

PIDs in CMIP6

Development: Merret, Tobias,
Katharina, Stephan

Merret Buurman
Deutsches Klimarechenzentrum (DKRZ)

Introduction

All CMIP6 datasets and files get PIDs.

PIDs keep track of...

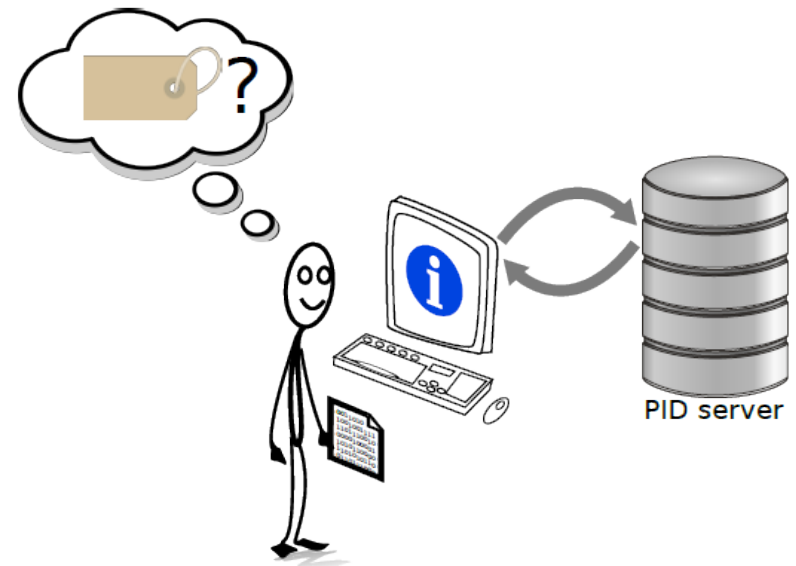
- Current copies (originals, replicas)
- Older/newer versions
- Which files contained in which datasets
- Checksums
- Errata Ids
- ...



...and will be available after an object's lifetime!

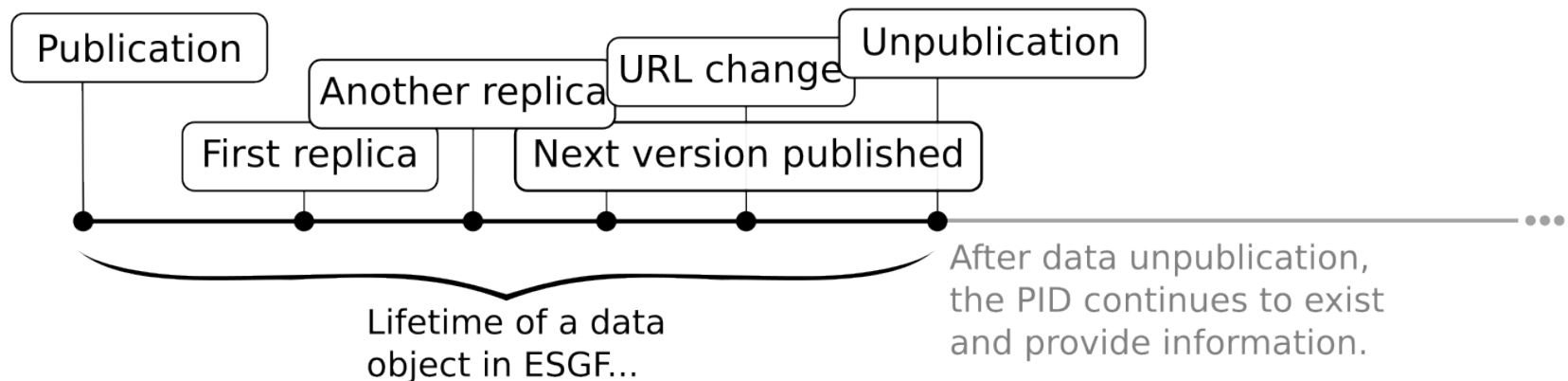
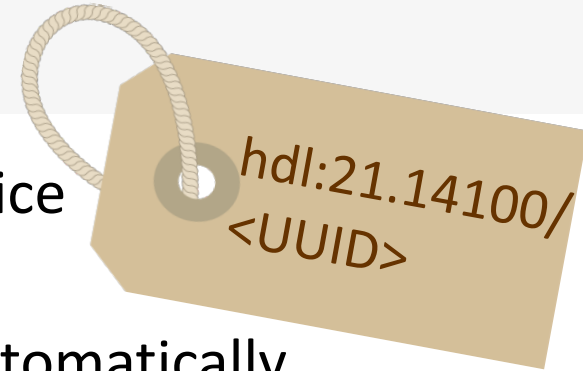
Users' perspective

- PIDs are found in CoG and in netcdf headers.
- PIDs lead to a Landing Page which allows to browse hierarchies, versions ...

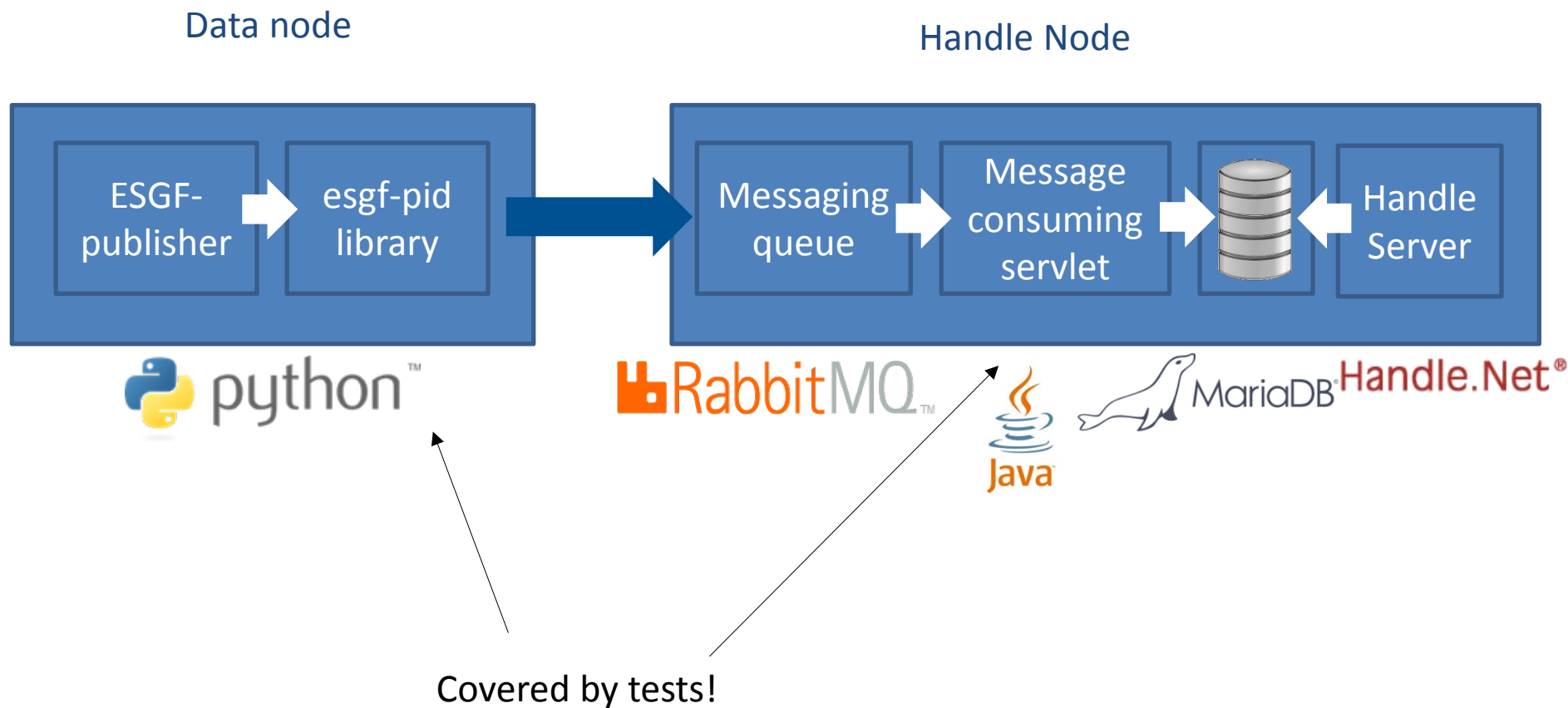


How? – Concept

- ESGF-publisher sends messages to PID service whenever it (un)publishes a dataset
- Previous/following versions are updated automatically
- File PIDs: UUIDs written by CMOR into netcdf file (former tracking ids)
- Dataset PIDs: UUIDs created by esgf-pid library

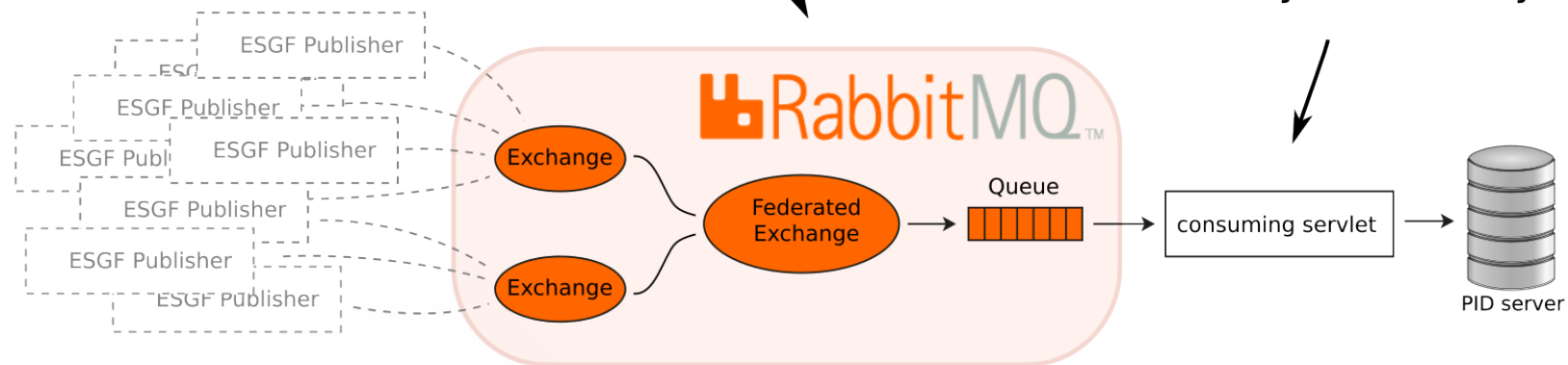


How? – Publishing process



How? – RabbitMQ federation

The PID registration and metadata update tasks are pushed to a message queueing system facilitating high availability and scalability...



Main concern: No blocking/slowing down the publication!
(e.g. down times, publication peaks...)

What does the service rely on?

Data provider:

- Use CMOR (creates file PIDs)
- Changed files → changed checksum
→ Use CMOR again → get new PID
- Change in dataset (e.g. forgotten files) → New version
(Same dataset_id → Same content)

Data publisher

- All must go through publisher (or call esgfpid library)
(incl. replicas and unpublication)
- Config: RabbitMQ URLs and credentials, etc. ...

Operative processes and CDNOT (current suggestion)

- Tier-1 nodes know credentials for RabbitMQ nodes & distribute them to publisher nodes
- Tier-1 nodes maintain list of publisher nodes (e.g. to notify them if RabbitMQ credentials change)
- Publishers must be configured for CMIP6
 - will not work if RabbitMQ configuration is missing
 - must contact Tier-1 nodes to acquire credentials for multiple RabbitMQ endpoints
 - if configured correctly, but all RabbitMQ nodes are offline, publication will still work
- CDNOT to oversee this

Other PID applications

Errata ids

- Errata addition and removal (by IPSL)
→ Talks by Guillaume and Atef

<https://acme-climate.atlassian.net/wiki/display/ESGF/QCWT+Errata+Service>

Other PID applications

Data carts

- PIDs for collections of data sets
- Anyone can stick a PID to such a collection, using the CoG
- How? CoG sends a PID request to the RabbitMQ queue using the same library

My DataCart

Number of Items (4)

Collective Services for All Selected Datasets: [[WGET Script](#)] [[LAS Visualization](#)] [[Globus Download](#)] [[Collection PID](#)]

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

☐ **Select All Datasets**

 [Remove All](#)

cmip6.CMIP.CSIRO-BOM.ACCESS-1-0.piControl.r1i1p1f1.6hrLev.ua.gn

Data Node: esgf-dev.dkrz.de

Version: 20160714

Total Number of Files (for all variables): 1

[[Show Metadata](#)] [[Show Files](#)] [[THREDDS Catalog](#)] [[WGET Script](#)] [[LAS Visualization](#)] [[PID](#)] [[Show Citation](#)]

 [Remove](#)

obs4mips.SU.ATSR2-AATSR.od550aer.mon

Data Node: eridanus.eoc.dlr.de

Version: 42

Total Number of Files (for all variables): 7

[[Show Metadata](#)] [[Show Files](#)] [[THREDDS Catalog](#)] [[WGET Script](#)] [[LAS Visualization](#)] [[Tech Note](#)]

 [Remove](#)

cmip6.CMIP.CSIRO-BOM.NICAM.piControl.r1i1p1f1.Amon.hfls.gn

Data Node: esgf-dev.dkrz.de

Version: 20160714

Total Number of Files (for all variables): 2

[[Show Metadata](#)] [[Show Files](#)] [[THREDDS Catalog](#)] [[WGET Script](#)] [[LAS Visualization](#)] [[PID](#)] [[Show Citation](#)]

 [Remove](#)

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](#)] [[Globus Download](#)]

 [Remove](#)

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:

Collection PID for the selected datasets:
hdl:21.14100/1f9d6b80-6f2e-395b-aeb4-b00157263733

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

cmip6.CMIP.CSIRO-BOM.ACCESS-1-0.piControl.r1i1p1f1.6hrLev.ua.gn

Data Node: esgf-dev.dkrz.de

☒ Version: 20160714

Total Number of Files (for all variables): 1

[[Show Metadata](#)] [[Show Files](#)] [[THREDDS Catalog](#)] [[WGET Script](#)] [[LAS Visualization](#)] [[PID](#)] [[Show Citation](#)]

 Remove

obs4mips.SU.ATSR2-AATSR.od550aer.mon

Data Node: eridanus.eoc.dlr.de

☒ Version: 42

Total Number of Files (for all variables): 7

[[Show Metadata](#)] [[Show Files](#)] [[THREDDS Catalog](#)] [[WGET Script](#)] [[LAS Visualization](#)] [[Tech Note](#)]

 Remove

cmip6.CMIP.CSIRO-BOM.NICAM.piControl.r1i1p1f1.Amon.hfls.gn

Data Node: esgf-dev.dkrz.de

☒ Version: 20160714

Total Number of Files (for all variables): 2

[[Show Metadata](#)] [[Show Files](#)] [[THREDDS Catalog](#)] [[WGET Script](#)] [[LAS Visualization](#)] [[PID](#)] [[Show Citation](#)]

 Remove

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](#)] [[Globus Download](#)]

 Remove

File Edit View History Bookmarks Tools Help

Data Information View

<https://handle-esgf.dkrz.de/lp/datacart/21.14100/53386ee1-da0f-309b-9417-8566d800209e>


ESGF
Earth System Grid Federation

Data Cart Information View

Data cart
Collection

General Information


| | |
|-----------------------|---|
| DRS name | Collection |
| Persistent identifier | 21.14100/53386ee1-da0f-309b-9417-8566d800209e |
| TimeStamp | 2016-11-09T11:49:54.711+00:00 |

Handles belonging to this Collection

| | |
|---|--|
| cmip6.CMIP.CSIRO-BOM.NICAM.piControl.r1i1p1f1.Amon.rsut.gn (version 20160714) (dataset) | hdl:21.14100/e8d9b430-4369-3cd7-a787-d2138e5691c5 |
| cmip6.CMIP.CSIRO-BOM.NICAM.piControl.r1i1p1f1.Amon.hfls.gn (version 20160714) (dataset) | hdl:21.14100/f21e70e8-beda-3561-85e7-af42440026ac Old |
| obs4MIPs.FUB-DWD.SSMI-MERIS.prw.mon (version 20140616) (dataset) | hdl:10876/ESGF/4ee9d37b-6454-44bf-b3ef-e738b2ecedb4 |

Dataset ids belonging to this Collection

obs4mips.SU.ATSR2-AATSR.od550aer.mon.v42

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

Publication demo

(Video, 1 minute, please use Chrome)

<https://esgf-fedtest.dkrz.de/api/f2fvideo/>

1. Watch the publisher assign PIDs to datasets and send messages to RabbitMQ (console output)
2. Watch RabbitMQ receive the messages (monitoring GUI)
3. Watch a user get a PID for a custom set of datasets (CoG)

Next steps...

- Command line tool to check netcdf files / content of a directory:
 - Any outdated versions?
 - Any corrupted files?

Any questions or comments?

Please contact me: buurman@dkrz.de

or have a look at the poster →

To play with the landing page, please check out the test handle:

hdl.handle.net/21.14100/f21e70e8-beda-3561-85e7-af42440026ac

To view the contents of a handle record, please check out

hdl.handle.net/21.14100/f21e70e8-beda-3561-85e7-af42440026ac?noredirect

Persistent Identifiers (PIDs) for CMIP6 data
M. Buurman, T. Weigel, K. Berger, S. Kindermann
German Climate Computing Centre (DKRZ)

What are Persistent Identifiers...

All files and datasets objects get persistent identifiers (PIDs). PIDs are unique identifiers that are globally accessible, along with some metadata (the PID record). This information can be viewed on a landing page, even after unpublication of the data from the ESGF.

Implementation

The PID creation is embedded in the ESGF data publication process, but carried out asynchronously to prevent slowing down the data publication during publication peaks.

(1) The ESGF publisher uses the `esgfpid` python library to push the PID creation tasks to a distributed message queuing system.

(2) The tasks are processed by a `message consuming service` that creates or updates the PID records.

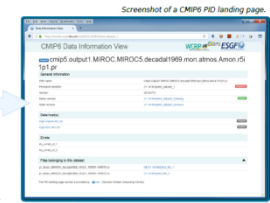
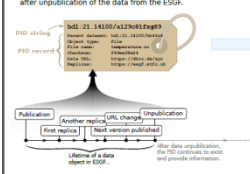
Handles.net*

The persistent identifiers used in ESGF/CMIP6 are called handles. The Handle System is an implementation of the PID concept, which is also underlying the Digital Object Identifiers (DOIs).

While technically the same, handles can be employed more flexibly than DOIs.

CMIP6 Example Handle: 21.14100/efeca96c-3220-3c2b-be7f-9f6cad083a05 ESGF...

Contact: buurman@dkrz.de
<http://www.dkrz.de>

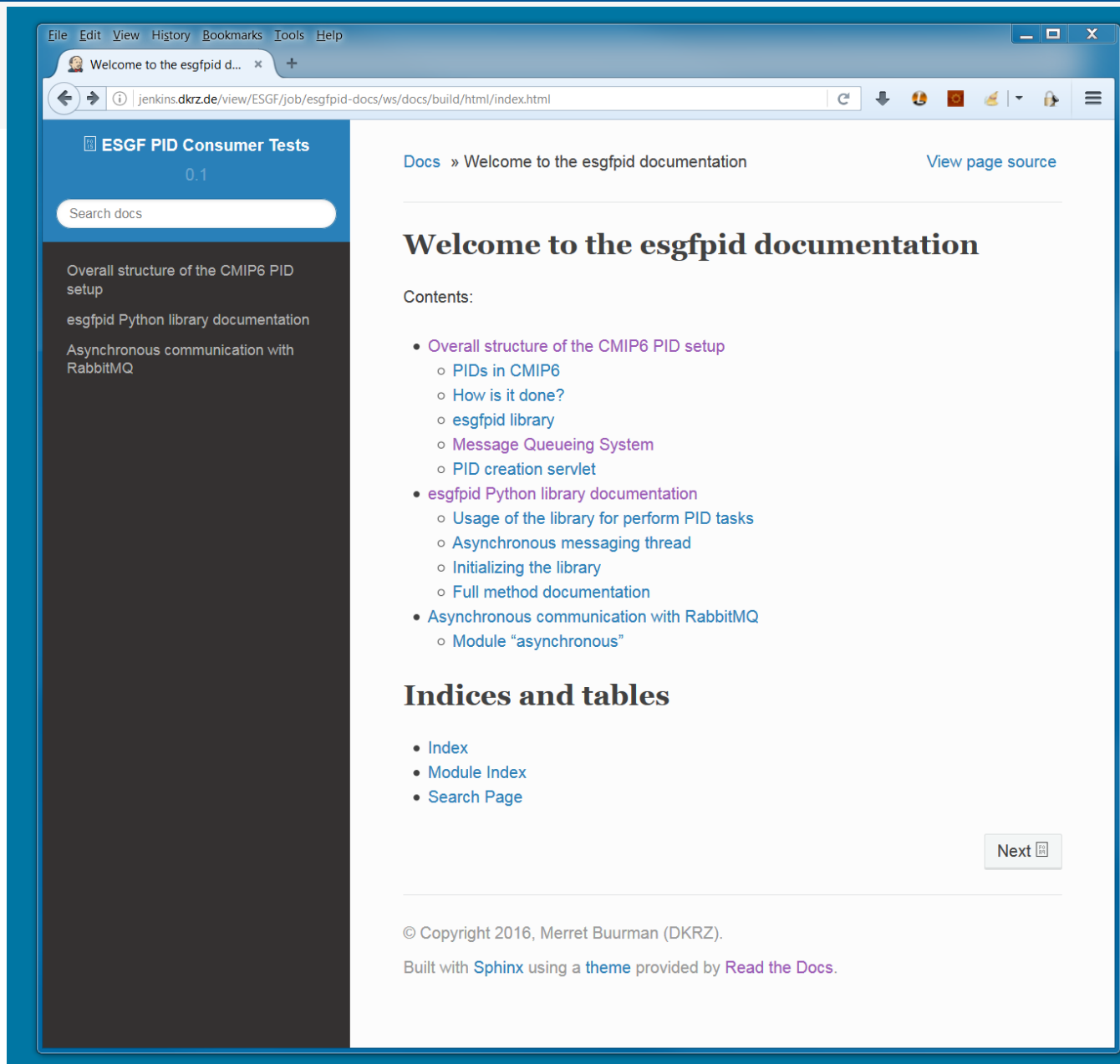
Additional slides, just in case...

esgfpid python library

- Source code on IS-ENES-GitHub:
<https://github.com/IS-ENES-Data/esgf-pid>
- Installable via pip from Pypi:
<https://pypi.python.org/pypi/esgfpid/>
- Unit-test covered (automatic CI using Jenkins):
<http://jenkins.dkrz.de/view/ESGF/job/esgf-pid/>
- Integration tests exist, but have to be run manually, because it is difficult to check the desired behaviour (as the library is designed to keep quiet).
- Documentation built by sphinx-docs:
<https://doc.redmine.dkrz.de/esgfpid/html/>

Documentation of esgfpid

- Documentation of the python library
 - Overview over whole CMIP6-PID-Setup (WIP)
 - General usage of esgfpid python library, with examples
 - Documentation of all API methods and their parameters
 - Built using sphinxdocs; automated by jenkins (WIP)
 - Have a look:
<https://doc.redmine.dkrz.de/esgfpid/html/>



File Edit View History Bookmarks Tools Help

Welcome to the esgfpid d...

jenkins.dkrz.de/view/ESGF/job/esgfpid-docs/ws/docs/build/html/index.html

ESGF PID Consumer Tests

0.1

Search docs

Overall structure of the CMIP6 PID setup

esgfpid Python library documentation

Asynchronous communication with RabbitMQ

Docs » Welcome to the esgfpid documentation [View page source](#)

Welcome to the esgfpid documentation

Contents:

- Overall structure of the CMIP6 PID setup
 - PIDs in CMIP6
 - How is it done?
 - esgfpid library
 - Message Queueing System
 - PID creation servlet
- esgfpid Python library documentation
 - Usage of the library for perform PID tasks
 - Asynchronous messaging thread
 - Initializing the library
 - Full method documentation
- Asynchronous communication with RabbitMQ
 - Module “asynchronous”

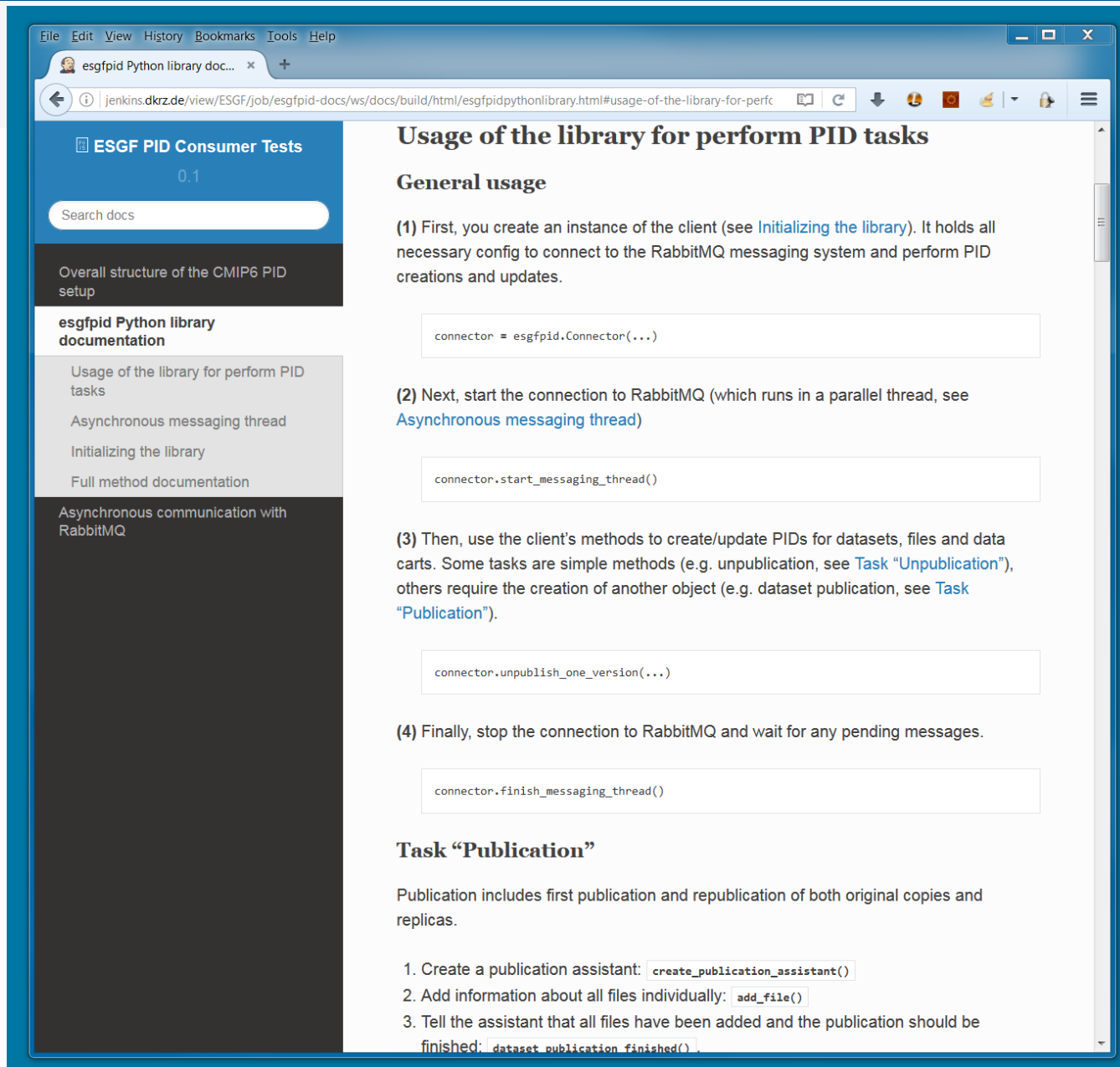
Indices and tables

- [Index](#)
- [Module Index](#)
- [Search Page](#)

[Next](#)

© Copyright 2016, Merret Buurman (DKRZ).

Built with [Sphinx](#) using a [theme](#) provided by [Read the Docs](#).



File Edit View History Bookmarks Tools Help

esgfpid Python library doc... *

jenkins.dkrz.de/view/ESGF/job/esgfpid-docs/ws/docs/build/html/esgfpidpythonlibrary.html#usage-of-the-library-for-perfc

ESGF PID Consumer Tests

0.1

Search docs

Overall structure of the CMIP6 PID setup

esgfpid Python library documentation

- Usage of the library for perform PID tasks
- Asynchronous messaging thread
- Initializing the library
- Full method documentation

Asynchronous communication with RabbitMQ

Usage of the library for perform PID tasks

General usage

(1) First, you create an instance of the client (see [Initializing the library](#)). It holds all necessary config to connect to the RabbitMQ messaging system and perform PID creations and updates.

```
connector = esgfpid.Connector(...)
```

(2) Next, start the connection to RabbitMQ (which runs in a parallel thread, see [Asynchronous messaging thread](#))

```
connector.start_messaging_thread()
```

(3) Then, use the client's methods to create/update PIDs for datasets, files and data carts. Some tasks are simple methods (e.g. unpublication, see [Task "Unpublication"](#)), others require the creation of another object (e.g. dataset publication, see [Task "Publication"](#)).

```
connector.unpublish_one_version(...)
```

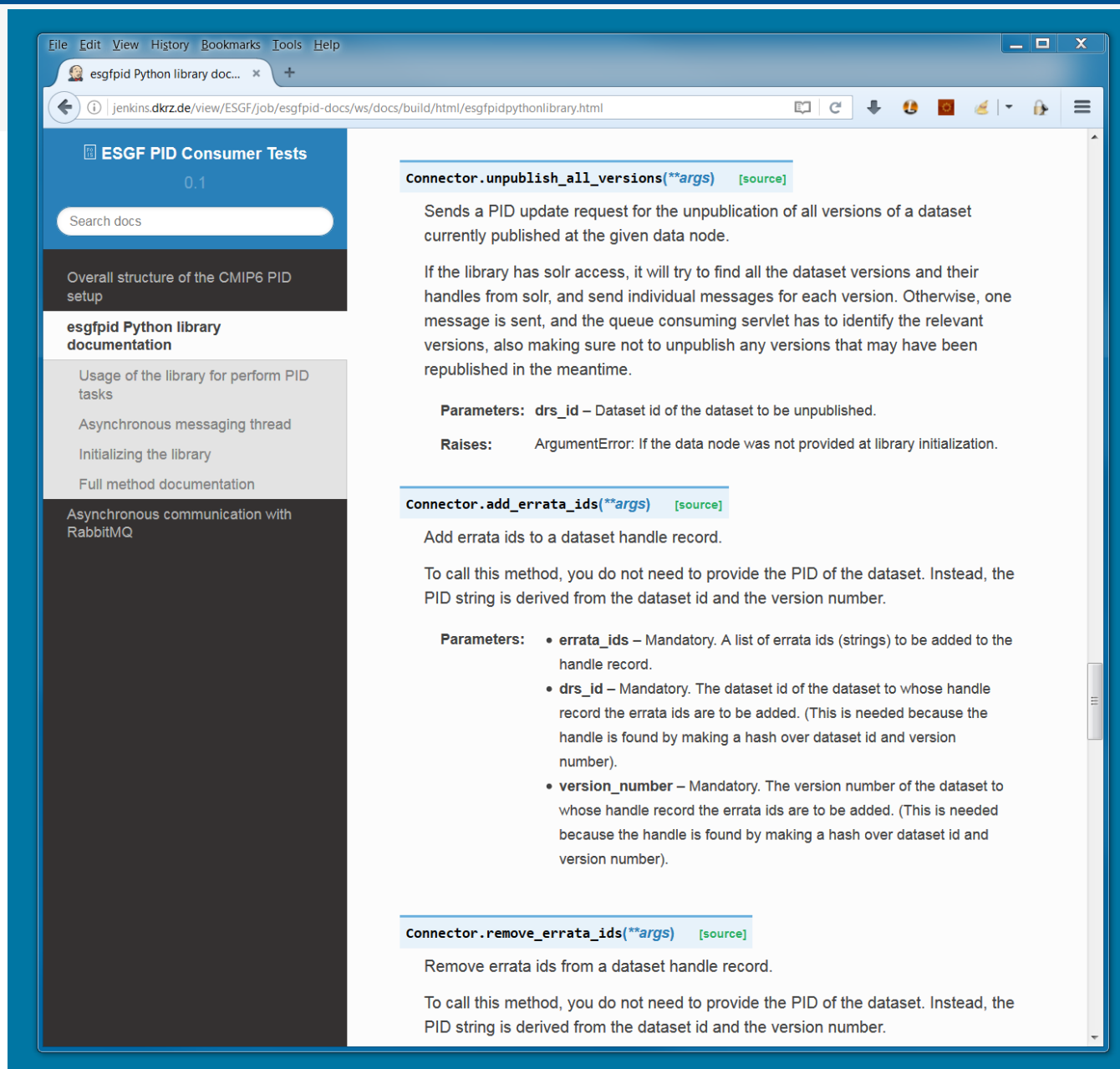
(4) Finally, stop the connection to RabbitMQ and wait for any pending messages.

```
connector.finish_messaging_thread()
```

Task "Publication"

Publication includes first publication and republication of both original copies and replicas.

1. Create a publication assistant: `create_publication_assistant()`
2. Add information about all files individually: `add_file()`
3. Tell the assistant that all files have been added and the publication should be finished: `dataset_publication_finished()`



File Edit View History Bookmarks Tools Help

esgfpid Python library doc... x +

jenkins.dkrz.de/view/ESGF/job/esgfpid-docs/ws/docs/build/html/esgfpidpythonlibrary.html

ESGF PID Consumer Tests

0.1

Search docs

Overall structure of the CMIP6 PID setup

esgfpid Python library documentation

- Usage of the library for perform PID tasks
- Asynchronous messaging thread
- Initializing the library
- Full method documentation

Asynchronous communication with RabbitMQ

Connector.unpublish_all_versions(**args) [\[source\]](#)

Sends a PID update request for the unpublication of all versions of a dataset currently published at the given data node.

If the library has solr access, it will try to find all the dataset versions and their handles from solr, and send individual messages for each version. Otherwise, one message is sent, and the queue consuming servlet has to identify the relevant versions, also making sure not to unpublish any versions that may have been republished in the meantime.

Parameters: `drs_id` – Dataset id of the dataset to be unpublished.

Raises: `ArgumentError`: If the data node was not provided at library initialization.

Connector.add_errata_ids(**args) [\[source\]](#)

Add errata ids to a dataset handle record.

To call this method, you do not need to provide the PID of the dataset. Instead, the PID string is derived from the dataset id and the version number.

Parameters:

- `errata_ids` – Mandatory. A list of errata ids (strings) to be added to the handle record.
- `drs_id` – Mandatory. The dataset id of the dataset to whose handle record the errata ids are to be added. (This is needed because the handle is found by making a hash over dataset id and version number).
- `version_number` – Mandatory. The version number of the dataset to whose handle record the errata ids are to be added. (This is needed because the handle is found by making a hash over dataset id and version number).

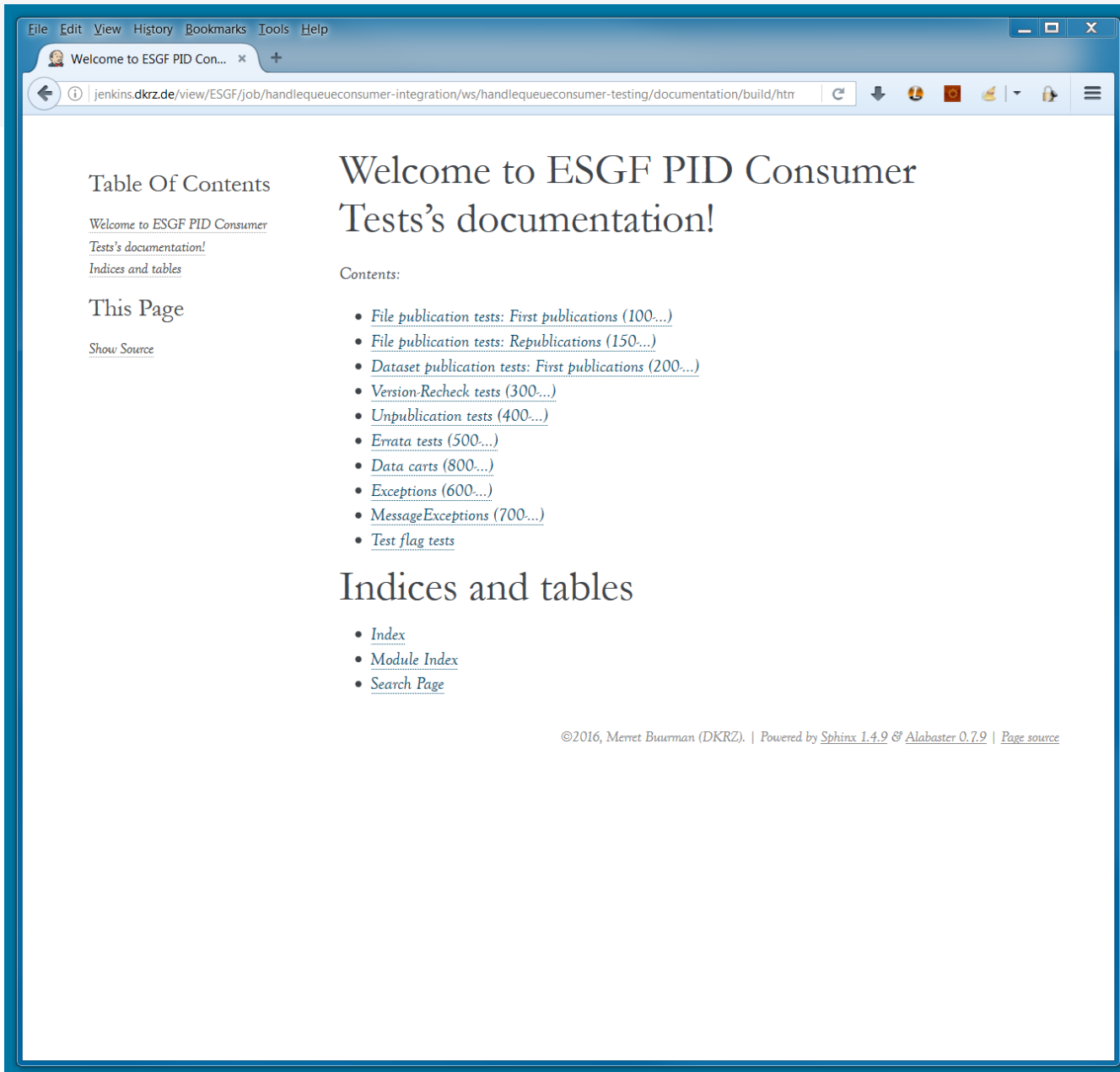
Connector.remove_errata_ids(**args) [\[source\]](#)

Remove errata ids from a dataset handle record.

To call this method, you do not need to provide the PID of the dataset. Instead, the PID string is derived from the dataset id and the version number.

Queue-consumer (Java servlet)

- Consumer has thorough integration tests:
<http://jenkins.dkrz.de/view/ESGF/job/handlequeueconsumer-integration/>
- Consumer has some JUnit tests (same jenkins job)
- Documentation of integration tests (built by sphinx-docs):
<https://doc.redmine.dkrz.de/consumer-integration/html/>
- Documentation of consumer servlet (javadoc): DKRZ-internal
(<https://doc.redmine.dkrz.de/handlequeueconsumer/apidocs/>)



File Edit View History Bookmarks Tools Help

Welcome to ESGF PID Con...

jenkins.dkrz.de/view/ESGF/job/handlequeueconsumer-integration/ws/handlequeueconsumer-testing/documentation/build/htn

Table Of Contents

[Welcome to ESGF PID Consumer Tests's documentation!](#)

[Indices and tables](#)

This Page

[Show Source](#)

Welcome to ESGF PID Consumer Tests's documentation!

Contents:

- [File publication tests: First publications \(100...\)](#)
- [File publication tests: Republications \(150...\)](#)
- [Dataset publication tests: First publications \(200...\)](#)
- [Version-Recheck tests \(300...\)](#)
- [Unpublication tests \(400...\)](#)
- [Errata tests \(500...\)](#)
- [Data carts \(800...\)](#)
- [Exceptions \(600...\)](#)
- [MessageExceptions \(700...\)](#)
- [Test flag tests](#)

Indices and tables

- [Index](#)
- [Module Index](#)
- [Search Page](#)

©2016, Merret Buerman (DKRZ). | Powered by [Sphinx 1.4.9](#) & [Alabaster 0.7.9](#) | [Page source](#)

File Edit View History Bookmarks Tools Help

Errata tests (500-...) — ESG...

jenkins.dkrz.de/view/ESGF/job/handlequeueconsumer-integration/ws/handlequeueconsumer-testing/documentation/build

testcase501 (addition)

Adding several errata ids to (virgin) record.

The record did not have any errata info before.

Errata had never been added before, so the `_ADDED_ERRATA_IDS` field has to be created.

Before: No errata at all.

```
testcase501 = {
  "HANDLE": "21.T14996/TESTCASE501",
  "AGGREGATION_LEVEL": "DATASET",
  "FIXED_CONTENT": "TRUE",
  "DRS_ID": "foo/bar/baz/drs/tc501",
  "VERSION_NUMBER": "2015",
  "URL": "https://handle-esgf.dkrz.de/lp/21.t14996/testcase501",
  "HOSTING_NODE": "<locations><location publishedOn=\\\"2000-10-10T10:00:00.000000+00:00\\\">",
  "HAS_PARTS": "hdl:21.t14996/snake;hdl:21.t14996/monkey"
}
```

Message: Adding two errata ids (errata_blue, errata_red)

```
testcase501 = {
  "handle": "21.t14996/testcase501",
  "message_timestamp": "2002-10-10T10:00:00.000000+00:00",
  "errata_ids": ["errata_blue", "errata_red"],
  "operation": "add_errata_ids",
  "drs_id": "foo/bar/baz/drs/tc501",
  "version_number": "2016",
  "ROUTING_KEY": "cmip6.publisher.HASH.errata.add"
}
```

After: Field for errata is created, the errata ids are added (errata_blue, errata_red)

```
testcase501 = {
  "HANDLE": "21.T14996/TESTCASE501",
  "AGGREGATION_LEVEL": "DATASET",
  "FIXED_CONTENT": "TRUE",
  "DRS_ID": "foo/bar/baz/drs/tc501",
  "VERSION_NUMBER": "2015",
  "URL": "https://handle-esgf.dkrz.de/lp/21.t14996/testcase501",
  "HOSTING_NODE": "<locations><location publishedOn=\\\"2000-10-10T10:00:00.000000+00:00\\\">",
  "HAS_PARTS": "hdl:21.t14996/snake;hdl:21.t14996/monkey",
  "ERRATA_IDS": "errata_blue;errata_red",
  "_ADDED_ERRATA_IDS": "errata_blue##2002-10-10T10:00:00.000+00:00;errata_red##2002-10-10T10:00:00.000+00:00"
}
```

testcase502 (addition)

Adding another two errata ids to existing

Introduction

Interconnection/Browsing of data objects

The PIDs are interlinked to keep track of the relationships between data objects:

