Metadata aspects at DKRZ
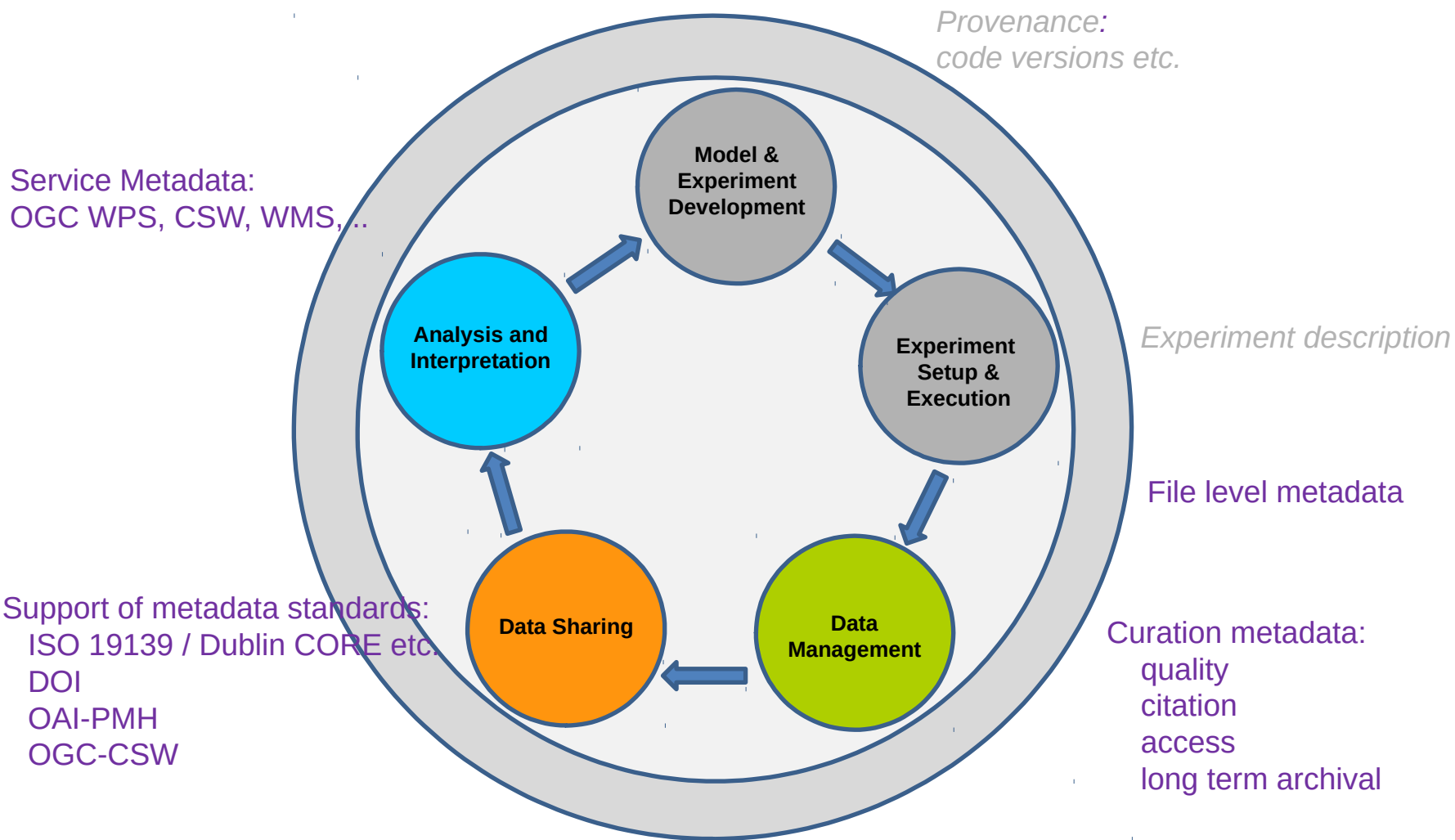
LSDMA Metadata Topical Meeting
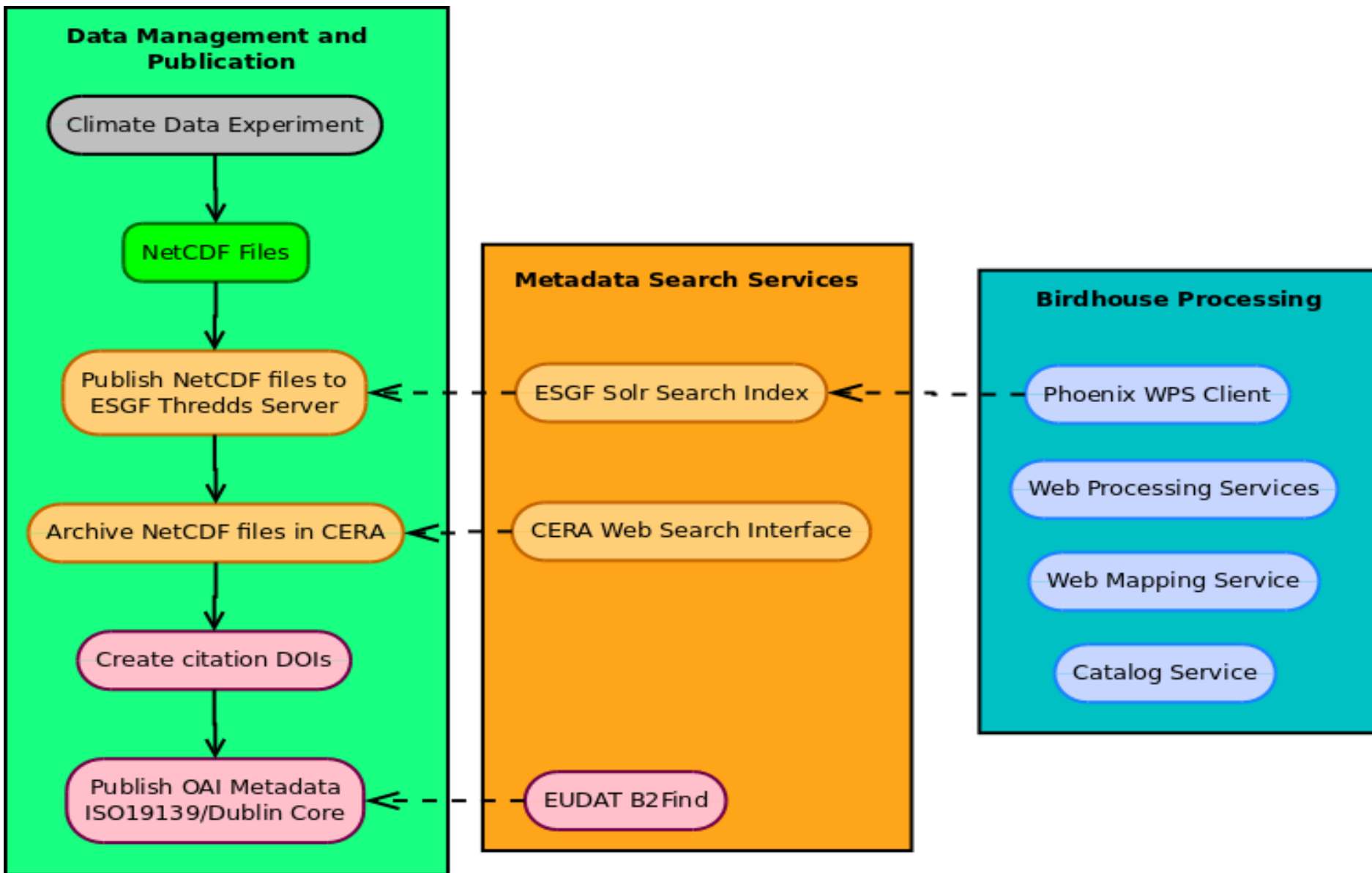## October 2014

Carsten Ehbrecht
Stephan Kindermann

Overview

- Metadata aspects in the Climate Data Life Cycle
- From experiment outputs over data management to data processing ...
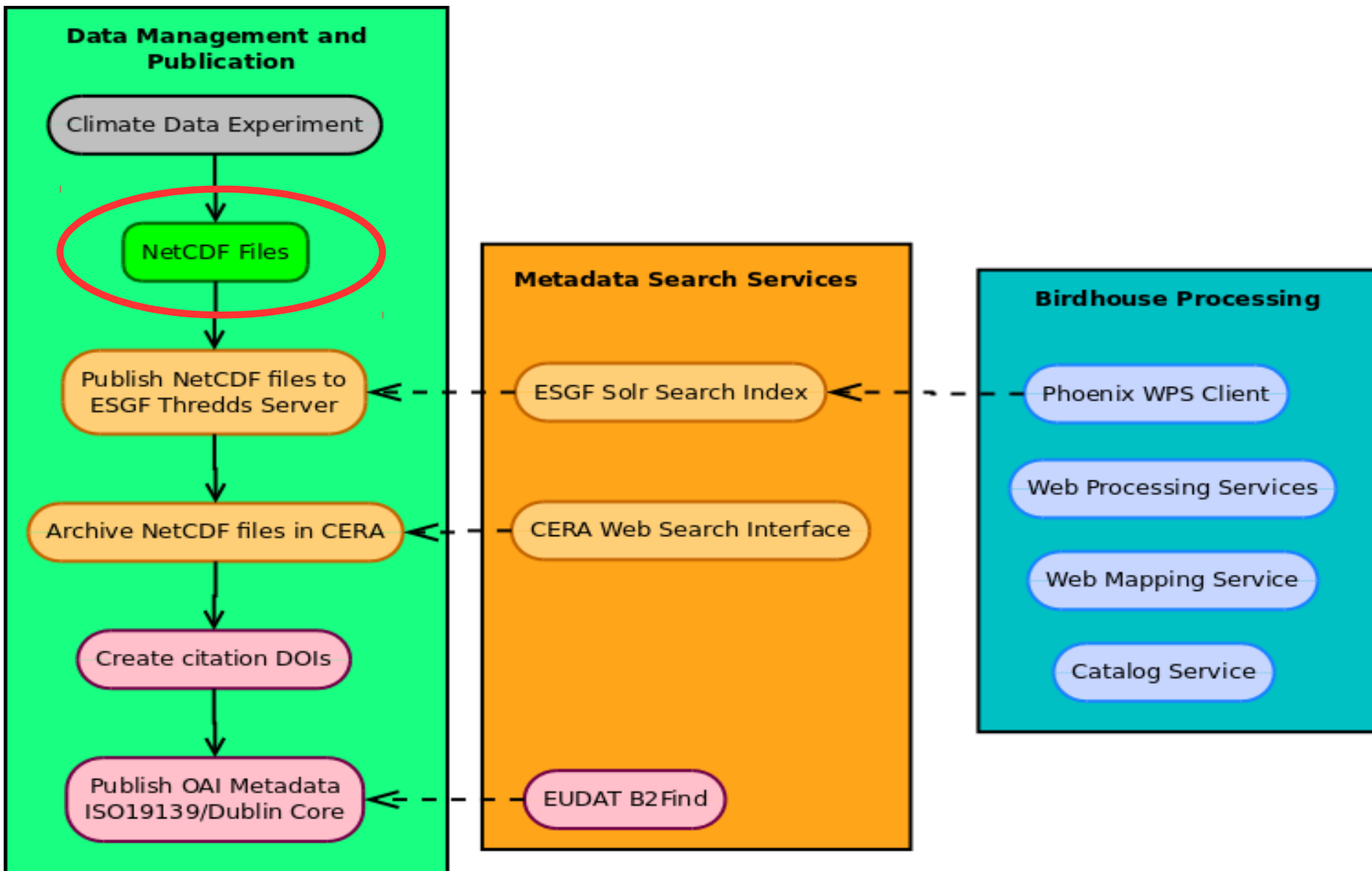- Example Processing Application: LSDMA Birdhouse Project
- Summary / Discussion

Climate Data Life Cycle: Metadata

Provenance:
code versions etc.

Service Metadata:
OGC WPS, CSW, WMS,..

Experiment description

**Model &
Experiment
Development**

**Analysis and
Interpretation**

**Experiment
Setup &
Execution**

Data Sharing

**Data
Management**

File level metadata

Support of metadata standards:
   ISO 19139 / Dublin CORE etc.
   DOI
   OAI-PMH
   OGC-CSW

Curation metadata:
   quality
   citation
   access
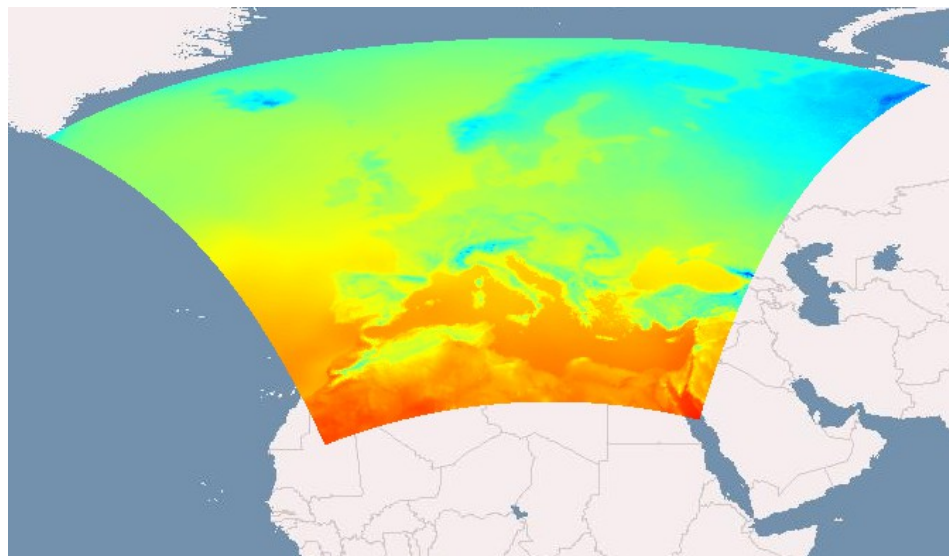   long term archival

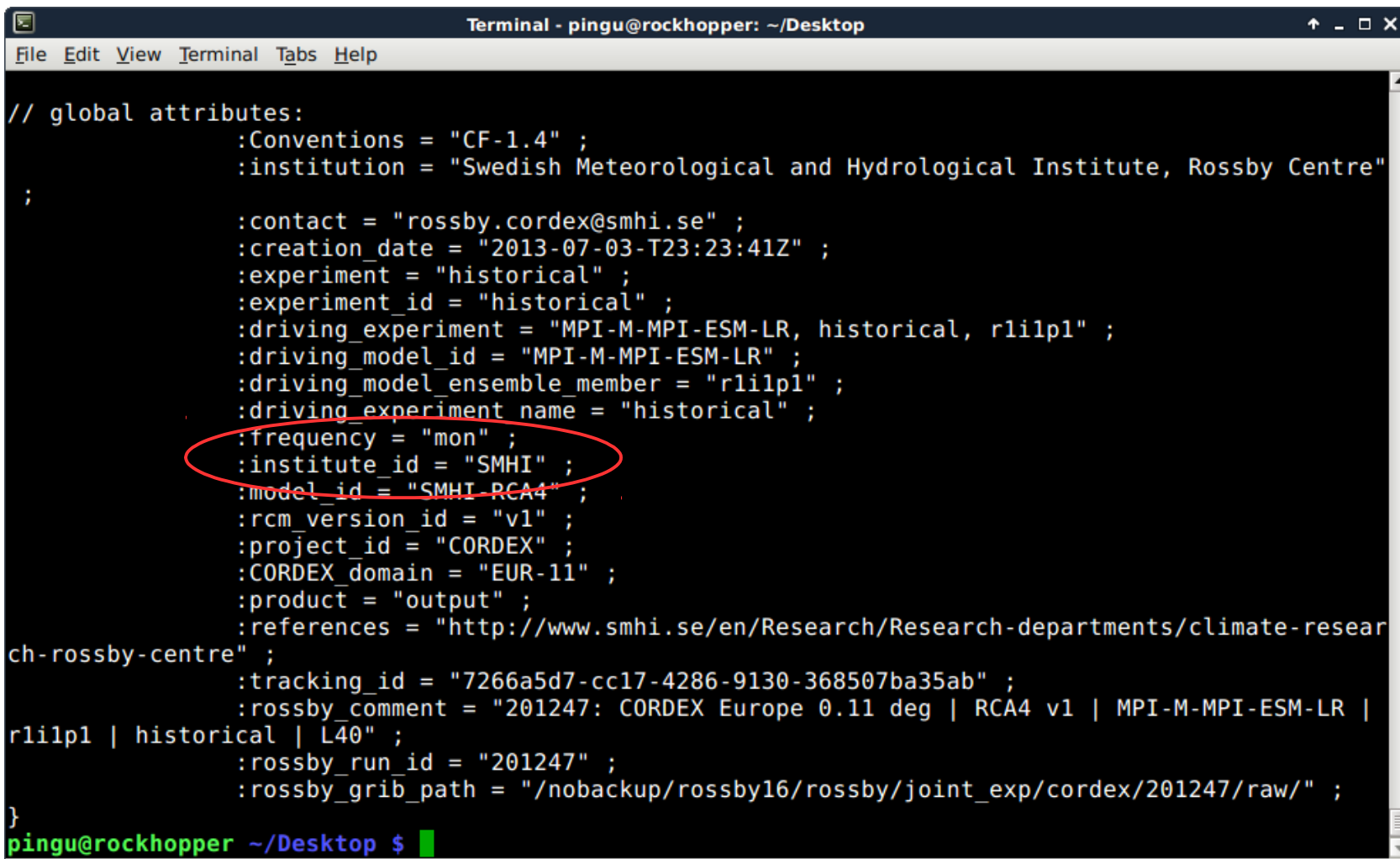# Leitmotif: the red yarn ...

Leitmotif: the red yarn … NetCDF

Experiment Output: NetCDF files

- NetCDF = Network Common Data Form
- Open standard file format for array-oriented scientific data
- Self-describing: contains metadata as key/value attributes
  - Additional metadata: author, institute, variables names (CF conventions), …
- Other formats used in climate science:
  - GRIB (contains no key/value metadata)
- http://en.wikipedia.org/wiki/NetCDF

Example: Metadata of NetCDF file

$ ncdump –h cordex.nc

Climate and Forecast Metadata Conventions

[..] designed to promote processing and sharing of NetCDF files (wikipedia).

Project Page: http://cfconventions.org

Example: Standard name for surface temperature
(http://cfconventions.org/standard-names.html)

```
-<entry id="surface_temperature">
    <canonical_units>K</canonical_units>
    <grib>E139</grib>
    <amip>ts</amip>
  -<description>
    The surface called "surface" means the lower boundary of the atmosphere.
    The surface temperature is the temperature at the interface, not the bulk
    temperature of the medium above or below. Unless indicated in the
    cell_methods attribute, a quantity is assumed to apply to the whole area of
    each horizontal grid box. Previously, the qualifier where_type was used to
    specify that the quantity applies only to the part of the grid box of the named
    type. Names containing the where_type qualifier are deprecated and newly
    created data should use the cell_methods attribute to indicate the horizontal
    area to which the quantity applies.
    </description>
</entry>
```
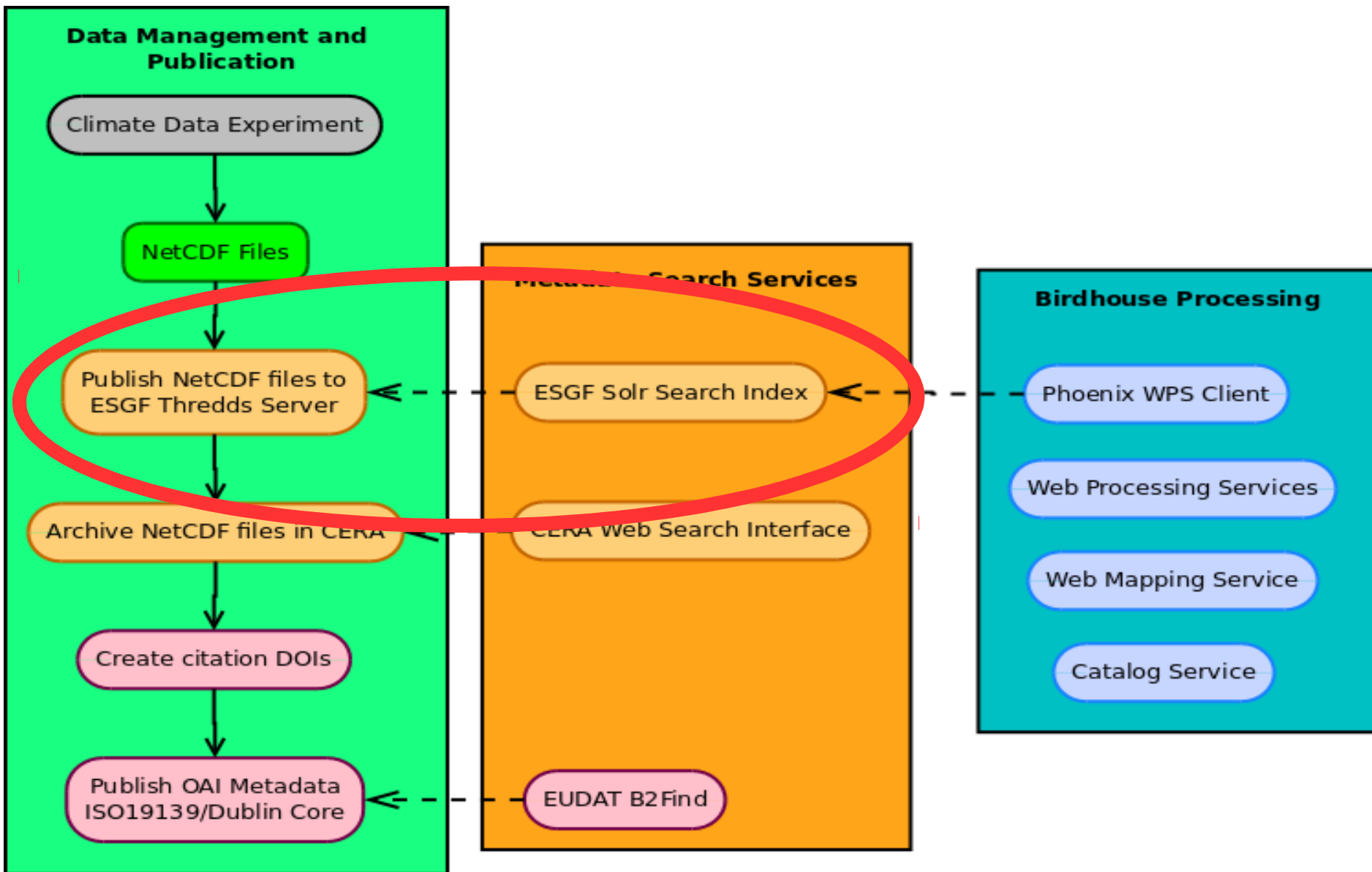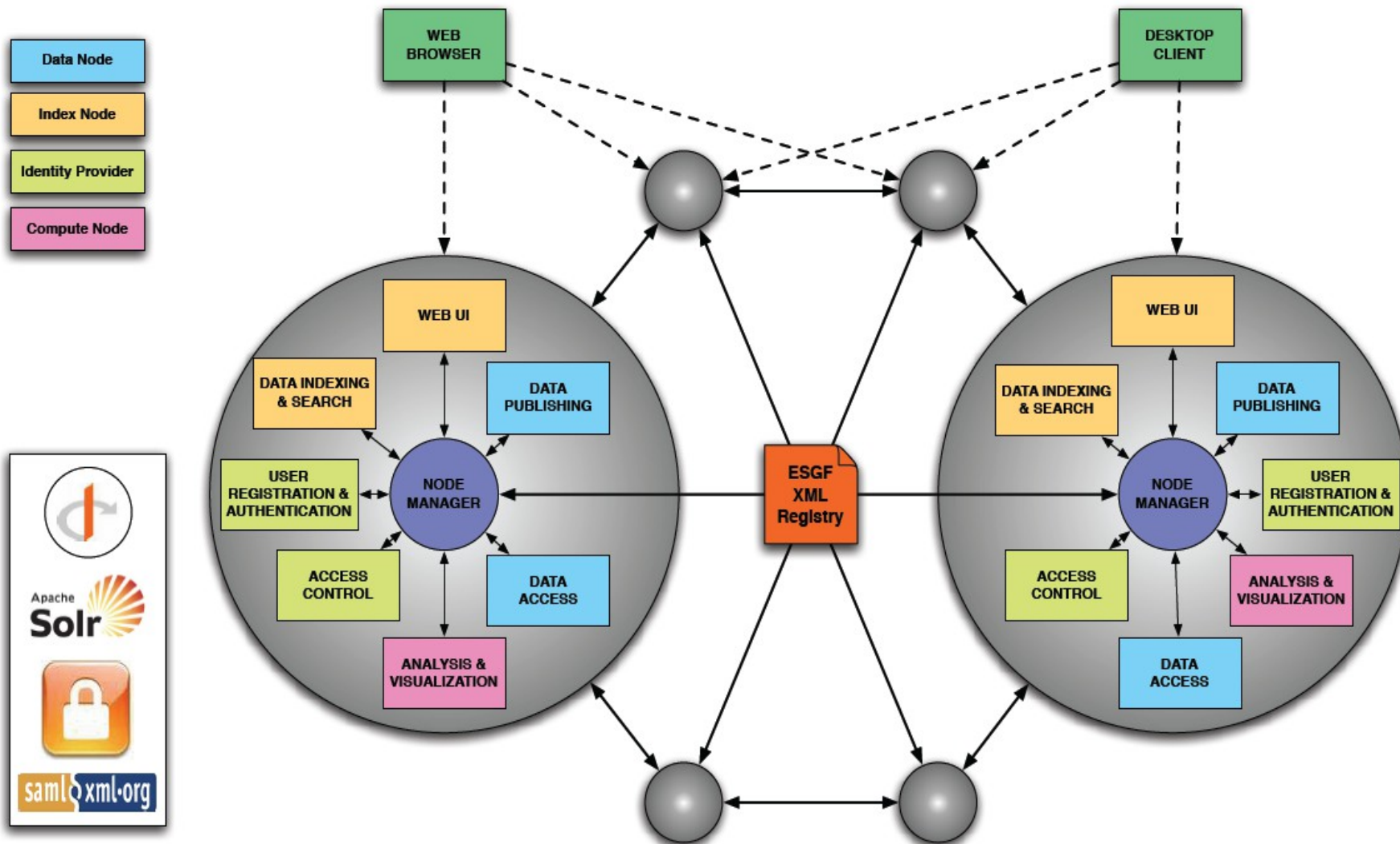
Leitmotif: the red yarn … ESGF

ESGF: Federated Data Access

ESGF faceted search: http://esgf-data.dkrz.de

Thredds data server: http://carbon.dkrz.de/thredds/

Dataset: cordex.output.AFR-44.CLMcom.CNRM-CERFACS-CNRM-CM5.historical.r0i0p0.CCLM4-8-17.v1.fx.orog/orog_AFR-44_CNRM-CERFACS-CNRM-CM5_historical_r0i0p0_CLMcom-CCLM4-8-17_v1_fx.nc

- *Data format:* netCDF
- *Data size:* 262.3 Kbytes
- *Data type:* GRID
- *ID:* cordex.output.AFR-44.CLMcom.CNRM-CERFACS-CNRM-CM5.historical.r0i0p0.CCLM4-8-17.v1.fx.orog.v20140401.orog_AFR-44_CNRM-CERFACS-CNRM-CM5_historical_r0i0p0_CLMcom-CCLM4-8-17_v1_fx.nc
- *RestrictAccess:* esg-user

Access:

1. **HTTPServer:** /thredds/fileServer/cordex/output/AFR-44/CLMcom/CNRM-CERFACS-CNRM-CM5/historical/r0i0p0/CLMcom-CCLM4-8-17/v1/fx/orog/v20140401/orog_AFR-44_CNRM-CERFACS-CNRM-CM5_historical_r0i0p0_CLMcom-CCLM4-8-17_v1_fx.nc
2. **GridFTP:** gsiftp://carbon.dkrz.de:2811//cordex/output/AFR-44/CLMcom/CNRM-CERFACS-CNRM-CM5/historical/r0i0p0/CLMcom-CCLM4-8-17/v1/fx/orog/v20140401/orog_AFR-44_CNRM-CERFACS-CNRM-CM5_historical_r0i0p0_CLMcom-CCLM4-8-17_v1_fx.nc
3. **OPENDAP:** /thredds/dodsC/cordex/output/AFR-44/CLMcom/CNRM-CERFACS-CNRM-CM5/historical/r0i0p0/CLMcom-CCLM4-8-17/v1/fx/orog/v20140401/orog_AFR-44_CNRM-CERFACS-CNRM-CM5_historical_r0i0p0_CLMcom-CCLM4-8-17_v1_fx.nc

Variables:

- *Vocabulary* [CF-1.0]:
    - **orog (m)** = *Surface Altitude* = surface_altitude

Properties:

- file_id = "cordex.output.AFR-44.CLMcom.CNRM-CERFACS-CNRM-CM5.historical.r0i0p0.CCLM4-8-17.v1.fx.orog.orog_AFR-44_CNRM-CERFACS-CNRM-CM5_historical_r0i0p0_CLMcom-CCLM4-8-17_v1_fx.nc"
- file_version = "1"
- size = "262363"
- tracking_id = "060991c1-838a-41b7-9e0a-9212c071ade4"
- mod_time = "2014-03-14 10:52:15"
- checksum = "6f222ffb1ffa77ce82308386c87f74a6"
- checksum_type = "MD5"

Leitmotif: the red yarn … CERA
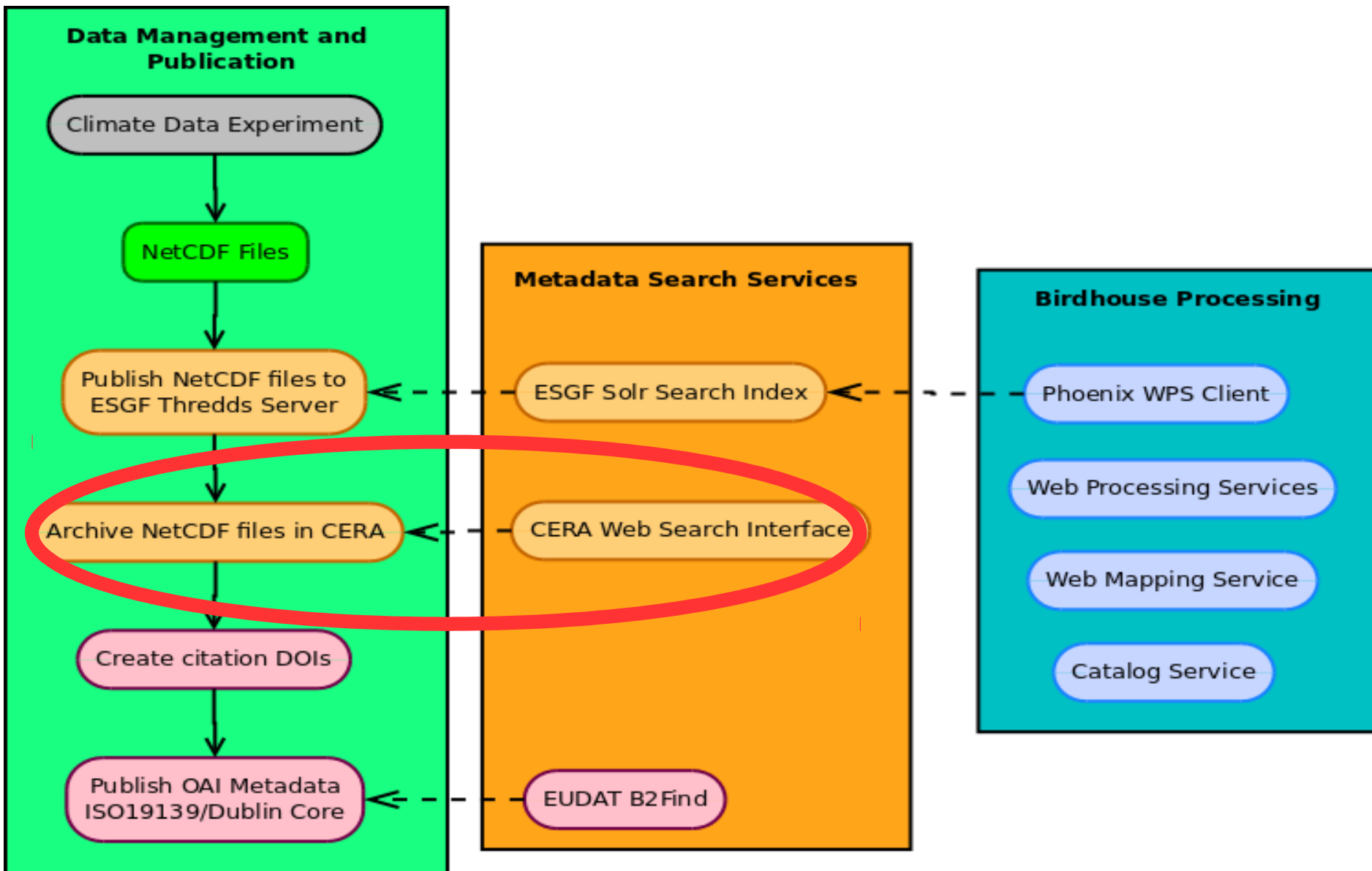
CERA: Long Term Archiving at DKRZ

- Long-term archiving is requested by the founding agencies in the context of "Rules for Good Scientific Practices" as a prerequisite to enable future research activities
- Storage period of at least 10 years
- Oracle database for metadata with pointers to tape archive with disk cache
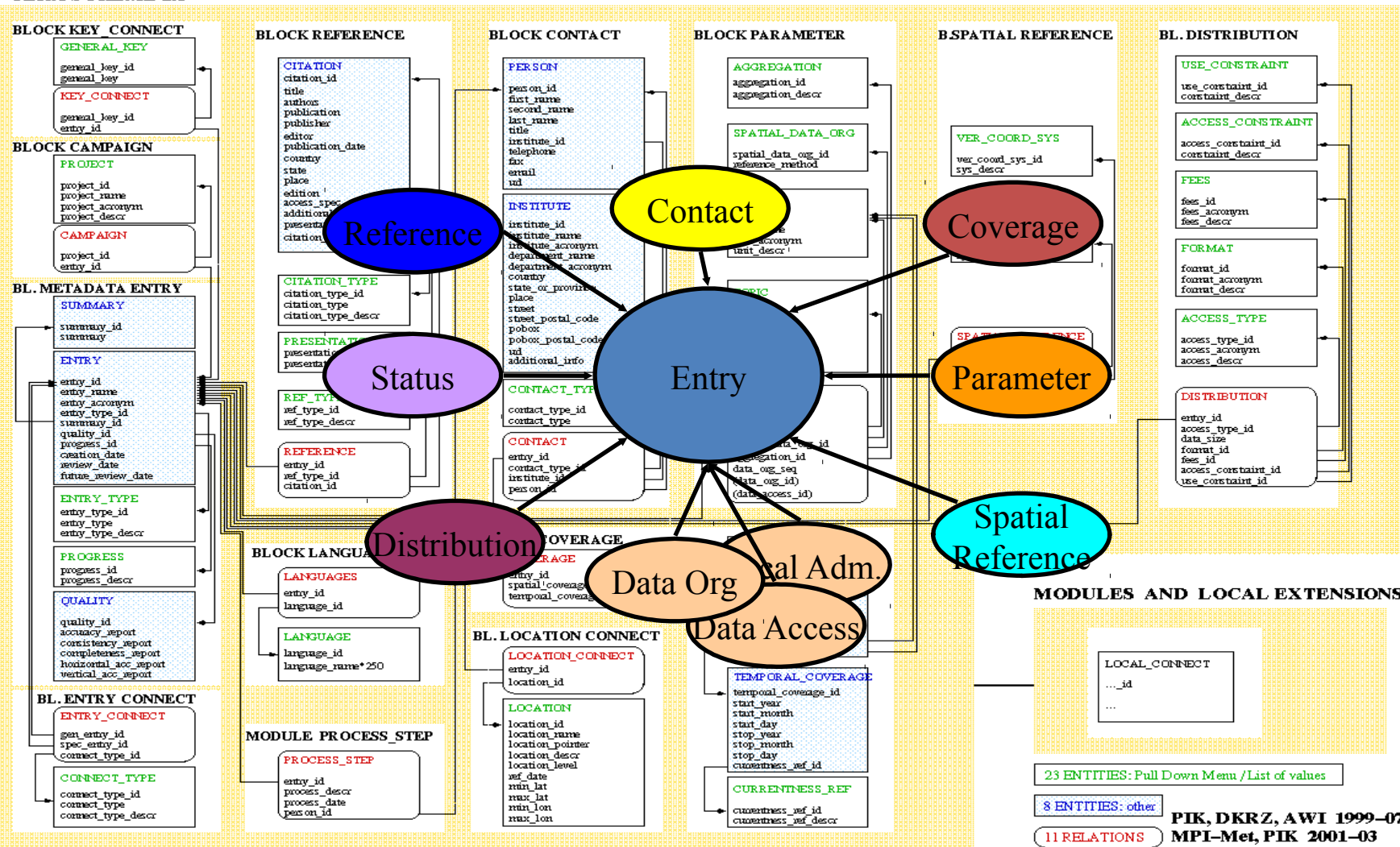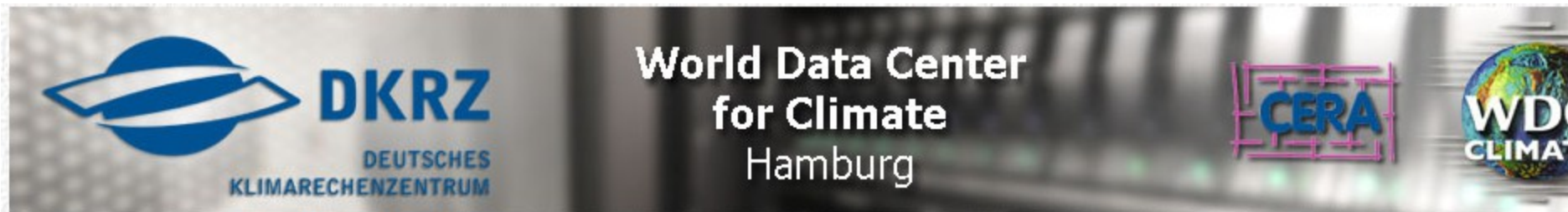
… strong and willed Cera (The Land Before Time)
http://landbeforetime.wikia.com/wiki/Cera

CERA: Data documentation

CERA SCHEME 2.5

CERA web access: http://cera-www.dkrz.de

Leitmotif: the red yarn … Data Citation

Data Management and Publication

- Climate Data Experiment
- NetCDF Files
- Publish NetCDF files to ESGF Thredds Server
- Archive NetCDF files in CERA
- Create citation DOIs
- Publish OAI Metadata ISO19139/Dublin Core

Metadata Search Services

- ESGF Solr Search Index
- CERA Web Search Interface
- EUDAT B2Find

Birdhouse Processing

- Phoenix WPS Client
- Web Processing Services
- Web Mapping Service
- Catalog Service

Data citation with DOI (digital object identifier)

- DOIs are used for citation of published climate experiment data archived in the CERA database at DKRZ

  Example: http://dx.doi.org/DOI:10.1594/WDCC/CLM_B1_2_D3
- DOI remains fixed over lifetime, whereas the location of the document may change (wikipedia)
- "Cream of the Cake" for climate data scientists

CERA: DOI landing page

DKRZ

World Data Center for Climate
Hamburg

Not logged in (**Login**) | **Process List**   **CERA UI Home | WDCC Home | Impressum**

Always quote citation when using data!

DOI for Scientific and Technical Data
10.1594/WDCC/CLM_B1_2_D3

Citation elements

Creator
(person(s) or institute(s) responsible for this assemblage of data: e.g. author, data collector, editor...)
Lautenschlager,Michael; Keuler,Klaus; Wunram,Claudia; Keup-Thiel,Elke; Schubert, Martina; Will,Andreas; Rockel,Burkhardt; Boehm,Uwe

Publication Year
2009

Title
Climate Simulation with CLM, Scenario B1 run no.2, Data Stream 3: European region MPI-M/MaD

DOI Publisher
WDCC at DKRZ

Identifier
DOI:10.1594/WDCC/CLM_B1_2_D3

*The **DataCite** consortium proposes a **citation format** (Creator (PublicationYear): Title. DOI Publisher. Identifier) and also offers citation export in different formats.*
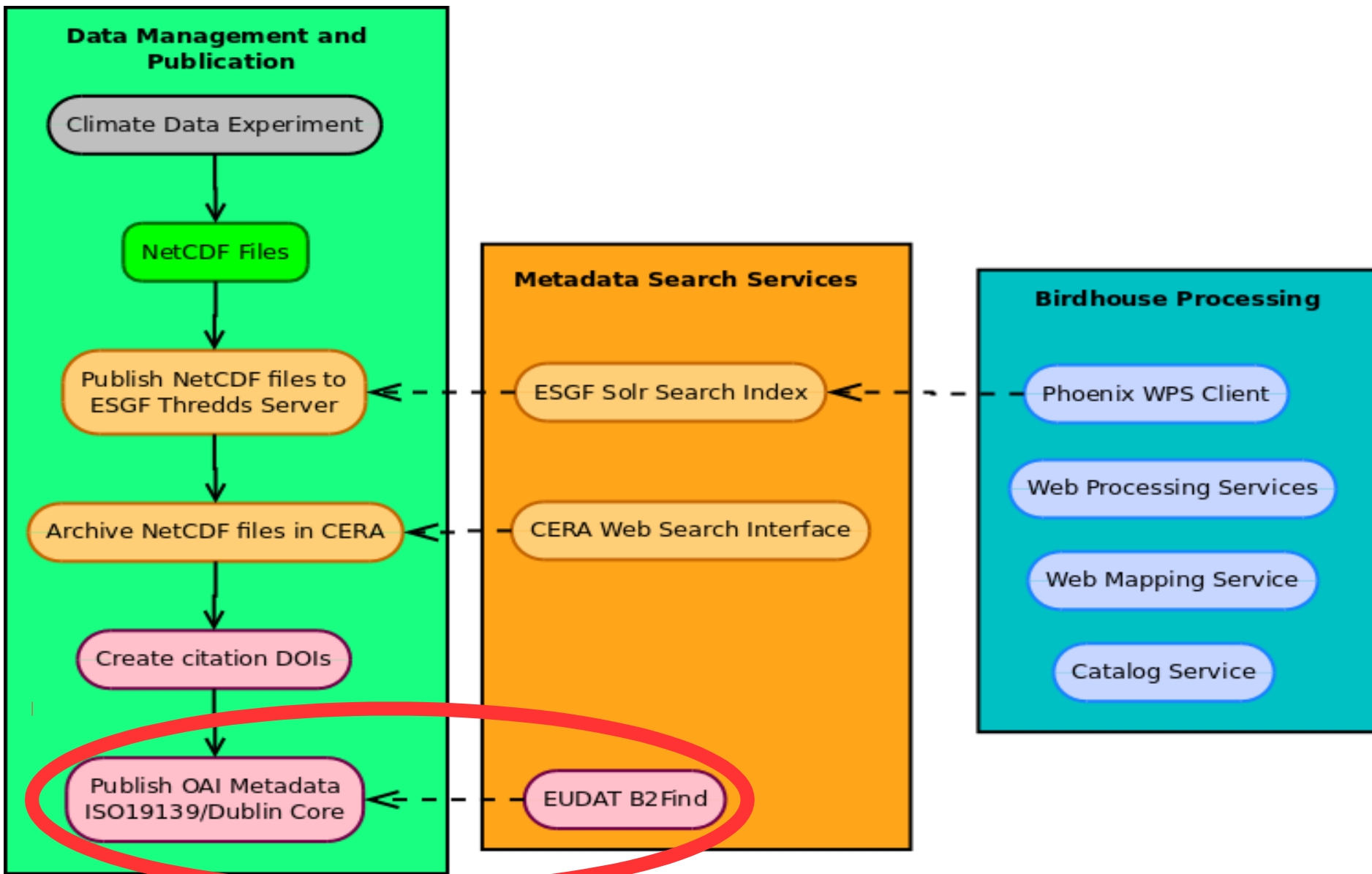
Detailed Metadata
http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=CLM_B1_2_D3

Data Access
http://cera-www.dkrz.de/WDCC/ui/EntryList.jsp?acronym=CLM_B1_2_D3

Leitmotif: the red yarn … Publish OAI/B2Find

OAI Metadata Server: ISO 19139 document

```xml
-<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0
 /OAI-PMH.xsd">
   <responseDate>2014-10-02T09:17:12Z</responseDate>
   <request verb="GetRecord" identifier="oai:c3grid.dkrz:de.dkrz.wdcc.iso_doi2110948"
   metadataPrefix="iso">http://c3grid1.dkrz.de:8080/oai/provider</request>
 -<GetRecord>
  -<record>
   -<header>
      <identifier>oai:c3grid.dkrz:de.dkrz.wdcc.iso_doi2110948</identifier>
      <datestamp>2014-03-17T08:20:34Z</datestamp>
      <setSpec>iso-wdcc-doi</setSpec>
   </header>
  -<metadata>
   -<gmi:MI_Metadata xsi:schemaLocation="http://www.isotc211.org/2005/gmi http://www.ngdc.noaa.gov
    /metadata/published/xsd/schema.xsd" id="de.dkrz.mpim.iso2110948">
    -<gmd:fileIdentifier>
       <gco:CharacterString>de.dkrz.mpim.iso2110948</gco:CharacterString>
     </gmd:fileIdentifier>
    -<gmd:hierarchyLevel>
       <gmd:MD_ScopeCode codeList="http://wis.wmo.int/2006/catalogues
       /gmxCodelists.xml#MD_ScopeCode" codeListValue="series">series</gmd:MD_ScopeCode>
     </gmd:hierarchyLevel>
    -<gmd:hierarchyLevelName>
       <gco:CharacterString>CLM_regional_climate_model_runs:CLM_B1_2_D3</gco:CharacterString>
     </gmd:hierarchyLevelName>
    -<gmd:contact>
     -<gmd:CI_ResponsibleParty id="data_provider">
      -<gmd:individualName>
         <gco:CharacterString>Keup-Thiel, Elke</gco:CharacterString>
       </gmd:individualName>
```

EUDAT B2Find (CKAN): http://b2find.eudat.eu/

EUDAT B2Find: Additional Info for Dataset

Additional Info

| Field	Value
Source	http://dx.doi.org/DOI:10.1594/WDCC/CLM_B1_2_D3
Author	Dr. Michael Lautenschlager;Dr. rer. nat. Klaus Keuler;Dr. Burkhardt Rockel; Martina Schubert;Dr. Claudia Wunram
Maintainer	Max-Planck-Institut fuer Meteorologie (MD), Model and Data
Discipline	series
Format	network Common Data Format
Language	German
MetaDataAccess	http://c3grid1.dkrz.de:8080/oai/provider?verb=GetRecord&metadataPrefix=iso&identifier=oai:c3grid.dkrz:de.dkrz.wdcc.iso_doi_iso_old2110948
Origin	Max-Planck-Institut fuer Meteorologie (MD), Model and Data
PublicationYear	2006
TempCoverageBegin	63116496000
TempCoverageEnd	66269491200
TemporalCoverage:BeginDate	2001-01-31T00:00:00Z
TemporalCoverage:EndDate	2100-12-31T00:00:00Z
oai_identifier	oai:c3grid.dkrz:de.dkrz.wdcc.iso_doi_iso_old2110948

# Leitmotif: the red yarn … OGC Services
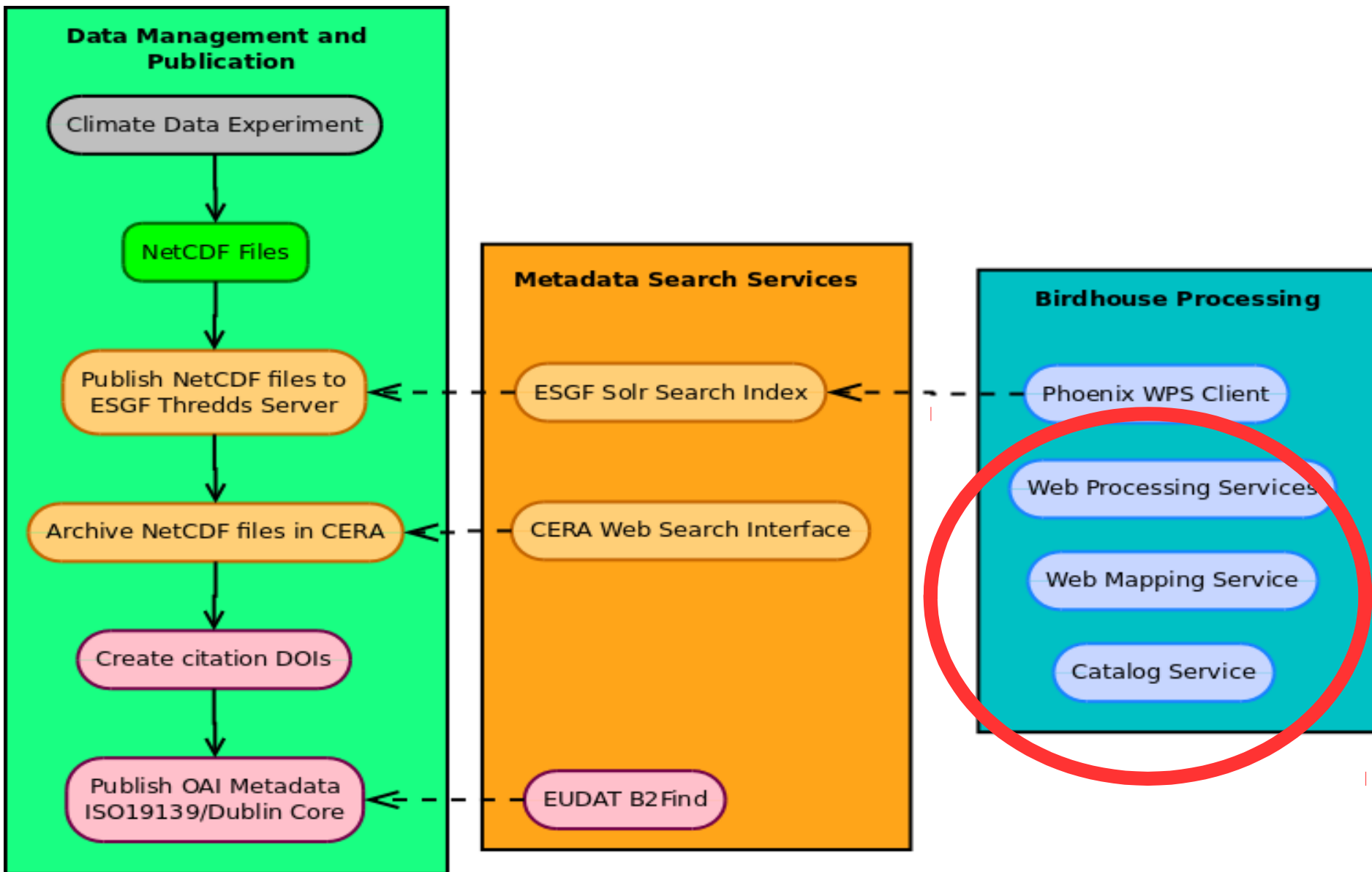
The Web Processing Service Standard

- "The OGC **Web Processing Service** (**WPS**) Interface Standard provides rules for standardizing inputs and outputs (requests and responses) for invoking geospatial processing services, ..., as a Web service" (wikipedia)

- "WPS makes it possible to publish, find, and bind to processes in a standardized and thus interoperable fashion" (wikipedia)

- Precondition for workflow composition
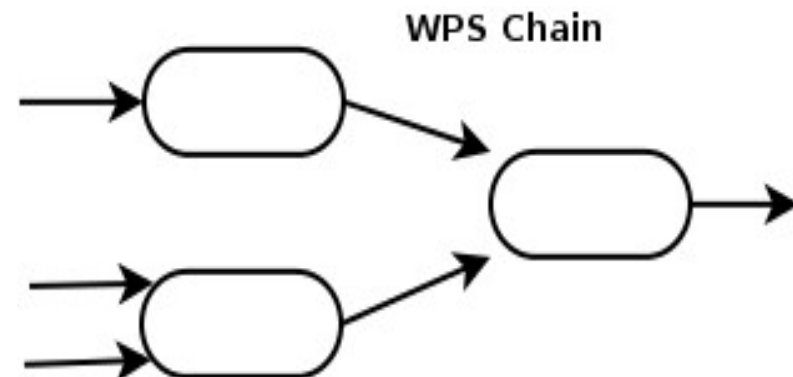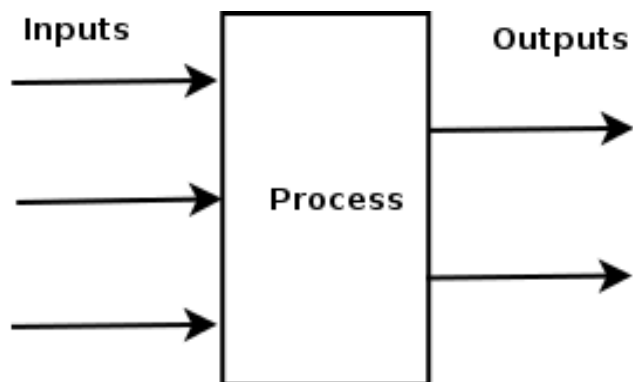
Example: WPS capabilities

IP[y]: Notebook wps-caps (autosaved)

File   Edit   View   Insert   Cell   Kernel   Help   O

Code ▾   Cell Toolbar: None ▾
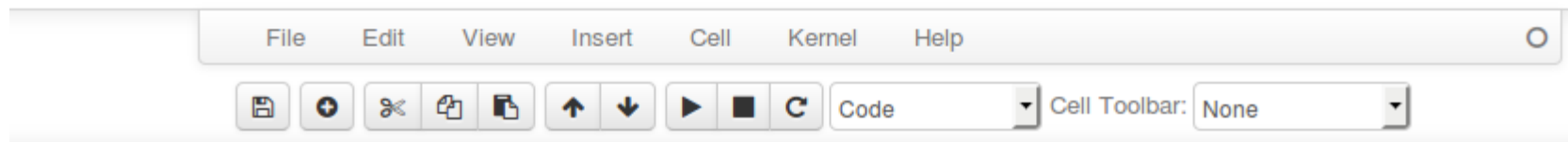
```
In [1]:  from owslib.wps import WebProcessingService
         wps = WebProcessingService("http://localhost:8094/wps", skip_caps=True)
         wps.version
```

Out[1]: '1.0.0'

```
In [2]:  wps.getcapabilities()
         wps.identification.abstract
```

Out[2]: 'WPS processes for testing and demos'

```
In [3]:  [op.name for op in wps.operations]
```
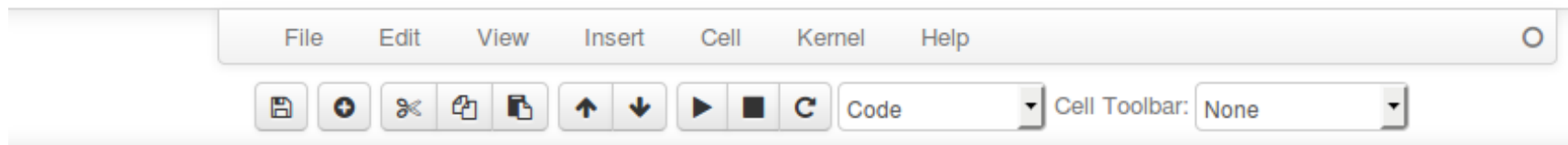
Out[3]: ['GetCapabilities', 'DescribeProcess', 'Execute']

```
In [4]:  for process in wps.processes:
             print process.title, "/", process.identifier
```

```
         Chomsky test generator / chomsky
         Hello World / helloworld
         Testing all Data Types / inout
         Answer to Life, the Universe and Everything / ultimatequestionprocess
         Word Counter / wordcount
```

Example: WPS describeprocess

IP[y]: Notebook wps-describe *(autosaved)*

File	Edit	View	Insert	Cell	Kernel	Help

In [1]:
```python
from owslib.wps import WebProcessingService
wps = WebProcessingService("http://localhost:8094/wps", skip_caps=True)
wps.version
```

Out[1]: '1.0.0'

In [2]:
```python
wps.getcapabilities()
wps.identification.abstract
```

Out[2]: 'WPS processes for testing and demos'

In [3]:
```python
process = wps.describeprocess(identifier='helloworld')
print process.title, "/", process.abstract
```

Hello World / Welcome user and say hello ...

In [4]:
```python
for param in process.dataInputs:
    print param.title, "/", param.dataType
```

Your name / //www.w3.org/TR/xmlschema-2/#string

In [5]:
```python
for param in process.processOutputs:
    print param.title, "/", param.dataType
```

Welcome message / //www.w3.org/TR/xmlschema-2/#string

Example: CSW catalog service


```
IP[y]: Notebook   csw-example (autosaved)

File   Edit   View   Insert   Cell   Kernel   Help

Code              Cell Toolbar: None
```
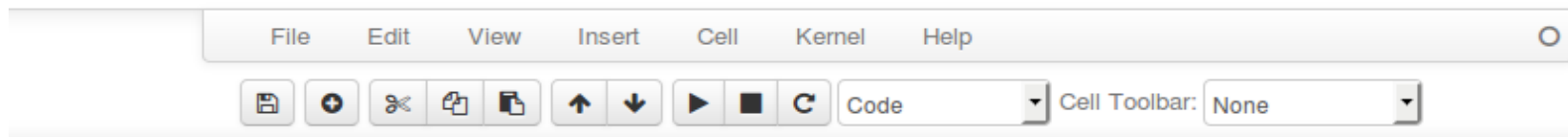
```
In [1]: from owslib.csw import CatalogueServiceWeb
        csw = CatalogueServiceWeb('http://localhost:8082/csw')
        csw.version

Out[1]: '2.0.2'
```

```
In [2]: [op.name for op in csw.operations]

Out[2]: ['GetCapabilities',
         'Transaction',
         'GetRepositoryItem',
         'DescribeRecord',
         'GetDomain',
         'GetRecordById',
         'GetRecords',
         'Harvest']
```

```
In [5]: csw.getrecords(keywords=['pywps'], maxrecords=5)
        csw.results

Out[5]: {'matches': 2, 'nextrecord': 0, 'returned': 2}
```

```
In [6]: for rec in csw.records:
            item = csw.records[rec]
            print item.title, "/", item.abstract

        Malleefowl / Malleefowl Processes (esgf, workflow, publish, security, ...)
        Flyingpigeon / Processes for climate data, indices and extrem events
```

Leitmotif: the red yarn … Birdhouse Processing
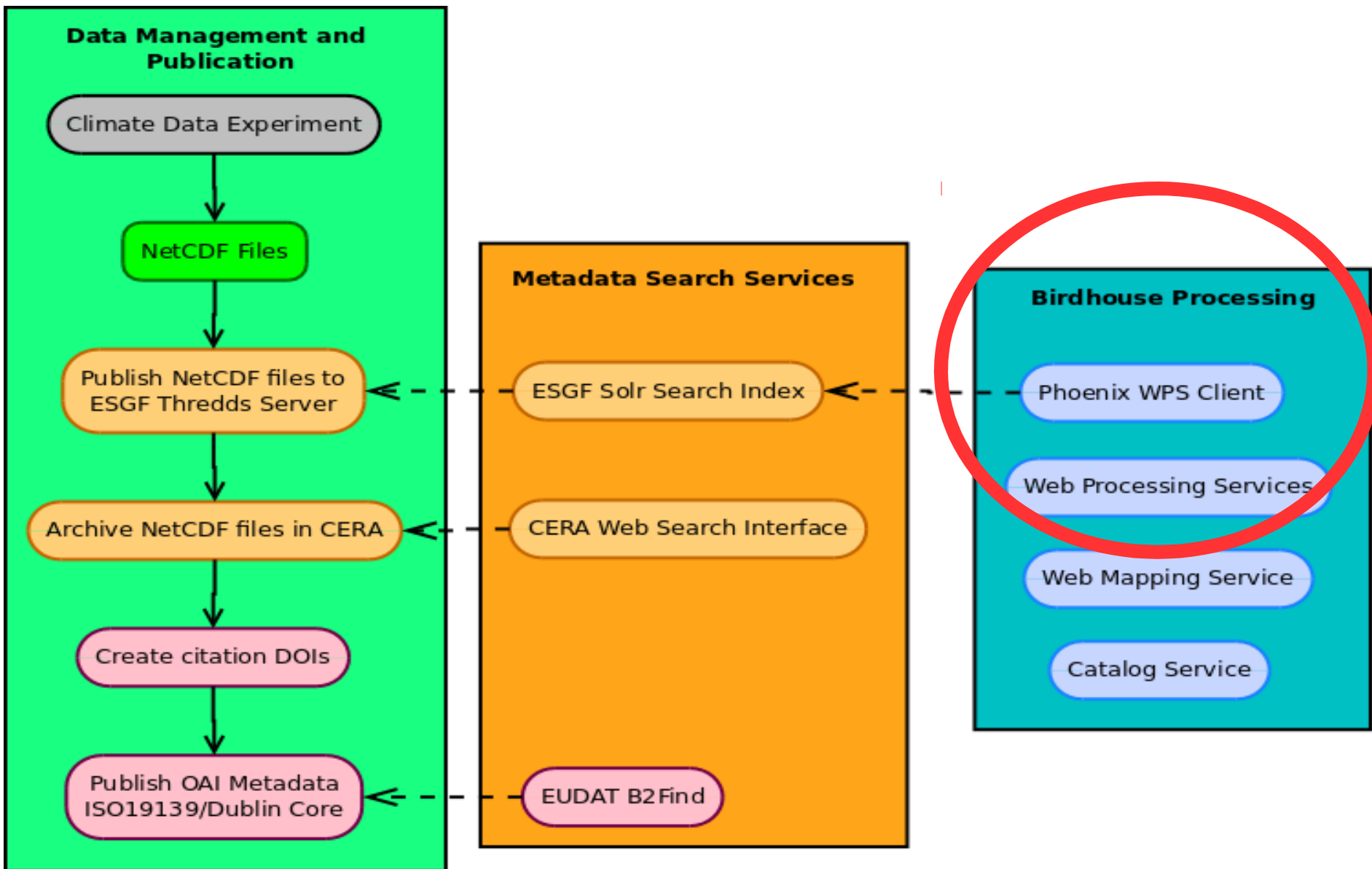
The Birdhouse WPS component system

- Currently being developed in the LSDMA „Erde und Umwelt" life cycle lab
- Provides generic components supporting standards based earth data access, discovery, management and processing
- Uses OGC services like WPS, CSW and WMS
- Available on GitHub: https://github.com/bird-house

LSDMA

Birdhouse Overview

WPS process selection

WPS execute process with parameters

WPS outputs

Catalog Service for WPS registration and published files

ESGF as data source

Visualizing with WMS
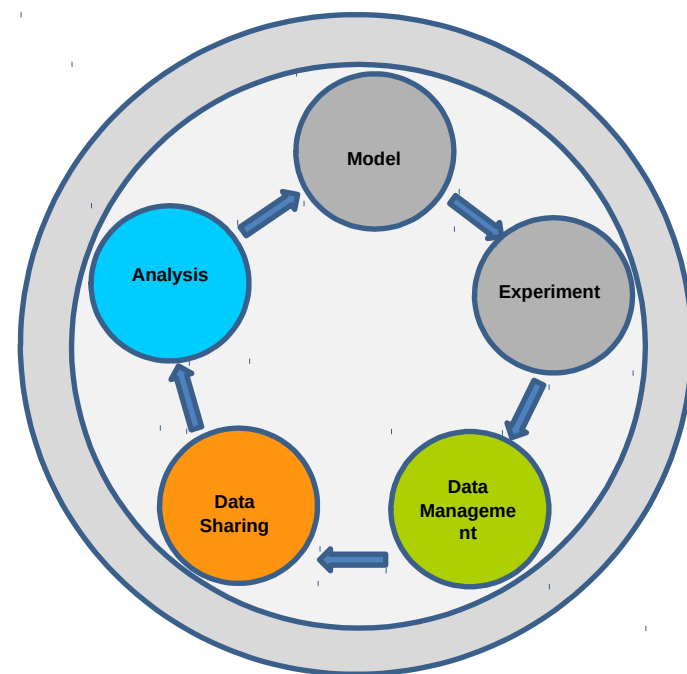
Summary

- Climate Data Life Cycle involves a heterogeneous set of tools over institutional borders
- Open Standards are important to …
  - provide interoperable services,
  - make data searchable in a common way
- Currently used Open Standards
  - OGC Services, NetCDF format, CF name conventions, ISO 19139, Dublin Core, …

Thank You