Leveraging PIDs for object management in data infrastructures

RDA UK Node Workshop, July 16 2019

Tobias Weigel (DKRZ)

doi:10.5281/zenodo.3361619

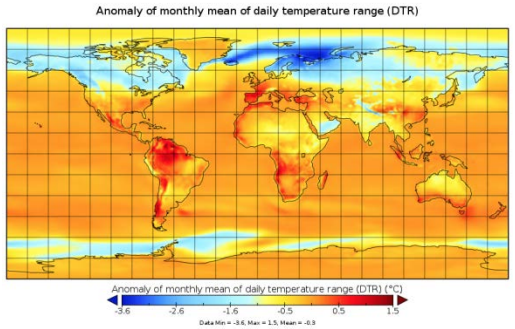
Motivational use case

CMIP6: Coupled Model Intercomparison Project, phase 6

Data infrastructure support: Earth System Grid Federation (ESGF)

Automated assignment of Handle PIDs as part of e-infrastructure workflow

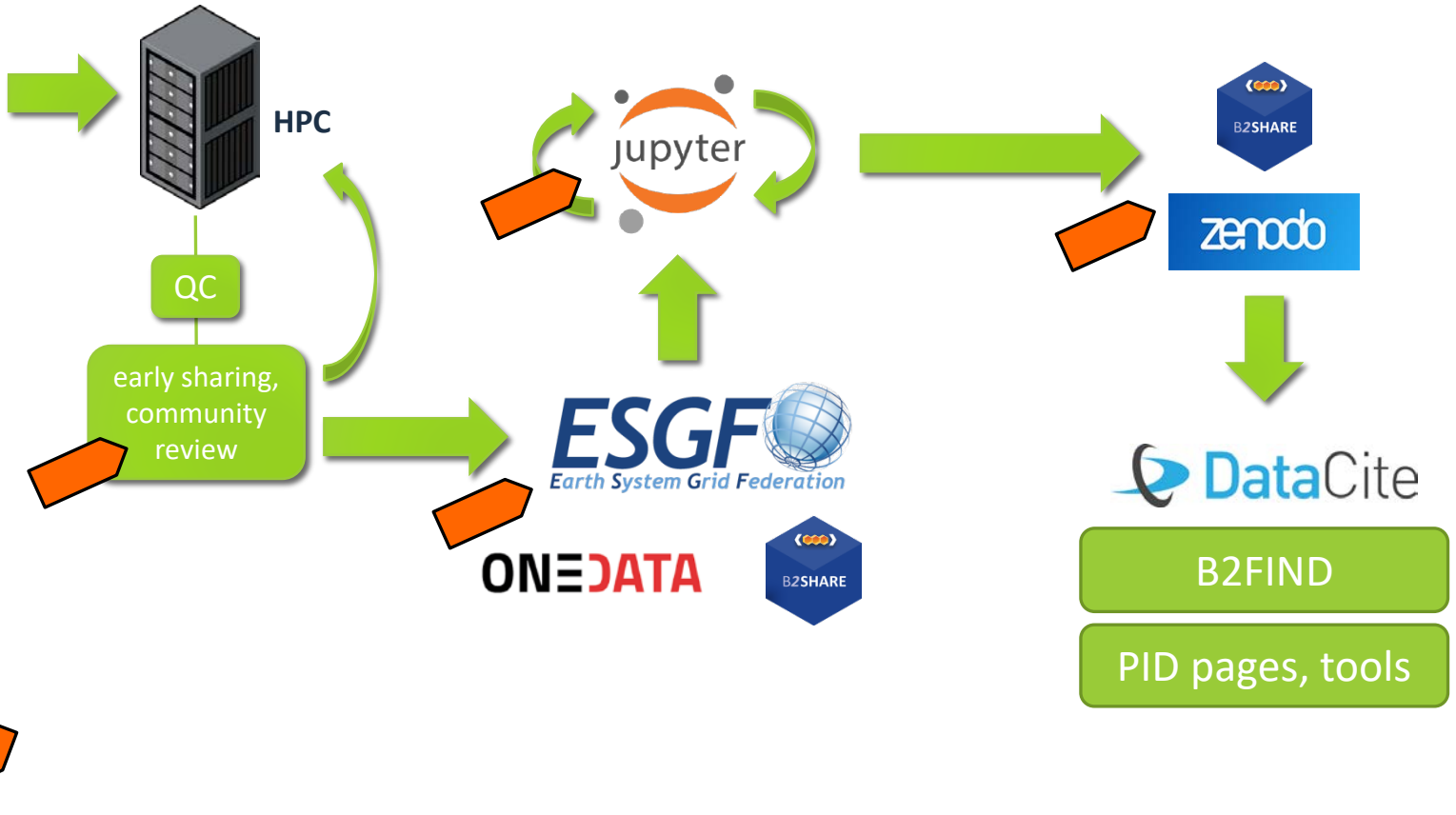
Specific feature requested by users to improve
- tracking of data versions and replicas
- referenceability before formal publication



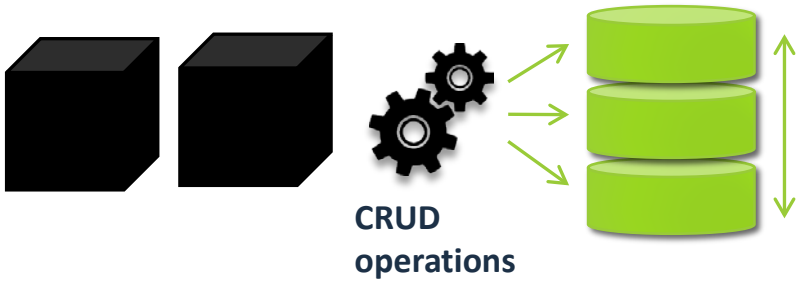
ESGF nodes

Balaji, V., Taylor, K., Juckes, M. et al. (2018): Requirements for a global data infrastructure in support of CMIP6. Geosci. Model Dev. doi:10.5194/gmd-11-3659-2018

Motivational use case: basic workflow



FAIR Digital Object Management



What shall we do if number, variety and complexity of objects increase?

- Make tools and services more efficient and effective – create the Intelligent Data Fabric

Automated management of data objects and their metadata

- early-to-mid data life cycle – pre-publication
- Record connections between (meta)data objects, software, workflows, ...

Combining multiple RDA groups and recommendations

Data Fabric IG:

- Architectural vision for DO management
- PIDs as key element: e-infrastructure layer of stable references

PID Kernel Information WG:

- Required metadata to enable machine actions
- Based on work of earlier PID Information Types WG

Data Type Registries WG / Data Typing WG:

- Register (meta)data types and provide machine-oriented API
- Enable service chaining/orchestration

Research Data Collections WG:

- CRUD operations for managing groups of objects

Data Fabric IG

- Group has existed formally since P4
- Concerned with cross-group/discipline infrastructural challenges
- Co-chairs: LI Jianhui (CNIC, CAS), Robert Quick (IU), Tobias Weigel (DKRZ)

- Aims to promote the creation of an ecosystem of reusable components and approaches built with them
- Recognizes that challenges are driven from community demand, but overarching solutions provide larger benefits
- Based on an understanding of Digital Objects and PIDs as foundational technologies

https://www.rd-alliance.org/group/data-fabric-ig.html

RDA Recommendation on PID Kernel Information

7 Guiding Principles for PID Kernel Information

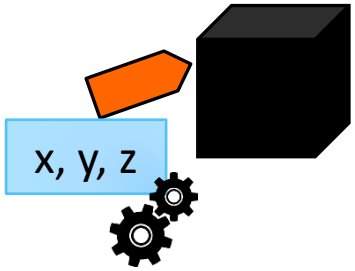
- Independent of specific infrastructure or technologies
- Geared towards minimizing human interaction, long-term stability of processes relying on Kernel Information

Draft Kernel Information profile

- 15 elements, 6 aligned with W3C PROV

Exemplary high-level architecture

Use cases and community adoption



https://doi.org/10.15497/rda00031

Unified management of Research Data Collections
The RDA Research Data Collections WG Recommendations

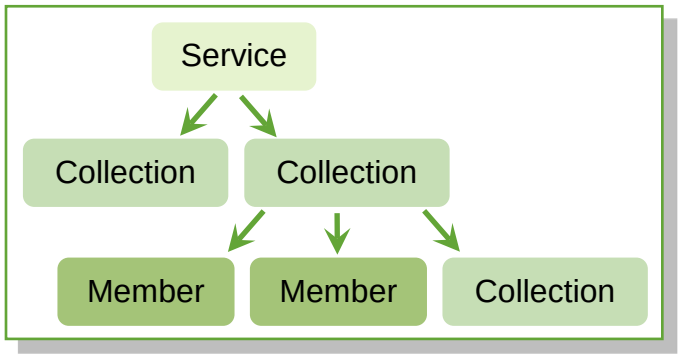
(Research) data management beyond single objects

Not just describe collections, but enable **actions** on them

- Create, Read, Update, Delete, List plus some others
- Machine agents as primary users: built-in scalability and automation
- Flexible hierarchy of **Services**, **Collections**, and **Members**
- Manage a collection of objects transparently just as another object

API specification against which tools and services can be built across community boundaries

- RESTful API specification and elemental metadata model
- Pilot implementations in several application areas exist
- Collection storage agnostic: Able to work with multiple backends
- Multiple points for custom extensions, such as ontologies, typing and specialized operations



Further information: http://dx.doi.org/10.15497/RDA00022

Upcoming: PID Kernel Information Profiles WG

Basic metadata profile defined – community profiles encouraged

Proliferation of profiles is challenge for long-term adoption

How do we manage profiles?

What are good life cycle models for profiles?

Are additional techncial interfaces necessary?

Follow via the group:
https://www.rd-alliance.org/groups/pid-kernel-information-wg

Thank you for your attention.