Data and Metadata Quality Assurance Criteria and Aspects

		Consistency			Completeness		Accessibility		Accuracy	
Maturity/Score	**Grade**	**Data Organisation and Data Object**	**Versioning and Controlled Vocabularies (CVs)**	**Data-Metadata Consistency**	**Existence of Data**	**Existence of Core Metadata and Provenance**	**Technical Data Access by Identifier/Lineage**	**Core Metadata and Provenance Access by Identifier**	**Plausibility**	**Statistical Anomalies**
1	conceptual	conceptual development	conceptual development	not evaluated	not evaluated	not evaluated	not evaluated	not evaluated	not evaluated	not evaluated
2	research	-informal data organisation -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	collaboration	-data organisation 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/contact 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 (bijectiv) 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	exchange	-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	re-use & impact	-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

conceptual	**under conceptual development or not evaluated**
research	**aspects characterized for data production and processing with individual checks**
collaboration	**aspects characterized for project collaboration with systematic main component checks provide legal to project requirements e.g. as defined in data management plan**
exchange	**aspects characterized for data and metadata preservation with systematic checks provide legal to community standards and long-term archive requirements**
re-use & impact	**aspects characterized for a distributed virtual research environment with systematic checks provide legal to general/international standards including cross-references to journal article and altmetric**

[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

CVs= Controlled Vocabularies

PID= persistent identifier

Heinke Höck, Martina Stockhause & Frank Toussaint / DKRZ