Versioning / Replication WG

„Kick-Off“

Versioning / Replication

- Basic Requirements / Assumptions

- Initial prototype
- Extension of publication procedure
- Publication policies and their enforcement
- Replication

- Next Steps / Roadmap

Initial Requirements

- Bring versioning related information to end-users
- No replacement of current publishing process
- Definition of stable („core“) APIs
- Enable automatic replication procedures
- Define „human ressource aware“ roadmap

→ Close relation to publishing and QC working groups !

Requirement 1: Bring versioning info to end users

Needed:

- Persistent identifier associated to file
- „core“ metadata attached to identifier
  - ref to file, newer/older version, replica, checksum, date, ..
- Stable REST API to register/change/resolve PIDs and PID metadata
- Operational/scalable resolver system for PIDs

Requirement 1: Bring versioning info to end users

Implementation options:

A) Develop own solution and integrate with ESGF publisher

B) Take existing solution and integrate with ESGF publisher

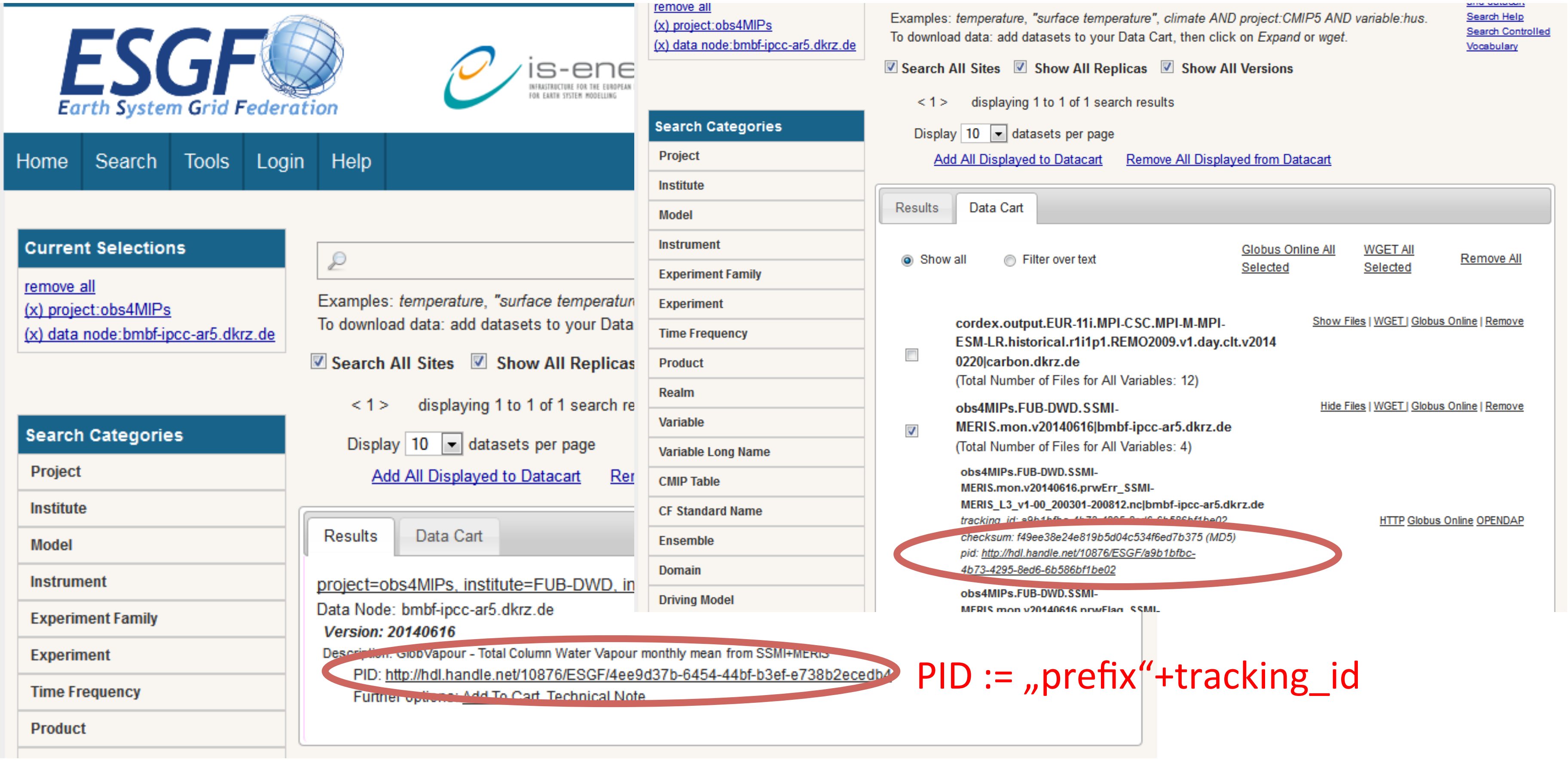
→ Initial prototyping done following B)

handle.net PID system:

- stable API
- production ready, distributed, scalable resolution system
- existing large scale deployments (e.g. DOI system)

Initial prototyping

„by hand“ PID assignment for a smaller obs4MIPs project published at DKRZ:


ESGF
Earth System Grid Federation
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING
Home
Search
Tools
Login
Help
Current Selections
Search Categories
Project
Institute
Model
Instrument
Experiment Family
Experiment
Time Frequency
Product
Realm
Variable
Variable Long Name
CMIP Table
CF Standard Name
Ensemble
Domain
Driving Model
Search Help
Search Controlled Vocabulary
Results
Data Cart
Add All Displayed to Datacart
Remove All Displayed from Datacart
Globus Online All Selected
WGET All Selected
Remove All
Show Files | WGET | Globus Online | Remove
Hide Files | WGET | Globus Online | Remove
HTTP Globus Online OPENDAP

remove all
(x) project:obs4MIPs
(x) data node:bmbf-ipcc-ar5.dkrz.de

Examples: *temperature*, *"surface temperature"*, *climate AND project:CMIP5 AND variable:hus*.
To download data: add datasets to your Data Cart, then click on *Expand* or *wget*.

☑ Search All Sites ☑ Show All Replicas ☑ Show All Versions

< 1 > displaying 1 to 1 of 1 search results

Display 10 datasets per page

project=obs4MIPs, institute=FUB-DWD, in

Data Node: bmbf-ipcc-ar5.dkrz.de

Version: 20140616

Description: GlobVapour - Total Column Water Vapour monthly mean from SSMI+MERIS

PID: http://hdl.handle.net/10876/ESGF/4ee9d37b-6454-44bf-b3ef-e738b2ecedb4

Further options: Add To Cart Technical Note

● Show all ○ Filter over text

☐ **cordex.output.EUR-11i.MPI-CSC.MPI-M-MPI-ESM-LR.historical.r1i1p1.REMO2009.v1.day.clt.v20140220|carbon.dkrz.de**
(Total Number of Files for All Variables: 12)

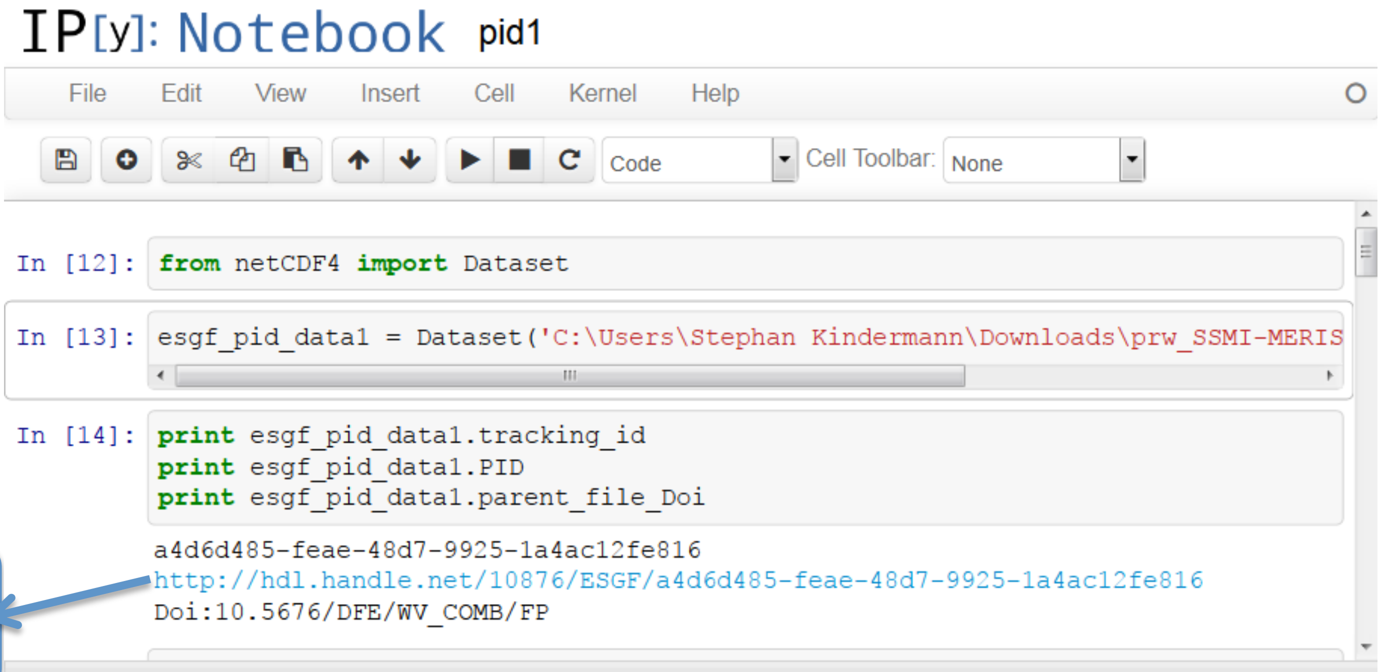
☑ **obs4MIPs.FUB-DWD.SSMI-MERIS.mon.v20140616|bmbf-ipcc-ar5.dkrz.de**
(Total Number of Files for All Variables: 4)

obs4MIPs.FUB-DWD.SSMI-MERIS.mon.v20140616.prwErr_SSMI-MERIS_L3_v1-00_200301-200812.nc|bmbf-ipcc-ar5.dkrz.de
tracking_id: a9b1bfbc-4b73-4295-8ed6-6b586bf1be02
checksum: f49ee38e24e819b5d04c534f6ed7b375 (MD5)
pid: http://hdl.handle.net/10876/ESGF/a9b1bfbc-4b73-4295-8ed6-6b586bf1be02

obs4MIPs.FUB-DWD.SSMI-MERIS.mon.v20140616.prwFlag_SSMI-

PID := „prefix“+tracking_id

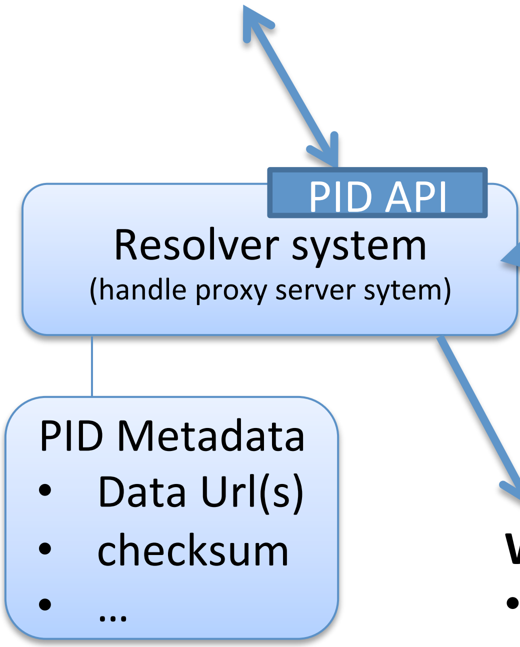
Initial protoyping



```
In [12]: from netCDF4 import Dataset

In [13]: esgf_pid_data1 = Dataset('C:\Users\Stephan Kindermann\Downloads\prw_SSMI-MERIS

In [14]: print esgf_pid_data1.tracking_id
         print esgf_pid_data1.PID
         print esgf_pid_data1.parent_file_Doi

         a4d6d485-feae-48d7-9925-1a4ac12fe816
         http://hdl.handle.net/10876/ESGF/a4d6d485-feae-48d7-9925-1a4ac12fe816
         Doi:10.5676/DFE/WV_COMB/FP
```

Tools

- Publisher
- Replication service



PID API

Resolver system

(handle proxy server sytem)

PID Metadata

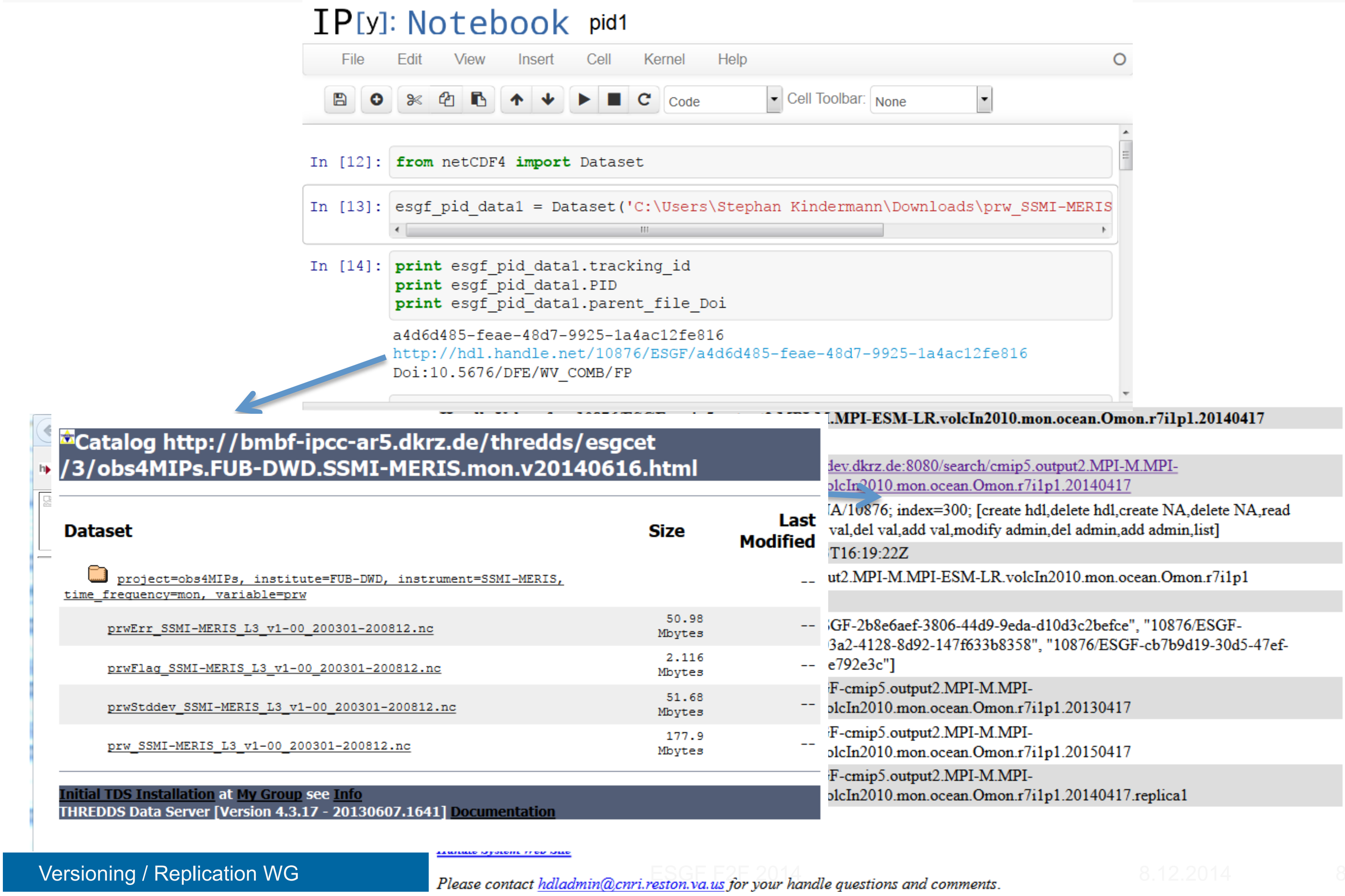
- Data Url(s)
- checksum
- …

redirect

Web Landing Pages

- Data download page
- Versioning Info
- Replica Info

Initial protoyping



```
In [12]: from netCDF4 import Dataset

In [13]: esgf_pid_data1 = Dataset('C:\Users\Stephan Kindermann\Downloads\prw_SSMI-MERIS

In [14]: print esgf_pid_data1.tracking_id
         print esgf_pid_data1.PID
         print esgf_pid_data1.parent_file_Doi

         a4d6d485-feae-48d7-9925-1a4ac12fe816
         http://hdl.handle.net/10876/ESGF/a4d6d485-feae-48d7-9925-1a4ac12fe816
         Doi:10.5676/DFE/WV_COMB/FP
```

Catalog http://bmbf-ipcc-ar5.dkrz.de/thredds/esgcet/3/obs4MIPs.FUB-DWD.SSMI-MERIS.mon.v20140616.html

Dataset	Size	Last Modified
project=obs4MIPs, institute=FUB-DWD, instrument=SSMI-MERIS, time_frequency=mon, variable=prw		--
prwErr_SSMI-MERIS_L3_v1-00_200301-200812.nc	50.98 Mbytes	--
prwFlag_SSMI-MERIS_L3_v1-00_200301-200812.nc	2.116 Mbytes	--
prwStddev_SSMI-MERIS_L3_v1-00_200301-200812.nc	51.68 Mbytes	--
prw_SSMI-MERIS_L3_v1-00_200301-200812.nc	177.9 Mbytes	--

Initial TDS Installation at My Group see Info
THREDDS Data Server [Version 4.3.17 - 20130607.1641] Documentation

.MPI-ESM-LR.volcIn2010.mon.ocean.Omon.r7i1p1.20140417

dev.dkrz.de:8080/search/cmip5.output2.MPI-M.MPI-
olcIn2010.mon.ocean.Omon.r7i1p1.20140417

IA/10876; index=300; [create hdl,delete hdl,create NA,delete NA,read val,del val,add val,modify admin,del admin,add admin,list]

T16:19:22Z

ut2.MPI-M.MPI-ESM-LR.volcIn2010.mon.ocean.Omon.r7i1p1

GF-2b8e6aef-3806-44d9-9eda-d10d3c2befce", "10876/ESGF-
3a2-4128-8d92-147f633b8358", "10876/ESGF-cb7b9d19-30d5-47ef-
e792e3c"]

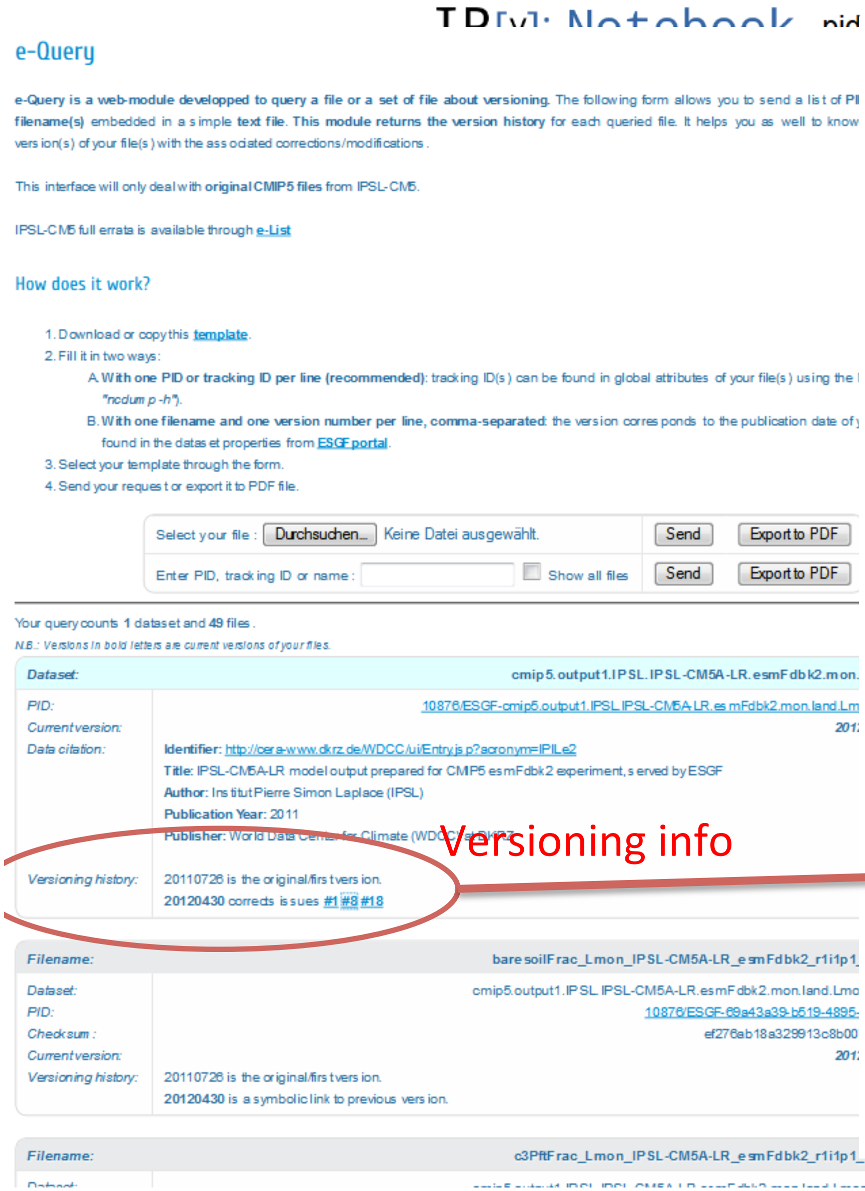
F-cmip5.output2.MPI-M.MPI-
olcIn2010.mon.ocean.Omon.r7i1p1.20130417

F-cmip5.output2.MPI-M.MPI-
olcIn2010.mon.ocean.Omon.r7i1p1.20150417

F-cmip5.output2.MPI-M.MPI-
olcIn2010.mon.ocean.Omon.r7i1p1.20140417.replica1

Please contact hdladmin@cnri.reston.va.us for your handle questions and comments.

Initial protoyping



TR[v]: Notebook pid

e-Query

e-Query is a web-module developped to query a file or a set of file about versioning. The following form allows you to send a list of PI **filename(s)** embedded in a simple **text file. This module returns the version history** for each queried file. It helps you as well to know version(s) of your file(s) with the associated corrections/modifications.

This interface will only deal with **original CMIP5 files** from IPSL-CM5.

IPSL-CM5 full errata is available through e-List

How does it work?

1. Download or copy this template.
2. Fill it in two ways:
   A. **With one PID or tracking ID per line (recommended):** tracking ID(s) can be found in global attributes of your file(s) using the *"ncdump -h"*).
   B. **With one filename and one version number per line, comma-separated:** the version corresponds to the publication date of found in the dataset properties from ESGF portal.
3. Select your template through the form.
4. Send your request or export it to PDF file.

Select your file : Durchsuchen... Keine Datei ausgewählt.
Send
Export to PDF
Enter PID, tracking ID or name :
Show all files
Send
Export to PDF

Your query counts **1** dataset and **49** files.

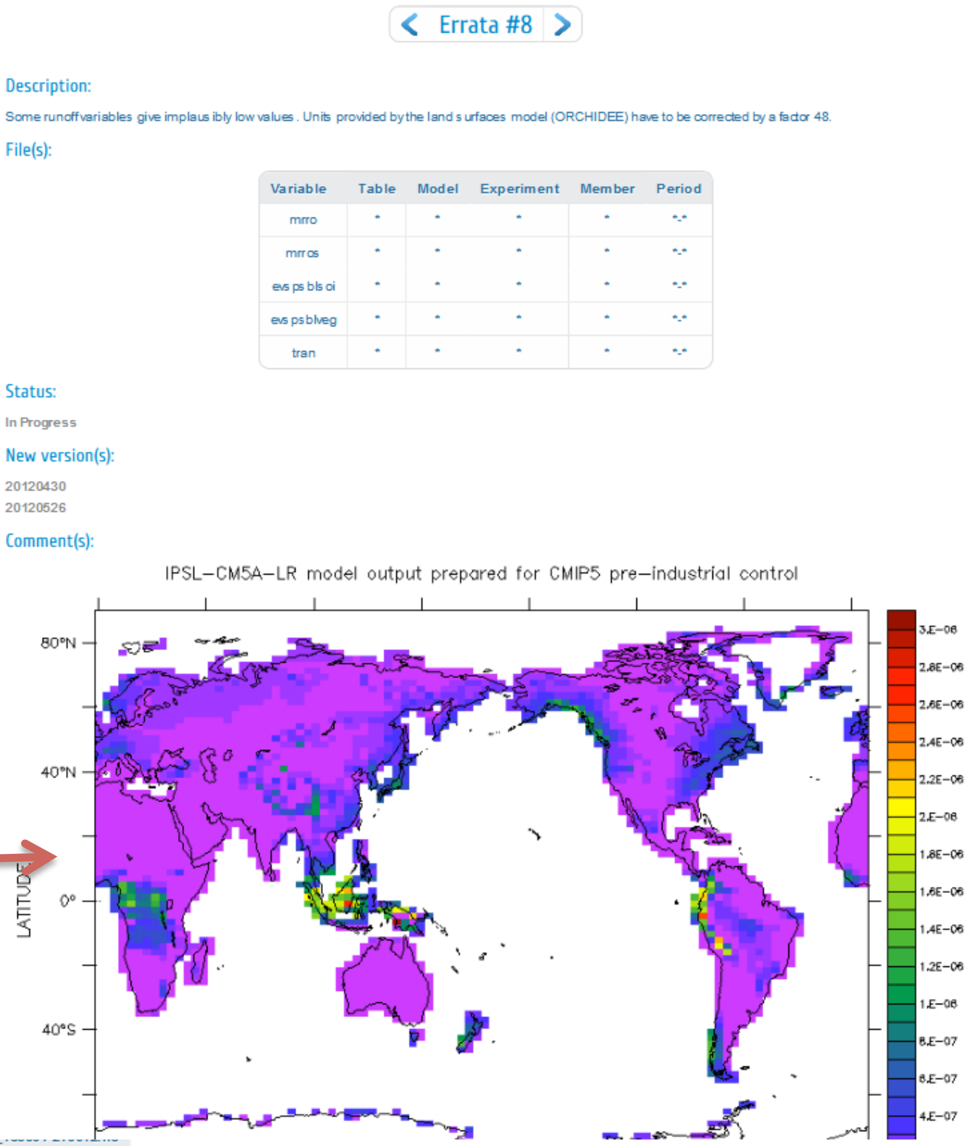
N.B.: Versions in bold letters are current versions of your files.

Dataset:	**cmip5.output1.IPSL.IPSL-CM5A-LR.esmFdbk2.mon.**
PID:	10876/ESGF-cmip5.output1.IPSL.IPSL-CM5A-LR.esmFdbk2.mon.land.Lm
Current version:	***201***
Data citation:	**Identifier:** http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=IPILe2 **Title:** IPSL-CM5A-LR model output prepared for CMIP5 esmFdbk2 experiment, served by ESGF **Author:** Institut Pierre Simon Laplace (IPSL) **Publication Year:** 2011 **Publisher:** World Data Center for Climate (WDCC) at DKRZ
Versioning history:	20110726 is the original/first version. **20120430** corrects issues #1 #8 #18

Filename:	**baresoilFrac_Lmon_IPSL-CM5A-LR_esmFdbk2_r1i1p1_**
Dataset:	cmip5.output1.IPSL.IPSL-CM5A-LR.esmFdbk2.mon.land.Lmo
PID:	10876/ESGF-69a43a39-b519-4895-
Checksum :	ef276ab18a329913c8b00
Current version:	***201***
Versioning history:	20110726 is the original/first version. **20120430** is a symbolic link to previous version.

Filename:	**c3PftFrac_Lmon_IPSL-CM5A-LR_esmFdbk2_r1i1p1_**
Dataset:	cmip5.output1.IPSL.IPSL-CM5A-LR.esmFdbk2.mon.land.Lmon.r1i1p1.v20120430

Versioning info



Description:

Some runoff variables give implausibly low values. Units provided by the land surfaces model (ORCHIDEE) have to be corrected by a factor 48.

File(s):

Variable	Table	Model	Experiment	Member	Period
mrro	*	*	*	*	*-*
mrros	*	*	*	*	*-*
evspsblsoi	*	*	*	*	*-*
evspsblveg	*	*	*	*	*-*
tran	*	*	*	*	*-*

Status:

In Progress

New version(s):

20120430
20120526

Comment(s):

IPSL–CM5A–LR model output prepared for CMIP5 pre–industrial control

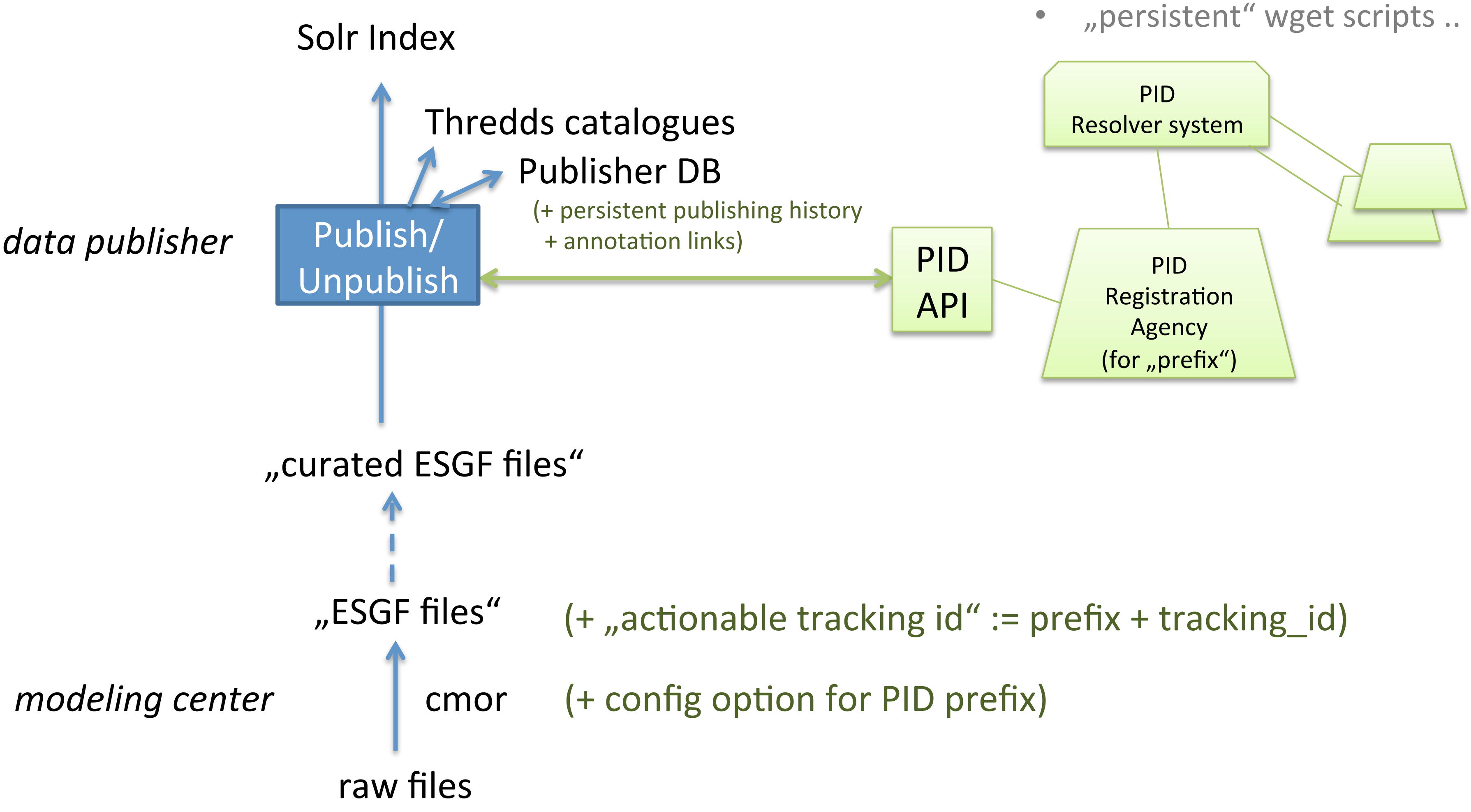
Next steps

Integration in publishing process

The (modified) publication process

„intelligent“ file services:

- newer versions
- nearest replica
- „persistent“ wget scripts ..



Solr Index

Thredds catalogues

Publisher DB

(+ persistent publishing history
+ annotation links)

data publisher

Publish/
Unpublish

PID
API

PID
Resolver system

PID
Registration
Agency
(for „prefix“)

„curated ESGF files“

„ESGF files“

(+ „actionable tracking id“ := prefix + tracking_id)

modeling center

cmor

(+ config option for PID prefix)

raw files

Next steps: Agreements / Policies

- Restrict user definable publication options
  - publish/unpublish only, automatic versioning and replication option settings
- PID assignment is part of an atomic publication: pid assignment failure – publication failure
- Assignment of PID prefix to modeling centers
- Commitment of some sites to run PID Handle servers or establish liaison with existing PID sites
  - Long term commitment → careful planing !

Next steps: Implementation

We propose to follow the prototype:

- Integrate handle PIDs in publication process
- PID metadata includes basic versioning and replication information
- Versioning and publication history storage at data nodes → see prototype in QC WG

Enforcement of PIDs for data entities inline with other intitiatives:

- DataOne (https://mule1.dataone.org/ArchitectureDocs-current/design/PIDs.html)
- EarthCube
- RDA (http://rda.org)
- ANDS (http://ands.org.au/guides/persistent-identifiers-working.html)

Replication: Next

→ Agree on one replication tool to work on

- syncrodata !?

→ Requirements list and priorities

- supported tranfer mechanisms (+ globus!?)
- monitoring (←→ icnwg working group)
- notification hooks and mechanism to enable automatic procedures

→ Definition of roadmap

Summary

- Start from end user perspective: How to bring versioning (and replica) info to end users (and later data-evaluation wflows - for provenance tracking)

~ Plan:

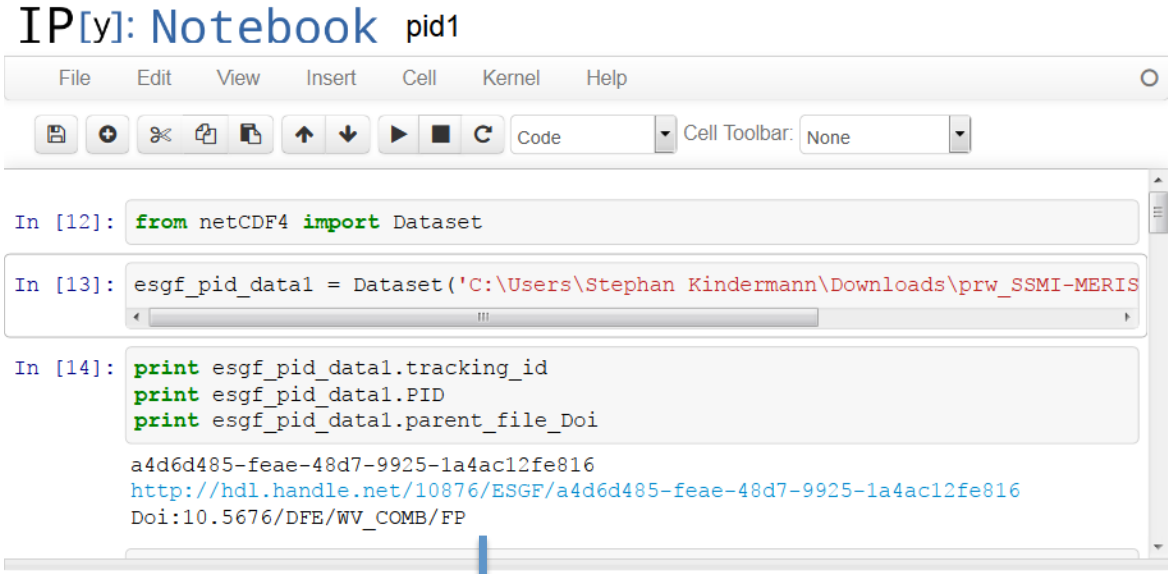
- February/March detailed work plan with ressource estimation
- March: ESGF PID scenario presentation at RDA meeting in San Diego

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE

- June: working publication add on prototype – additional sites running Handle service
- August/September: Intensive testing – tuning
- November / December: integrate as optional part in ESGF publisher

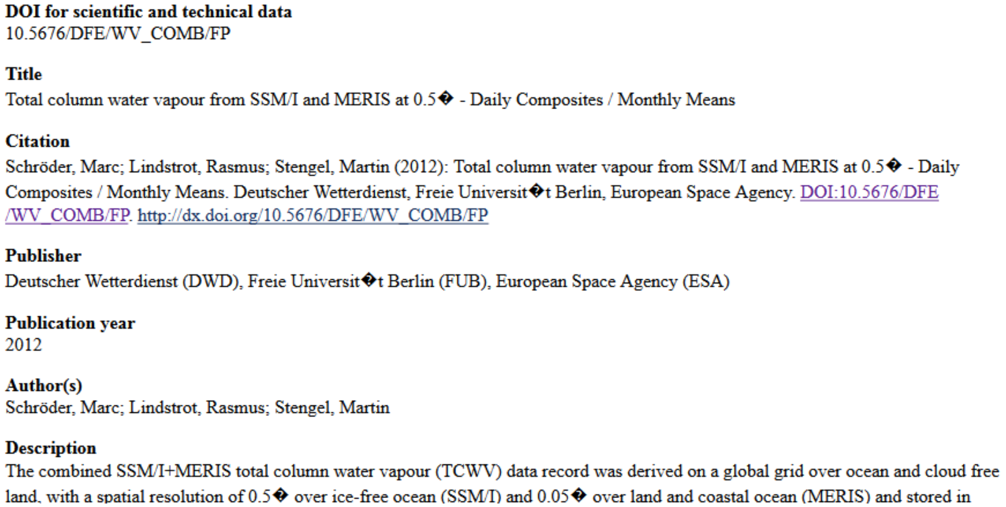
  (requirement for ESGF projects with strong data curation requirements)

Initial protoyping



```
In [12]: from netCDF4 import Dataset

In [13]: esgf_pid_data1 = Dataset('C:\Users\Stephan Kindermann\Downloads\prw_SSMI-MERIS

In [14]: print esgf_pid_data1.tracking_id
         print esgf_pid_data1.PID
         print esgf_pid_data1.parent_file_Doi

         a4d6d485-feae-48d7-9925-1a4ac12fe816
         http://hdl.handle.net/10876/ESGF/a4d6d485-feae-48d7-9925-1a4ac12fe816
         Doi:10.5676/DFE/WV_COMB/FP
```

PID → DOI transition strategy: Later step



DOI for scientific and technical data
10.5676/DFE/WV_COMB/FP

Title
Total column water vapour from SSM/I and MERIS at 0.5� - Daily Composites / Monthly Means

Citation
Schröder, Marc; Lindstrot, Rasmus; Stengel, Martin (2012): Total column water vapour from SSM/I and MERIS at 0.5� - Daily Composites / Monthly Means. Deutscher Wetterdienst, Freie Universit�t Berlin, European Space Agency. DOI:10.5676/DFE/WV_COMB/FP. http://dx.doi.org/10.5676/DFE/WV_COMB/FP

Publisher
Deutscher Wetterdienst (DWD), Freie Universit�t Berlin (FUB), European Space Agency (ESA)

Publication year
2012

Author(s)
Schröder, Marc; Lindstrot, Rasmus; Stengel, Martin

Description
The combined SSM/I+MERIS total column water vapour (TCWV) data record was derived on a global grid over ocean and cloud free land, with a spatial resolution of 0.5� over ice-free ocean (SSM/I) and 0.05� over land and coastal ocean (MERIS) and stored in