RESEARCH DATA ALLIANCE

Repository Platforms for Research Data IG

KIT Data Manager

Thomas Jejkal

research data sharing without barriers

rd-alliance.org

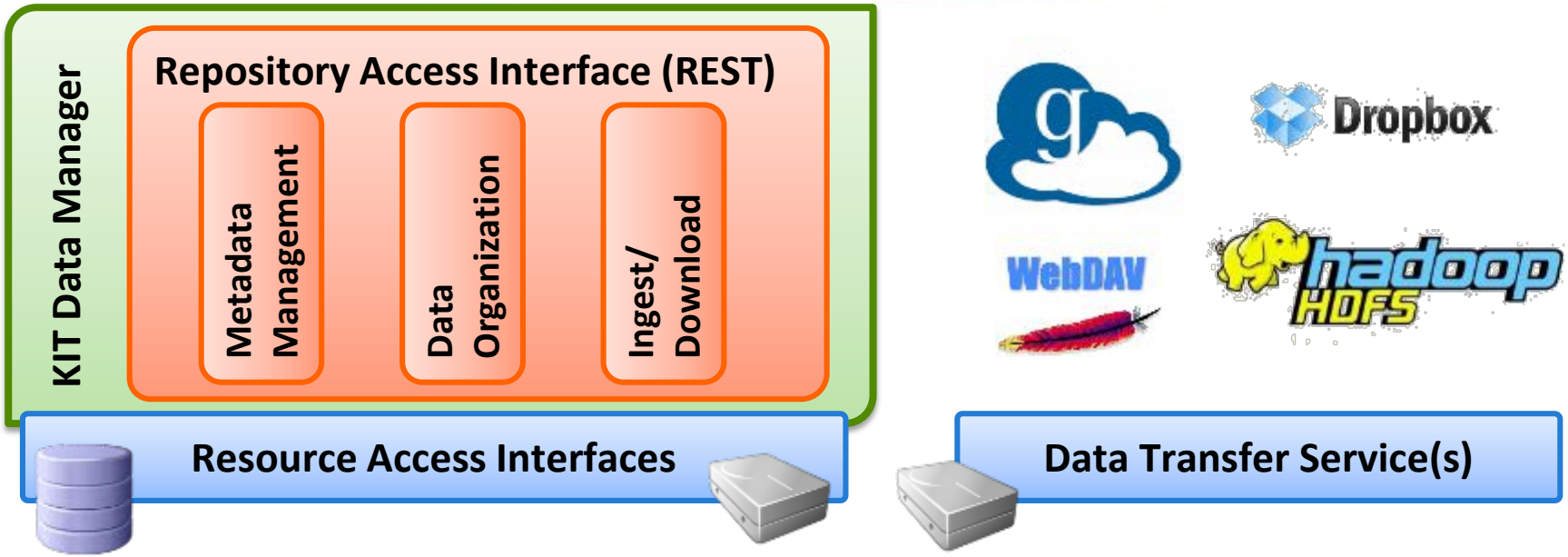
Introduction

- Challenge: project diversity
  - High throughput microscopy (16 TB per day w/o processing)
  - ANKA synchrotron facility at KIT (MB – GB per daq, authorization)
  - Data platform for nanoscience (distributed repository platform, www.nffa.eu)
  - Broad range of digital humanities projects (e.g. DARIAH, eCodicology, Archaeology)
- Extensible, out-of-the-box base platform for current and future projects
- Started development in 2011

Goals

- Focus on research data
  - High demands in terms of variety, velocity, volume (+ security)
  - High flexibility with regard to ingest workflows and supported content
  - High level of extensibility, e.g. for data organization, access and storage, and metadata extraction
- Step in directly after data acquisition
  - Registration of basic metadata as early as possible
  - Ingest of raw (unprocessed) data
  - Process data near/by repository and ingest processing results
- Cover entire research data lifecycle

Ingest (and Download)



- Ingest divided into pre-ingest and repository ingest
- Repository can provide multiple access points
- High-performance data transfer and flexibility
- Asynchronous workflows (validation, metadata extraction, format verification)

Operations

- Fine-grained security module
  - Access restrictions to content and functionalities
  - Permission hierarchy on user/group level
  - Custom workflows depending on group issuing ingest/download
- Integration via RESTful service interfaces providing access to…
  - …basic (administrative) metadata
  - …structural metadata (data organization)
  - …ingest/download functionalities or transparent data access
  - …deployed data processing tasks
- Basic administrative UI and generic end-user UI demonstrator

Sustainability

- Setup and integration of repository instance during project lifetime
  - Teaching of users and developers
  - Hand-over to community as self-operated service
- Core components open source
  - Basic out-of-the-box repository platform
  - Generic functionalities are taken over into code base
- Keeping track of landscape changes
  - Continuous integration of established standards
  - Adoption of RDA results (RDA Collections WG, Data Fabric IG, Research Data Repository Interoperability WG)

Summary

- Generic repository platform focussed (but not restricted) to research data
- Generic services as entry point for setting up research data repositories from scratch
- Community-specific customizations still necessary
  - Custom ingest workflows including metadata extraction
  - Custom tools and services, e.g. ingest tools, processing, search
  - Tailored user interfaces, e.g. for visualization
- Currently working on consolidation and modularization of services