ESGF PID Services

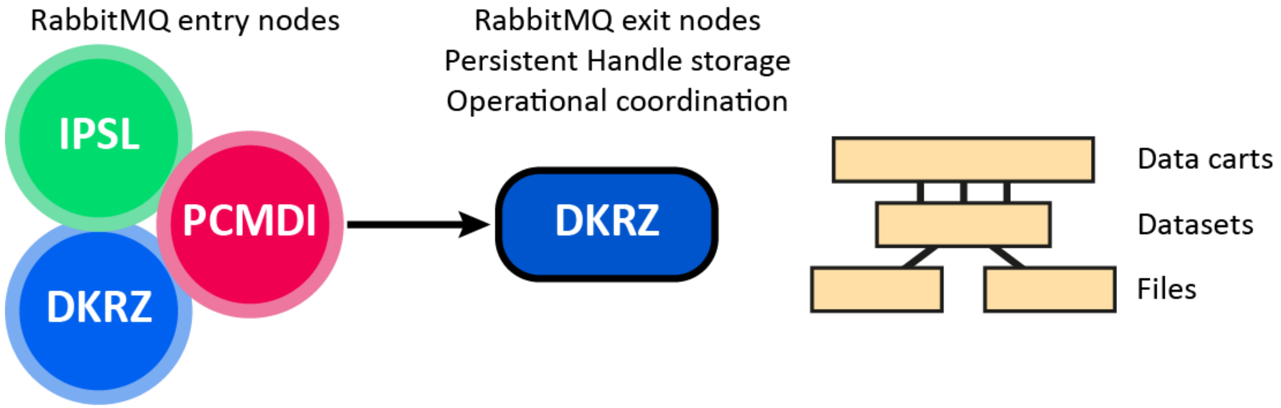
ESGF F2F Conference 2017

San Francisco

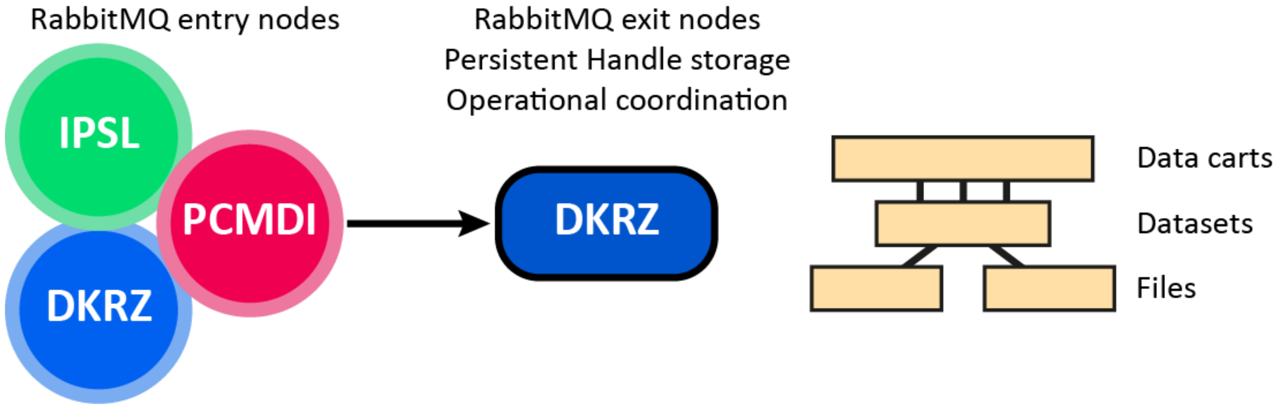
Tobias Weigel
Deutsches Klimarechenzentrum (DKRZ)

ESGF PID System: Current status

- Current status of nodes (DKRZ, IPSL, PCMDI)
- Core software components are stable
  - publisher library, consumer, viewer
  - monitoring, fallbacks, first aid documentation
- Evolving practice based on what works and what doesn‘t (e.g. credentials distribution)



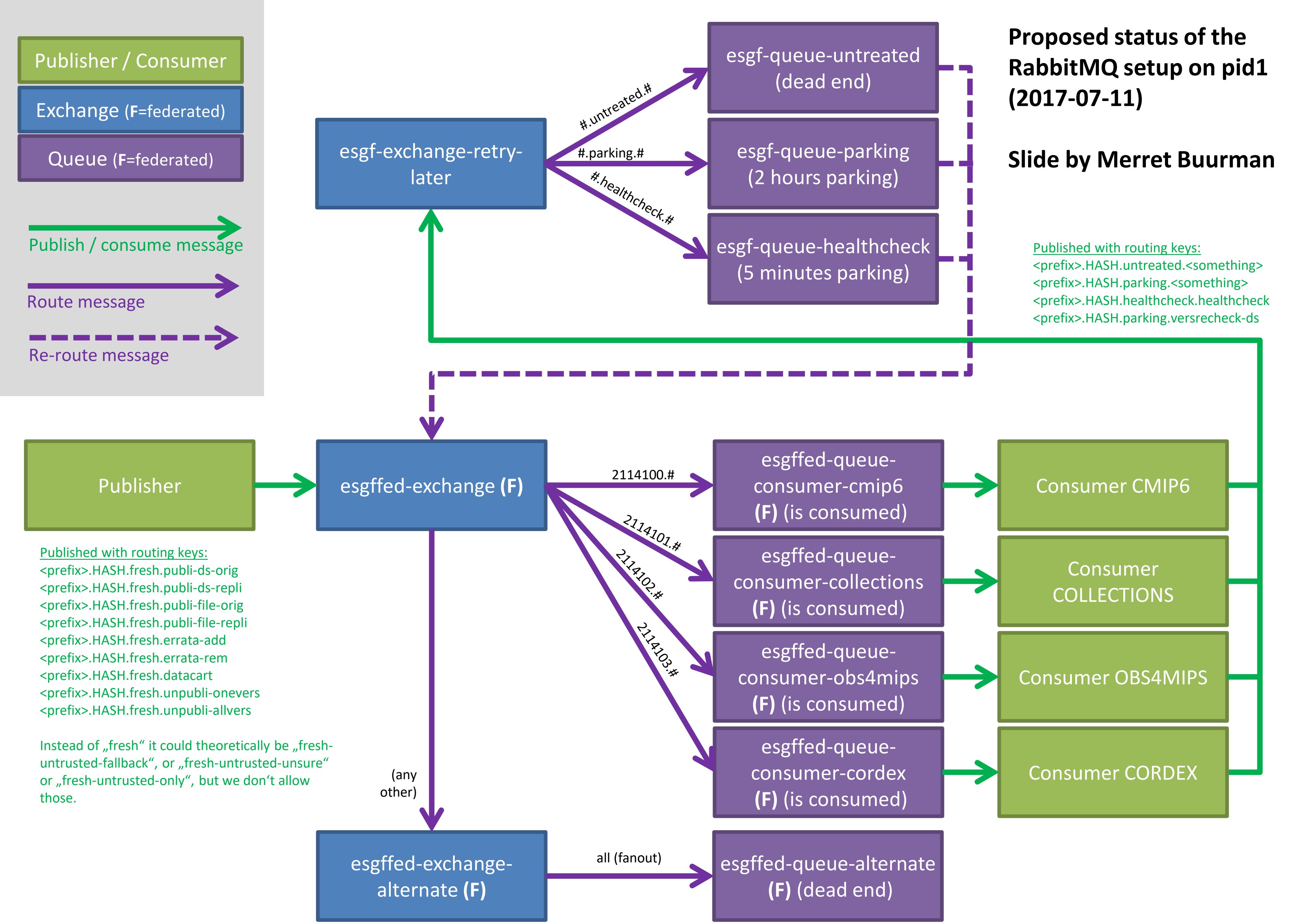
How does it work in detail?



- Publisher -> RabbitMQ -> Consumer
- Most recent developments:
  - All channels to RabbitMQ entry nodes and within the federation secured by TLS
  - Redefinition of queues to accommodate multiple prefixes/projects and make extensions easier
    - As part of this: Separation of collections from CMIP6 (dedicated prefix)
  - Information pages explaining data cart (collection) PIDs

Proposed status of the RabbitMQ setup on pid1 (2017-07-11)

Slide by Merret Buurman



Publisher / Consumer

Exchange (**F**=federated)

Queue (**F**=federated)

Publish / consume message

Route message

Re-route message

esgf-exchange-retry-later

#.untreated.#

#.parking.#

#.healthcheck.#

esgf-queue-untreated (dead end)

esgf-queue-parking (2 hours parking)

esgf-queue-healthcheck (5 minutes parking)

Published with routing keys:
<prefix>.HASH.untreated.<something>
<prefix>.HASH.parking.<something>
<prefix>.HASH.healthcheck.healthcheck
<prefix>.HASH.parking.versrecheck-ds

Publisher

esgffed-exchange **(F)**

2114100.#

2114101.#

2114102.#

2114103.#

esgffed-queue-consumer-cmip6 **(F)** (is consumed)

esgffed-queue-consumer-collections **(F)** (is consumed)

esgffed-queue-consumer-obs4mips **(F)** (is consumed)

esgffed-queue-consumer-cordex **(F)** (is consumed)

Consumer CMIP6

Consumer COLLECTIONS

Consumer OBS4MIPS

Consumer CORDEX

Published with routing keys:
<prefix>.HASH.fresh.publi-ds-orig
<prefix>.HASH.fresh.publi-ds-repli
<prefix>.HASH.fresh.publi-file-orig
<prefix>.HASH.fresh.publi-file-repli
<prefix>.HASH.fresh.errata-add
<prefix>.HASH.fresh.errata-rem
<prefix>.HASH.fresh.datacart
<prefix>.HASH.fresh.unpubli-onevers
<prefix>.HASH.fresh.unpubli-allvers

Instead of „fresh“ it could theoretically be „fresh-untrusted-fallback“, or „fresh-untrusted-unsure“ or „fresh-untrusted-only“, but we don‘t allow those.

(any other)

esgffed-exchange-alternate **(F)**

all (fanout)

esgffed-queue-alternate **(F)** (dead end)

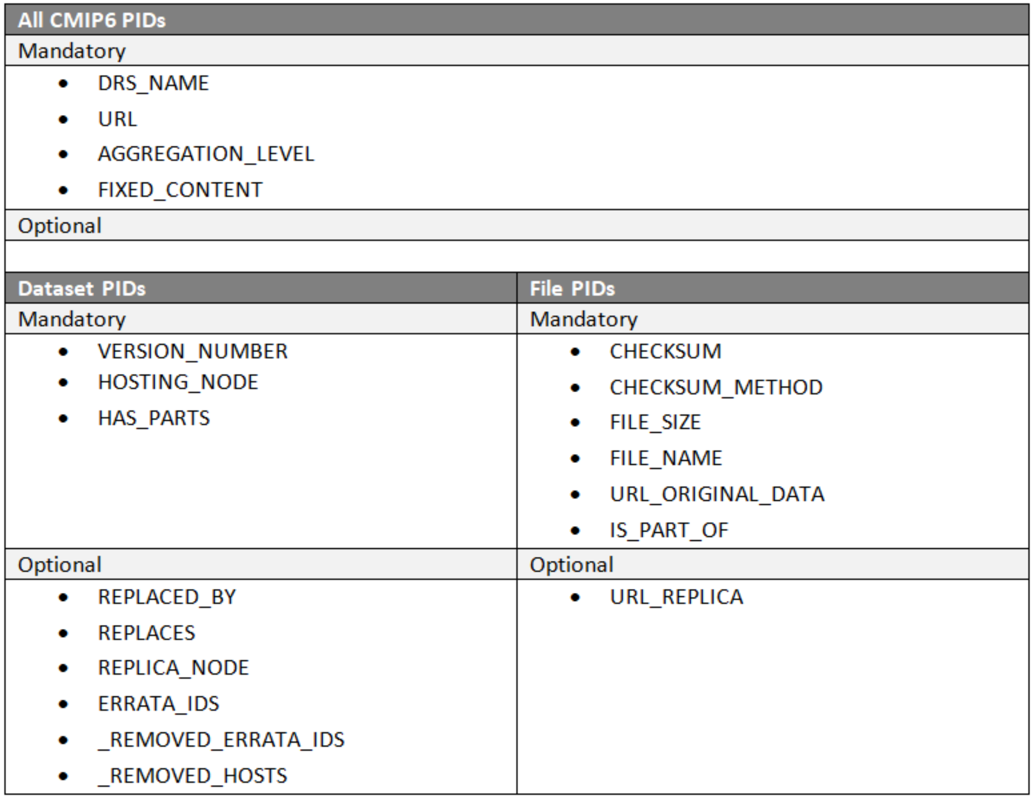
All routing keys have the mandatory pattern <prefix>.<HASH>.<rabbit-instruction>.<operation>

CMIP6 et al.: Current prefix table

Prefix	Scope
21.14100	CMIP6 files and datasets
21.14101	Custom collections
21.14102	Obs4MIPs
21.14103	CORDEX

- All registered and maintained by DKRZ
- All operational (RabbitMQ, Handle Servers, publisher lib support)
- Prepared for possible future extensions
  - preferably with same ESGF workflow

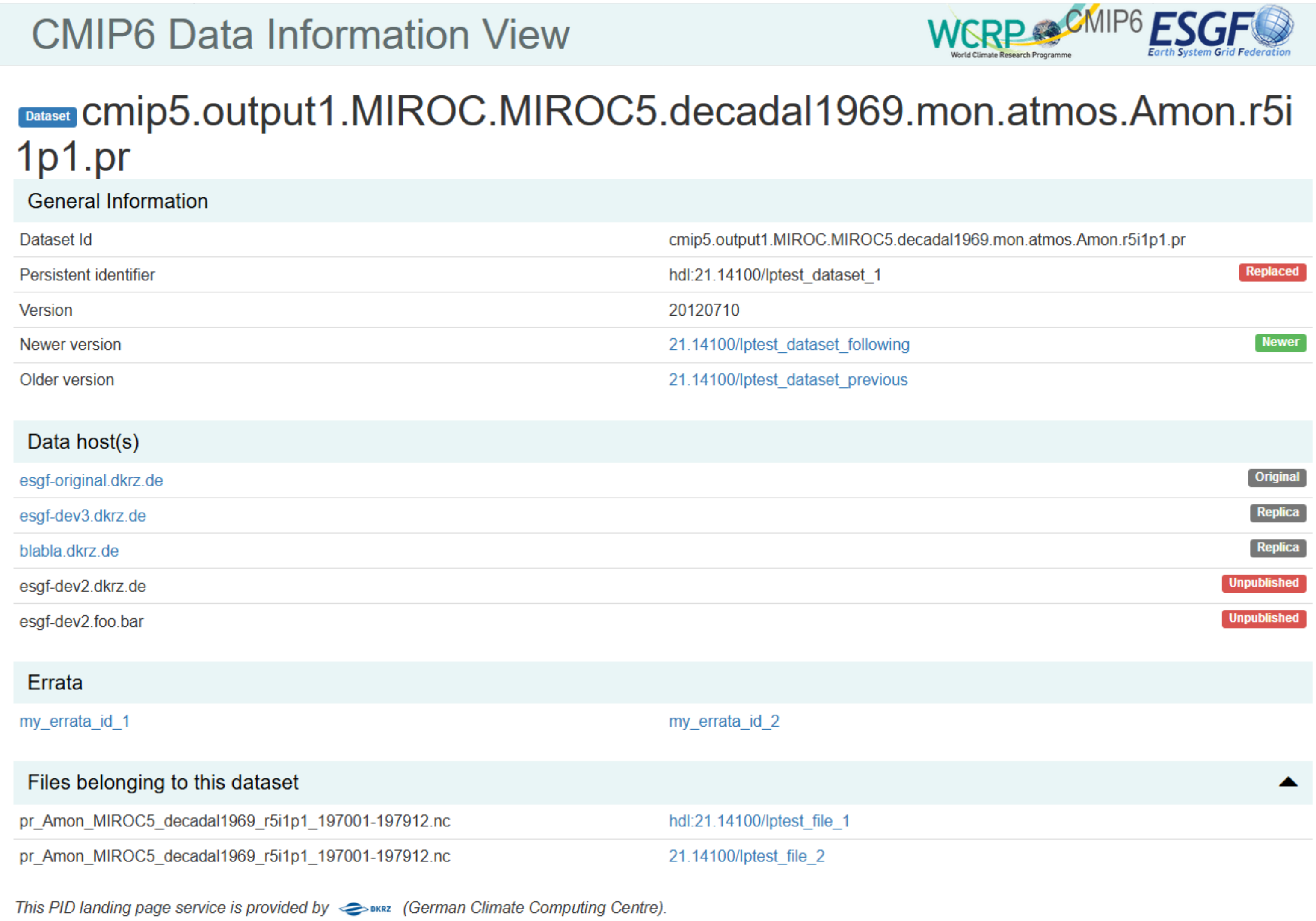
Kernel Information



All CMIP6 PIDs	
Mandatory	
• DRS_NAME • URL • AGGREGATION_LEVEL • FIXED_CONTENT	
Optional	
Dataset PIDs	File PIDs
Mandatory	Mandatory
• VERSION_NUMBER • HOSTING_NODE • HAS_PARTS	• CHECKSUM • CHECKSUM_METHOD • FILE_SIZE • FILE_NAME • URL_ORIGINAL_DATA • IS_PART_OF
Optional	Optional
• REPLACED_BY • REPLACES • REPLICA_NODE • ERRATA_IDS • _REMOVED_ERRATA_IDS • _REMOVED_HOSTS	• URL_REPLICA

- In sync with evolving RDA best practices on PID Kernel Information
- Possible changes to profile only for conformance with these
  - large-scale migration of PID records was done within EUDAT; tools and workflows exist – once recommendation emerges, we‘ll migrate

Viewer example



CMIP6 Data Information View

WCRP
World Climate Research Programme
CMIP6
ESGF
Earth System Grid Federation

Dataset cmip5.output1.MIROC.MIROC5.decadal1969.mon.atmos.Amon.r5i1p1.pr

General Information

Dataset Id	cmip5.output1.MIROC.MIROC5.decadal1969.mon.atmos.Amon.r5i1p1.pr	
Persistent identifier	hdl:21.14100/lptest_dataset_1	Replaced
Version	20120710	
Newer version	21.14100/lptest_dataset_following	Newer
Older version	21.14100/lptest_dataset_previous	

Data host(s)

esgf-original.dkrz.de	Original
esgf-dev3.dkrz.de	Replica
blabla.dkrz.de	Replica
esgf-dev2.dkrz.de	Unpublished
esgf-dev2.foo.bar	Unpublished

Errata

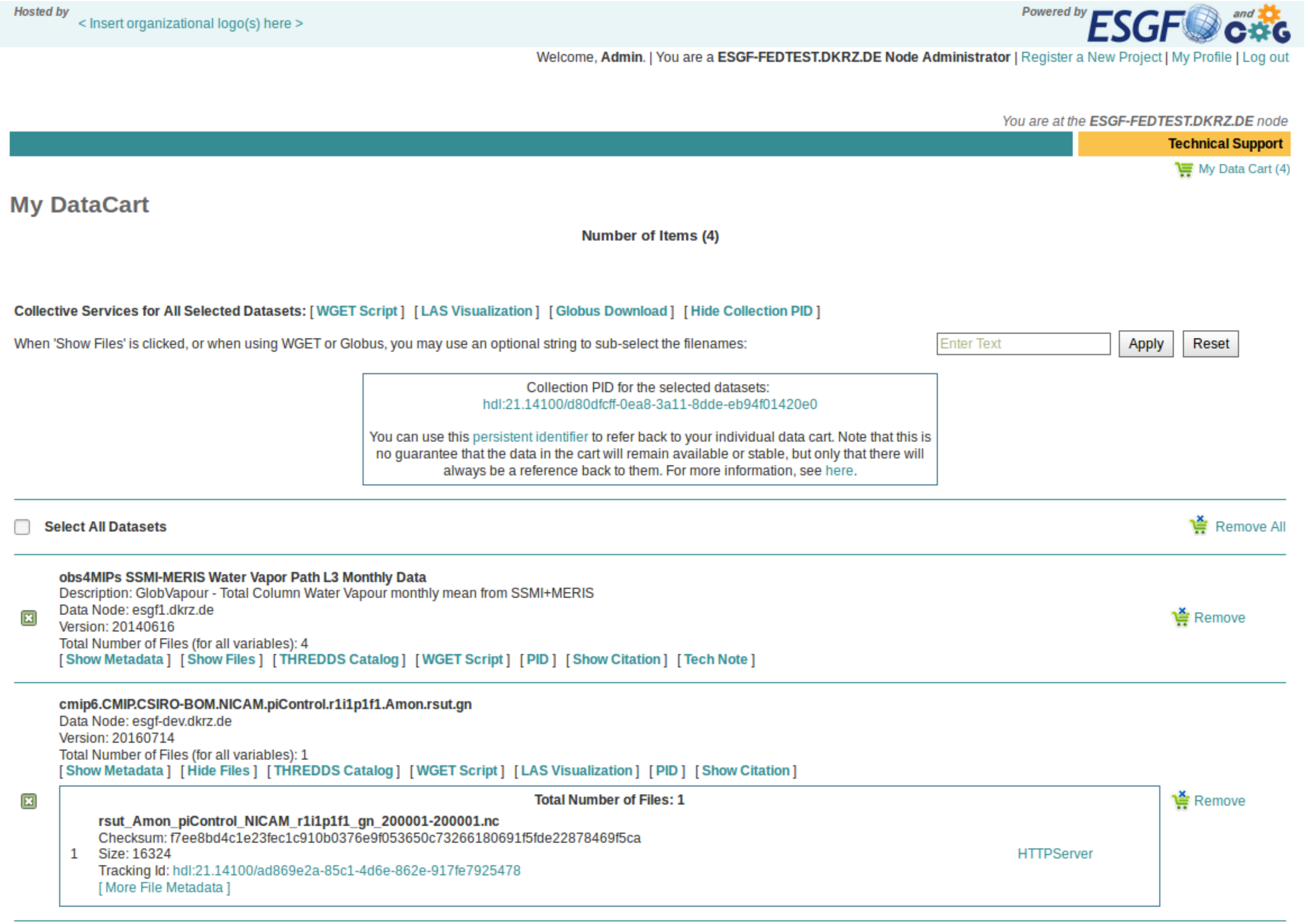
my_errata_id_1	my_errata_id_2

Files belonging to this dataset ▲

pr_Amon_MIROC5_decadal1969_r5i1p1_197001-197912.nc	hdl:21.14100/lptest_file_1
pr_Amon_MIROC5_decadal1969_r5i1p1_197001-197912.nc	21.14100/lptest_file_2

This PID landing page service is provided by DKRZ (German Climate Computing Centre).

Collection service



Hosted by < Insert organizational logo(s) here >

Powered by ESGF and CoG

Welcome, **Admin**. | You are a **ESGF-FEDTEST.DKRZ.DE Node Administrator** | Register a New Project | My Profile | Log out

*You are at the **ESGF-FEDTEST.DKRZ.DE** node*

Technical Support

My Data Cart (4)

My DataCart

Number of Items (4)

Collective Services for All Selected Datasets: [**WGET Script**] [**LAS Visualization**] [**Globus Download**] [**Hide Collection PID**]

When 'Show Files' is clicked, or when using WGET or Globus, you may use an optional string to sub-select the filenames: Enter Text Apply Reset

Collection PID for the selected datasets:
hdl:21.14100/d80dfcff-0ea8-3a11-8dde-eb94f01420e0

You can use this persistent identifier to refer back to your individual data cart. Note that this is no guarantee that the data in the cart will remain available or stable, but only that there will always be a reference back to them. For more information, see here.

☐ **Select All Datasets** Remove All

☒ **obs4MIPs SSMI-MERIS Water Vapor Path L3 Monthly Data**
Description: GlobVapour - Total Column Water Vapour monthly mean from SSMI+MERIS
Data Node: esgf1.dkrz.de
Version: 20140616
Total Number of Files (for all variables): 4
[**Show Metadata**] [**Show Files**] [**THREDDS Catalog**] [**WGET Script**] [**PID**] [**Show Citation**] [**Tech Note**]
Remove

☒ **cmip6.CMIP.CSIRO-BOM.NICAM.piControl.r1i1p1f1.Amon.rsut.gn**
Data Node: esgf-dev.dkrz.de
Version: 20160714
Total Number of Files (for all variables): 1
[**Show Metadata**] [**Hide Files**] [**THREDDS Catalog**] [**WGET Script**] [**LAS Visualization**] [**PID**] [**Show Citation**]
Remove

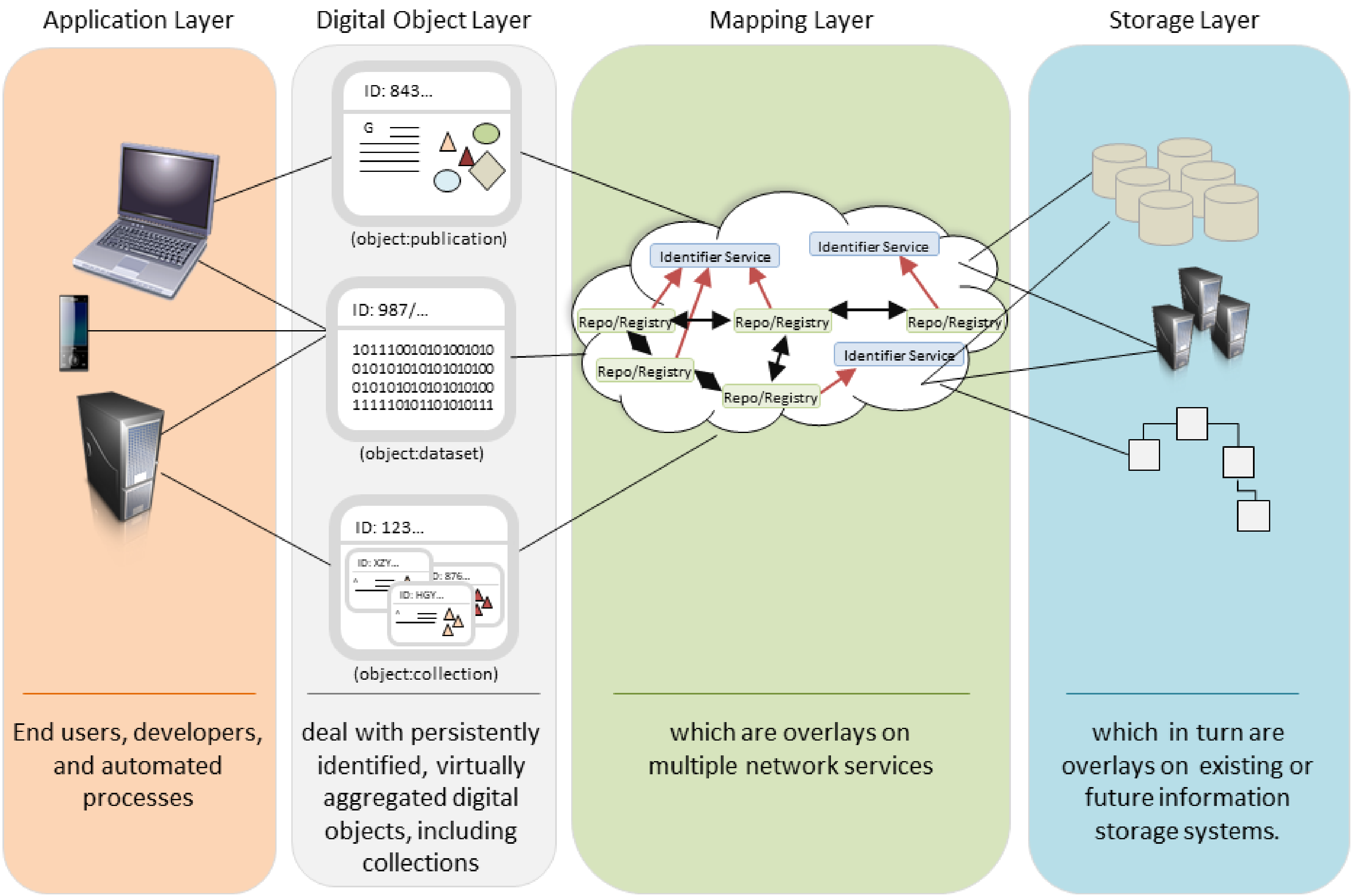
Total Number of Files: 1

1 **rsut_Amon_piControl_NICAM_r1i1p1f1_gn_200001-200001.nc**
Checksum: f7ee8bd4c1e23fec1c910b0376e9f053650c73266180691f5fde22878469f5ca
Size: 16324
Tracking Id: hdl:21.14100/ad869e2a-85c1-4d6e-862e-917fe7925478
[More File Metadata]
HTTPServer

Next steps

- Handle operations under EUDAT B2HANDLE umbrella
  - Backup/mirroring, possibly for all prefixes
- Data usage tracking tool/service – after the fact
  - multiple ways to do it, but user must always specify filters
  - would complement collection service (data cart)
- Handle mass management tools (EUDAT/EOSChub)

RDA Data Fabric embedding



Application Layer

End users, developers, and automated processes

Digital Object Layer

deal with persistently identified, virtually aggregated digital objects, including collections

Mapping Layer

which are overlays on multiple network services

Storage Layer

which in turn are overlays on existing or future information storage systems.

Long-term perspective

- PIDs as binding element to evolve from file-based management
  - following in line with RDA recommendations
    - Data Fabric, PID Kernel Information, Data Type Registries, ...
    - working towards cross-community practice (biodiversity, materials, health, astro, ...)
  - enable basic tracking (versions, prov, usage), enhanced at higher layers
- Static/dynamic collections, connection to LTA workflows (DOIs)
  - IS-ENES3?
  - How this works for authors is still unclear (citation policies): citation for credit & citation for provenance

Thank you for your attention.