Effective Management of Raw Sequencing Data and Corresponding Metadata

GERMAN CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

Omics IT and Data Management Core Facility (ODCF)
Mathematikon — Berliner Str. 41, 69120 Heidelberg, Germany

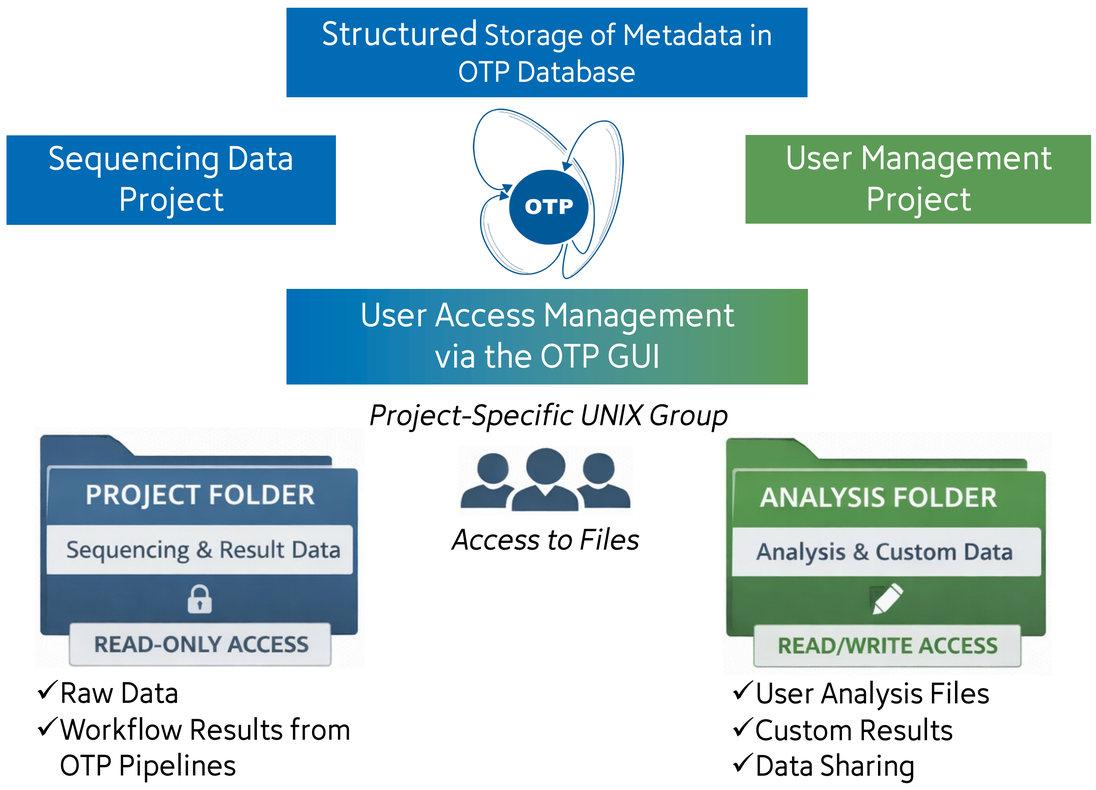
OMICS IT AND DATA MANAGEMENT CORE FACILITY (ODCF) IN A NUTSHELL

Recent advances have led to an explosion in the use of high-throughput, large-scale experiments in life sciences, including increasing use of omics technologies (genomics, proteomics, epigenomics). These generate huge quantities of data that needs to be stored, shared and analyzed effectively. Therefore, ODCF provides high-throughput computing infrastructure, data management for sequencing and other research data and essential bioinformatic services to DKFZ groups and their collaborating partners.

DATA MANAGEMENT STRATEGY

ODCF manages research data through One-Touch Pipeline (OTP) projects, which serve as the core units for sequencing data lifecycle management.

- **Project Types:**
  - **Sequencing project:** Standardized filesystem hierarchy with direct NGS Core Facility (NGSCF) data transfer, dual access via OTP graphical user interface (GUI) and DKFZ cluster, sequencing-specific workflows, and full metadata tracking in the OTP database.
  - **User management project:** Data storage for collaboration, analysis results, and non-sequencing research data.
- **Project Lifecycle:** From high-performance project space to analysis storage and long-term archival, ensuring efficient resource usage and structured data management.
- **Project Access:** Managed in the OTP GUI on a per-project basis via a dedicated UNIX group. Group members with file access can access two folders:
  - **Project Folder:** Structured directory containing raw sequencing data and workflow results; read-only access.
  - **Analysis Folder:** Space for additional project data and user-generated results; read/write access for project members.

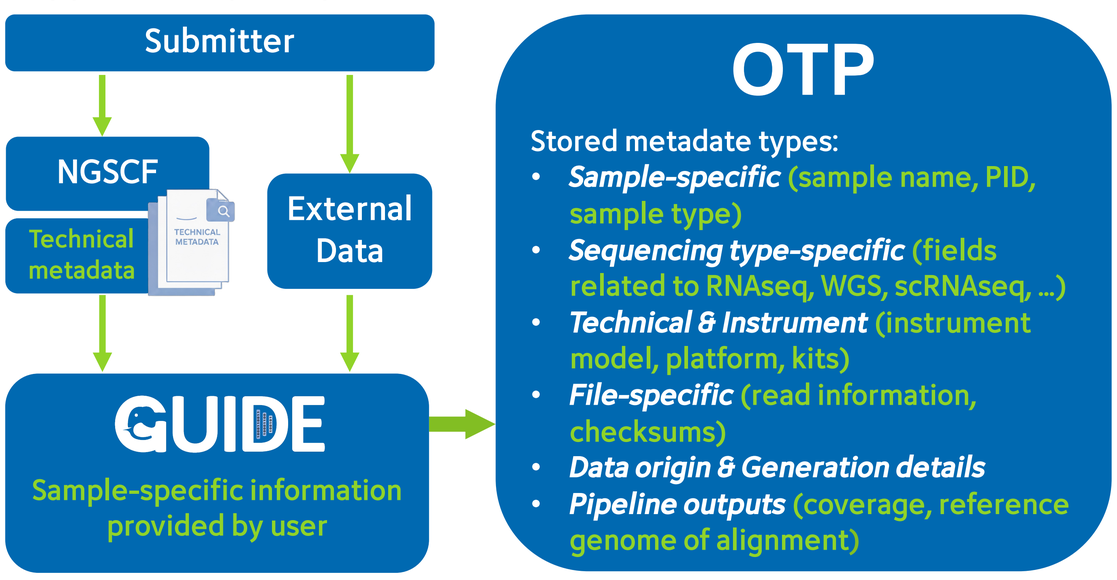


Overview about project types, the user management and the managment of raw data, the related metadata and possible analyses.

View-by-PID Structure:
/omics/odcf/project/{OE####}/{ProjectName}/sequencing/{SequencingType}/view-by-pid/PID}/{SampleType}/{LibraryLayout}/run{RunID}/sequence/{FileName}

METADATA COLLECTION AND IMPORT

For robust, long-term storage and reproducible analysis, the ODCF requires structured metadata for every sample submitted through the OTP. This systematic approach ensures that all research data is well-structured and accompanied by the rich context necessary to support FAIR principles.

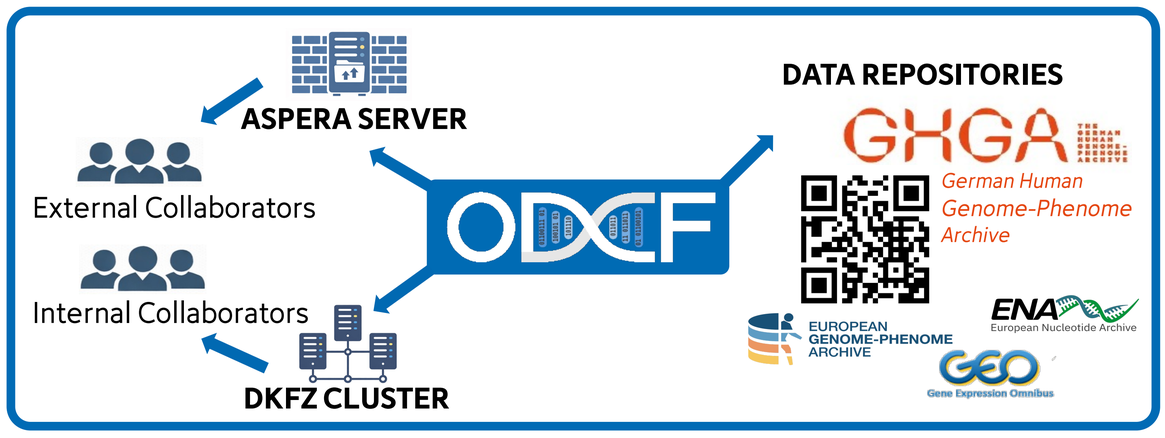


The ODCF GUIDE serves as the central platform for the streamlined submission of metadata, ensuring consistency with existing project and division conventions.

- **Incoming Data:** It can originate from two sources: the DKFZ NGSCF, where sequencing data and most technical run metadata are captured automatically and users provide only sample- and experiment-specific fields; or external sequencing providers, where data is transferred into ODCF and users submit the complete metadata set in ODCF GUIDE, including technical parameters and file-level details.
- **Automated Extraction:** An optional Sample Name Parser can be configured to automatically extract Patient IDs (PID) and sample types from structured sample names, significantly reducing manual entry errors.
- **OTP Registration:** Data is officially registered in the OTP project once two conditions are met: the sequencing data is available (via NGSCF or external transfer) and the ODCF GUIDE metadata form is finalized by the user.

TRANSFERRING DATA

ODCF provides multiple data transfer capabilities: for internal collaborators via the cluster infrastructure, for external collaborators through the ASPERA server, and support for uploads to external data repositories, particularly platforms such as GHGA.



www.odcf.dkfz.de

www.otp.dkfz.de

www.odcf-guide.dkfz.de

www.denbi.de