EGU2015-10158

Maturity Matrices
for Quality of Model- and Observation-Based Data Records in Climate Science

H. Höck (1), V. John (3), A. Kaiser-Weiss (2), F. Kaspar (2), J. Schulz (3), M. Stockhause (1), F. Toussaint (1), M. Lautenschlager (1)
(1) German Climate Computing Centre (DKRZ, www.dkrz.de), (2) Deutscher Wetterdienst (DWD), (3) European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT)

funded by:

Introduction:

In the field of Software Engineering the Capability Maturity Model is used to evaluate and improve software development processes by assessing the so-called maturity level of a software. Recently, this method was adapted to assess the maturity of research data in the earth system sciences.

CORE-CLIMAX System Maturity Matrix[3]

designed for monitoring the process to generate Climate Data Records

The Maturity Matrix concept had been transferred to satellite climate data record generation[1]. It monitors the adherence to best practices in climate data record generation that have emerged from science and engineering over the last decades. The FP7 project CORE-CLIMAX widened and tested the concept for Climate Data Records (CDR) derived from in-situ observations and weather prediction model-based reanalyses [2].

The resulting **System Maturity Matrix** (**SMM)** presented has been used to assess the maturity of 37 European CDR production entities in preparation of the Copernicus Climate Change Service [3]. Self-assessments at Deutscher Wetterdienst (DWD) and EUMETSAT helped internal evaluation of the data production process. The CORE-CLIMAX project reached consensus that the application of the SMM helps data providers to assess the status of their production systems according to the state of the art, e.g. as provided by guidelines of the Global Climate Observing System (GCOS). Repeated application enables progress monitoring for ongoing developments.

1 & 2	Research Capability (RC)	3 & 4	Initial Operations Capability (IOC)	5 & 6	Full Operations Capability (FOC)

Maturity	SOFTWARE READINESS	METADATA	USER DOCUMENTATION	UNCERTAINTY CHARACTERISATION	PUBLIC ACCESS, FEEDBACK, UPDATE	USAGE
1	Conceptual development	None	Limited scientific description of the methodology available from PI	None	Restricted availability from PI	None
2	Research grade code	Research grade	Comprehensive scientific description of the methodology, report on limited validation, and limited product user guide available from PI; paper on methodology is submitted for peer-review	Standard uncertainty nomenclature is identified or defined; limited validation done; limited information on uncertainty available	Data available from PI, feedback through scientific exchange, irregular updates by PI	Research: Benefits for applications identified DSS: Potential benefits identified
3	Research code with partially applied standards; code contains header and comments, and a README file; PI affirms portability, numerical reproducibility and no security problems	Standards defined or identified; sufficient to use and understand the data and extract discovery metadata	**Score 2** + paper on methodology published; comprehensive validation report available from PI and a paper on validation is submitted; comprehensive user guide is available from PI; Limited description of operations concept available from PI	**Score 2** + standard nomenclature applied; validation extended to full product data coverage, comprehensive information on uncertainty available; methods for automated monitoring defined	Data and documentation publically available from PI, feedback through scientific exchange, irregular updates by PI	Research: Benefits for applications demonstrated. DSS: Use occurring and benefits emerging
4	**Score 3** + draft software installation/user manual available; 3rd party affirms portability and numerical reproducibility; passes data providers security review	**Score 3** + standards systematically applied; meets international standards for the data set; enhanced discovery metadata; limited location level metadata	**Score 3** + comprehensive scientific description available from data provider; report on inter comparison available from PI; paper on validation published; user guide available from data provider; comprehensive description of operations concept available from PI	**Score 3** + procedures to establish SI traceability are defined; (inter)comparison against corresponding CDRs (other methods, models, etc); quantitative estimates of uncertainty provided within the product characterising more or less uncertain data points; automated monitoring partially implemented	Data record and documentation available from data provider and under data provider's version control; Data provider establishes feedback mechanism; regular updates by PI	**Score 3** + Research: Citations on product usage in occurring DSS: societal and economical benefits discussed
5	**Score 4** + operational code following standards, actions to achieve full compliance are defined; software installation/user manual complete; 3rd party installs the code operationally	**Score 4**+ fully compliant with standards; complete discovery metadata; complete location level metadata	**Score 4** + comprehensive scientific description maintained by data provider; report on data assessment results exists; user guide is regularly updated with updates on product and validation; description on practical implementation is available from data provider	**Score 4** + SI traceability partly established; data provider participated in one inter-national data assessment; comprehensive validation of the quantitative uncertainty estimates; automated quality monitoring fully implemented (all production levels)	**Score 4** + source code archived by Data Provider; feedback mechanism and international data quality assessment are considered in periodic data record updates by Data Provider	**Score 4**+ Research: product becomes reference for certain applications DSS: Societal and economic benefits are demonstrated
6	**Score 5** + fully compliant with standards; Turnkey System	**Score 5** + regularly updated	**Score 5** + journal papers on product updates are and more comprehensive validation and validation of quantitative uncertainty estimates are published; operations concept regularly updated	**Score 5** + SI traceability established; data provider participated in multiple inter-national data assessment and incorporating feedbacks into the product development cycle; temporal and spatial error covariance quantified; Automated monitoring in place with results fed back to other accessible information, e.g. meta data or documentation	**Score 5** + source code available to the public and capability for continuous data provisions established (ICDR)	**Score 5** + Research: Product and its applications becomes references in multiple research field DSS: Influence on decision and policy making demonstrated

Quality Maturity Matrix[4]

designed for research data in the earth sciences

Quality Maturity Matrix Level	Level 1	Level 2	Level 3	Level 4	Level 5
Data Production Steps	concept	production processing first use	project collaboration	long-term archiving	re-use
Characteristic	not evaluated	informal	data management plan	community standard	general international standard

Based on the already existing CORE-CLIMAX SMM[3] and the CMM[5], the World Data Center for Climate-**WDCC** developed a generic Quality Assessment System for research data in the earth sciences because models and their related output have some additional characteristics that need specific consideration in such an approach. The Maturity Matrix at **DKRZ** was developed in collaboration with KomFor funded by DFG. A self-assessment is performed using a maturity matrix evaluating the data quality for five maturity levels with respect to the criteria and aspects. The Quality Maturity Matrix criteria are developed to support the phases of the data production steps. Use of QMM allows to compare and document the current maturity of data and metadata.

Maturity	Data and Metadata Quality Assurance Criteria and Aspects								
	Consistency			Completeness		Accessibility		Accuracy	
	Data Organisation and Data Object	Versioning and Controlled Vocabularies	Data-Metadata Consistency	Existence of Data/(Data Persistence)	Existence of Core Metadata and Provenance	Technical Data Access by Identifier/Lineage	Core Metadata and Provenance Access by Identifier	Plausibility	Statistical Anomalies
1	conceptual development	conceptual development	not evaluated	not evaluated	not evaluated	not evaluated	not evaluated	not evaluated	not evaluated
2	-informal data organization -file names to internal rules -file extensions are consistent	-informal versioning -CVs are consistent	creators are correct	data is in production and may be deleted or overwritten	-creators exist -data provenance is unsystematically documented	data is accessible by file names	-creators -data provenance unsystematically documented are accessible	documented procedure about technical sources of errors and deviation/inaccuracy exists	missing values are indicated e.g. with fill values
3	-data organization is documented -internal identifiers (with mapping to data objects) e.g. file names and formats correspond to project requirements -file extensions, size and checksum of main components are consistent	-systematic versioning correspond to project requirements -formal CVs of main components are consistent	creators/cont act are correct	datasets exist, not complete and may be deleted but not overwritten unless explicitly specified	-creators/contact exist -naming conventions for discovery exist -datasets provenance is basically documented[3]	-datasets are accessible by internal identifier and mapping (bijective) to objects are documented[3] -checksums are accessible	-creators/contact with naming conventions -datasets provenance are accessible	score2 + documented procedure about methodological sources of errors and deviation/inaccuracy exists	score 2 + documented procedure about rough anomalies are available e.g. outliers concerning limits.
4	-data organization is structured/conform according to well-defined rules -entry names and data formats are conform to community standards -datasets are re-usable with self-describing data objects which meet the community standards -file extension, size and checksum are consistent	-systematic versioning collection including documentation of enhancement is conform to community standards -old versions stored[1] -formal CVs of data are conform to community standards	main metadata components[4] are consistent	-data entities (conform to community standards) are complete[2] -number of data sets (aggregation) is consistent -data are persistent, as long as expiration date requires	main metadata components[4] exist	-complete datasets (conform to community standards) are accessible by permanent (minimum 10 years see rules of good scientific practice) identifier with resolving to data access as long as expiration date requires -checksums are accessible	-main metadata components[4] with data expiration date -detailed description of data production steps and methods are accessible by identifier	score 3	score 3 + -documented procedure about systematic deviations in time and space (e.g. changes in mean, variance and trends) and random errors exist -scientific consistency among multiple data sets and their relationships is documented[1]
5	-data organization is structured/conform according to standardized rules -data formats are conform to general/international standards -data objects are consistent to external scientific objects and up-to-date -file extension, size and checksum are consistent -data objects with general/international standards are self-describing -data objects are fully machine-readable with references to sources	score 4 + -documentation of not included newer versions is consistent -CVs are general/international standardized	score 4 + external metadata and data are consistent	-data entities (conform to general/international standards) are complete[2] -number of data sets (aggregation) is consistent -data are persistent, as long as expiration date requires	-metadata is conform to general/international standards -data provenance chain exists including internal and external objects e.g. software, articles, method and workflow description	-complete data (conform to general/international standards) is accessible by global resolvable identifier (PID) registered with resolving to data access including backup as long as expiration date requires -data is accessible within other data infrastructures including cross references -external PID references supported -provenance chain is accessible	-metadata with data expiration date including backup general/international standardized -data provenance chain including internal and external objects e.g. software, articles, methods and workflow description are accessible by global resolvable identifier	score 3 + -documented procedure with validation against independent data -references to evaluation results (data) and methods exists	score 4

Foot Notes
[1] if feasible
[2] dynamic datasets -data stream are not affected
[3] e.g. in data header
[4] -data source e.g. sensor
-creators/contact and publisher if feasible
-metadata for search and discovery e.g. keywords
-quality assurance procedure (approval and review)
-data citation
-detailed description of data production steps and method
-data expiration date
-access constraint
-contributor(s) if feasible

Conclusions:

1) Design purposes should be considered when interpreting Maturity Matrices.
2) Self-assessment with SMM at DWD and EUMETSAT and QMM at WDCC successfully helped identifying areas for improvement.

References:

[1] BATES, John J.; PRIVETTE, Jeffrey L. A maturity model for assessing the completeness of climate data records. Eos, 2012, 93. Jg., Nr. 44, S. 441-441.
[2] CORE-CLIMAX Project Deliverable D222 available from http://www.coreclimax.eu/
[3] CORE-CLIMAX CDR Assessment Report available from http://www.coreclimax.eu/
[4] http://www.komfor.net/qa.html
[5] http://en.wikipedia.org/wiki/Implementation_maturity_model_assessment (12.03.2015)