

Discussion of fitness for purpose and uncertainty assessment

Michael Borsche, Andrea Kaiser-Weiss, Frank Kaspar,
(DWD)

UERRA showcase, 24 Nov 2016, Reading, UK

Discussion of fitness for purpose

- ➔ Evaluation methods have been developed to assess uncertainties
- ➔ Follow up: how to use the data? What do we provide the users with?
- ➔ SMHI: analysis of wind power production
- ➔ Maturity Matrix:
 - ➔ System Maturity Matrix
 - ➔ Quality Maturity Matrix
- ➔ Post processing to local scale

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 international 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 international 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

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/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	-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