AP4 Datenqualitätsprüfung

WP 4.1 QA-Werkzeug zur Prüfung der Daten-konformität

WP 4.2 Unterstützung bei der lokalen Benutzung des QA-Werkzeugs

WP 4.3 Werkzeug von WP 4.1 im Internet (Spot-check-Dienst)

Quality Assurance Tool: QA-DKRZ

H.-D. Hollweg

DKRZ, hollweg@dkrz.de

Overview

- **QA-DKRZ Tool**
  - Work-flow
  - Dependencies
  - Usage
- **Annotation Model**
  - Structure of Results: Files and directories
  - YAML formatted log-file output
  - JSON formatted summary
- **QA-DKRZ: status**

Purpose:

Assure that every file entering ESGF/CERA complies to conventions and project rules.

If not, then issue annotations.

Supported Projects:

- CMIP5/6, CORDEX, HAPPI.
- User-defined deviations (option PROJECT_AS).
- CF Conventions.
- Data-check without meta-data.

Work-flow

General procedure

- **download sources/binaries**
- **installation | update: QA_DKRZ/install**
  note: automatic update by option --auto
- **operation** (1. and 2. alternatively)

  1. **qa-dkrz (based on bash scripts)**
  2. **qa-dkrz.py (based on python scripts)**

Work-flow

- **Installation** (1. and 2. alternatively)

  1. git clone https://github.com/IS-ENES-Data/QA-DKRZ
  2. conda create -n qa-dkrz -c conda-forge -c h-dh qa-dkrz
     - conda create -n cmor -c conda-forge -c pcmdi cmor

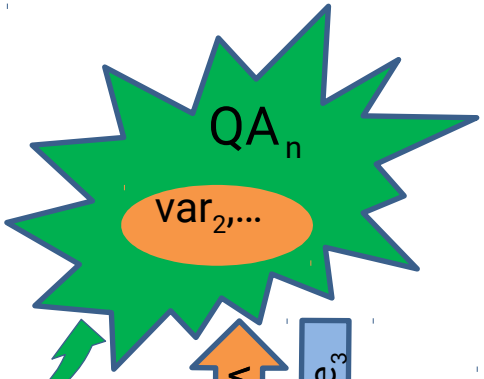
- **Finalisation: QA-DKRZ/install** *[up] PROJECT*
  - download external tables|packages → QA_Tables
  - compile C++ program (only for GitHub installation)
  - automatic/requested update search for tables (if enabled by option)
  - create|update HOME/.qa-dkrz/config.txt

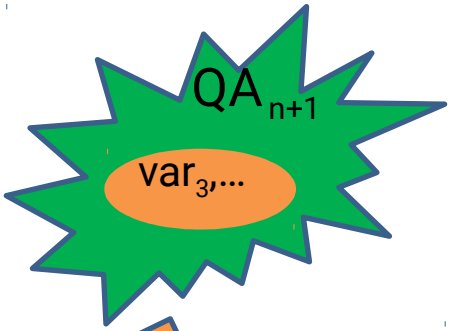
Work-flow

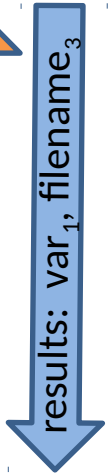
- **Operation: Steady instance (session)**
  - data access
  - check tables and options
  - launch of async. back-ground processes (bash) or multi-threading (python)
  - produce summary
- **QA Check**
  - organise parameters and start C++-exe
  - start PrePARE, ouput → annotation format
  - stream check results to log-files

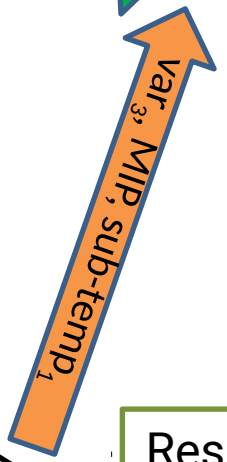
Tables:

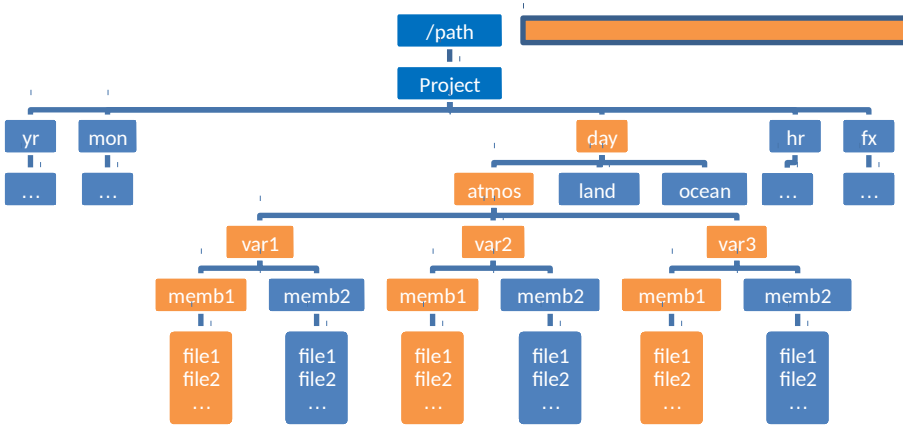
- Conventions
- Check-lists
- CV
- DRS
- Variable Requ.

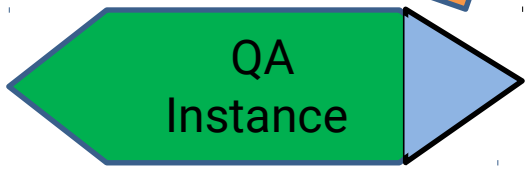








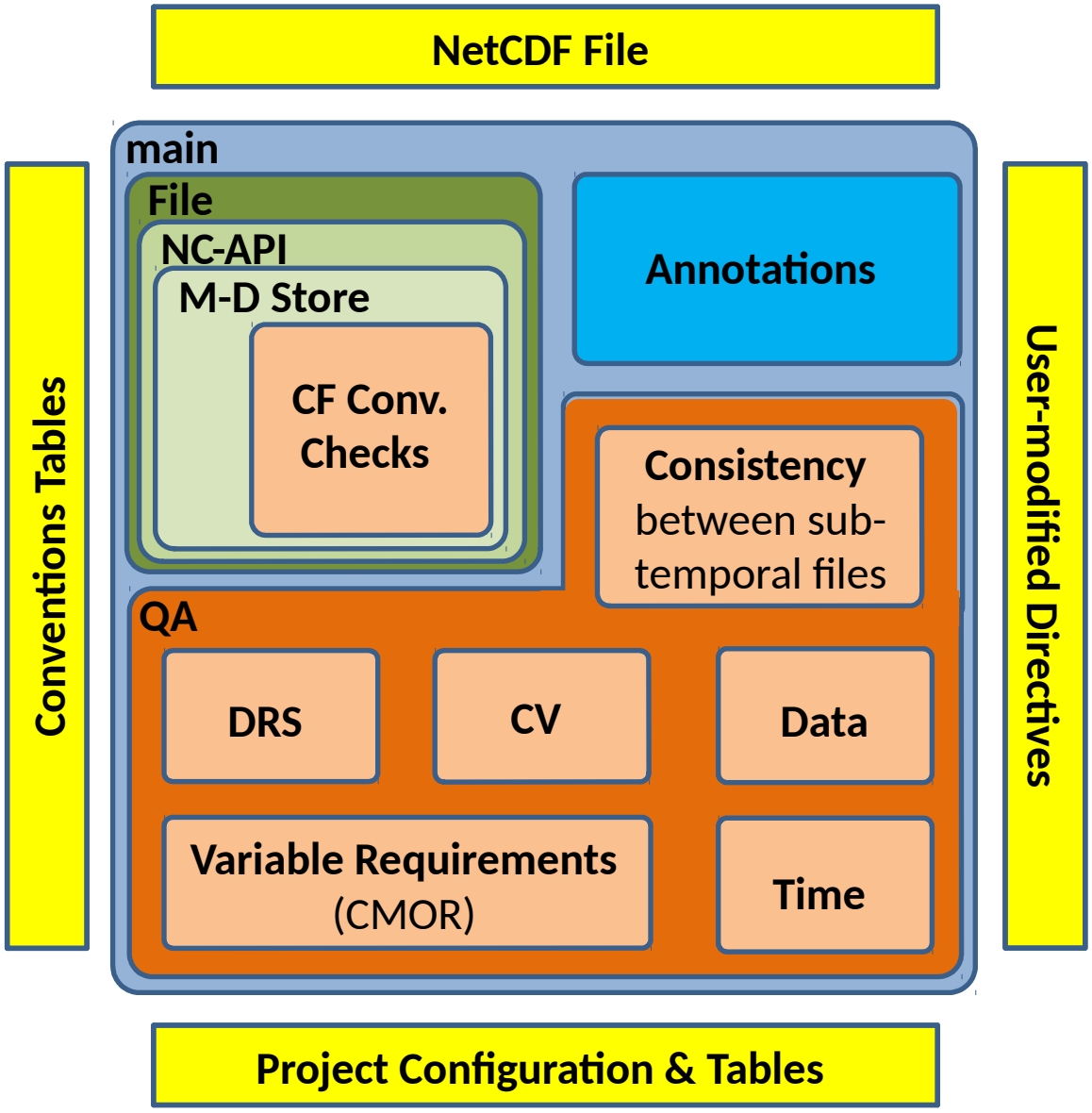




Results:

- log-file
- sum.json
- tag-wise
- atomic Δt

QA Program (C++)



Quality Assurance (QA)

- **Data Reference Syntax (DRS)**
- **Controlled Vocabulary (CV)**
- **Variable Requirements** (CMIP Model Output Requir.)
- **Time Properties**
- **Consistency** between parent - child files (atomic and experiments)
- **Data Checks**
  infinity and not-a-number
  outlier tests
  replicated record detection

Note:
every check may be disabled

Dependencies

Libraries

- zlib www.zlib.net
- hdf5 www.hdfgroup.org/HDF5
- netcdf www.unidata.ucar.edu/netcdf
- udunits2 www.unidata.ucar.edu/software/udunits

Tables

- CF Conv. http://cfconventions.org
- CMIP6_MIP http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/CMIP6_MIP_tables.xlsx
- CMIP6_CV https://github.com/WCRP-CMIP/CMIP6_CVs

Externals

- PrePARE http://cmor.llnl.gov
- xlsx2csv http://github.com/dilshod/xlsx2csv
- Jsoncpp https://github.com/open-source-parsers/jsoncpp

Use-case: mistral (preliminary!)

Provided installation:

install|update: disabled

GitHub Repository:

X=/work/kd0956/sw/QA-DKRZ

1. X/scripts/qa-dkrz [options] [file.nc]
2. X/bin/python X/python/qa-dkrz/qa-dkrz.py [options] [file.nc]

Use-case: mistral (preliminary!)

conda: Y=/work/kd0956/sw/miniconda2

1. source Y/bin/activate qa-dkrz
   qa-dkrz.py [options] [file.nc]
   source deactivate

2. Y/envs/qa-dkrz/bin/qa-dkrz.py [options] [file.nc]
   (note: setting a symbolic link or alias works, too)

3. Y/envs/qa-dkrz/bin/qa-dkrz [options] [file.nc]

Use-case: mistral (preliminary!)

Test case:

X=/work/kd0956/sw/QA-DKRZ

X/scripts/qa-dkrz -P CMIP6 tas_Amon_1pctCO2_MPI-ESM-LR_r1i1p1f2_gn_200601-210012.nc

Operation:

X/scripts/qa-dkrz -f task.txt [options]

Task.txt:

collection of frequently changing options, e.g.

- PROJECT_DATA
- QA_RESULTS
- SELECT
- CHECK_MODE
- PROJECT or alterenatively QA_CONF=CMIP6_qa.conf

Structure of QA-Results: Files and Directories

check_logs (root-directory)

log-files (**files**: DRS-based name.log, YAML)
entry for each checked file; possibly with annotations.

Period (**files**: DRS-based-name.period, YAML)
time range of atomic variables. If too short, then marked.

Annotations (**files**: unique DRS-based-name.json, JSON)
summary extracted from a log-file.

Tags
DRS-based-name (**directories**)
a file for each annotation found in the corresponding log-file.

Structure of QA-Results: Files and Directories

```
check_logs/
├── AFR-44_CNRM_ECMWF-ERAINT_evaluation_r1i1p1_v1.log
│
├── Summary
│   └── AFR-44_CNRM_ECMWF-ERAINT_evaluation_r1i1p1_v1.json
│
├── Period
│   └── AFR-44_CNRM_ECMWF-ERAINT_evaluation_r1i1p1_v1.period
│
└── Tags
    └── AFR-44_CNRM_ECMWF-ERAINT_evaluation_r1i1p1_v1
        ├── L1_1_1
        ├── L1_1_2
        ├── ...
        ├── L2_SF
        └── time_range.txt
```

4 directories, 27 files

QA-DKRZ

- **Sources: GitHub**

  https://github.com/IS-ENES-Data/QA-DKRZ

- **Binaries**

  conda create -n qa-dkrz -c conda-forge -c h-dh qa-dkrz

- **Spot-Check-Dienst (WPS)**

  https://bovec.dkrz.de (C. Ehbrecht)

- **Documentation: ReadTheDocs.org**

  http://qa-dkrz.readthedocs.io/en/latest

Annotation Model

- Check-list file
- Log-file (YAML)
- Summary (JSON)

Check-list File

Format: [text] & tag [,level] [,task] [,variable] [,constraint]

Brace grouping {}:
Example: given: a,b{v{D(z),x,b=2}},{u,v},w
result: 'a,b,w', 'a,v,x,b=2,w', 'a,b,u,v, w'

Key words of actions: {Ln, D, EM, tag, var, V=value, R=record}

- level: L1 – L4 (warning – emergency stop)
- D: Discard
- tag: Identifier.
- EM: Email notification (EM)
- var: Comma-separated acronyms of variables; directive is only applied to these variable(s).
- value: Constraining value, e.g {tag,D,V=0,var} discards test for variable var only if value=0
- record: Apply to time value(s) r_0 [- r_1]

Examples (from `CORDEX_check-list.conf`):

Height requires units=m & 55_1,L1

every height variable is checked for units [m]

Near-surface height must be 0 - 10m
& 55_2,L1,{D,rlut,rsdt,rsut}

variables discarded from check: rlut, rsdt, rsut

Suspecting replicated records
& R3200,L1{D,sund},{D,V=0,clivi,mrfso,prsn,sftgif}

sund discarded,
clivi ... discarded for records with constant value=0.

Log-file (YAML)

```yaml
---
# Log-file of a QA session started by qa-DKRZ
configuration:
  command-line: -m -f task.CMIP6 -e_check_mode=-CNSTY -e_next
  options:
    APPLY_MAXIMUM_DATE_RANGE:
    ...
    SELECT_VAR_LIST: .*
start:
  date: 2016-12-02T11:23:38
  qa-revision: master-66ca331
items:
  - date: 2016-12-02T11:23:40
    file: tas_Amon_1pctCO2_MPI-ESM-LR_r1i1p1f2_gn_200601-210012.nc
    data_path: /path/CMIP6/CMIP/MPI-M/.../r1i1p1f2/Amon/tas/gn/v20161130
    conclusion: 'CF: FAIL, CV: FAIL, DATA: PASS, DRS(F): PASS, DRS(P): FAIL, TIME: PASS
    checksum:       ce5e24ffeb5c38665a17570f4a564f0e.md5
    creation_date:  2016-12-02T12:40:29Z
    tracking_id:    06cfd581-917a-4888-9b92-a07a726469d0
```

```
events:
  - event:
      caption: 'DRS path: path component member_id=<r1i1p1f2> does not
                match global attribute value <r1i1p1f1>.'
      impact: L1
      tag: '1_2'
  - event:
      caption: 'Attribute institution:
                found <Max Planck Institute for Meteorology>,
                expected from CMIP6_institution_id.json
                <Max Planck Institute for  Meteorology, Hamburg 20146,
                 Germany>.'
      impact: L1
      tag: '2_4'
  - event:
      caption: 'Coordinate variable <height>: No data.'
      impact: L2
      tag: 'CF_0d,
status: 2
```

Summary (JSON)


```
{
   "QA_conclusion": [ PASS | FAIL ]  ",
   "project": "CORDEX",
   "DRS_0": "cordex",
   "DRS_1": "output",
   "DRS_2": "AFR-44",
   ...
   "DRS_8": "v1",
   "DRS_9": "SHARED",
   "DRS_10": "SHARED",
   "annotation":
   [
      {
         "DRS_9": ["day", "mon"],
         "DRS_10": ["tauv",“tauu“],
         "caption": "DRS CV path: global attribute RCMModelName = <QWER> vs. <ASDF>.",
         "severity": "L2"
      },
```


… to be continued

Summary (JSON)

Continuation with inclusion of PrePARE output ...


```
      {
         "caption": "CMOR Warning: Your input attribute institution >Max Planck Institute for Meteorology< will be replaced with >Max Planck Institute for Meteorology, Hamburg 20146, Germany< as defined in your Control Vocabulary file.
      }
   ]
}
```

QA-DKRZ: status

		CMIP5	CORDEX	CMIP6	Comment
Conv	CF	v1.4	v1.4	v1.7	www.cfconventions.org
	UGRID	-	-	v1.0	ugrid-conventions.github.io
DRS	(Path)				
	(File)				
CV		[1]			[1]) CMOR guide → machine read.
Var. Requir.				[2]	[2]) CMIP6_MIP_tables.xlsx
Consistency					across atomic var. & experiment
Time					
Data					NaN, Inf, replications, outlier
CMOR		-	-	PrePARE	http://cmor.llnl.gov
WPS					Co-work: C. Ehbrecht
OpenDAP					