QA-DKRZ:
The Annotation Model

H.-D. Hollweg, DKRZ, hollweg@dkrz.de

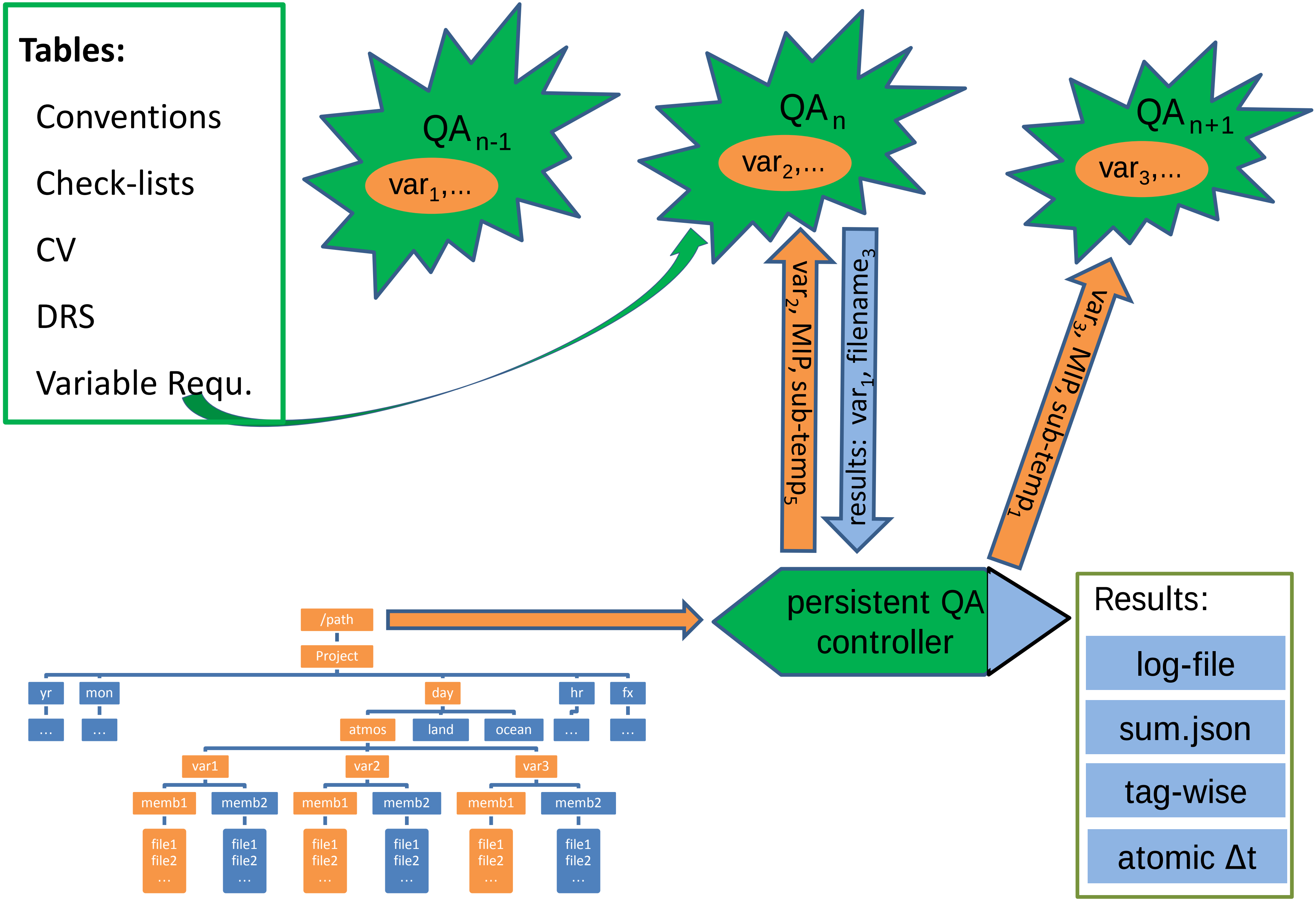
Overview

- **QA-DKRZ Tool**
  - Work-flow
  - Dependencies
- **Annotation Model**
  - Specification of actions tagged to checks
  - Structure of Result Files and Directories
  - YAML formatted log-file output
  - JSON formatted summary
- **QA-DKRZ: status**

Purpose:

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

If not, then issue annotations.



Tables:

Conventions

Check-lists

CV

DRS

Variable Requ.

QA$_{n-1}$ var$_1$,...

QA$_n$ var$_2$,...

QA$_{n+1}$ var$_3$,...

var$_2$, MIP, sub-temp$_5$

results: var$_1$, filename$_3$

var$_3$, MIP, sub-temp$_1$

persistent QA controller

Results:

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

/path

Project

yr mon day hr fx

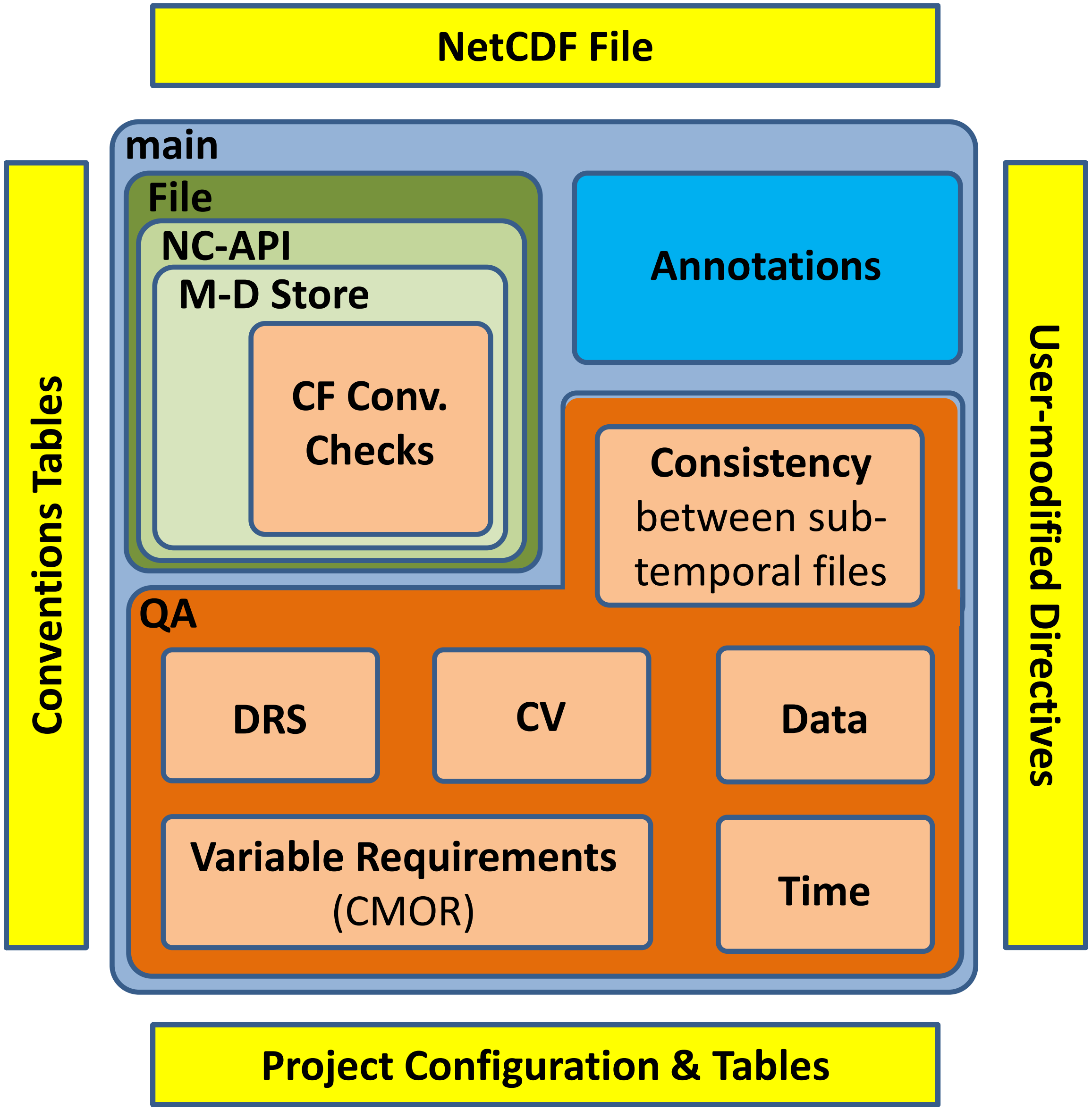
… … atmos land ocean … …

var1 var2 var3

memb1 memb2 memb1 memb2 memb1 memb2

file1 file2 … file1 file2 … file1 file2 … file1 file2 … file1 file2 … file1 file2 …

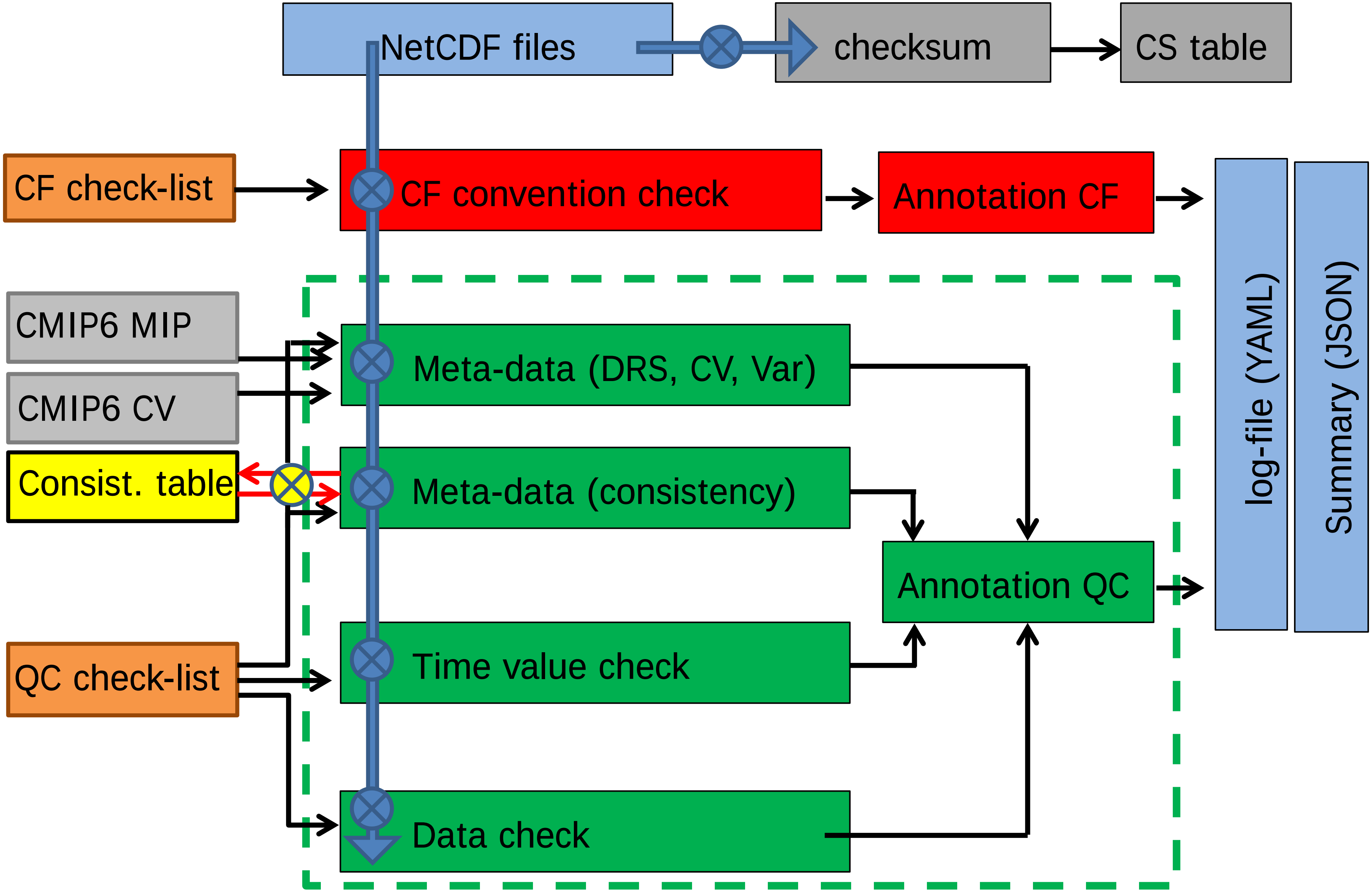
QA Program (C++)



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


NetCDF files
checksum
CS table
CF check-list
CF convention check
Annotation CF
CMIP6 MIP
CMIP6 CV
Consist. table
QC check-list
Meta-data (DRS, CV, Var)
Meta-data (consistency)
Time value check
Data check
Annotation QC
log-file (YAML)
Summary (JSON)

Libraries

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

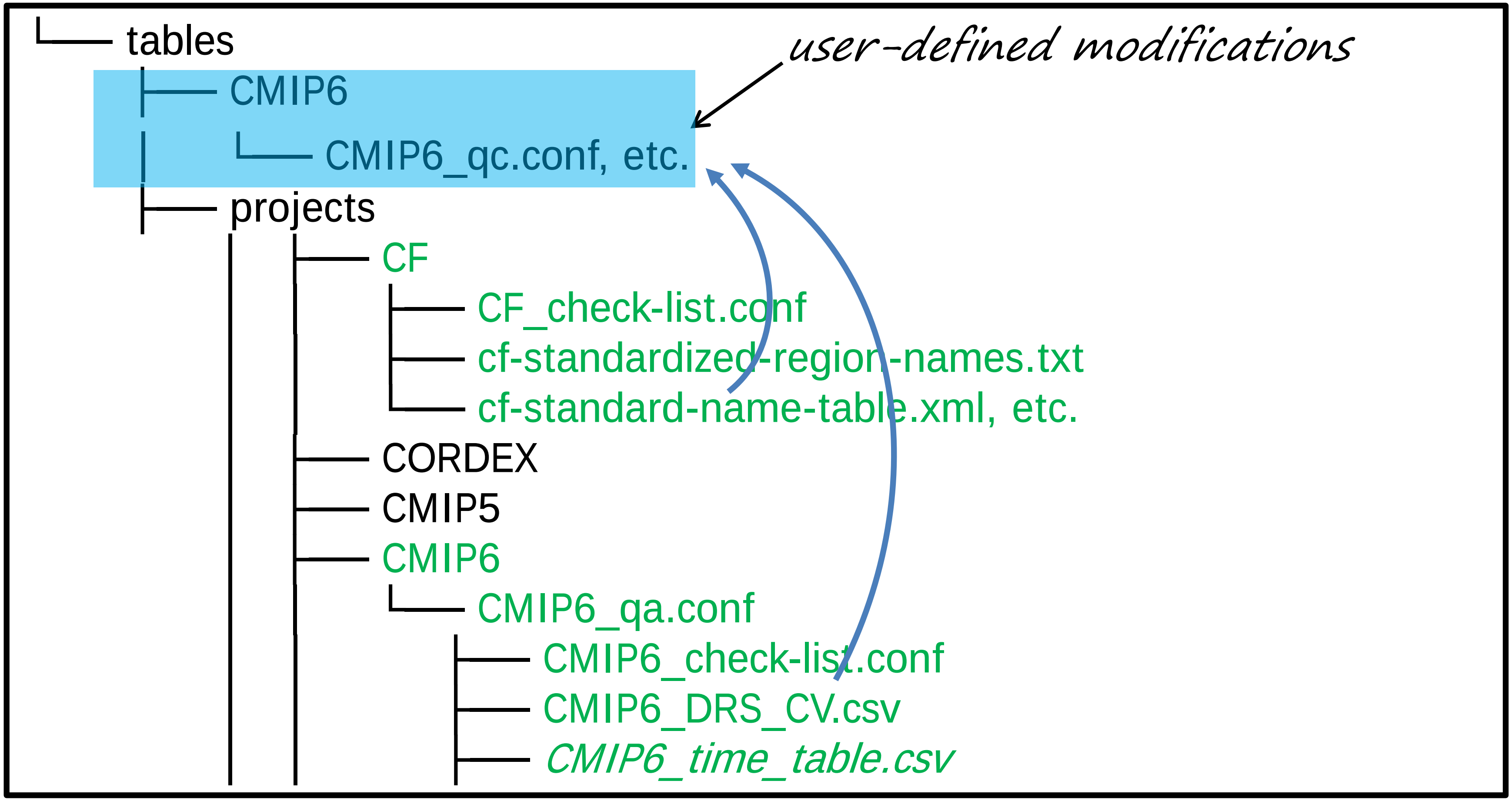
Tables

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

xlsx2csv	http://github.com/dilshod/xlsx2csv
jsoncpp	https://github.com/open-source-parsers/jsoncpp

Path: /home/user/.qa-dkrz



```
└── tables
    ├── CMIP6
    |    └── CMIP6_qc.conf, etc.
    ├── projects
    |    |    ├── CF
    |    |    |    ├── CF_check-list.conf
    |    |    |    ├── cf-standardized-region-names.txt
    |    |    |    └── cf-standard-name-table.xml, etc.
    |    |    ├── CORDEX
    |    |    ├── CMIP5
    |    |    ├── CMIP6
    |    |    |    └── CMIP6_qa.conf
    |    |    |         ├── CMIP6_check-list.conf
    |    |    |         ├── CMIP6_DRS_CV.csv
    |    |    |         ├── CMIP6_time_table.csv
```

user-defined modifications

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.

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

Tags

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

QA-DKRZ

- **Sources: GitHub**

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

- **Binaries**

  ```
  conda install -c birdhouse -c conda-forge qa-dkrz
  ```

  ehbrecht@dkrz.de

- **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)

```
---
# 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: L2
       tag: '2_4'
  - event:
       caption: 'Coordinate variable <height>: No data.'
       impact: L1
       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"],
         "caption": "DRS CV path: global attribute RCMModelName = <QWER> vs. <ASDF>.",
         "severity": "x"
      }
  ]
}
```

QA-DKRZ: status

		CMIP5	CORDEX	CMIP6	Comment
Conv	CF	🟩	🟩	🟩	version 1.4 – 1.7draft
	UGRID	-	-	🟥	
DRS	(Path)	🟩	🟩	🟩	
	(File)	🟩	🟩	🟩	
CV		1) 🟩	🟩	🟨	1) CMOR guide → machine read.
Var. Requir.		🟩	🟩	🟨	xlsx → csv table
Consistency		🟩	🟩	🟨	files across atomic & exp. scope
Time		🟩	🟩	🟨	
Data		🟩	🟩	🟩	
CMOR Run		-	-	🟨	expects provided CMOR instance
WPS		🟨	🟨	🟨	
OpenDAP		🟥	🟥	🟥	

QA for CMIP6 files before entering ESGF

- Check (only) DRS of paths
- Running CMIP6Validator in QA-DKRZ

QA-DKRZ: DRS Check

- event:

  capt: DRS path: path component member_id=<r1i1p1f2> does not match global attribute value <r1i1p1f1>.

  impact: L1

  tag: 1_2

CMIP6Validator Run:

- #! /bin/bash

- export PATH=/hdh/local/anaconda2/bin:${PATH}
- export UDUNITS2_XML_PATH=/hdh/local/anaconda2/share/udunits/udunits2.xml
- source activate env

- d1=/hdh/hdh/CMOR/cmip6-cmor-tables/Tables/CMIP6_Amon.json
- d2=/data/CMIP6/CMIP/MPI-M/MPI-ESM-LR/1pctCO2/r1i1p1f2/Amon/tas/gn/v20161130/tas_Amon_1pctCO2_MPI-ESM-LR_r1i1p1f2_gn_200601-210012.nc

- d3=cmor_out_tas.out2

- python /hdh/local/anaconda2/envs/env/lib/python2.7/site-packages/cmip6_cv/**CMIP6Validator.py** $d1 $d2

- Traceback:
- ! In function: cmor_get_cur_dataset_attribute

- !!!!!!!!!!!!!!!!!!!!!!!!!
- !
- ! Error: Dataset: current dataset does not have attribute : _AXIS_ENTRY_FILE
- !
- !!!!!!!!!!!!!!!!!!!!!!!!!⍰[0m

- Traceback:
- ! In function: cmor_get_cur_dataset_attribute
- !!!!!!!!!!!!!!!!!!!!!!!!!
- !
- ! Error: Dataset: current dataset does not have attribute : _FORMULA_VAR_FILE
- !
- !!!!!!!!!!!!!!!!!!!!!!!!!⍰[0m

- Traceback:
- ! In function: cmor_load_table_internal

- ! Error: Dataset: current dataset does not have attribute : _AXIS_ENTRY_FILE

- ! Error: Dataset: current dataset does not have attribute : _FORMULA_VAR_FILE

- ! Error: Could not find file: /hdh/hdh/CMOR/cmip6-cmor-tables/Tables/cur_dataset_attribute

- ! Error: Could not find file: /hdh/hdh/CMOR/cmip6-cmor-tables/**Tables/ibute**

- ! Error: Reading table Amon: axis name: 'time' for variable: 'ccb' is not defined in table. Table defines dimensions: 'longitude latitude time' for this variable

- ! Error: Reading table Amon: axis name: 'time' for variable: 'cct' is not defined in table. Table defines dimensions: 'longitude latitude time' for this variable

- ....