DKRZ Data Management Report: Quality Maturity Matrix Checklist for Levels 4 and 5 with Protocols

Application example at long-term archive of World Data Center for Climate (WDCC) at DKRZ

Revision	Authors	Scope
Mar-2019	Heinke Höck, Frank Toussaint	Workflow check

Contents

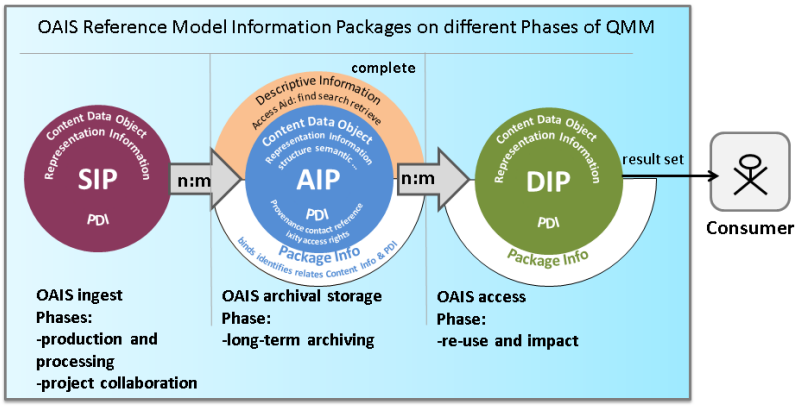
1. Why use Maturity Matrix for Data and Metadata Quality Assessment at DKRZ-LTA

'The term "maturity" relates to the degree of formality and optimization of processes, from *ad hoc* practices, to formally defined steps, to managed result metrics, to active optimization of the processes.' (wikipedia Capability Maturity Model)

The implementation of the QMM should lead to level 4 data and metadata quality level at LTA-WDCC (Long-Term-Archive at the World Data Center for Climate) (figure 1.)

Optimization of processes means to establish a system of checking critical known errors at the best time to avoid unnecessary work, e.g. format validation should take place before long-term archiving of the data.

The accuracy criteria, which contains the provision of documents for the evaluation of the data, does not have to comply with standardisation formalities. These documents can also be provided as cross-reference entries as part of the metadata. All other criteria use the OAIS standard (CCSDS 2012) to define the data objects, see figure below:



2. QMM level at DKRZ-LTA

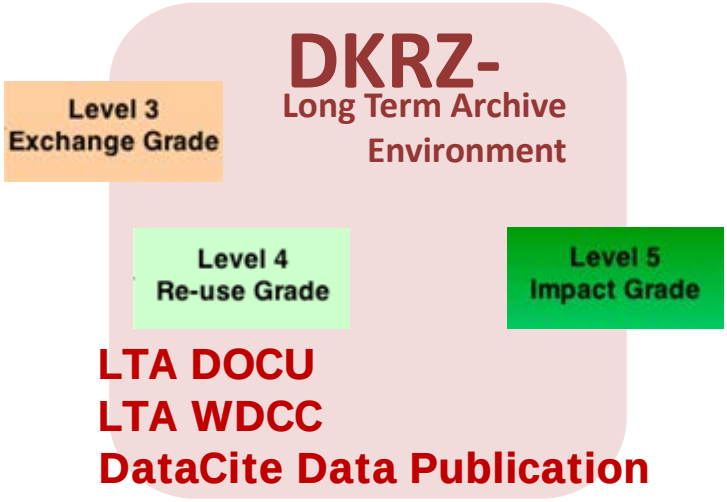


Figure 1: Three QMM grades in the DKRZ-LTA environment[1]

The DKRZ-LTA ingest Workflow is described with figure 2. The two possibilities for step 3 are:

3a) Metadata ingest sources are the Project Content Object Store in the case of self-describing SIPs e.g. ESGF and the Project Metadata Repository.

3b) Metadata ingest source is the metadata insert GUI - MetaXA[2]

In the following tables we use these three colours to mark the description of checks:

Metadata Workflow (WF) steps: WF3b MetaXa GUI, WF4 cera2_temp and WF5 upload into CERA2 (figure 2)

This checklist deals with checks carried out before, during and immediately after data is inserted into the CERA database WF8 (figure 2). The checks are performed both on the files in the file system and on the metadata.

Task list procedure[3]

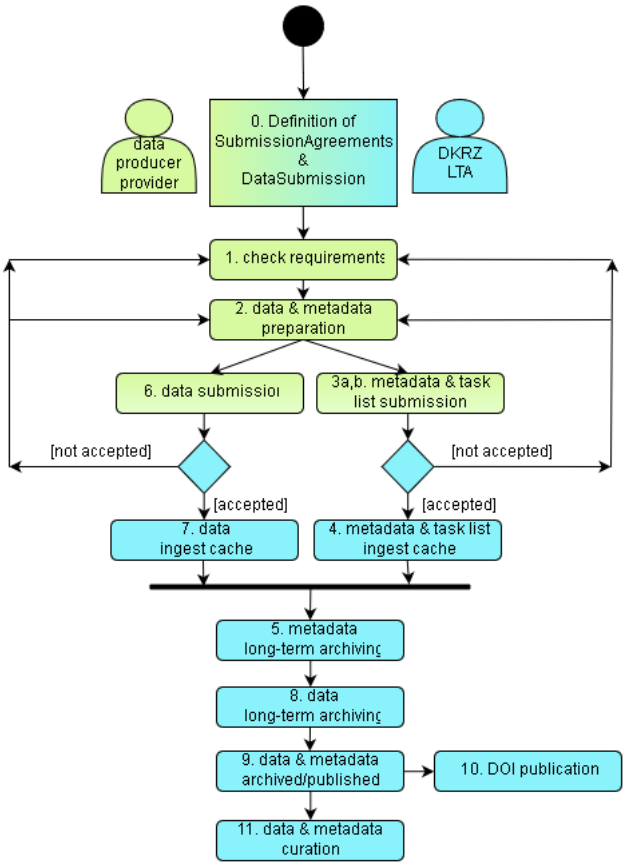


Figure 2: DKRZ-LTA Submission WF

3. QMM 4 with Checklist Version 15/05/2019

		Level 4	Checklist	Implementation with check protocol, WF description and link to documentation of rules – sufficient but not necessary
Consistency	Data Organisation and Data Object	***data organisation is structured/conform to***		
		well-defined rule e.g. discipline-specific standards and long-term archive requirements (OAIS Package Info - binds)	✓ 1) In what way are well-defined rules about data organization documented?	Definition of Submission Agreements: Protocols https://madwiki.dkrz.de/CERA%20Contents project name:<> Link to discipline-specific standards or data management plan Request Tracker: https://dm-rt.dkrz.de/ RT id:<> Paper published online: *http://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf* *CERA2 documentation:* *:https://www.dkrz.de/daten-en/cera/cera?set_language=en&cl=en* *Technical Report* 'The CERA-2 Data Model' doi:10.2312/WDCC/DKRZ_Report_No15 Cera Hierarchical Rules for WF11: https://madwiki.dkrz.de/CERA_Guidance
			✓ 2) In what way is the data organization implementation documented?	Workflows documented in paper published online: *http://cera-www.dkrz.de/docs/DataSubmissionPreparationGuide.pdf* http://cera-www.dkrz.de/docs/Archiving_Task_List.pdf
			✓ 3) Is the data organization structured in a consis manner in accordance with the data granularity and are files stored (e.g.in correct directory) in a consistent manner?	Comment: Mapping of project data structure in accordance with data granularity to the CERA2 structure is correct. The data are connected to the correct entries see figure 3.
			✓ 4) Is the data organization named in a consistent and documented manner?	WF4: Labelling are checked in cera2_temp (Block entry_connect and entry.entry_names) Comment: Only the structure parts of labelling (entry names) are affected. WF10:Entry_names and entry_acronym will be checked.
			✓ 5) Are AIPs structured in a consistent and documented manner?	Comment: Description of AIP structure storage with workflows. **PDI**: see Figure 4 for PDI structure in CERA2. **Content Data Object**: see CERA2 module data_access. If the data structure is specified in the data management plan of the project, this has to be checked e.g. time series for every parameter. **Representation Information**: see CERA2 module data_org, block spatial reference (structure) and block parameter and entry (semantic) **Descriptive Information**: see CERA2 blocks key_connect, campaign, entry, parameter and coverage (find, search) and CERA2 module data_access (retrieve). **Package Info**: see block entry_connect (binds), entry, reference (identifies) and module data_access (relates Content Info and PDI). The AIPs (figure 5.) are structured consistent in CERA2. WF4 and WF5 guaranteed with ingest cache cera2_temp and upload into cera2 consistency of the AIP metadata structure. WF7: Check of Content Data Object structure.

		✓ 6) Are files structured in a consistent manner to the well-defined rules and the Representation Information of the AIPs.	Comment: well defined rules are documented in the data header (e.g. netCDF CF CMOR). Checks are done at WF 2.
		✓ 7) Are access files DIPs structured in a consistent and documented manner?	Documentation of data access see: http://cera-www.dkrz.de/docs/FAQ/CERA-UI.html The file format remains unchanged.
		✓ 8) Is your data dynamic in time?	Assignment Progress='completely archived, will be continued' and *'dynamic'. https://madwiki.dkrz.de/CERA/CERAWorkflows?action=AttachFile&do=view&target=workflow_progress_g1_v2.jpg
		✓ 9) If it is required to follow a specific sequence in archiving this has to indicated as well.	WF4: Assignment in task list http://cera-www.dkrz.de/docs/Archiving_Task_List.pdf see figure 3.
	data objects (OAIS) are		
	AIPs **conform to well-defined rules e.g. discipline-specific standards and long-term archive requirements**	✓ 10) In what way are discipline-specific standards and long-term archive requirements for AIPs and DIPs documented?	Comment: Documentation of AIP and DIP requirement. Paper published online: *http://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf* *CERA2 documentation:*https://www.dkrz.de/daten-en/cera/cera?set_language=en&cl=en *Technical Report* 'The CERA-2 Data Model' doi:10.2312/WDCC/DKRZ_Report_No15 MetaXa Tutorial *cera-www.dkrz.de/LTA_metadata Help/Introduction* Indicated in task list: http://cera-www.dkrz.de/docs/Archiving_Task_List.pdf
		✓ 11) Are AIPs and DIPs consistent to well defined rules e.g. labelling? o Are access files DIPs named in a consistent and documented manner? Comment: Special characters and spaces in labelling should be avoided.	WF3a: dataset names and acronyms are constructed with repository entry names. WF3b: dataset names and acronyms are constructed with filenames. WF4: Check of dataset names and acronyms in cera2_temp by WDCC reviewer. WF10: Check of experiment and ds_group labelling.(name, acronym, title) DIPs are labelled with entry_acronyms.
	DIPs datasets are self-describing	✓ 12) Is the data self-describing data objects which meet the discipline-specific standards?	Correct for data with netCDF CF und grib standard format.
	data formats – Content Data Object (OAIS)		
	conform to well-defined rules e.g. discipline-specific standards and long-term archive requirements	✓ 13) In what way are the requirements for data formats conform to discipline-specific standards documented?	Paper published online: http://cera-www.dkrz.de/docs/DKRZ-LTA-Formats.pdf
		✓ 14) Are the formats correct (checked with format checker)?	Comment: WF2 User support of test files checking or support for converting to required formats e.g. format checker provision. http://cfconventions.org/compliance-checker.html WF5: Support with check in data ingest cache with format checker. After copy of cera2_temp into cera2 and with container infos WF8 Does each file have the expected file format? Comparison between expected format and the file command on console (format known by previous knowledge/or stored in the DB). File command checks the magic number of the file. Supported formats: NetCDF, GRIB, ASCII, TAR. GZ File Format is used from cera2. distrubution for the adjustment.. WF8: final For NetCDF and GRIB: Calculation of a header (For NetCDF: each file, for GRIB: The first record of the container). Calculation with cdo sinfo (grib) and ncdump -h (NetCDF). Implicit check if the files are valid GRIB/NetCDF files. Calculated header is stored in cera2. headerinfo. WF 10:Quality.Specification: [Technical Quality Assurance: The format is verified and correct][done]

		✓ 15) Are data header and AIP consistent?	The header itself is part of the metadata in some cases. WF10: Check of standard and long_name in parameter against header.
	data sizes are consistent **file extensions are consistent**	✓ 16) In what way are the requirements for file extensions documented?	DKRZ-LTA requirements: http://cera-www.dkrz.de/docs/DKRZ-LTA-Formats.pdf Project requirement documentation: https://madwiki.dkrz.de/CERA%20Contents
		✓ 17) Are the file extensions correct?	WF8: Check of file extensions consistency to block distribution. Documented in cera2_adm.slave_eventhistory.
		✓ 18) Are the data sizes checked and correct, the size of data set is not equal 0 *if feasible*?	WF8: data size calculation for table distribution is part of the data archiving process with bit stream preservation. WF10: The data sizes are controlled and correct o Quality.Specification: [Technical Quality Assurance: The data size is checked and correct][done] Test: @calc_size Size of every data set is > 0 o Quality.Specification:[Technical Quality Assurance: The size of every data set is not equal 0][done] Test: @data_size_0_test
		✓ 19) Are the file sizes feasible for access and transmission?	WF7 data file sizes < 2Giga see *http://cera-www.dkrz.de/docs/DataSubmissionPreparationGuide.pdf* The size of the data access can be reduced by selection of time period and by the usability of cdo operators with Jblob https://cera-www.dkrz.de/WDCC/ui/cerasearch/info?site=jblob
		✓ 20) Bit stream preservation is checked with checksum and date.	WF5,7: after copy of cera2_temp into cera2 Before insert data (requires entry in cera2. external_pointer) a. Does each target in External_Pointer have a corresponding file in the file system? b. Does each file in the file system have a corresponding target in External_Pointer? c. Does the size specified in External_Pointer match that of the real file? d. Is every file in the same directory? WF8: final after insert data: e. Does each file have the expected checksum? At this point in time, the original checksum is available in the Cera_Meta table. The entry there originates either from external_pointer or was calculated before insert data. Result is stored in cera_check. checks_performed After MD5-Checks (manuell): Does an entry for each record of the datasets exist in cera_check.checks_performed?
Versioning and Controlled Vocabularies (CVs)	***versioning follows/is***		
	systematic collection including documentation of enhancement conform to well-defined rules e.g. discipline-specific standards **old versions stored if feasible**	✓ 21) In what way are the long-term archive requirements or discipline-specific standards for versioning documented? Comment: *If no explicit versioning exists an implicit versioning could be established. But a good way of versioning implementation is the identification with the date of storage. As soon as more than one version exists versioning should be implemented.*	If no explicit versioning exists an implicit versioning will be established. As soon as more than one version exists versioning will be implemented in contact with data provider. project dependent documented <project_name>

		✓ 22) Is a versioning available if feasible?	<yes/no>
		✓ 23) How is the systematic versioning on AIPs , access files DIPs or data collection implemented and applied if feasible?	If new version is filled in then addendum v2 and entry_name, -acronym with reciprocal reference e. g. CMIP5 with date indication Systematic versioning is implemented and applied with the entry_names and entry_acronym.
		✓ 24) Do you labelled your data consistent with the versioning	WF4, 5: copy of old version metadata and check in cera2_temp by WDCC reviewer. WF10: check of versioning
		✓ *25) How is storage of old versioning implemented?*	WF8: DKRZ Storage Policy http://cera-www.dkrz.de/docs/DKRZ-LTA-PreservationAndStoragePolicy.pdf New entries will be created. No deleting of old versions.
	data labelled with CVs conform to		
	discipline-specific standards	✓ 26) In what way is the controlled vocabulary documented – conform to discipline-specific standards? Comment: e.g. cf standard names	project dependent documented <project_name> WF4: table topic with [cf-standard name] and add_infos if feasible and project requirements defined by submission agreements.
		✓ 27) Is a controlled vocabulary available if feasible?	WF4,5: <yes/no>
		✓ 28) Ensure you labelled your data consistent with the discipline-specific standard controlled vocabulary. The labeling of the AIPs is consistent with the controlled vocabulary (e.g. cf-standard names).	WF3a: dataset names and acronyms are constructed with community repository entry names. WF3b: dataset names and acronyms are constructed with filenames. WF4: checks in cera2_temp by WDCC reviewer
		✓ 29) Are labeling of the AIPs or internal identifiers with mapping to objects correct in accordance to CVs?	In cera2 the internal identifiers are entry_ids, entry_acronyms and entry_names. They are primary keys in the cera2 tables and connected to the CV with cf-standard names or topics. WF10: Check of CVs.
		✓ 30) Are naming conventions for discovery e.g. CVs correct?	WF4: netCDF CF standard names are used for discovery and access if applicable: https://cera-www.dkrz.de/WDCC/ui/cerasearch/topics WF10: Check of CVs.
Data-Metadata Consistency	***OAIS metadata components are consistent***		
	Complete PDI * **Provenance** **Context** **Reference - cross** **Fixity** **Access Rights** **and** **Representation Information** **Descriptive Information** **Package Info** ***maintenance and storage policy are not affected, since they belong to the repository certification.**	✓ 31) PDI are consistent/correct o **Provenance** ▪ data source e.g. sensor ▪ publisher if feasible ▪ detailed description of data production steps and method - quality assurance procedure (approval and review) is consistent ▪ contributor(s) if feasible –	▪ data source e.g. sensor Review DOI ▪ publisher if feasible WDCC: standard citation ▪ detailed description of data production steps and method Review DOI Project Summary, Experiment Summary, DS_group Summary and Quality. Inquirement of Add Infos (pdf) or reference to: model data methodology report or observational methodology report and data level classification and quality checking. or references ▪ contributor(s) if feasible – contact Required Investigator and metadata beim metadata upload, WF10: DOI review
		o **Context** ▪ project and experiment description	▪ project and experiment description WF10: DOI review
		o **Reference** ▪ data citation – e.g. creators ▪ contact	▪ data citation – e.g. creators WF10: DOI review
		o **Fixity** ▪ data expiration date	▪ data expiration date Creation date + 10 years minimum
		o **Access Rights** ▪ access constraint	▪ access constraint -access constraint WF10: access check by jblob download for special user

			✓ 32) Representation Information is consistent	Comment: see CERA2 module data_org, block spatial reference (structure) and block parameter and entry (semantic)
			✓ 33) Descriptive Information is consistent. ▪ metadata for search and discovery e.g. keywords	Comment: metadata for search, find and discovery e.g. keywords **is consistent** see CERA2 blocks key_connect, campaign, entry, parameter and coverage (find, search) and CERA2 module data_access (retrieve). WF5: After copy of cera2_temp into cera2 During Metadata Filling o [optional, currently not supported for large projects]: Determine start time, Min/Max/Mean values and NoOfTimesteps and write them into the CeraMeta table. Implicit test if files are valid. Values are determined with the UCAR GRIB1/2 libraries (for Grib) or the CDOs (for NetCDF) WF10:The spatial-temporal coverage description (metadata) is consistent to the data, time steps are correct and the time coordinate is continuous o Quality.Specification: [Technical Quality Assurance: The time description (metadata) and data are consistent][done start date, stop date checked] Test: cdo showdate, ncdump –v time
			✓ 34) Package Info is consistent	Comment: see block entry_connect (binds), entry, reference (identifies) and module data_access (relates Content Info and PDI). Required structure see figure 6. are fulfilled.
Completeness	Existence of Data (Completeness and Persistence)	**data entities (conform to discipline-specific standards) are complete dynamic datasets - data stream are not affected number of datasets (aggregation) is consistent data are persistent, as long as expiration date requires**	✓ 35) Do data sets exist - complete?	WF4: task list definition of completeness. WF8: Checks
			✓ 36) The data is persistent as long as expiration date (creation date plus minimum 10 years)	Storage policy: https://cera-www.dkrz.de/docs/DKRZ-LTA-PreservationAndStoragePolicy.pdf
			✓ 37) How is deleting and overwriting prevented?	WDCC workflow. It is not allowed to overwrite or delete data. Only a few persons at WDCC are allowed to do this.
			✓ 38) Is the number of data sets (aggregation) checked against the customer task list or project requirements?	WF8: Checks WF10: Number of data sets is correct and > 0 o Quality.Specification: [Technical Quality Assurance: The number of data sets is checked and not equal 0][done] Test: @entry_type_nDS_test
			✓ 39) The data is persistent as long as expiration date (creation date plus minimum 10 years) requires	WF9: completely archived documentation: https://madwiki.dkrz.de/CERA/CERAWorkflows
	Existence of Metadata	***OAIS metadata components exist***		
		Complete PDI * **Provenance** **Context** **Reference** **Fixity** **Access Rights** **and** **Representation Information** **Descriptive Information** **Package Info** ***maintenance and storage policy are not affected, since they belong to the repository certification.**	✓ 40) PDI exist o **Provenance** ▪ data source e.g. sensor ▪ publisher if feasible ▪ detailed description of data production steps and method - quality assurance procedure (approval and review) ▪ contributor(s) if feasible – contact	WF3b: metadata submission interface WF4: contact check in cera2_temp by WDCC reviewer WF10: Inquirement of detailed description of data production steps and method - quality assurance procedure (approval and review) and check
			o **Context** ▪ project and experiment description	WF3b: metadata submission interface WF4: entry, campaign check in cera2_temp by WDCC reviewer
			o **Reference** ▪ data citation – complete	WF3b: metadata submission interface WF4: reference standard citation check in cera2_temp by WDCC reviewer
			o **Fixity** ▪ data expiration date	Publication date + 10 years. There is no automatic deletion.
			o **Access Rights** ▪ access constraint	WF3b: metadata submission interface and WF 9. WF4: distribution check in cera2_temp by WDCC reviewer

			✓ 41) Representation Information exists	WF4: Metadata CERA2 module data_org, block spatial reference (structure) and block parameter and entry (semantic) exists
			✓ 42) Descriptive Information exists: ▪ metadata for search and discovery e.g. keywords	WF3b: metadata submission interface WF4: Metadata for search, find and discovery e.g. keywords **is consistent** see CERA2 blocks key_connect, campaign, entry, parameter and coverage (find, search) and CERA2 module data_access (retrieve) exist.
			✓ 43) Package Info exists	WF4: Metadata block entry_connect (binds), entry, reference (identifies) and module data_access (relates Content Info and PDI) exist.
			✓ 44) How is cross - reference update implemented?	MetaXA and part of DOI process
			✓ 45) How is citation persistency implemented?	Only a few WDCC persons have permit to update citation.
Accessibility	Data Access by Identifier	**data is accessible by**		
		permanent identifier (expiration is documented) (OAIS Package Info - identifies) **datasets have an expiration date and are accessible for at least 10 years (conform to rules of good scientific practice)**	✓ 46) Are complete data accessible by identifier for at least 10 years?	WF10: The data is accessible by Lobster, jblob and http://cera-www.dkrz.de/WDCC/ui/EntryList.jsp?acronym=<entry_acronym Storage at least 10 years: http://cera-www.dkrz.de/docs/DKRZ-LTA-PreservationAndStoragePolicy.pdf WF8: final After archiving: o o Can each file of the data set be retrieved from the archive? (NOTE: NOT JBlob! direct lobster connection) WF10: The data sets are accessible Quality.Specification: [Technical Quality Assurance: The data sets are all accessible via internet][done] Test: jblob, compact page, sqlplus select
			✓ 47) Are expiration dates of data available?	WF5: Creation Date + 10 years
		checksums are correct and accessible	✓ 48) Are *correct* checksums accessible? Comment: Checksums are created and checked	WF8: checksum creation at DKRZ-LTA
		a bijective mapping between identifier and datasets is documented e.g. in data header (OAIS Package Info - binds, identifies)	✓ 49) Does a bijective mapping between identifier and datasets exist?	Bijective mapping to objects is implemented with the CERA2 core schema and the data_access module and stored in the CERA database.
	Metadata Access by Identifier	**metadata is accessible by**		
		by permanent identifier (expiration is documented) (OAIS Package Info - identifies) **complete data citation is persistent**	✓ 50) PDI are accessible by identifier o **Provenance** ▪ data source e.g. sensor ▪ publisher if feasible ▪ detailed description of data production steps and method – quality assurance procedure (approval and review) ▪ contributor(s) if feasible – contact	WF5: The metadata is accessible by http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=<entry_acronym> WF10:The metadata are accessible Quality.Specification: [Technical Quality Assurance: The metadata are all accessible via internet][done] Test: ULR<entry_acronym

			✓ 51) PDI are accessible by identifier o **Context** ▪ project and experiment description o **Reference** ▪ data citation – e.g. creators o **Fixity** ▪ data expiration date o **Access Rights** ▪ access constraint	WF5: The metadata is accessible by http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=<entry_acronym> WF10: The metadata are accessible Quality.Specification: [Technical Quality Assurance: The metadata are all accessible via internet][done] Test: ULR<entry_acronym
			✓ 52) Representation, Descriptive Information and Package Info are accessible by identifier ▪ metadata for search and discovery e.g. keywords	
			✓ 53) Is complete data citation persistent?	WF10: The citation is persistent as long as expiration date (creation date plus minimum 10 years) requires. http://cera-www.dkrz.de/docs/DKRZ-LTA-PreservationAndStoragePolicy.pdf
		a mapping between data access identifier and metadata access identifier is implemented (OAIS Package Info relates Content Info and PDI)	✓ 54) Is the mapping between data access identifier and metadata access identifier accessible?	WF5: After copy of cera2_temp into cera2 Before insert (required entry in cera2.external_pointer) a. (Only for CMIP5!) Is the entry in CERA2. entry. entry_name (as file system path) the prefix of all targets in External_Pointer for this data set? b. cera2. parameter have a dummy entry to specify the meta table in cera2. parameter? If not, one is created.. WF8: Bijective mapping to metadata objects is implemented with the CERA2 entry name identifiers and the module data_access via table parameter. Case 3a) The task list defines the mapping between the entry_acronym and the content data object, which is then transferred to CERA. The Entry_acronym is the unique identifier that guarantees the bijective mapping between data object and metadata with the cera2 table parameter. For metadata references that are not located in cera2, this is done via Identifier block reference.
Accuracy	**Plausibility**	**Level 3**	55) Does a document of procedure about technical sources of errors and deviation/inaccuracy exist?	submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf
			56) Does a document of procedure about methodological sources of errors and deviation/inaccuracy exist	submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf
			57) Does a document of procedure with validation against independent data exit	submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf

			58) Does a document of evaluation results (data) and methods exist? Modell methods: It should include description of the models, components and their equations (link to model homepage). Detailed description of simulations with: •Resolution in time and space, dependencies of time and space resolutions. •Structure – grid description, extraction possibilities •Boundary conditions – forcing •Input data •Constants – for initialization and run e.g. orography, solar constant, drag coefficient, area leaf index •Information about benchmark tests and the reproducibility of simulation runs Description of family trees of models like: http://www.gfdl.noaa.gov/jrl_gcm or http://www.aip.org/history/climate/xAGCMtree.htm Observational methods: It should include description of campaigns, supersites, resolution in time and space, instruments and plat forms. For example see: https://icdc.zmaw.de/fileadmin/user_upload/HDCP2_Docs/h dcp2_obs_data_product_standard_v2.2.pdf Detailed descript ion of classification into a level systeme.g. http://www.godae.org/Data-definition.html.	submission papers add_info file: quality checking, model data methodology report, observational methodology report and datalevel in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf WF10: link to model description and detailed description requested for cera2.quality.accuracy report
Statistical Anomalies	**Level 3 + scientific consistency among multiple data sets and their relationships is documented if feasible.**	✓ 59) Are missing values indicated and how? ✓ 60) Is a document about procedure of statistical quality control available? The document should contain information on the procedure of data quality checking and its finding, e.g. details of the procedure, quality check protocols, images of the quality check findings, etc. It should include documentation about: Documented procedure of statistical quality control: Examples of statistical quality control tests a) Rough Errors Tests •LIM-test by Meek and Hatfield (The test checks every data point on whether it exceeds a predefined range of values.) •NOC-test by Meek and Hatfield (The test checks on whether data does not change for more than a predefined number of values. It can be used to detect errors of instrument.) •ROC-test by Meek and Hatfield (The test checks the rate of change. The difference between two consecutive elements is checked concerning limits.) b) Tests for systematic deviations in time and space (e.g. changes in mean, variance and trends) and random errors e.g.: Düsterhus, A. and Hense, A.: Advanced information criterion for environmental data quality assurance, Adv. Sci. Res., 8, 99 -104, doi:10.5194/asr-8- 99-2012 , 2012. Meek, D. Hatfield, J. (1994) Data quality checking for single station meteorological databases. Agricultural and Forest Meteorology - AGR FOREST METEOROL , vol. 69, no. 1-2, pp. 85-109, DOI: 10.1016/0168-1923(94)90083-3 ✓ 61) scientific consistency among multiple data sets and their relationships is documented if feasible.	WF10: DOI review ✓ submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf

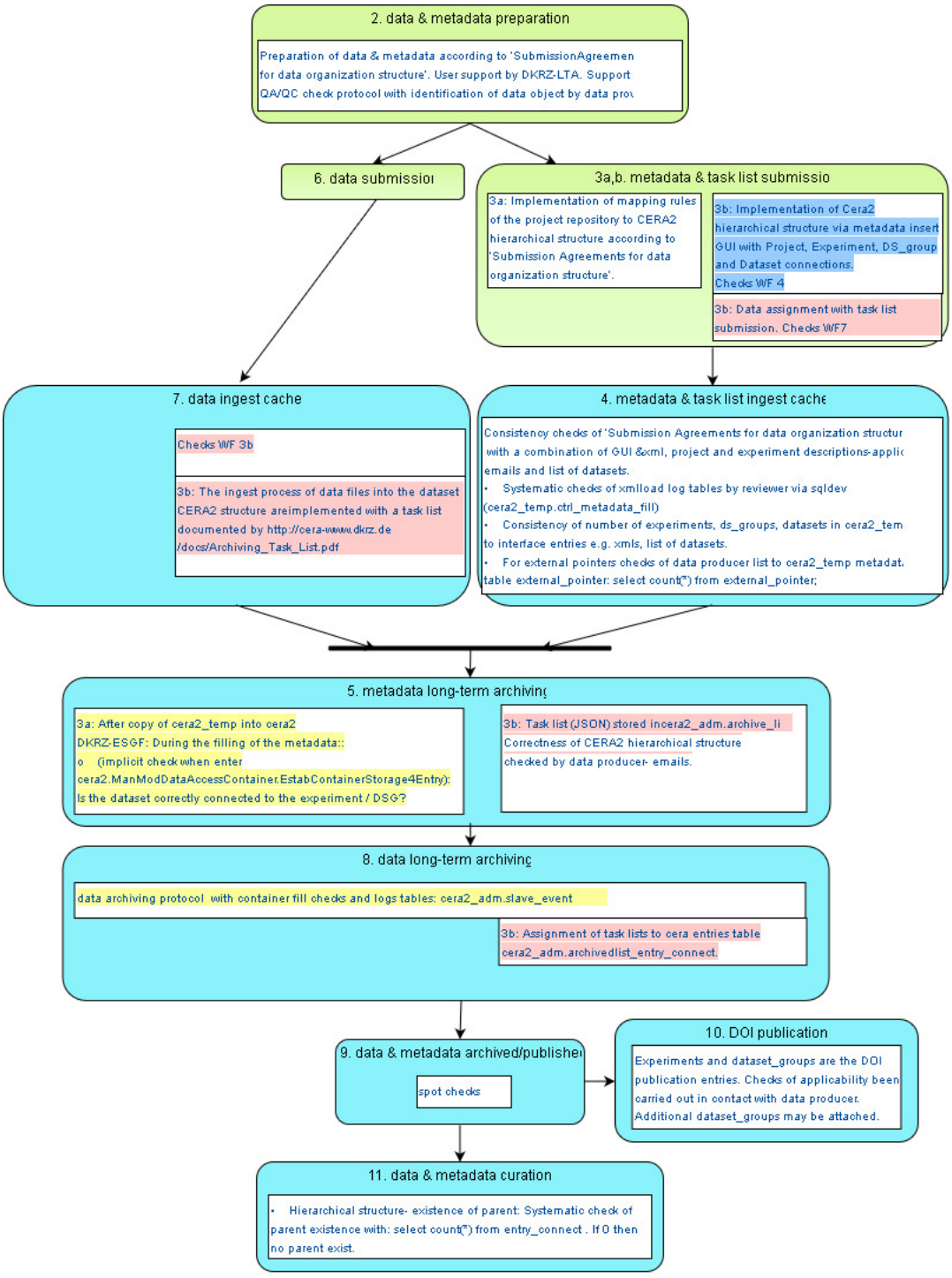


Figure 3: Data Organization WF with Checks

DKRZ

Table 1: DKRZ-LTA (OAIS Table 4-1) Example of PDI

Content Information Type	Reference	Provenance	Context	Fixity	Access Rights
DKRZ-LTA Data	Citation-Creators Publication Year Title DOI Publisher Persistent Identifier - DOI Journal reference	Data source Creators and contributors Description of data production steps and methods	Mission Funding history Pointer to related project description in original environment	Checksum and data object size protection against citation alteration	Access and use constraints Terms of use

Heinke Höck (DKRZ)

Figure 4: PDI at DKRZ-LTA

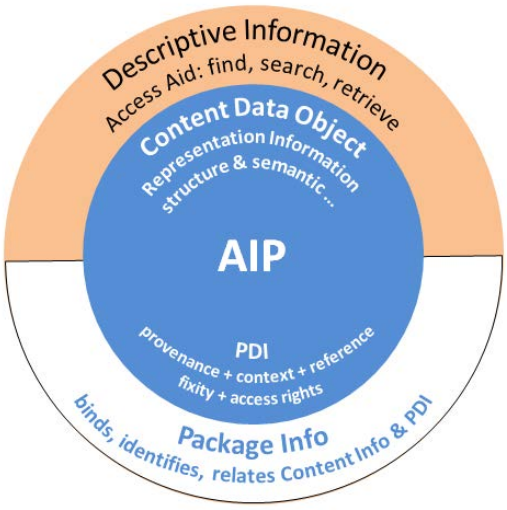


Figure 5: OAIS AIP

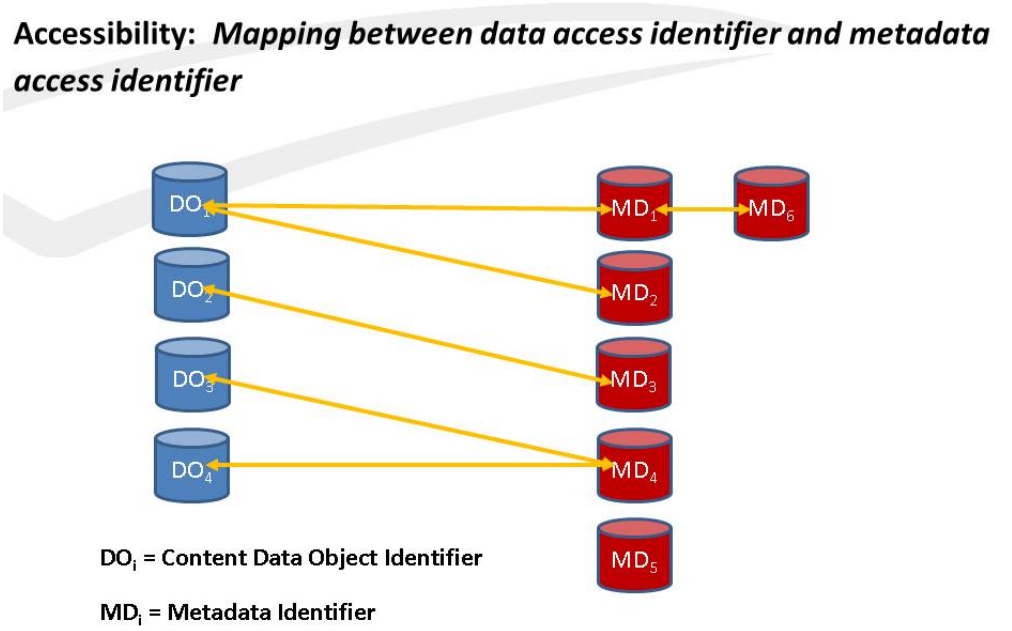


Figure 6: OAIS Package Info: binds and identifies

4. Protocol Template Level 4

Used metadata ingest case: <> to be filled in

Standards are used for data organization, data objects, data formats , CVs

Protocols with project name <> are available on the internal website
https://madwiki.dkrz.de/CERA%20Contents

	Checklist	Protocol – examples - sufficiently not necessary
✓	1) In what way are well-defined rules about data organization documented?	Definition of Submission Agreements: Protocols https://madwiki.dkrz.de/CERA%20Contents project name:<> Link to discipline-specific standards or data management plan Request Tracker: https://dm-rt.dkrz.de/ RT id:<>
✓	6) Are files structured in a consistent manner to the well-defined rules and the Representation Information of the AIPs.	Format = <>, CF is the default structure. Other formats must be checked. for NetCDF) Representation Information is part of the netCDF CF data header e.g. cdo-CMOR (https://cmor.llnl.gov/, https://www.dkrz.de/c6de)-> NetCDF-CF (has to be checked)
✓	8) Is your data dynamic in time?	Cera2 progress = <> - with all children
✓	12) Is the data self-describing data objects which meet the discipline-specific standards?	Format = <>, self-describing <yes/no> cv-list link<> e.g. NetCDF-CF with standard names
✓	14) Are the formats correct (checked with format checker)?	Format = <> Used format checker =<> Formats correct <yes/no>
✓	15) Are data header and AIP consistent?	Data header stored in cera2 <yes/no> DOI <yes/no> check of standard name, long_name, coverage
✓	16) In what way are the requirements for file extensions documented?	project name:<> e.g. cdo-CMOR=<yes/no> output .nc

✓ 18) Are the data sizes checked and correct, the size of data set is not equal 0 *if feasible*?	DOI <yes/no> - no does not lead to rejection
✓ 19) Are the file sizes feasible for access and transmission?	data file sizes < 2Giga <yes/no> With no is jblob cdo or time sections possible?
✓ 21) In what way are the long-term archive requirements or discipline-specific standard for versioning documented?	project name:<>
✓ 22) Is a versioning available if feasible?	Versioning = <yes/no> e.g. cdo CMOR <yes/no> -> has versioning
✓ 23) Description of versioning	project name:<> e.g. cdo CMOR documentation link (https://cmor.llnl.gov/, https://www.dkrz.de/c6de)
✓ 26) In what way is the controlled vocabulary documented – conform to discipline-specific standards? Comment: e.g. cf standard names	project dependent documented project_name <> cf-standard names <yes/no> < http://cfconventions.org/standard-names.html> e.g.cmip6 cv-list JSON online link https://github.com/WCRP-CMIP/CMIP6_CVs selection list of discipline-specific standard cf standard names http://cfconventions.org/
✓ 27) Is a controlled vocabulary available if feasible?	WF4,5: <yes/no> cf-standard names: <yes/no> cv-list link<> e.g. cdo-CMOR=<yes/no>
✓ 29) Are labeling of the AIPs or internal identifiers with mapping to objects correct in accordance to CVs?	DOI <yes/no> if no DOI is decided on a case-by-case basis.
✓ 30) Are naming conventions for discovery e.g. CVs correct?	DOI <yes/no>
✓ 31) PDI are consistent/correct content see 50) and 51)	DOI <yes/no>
✓ 33) Descriptive Information is consistent. ▪ metadata for search and discovery e.g. keywords ✓	DOI <yes/no>
✓ 35) Do data sets exist - complete?	Progress=completely archivied <yes/no>
✓ 40) PDI exist content see 50) and 51)	Progress=completely archivied <yes/no> Add_Infos – e.g. readme <yes/no> References <yes/no> Project link <yes/no> Model link <yes/no>
✓ 41) Representation Information exists	Progress=completely archivied <yes/no>
✓ 42) Descriptive Information exists: ✓ metadata for search and discovery e.g. keywords	Progress=completely archivied <yes/no>
✓ 43) Package Info exists	Progress=completely archivied <yes/no>
✓ 46) Are complete data accessible by identifier for at least 10 years?	Progress=completely archivied <yes/no> Access by https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=<entry_acronym>
✓ 50) PDI are accessible by identifier o **Provenance** ▪ data source e.g. sensor ▪ publisher if feasible ▪ detailed description of data production steps and method – quality assurance procedure (approval and review) ▪ contributor(s) if feasible – contact	Progress=completely archivied <yes/no> Access by https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=<entry_acronym>
✓ 51) PDI are accessible by identifier o **Context** ▪ project and experiment description o **Reference** ▪ data citation – e.g. creators o **Fixity** ▪ data expiration date o **Access Rights** ▪ access constraint	Progress=completely archivied <yes/no> Access by https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=<entry_acronym>

✓	
✓ 52) Representation, Descriptive Information and Package Info are accessible by identifier ✓ metadata for search and discovery e.g. keywords	Progress=completely archivied <yes/no> Access by https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=<entry_acronym>
✓ 55) Does a document of procedure about technical sources of errors and deviation/inaccuracy exist?	<yes/no>
✓ 56) Does a document of procedure about methodological sources of errors and deviation/inaccuracy exist	<yes/no>
✓ 57) Does a document of procedure with validation against independent data exit	<yes/no>
✓ 58) Does a document of evaluation results (data) and methods exist?	<yes/no>
✓ 59) Are missing values indicated and how?	<yes/no>
✓ 60) Is a document about procedure of statistical quality control available?	<yes/no>

Project name is only sufficient if the core information is filled (internal):
https://madwiki.dkrz.de/CERA%20Contents/VorlageKerninformationen

Discipline-specific standards

PIDs, NetCDF CF, DRS cmip6

interdisciplinary standards - general/international

ISO 19115/19139 (C3 grid), DataCite schema, GeoTIFF, netcdf, DOIs

Sufficient requirements since 01mar2018 (WF implementation) for level 4 at WDCC DKRZ-lta are:

1) Core documentation (internal) of project in **https://madwiki.dkrz.de/CERA%20Contents**
2) NetCDF CF format with standard names and netcdf format with usage of format checker
3) Cera2 progress = completely archived
4) DOI assignment with QA
5) Check of 19
6) Model link
7) Documentation of 55 – 60
8) Access constraint check

5. QMM Level 5 with Checklist Version 15/05/2019

		Level 5	Checklist	Implementation with check protocol, WF description and link to documentation of rules – sufficient but not necessary
Consistency	Data Organisation and Data Object	***data organisation is structured/conform to***		
		interdisciplinary standards	✓ In what way are well-defined rules about data organisation documented?	Interdisciplinary about data organization does not exist. Therefore no level 5 is available at the moment.
		data objects (OAIS) are		
		AIPs conform to interdisciplinary standard **up-to-date and consistent to external scientific objects if feasible**	✓ In what way are requirements for data objects (AIPs, DIPs) conform to interdisciplinary standards documented?	For the metadata there exists the standard ISO19115/19139. netcdf is available for the Content Data Object. However, the semantics are not fixed there. This is done with cf which is discipline specific climate and forecast. Therefore no level 5 is available at the moment.
		data formats – Content Data Object (OAIS)		
		conform to interdisciplinary standards		If the data objects are not conform to interdisciplinary standards, then it makes no sense to ask for the data formats. For example, netcdf is interdisciplinary, but only the subcriteria together provide a level for the criteria. Netcdf <yes/no> - used for completeness of data If level 4 and netcdf then level 5 is reached
		data sizes are consistent **file extensions are consistent**		If level 4 and netcdf then level 5 is reached
	Versioning and Controlled Vocabularies (CVs)	***versioning follows/is***		
		systematic collection including documentation of enhancement conform to well-defined rules **old versions stored if feasible** **old versions stored if feasible**		Level 4
		documentation of not included newer versions is consistent	✓ Does newer versions exist and where? ✓ Is documentation of not included newer version consistent to actual version?	<yes/no> <yes/no>
		data labelled with CVs conform to		
		interdisciplinary standards	✓ In what way are the controlled vocabulary conform to interdisciplinary standards documented?	Linked data is not available at WDCC. Therefore no level 5 is available at the moment.
	Data-Metadata Consistency	***OAIS metadata components are consistent***		
		Complete PDI * **Provenance** **Context** **Reference - cross** **Fixity** **Access Rights** **and** **Representation Information** **Descriptive Information** **Package Info** ***maintenance and storage policy are not affected, since they belong to the repository certification.**		**Level 4**
			o **Access Rights** ▪ access constraint	▪ access constraint -access constraint WF10: access check by jblob download for special user k*

		external metadata and data are consistent		
			✓ Do external metadata exist and where? ✓ How are they connected to the data objects e.g. link identifier in the metadata or OAI-PMH service	Yes Checks of OAI-PMH Services <yes/no>
	Existence of Data (Completeness and Persistence)	**data entities (conform to interdisciplinary standards) are complete dynamic datasets - data stream are not affected number of datasets (aggregation) is consistent data are persistent, as long as expiration date requires**		Level 4 and netcdf
Completeness	**Existence of Metadata**	***OAIS metadata components exist***		
		Complete PDI * Provenance Context Reference Fixity Access Rights and Representation Information Descriptive Information Package Info *maintenance and storage policy are not affected, since they belong to the repository certification.	▪ ▪	**Level 4**
		metadata is conform to interdisciplinary standards	✓ Does a mapping to interdisciplinary standards e.g. ISO of the AIP exist?	ISO 19115/19139 OAI-PMH DOI<yes/no> -> DataCite
		data provenance chain exists including internal and external objects e.g. software, articles, method and workflow description	✓ Is the provenance chain – lineage documented e.g. with PIDs	<yes/no>
			✓ Are the external objects connected to the lineage PIDs?	<yes/no>
Accessibility	**Data Access by Identifier**	**data is accessible by**		
		global resolvable identifier (PID persistent identifier) registered with resolving to data access including backup where it is commonly accepted that the identifier is persistently resolvable at least to information about fate of the object	✓ Are complete data accessible by global resolvable identifier (PID e.g. DOI)? ✓ Is PID with bijective mapping to objects available? ✓ Does a backup of data exist? ✓ Are complete data accessible by identifier at least to information about fate of the object? ✓ How is the information about fate of the object documented?	Level 4 and DOI
		data is accessible within other data infrastructures including cross references	✓ Do other data infrastructures exist with data access? ✓ What are the names of these infrastructures? ✓ Does cross references exist e.g. DataCite?	DOI <yes/no> EUDAT-B2FIND, DataCite, DWD-GISC, KomFor, WDS Cross references are inserted in CERA2 if available by data provider
		checksums are correct and accessible	✓ Are *correct* checksums accessible? Comment: Checksums are created and checked	Level 4
		a bijective mapping between identifier and datasets is documented e.g. in data header (OAIS Package Info - binds, identifies)	✓ Does a bijective mapping between identifier and datasets exist?	Level 4

Metadata Access by Identifier	**metadata is accessible by**		
	by permanent identifier (expiration is documented) (OAIS Package Info - identifies) complete data citation is persistent		Level 4
	external PID references are supported	How are external PIDs referenced e.g. DataCite RelatedIdentifier	Supported by CERA Block References and DataCite RelatedIdentifier
	a mapping between data access identifier and metadata access identifier is implemented (OAIS Package Info relates Content Info and PDI)		Level 4
Accuracy Plausibility	**Level 3**	Does a document of procedure about technical sources of errors and deviation/inaccuracy exist?	submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf
		Does a document of procedure about methodological sources of errors and deviation/inaccuracy exist	submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf
		Does a document of procedure with validation against independent data exit	submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf
		Does a document of evaluation results (data) and methods exist? Modell methods: It should include description of the models, components and their equations (link to model homepage). Detailed description of simulations with: •Resolution in time and space, dependencies of time and space resolutions. •Structure – grid description, extraction possibilities •Boundary conditions – forcing •Input data •Constants – for initialization and run e.g. orography, solar constant, drag coefficient, area leaf index •Information about benchmark tests and the reproducibility of simulation runs Description of family trees of models like: http://www.gfdl.noaa.gov/jrl_gcm or http://www.aip.org/history/climate/xAGCMtree.htm Observational methods: It should include description of campaigns, supersites, resolution in time and space, instruments and plat forms. For example see: https://icdc.zmaw.de/fileadmin/user_upload/HDCP2_Docs/hdcp2_obs_data_product_standard_v2.2.pdf Detailed descript ion of classification into a level systeme.g. http://www.godae.org/Data-definition.html.	submission papers add_info file: quality checking, model data methodology report, observational methodology report and datalevel in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf WF10: link to model description and detailed description requested for cera2.quality.accuracy report

Statistical Anomalies	**Level 3 + scientific consistency among multiple data sets and their relationships is documented if feasible.**	✓ Are missing values indicated and how? ✓ Is a document about procedure of statistical quality control available? The document should contain information on the procedure of data quality checking and its finding, e.g. details of the procedure, quality check protocols, images of the quality check findings, etc. It should include documentation about: Documented procedure of statistical quality control: Examples of statistical quality control tests a) Rough Errors Tests •LIM-test by Meek and Hatfield (The test checks every data point on whether it exceeds a predefined range of values.) •NOC-test by Meek and Hatfield (The test checks on whether data does not change for more than a predefined number of values. It can be used to detect errors of instrument.) •ROC-test by Meek and Hatfield (The test checks the rate of change. The difference between two consecutive elements is checked concerning limits.) b) Tests for systematic deviations in time and space (e.g. changes in mean, variance and trends) and random errors e.g.: Düsterhus, A. and Hense, A.: Advanced information criterion for environmental data quality assurance, Adv. Sci. Res., 8, 99 -104, doi:10.5194/asr-8- 99-2012 , 2012. Meek, D. Hatfield, J. (1994) Data quality checking for single station meteorological databases. Agricultural and Forest Meteorology - AGR FOREST METEOROL , vol. 69, no. 1-2, pp. 85-109, DOI: 10.1016/0168-1923(94)90083-3 ✓ 61) scientific consistency among multiple data sets and their relationships is documented if feasible.	WF10: DOI review ✓ submission papers add_info file: quality checking in work https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf

6. Protocol Template Level 5

Used metadata ingest case: <> to be filled in

Protocols with project name <> are available on https://madwiki.dkrz.de/CERA%20Contents

Aspect/Checklist	Protocol – examples - sufficiently not necessary/Comment
Data Organisation and Data Object	no level 5 is available at the moment. See Checklist above.
Versioning and Controlled Vocabularies (CVs)	no level 5 is available at the moment. See Checklist above.
Data-Metadata Consistency	Level 4 and
✓ Do external metadata exist and where? ✓ How are they connected to the data objects e.g. link identifier in the metadata or OAI-PMH service	Yes Checks of OAI-PMH Services <yes/no>
Existence of Data (Completeness and Persistence)	Level 4 and
✓ Format ist netCDF	<yes/no>
Existence of Metadata	Level 4 and
✓ Does a mapping to interdisciplinary standards e.g. ISO of the AIP exist?	ISO 19115/19139 OAI-PMH DOI<yes/no> -> DataCite
✓ Is the provenance chain – lineage documented e.g. with PIDs	<yes/no>
✓ Are the external objects connected to the lineage PIDs?	<yes/no>

Data Access by Identifier	Level 4 (DOI) and
✓ Do other data infrastructures exist with data access? ✓ What are the names of these infrastructures? ✓ Does cross references exist e.g. DataCite?	DOI <yes/no> EUDAT-B2FIND, DataCite, DWD-GISC, KomFor, WDS Cross references are inserted in CERA2 if available by data provider
Metadata Accessby Identifier	Level 4 (DOI) and
How are external PIDs referenced e.g. DataCite RelatedIdentifier	Supported by CERA Block References and DataCite RelatedIdentifier
Plausibility	Level 4 = Level 3
Statistical Anomalies	Level 4 = Level 3 and
scientific consistency among multiple data sets and their relationships is documented if feasible.	<yes/no>

Project name is only sufficient if the core information is filled (internal): https://madwiki.dkrz.de/CERA%20Contents/VorlageKerninformationen

7. Links:

[1] https://www.dkrz.de/up/services/data-management

[2] http://cera-www.dkrz.de/LTA_metadata

[3] http://cera-www.dkrz.de/docs/Archiving_Task_List.pdf