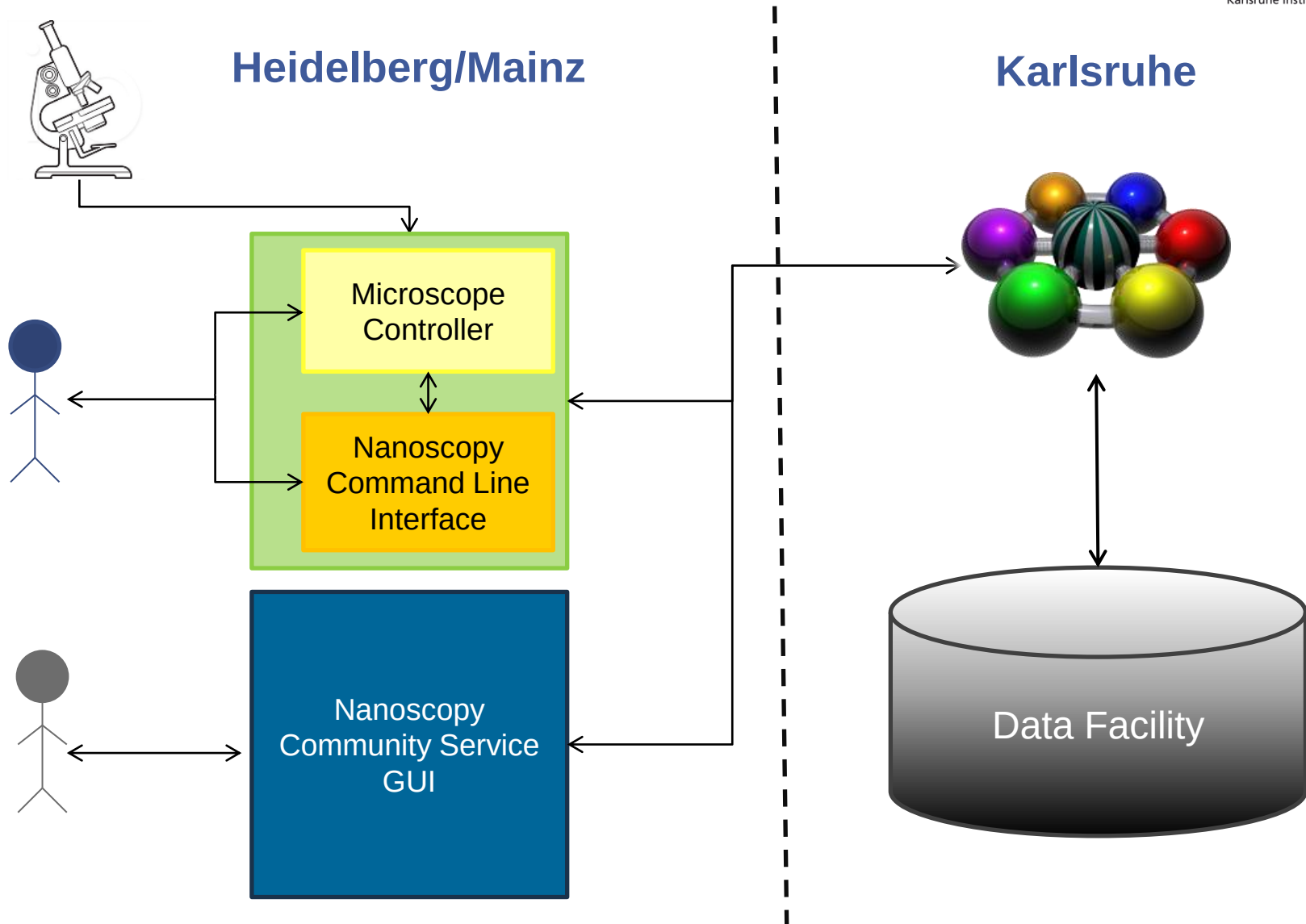
Architecture

The architecture as shown can be divided into four major components

1. Data Collection & **Ingest**
2. Reference Data Archive
3. Knowledge **Representation**
4. Large Scale Data Facility & High Performance Computing (HPC)



Project Setup





Current Status

- Easy to use Command Line Interface
- Access provisioning between Uni. Heidelberg and Karlsruhe (KIT)
- Initial features
 - Basic metadata management
 - Data ingest
 - Data download
 - Data representation
- Available clients:
 - Nanoscopy
 - eCodicology
 - Energy – BESS
 - Archaeology
 - (Biology)
- Implementation time – 1 week



Extending the Functions

- Content metadata definition
- Automated metadata extraction (HDF5, log & text files)
- Data discovery techniques (for e.g. Elastic search)
- Analysis workflow integration