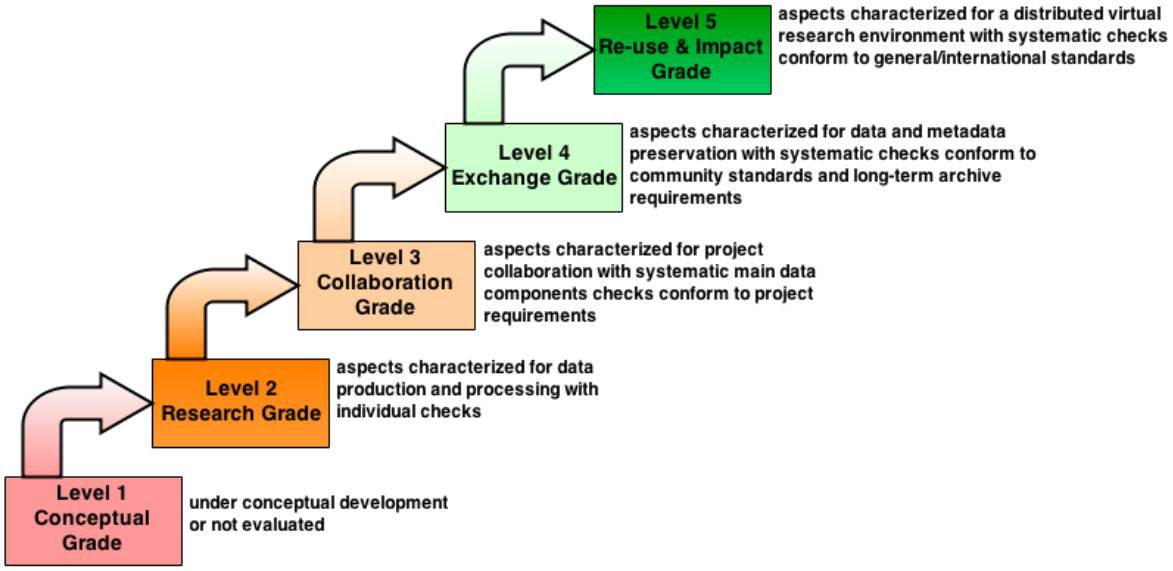
Quality Maturity Matrix (Version 06.11.2014)

Characteristics of Data and Metadata Quality Assurance Maturity Levels



The Criteria are:

Consistency

- **Data Organisation and Data Object**
- **Versioning and Controlled Vocabularies (CVs)**
- **Data-Metadata Consistency**

Completeness:

- **Existence of Data**
- **Existence of Core Metadata and Provenance**

Accessibility:

- **Technical Data Access by Identifier/Lineage**
- **Core Metadata and Provenance Access by Identifier**

Accuracy

- **Plausibility**
- **Statistical Anomalies**

Level 1 Conceptual Grade:

Characteristics: All aspects under conceptual development or not evaluated

Score 1 is always reached

Level 2 Research Grade:

Characteristics: aspects characterized for data production and processing with individual checks

Score 2 rating scales correspond to the following aspects.

Consistency:

- **Data Organisation and Data Object:**
  -informal data organization
  -file names to internal rules
  -file extensions
  are consistent

- **Versioning and Controlled Vocabularies (CVs)**:
  -informal versioning
  -CVs
  are consistent

- **Data-Metadata Consistency:** creators are correct

Completeness:

- **Existence of Data:** data is in production and may be deleted or overwritten

- **Existence of Core Metadata and Provenance:**
  -creators exist
  -data provenance is unsystematically documented

Accessibility:

- **Technical Data Access by Identifier/Lineage:** data is accessible by file names

- **Core Metadata and Provenance Access by Identifier:**
  -creators
  -data provenance unsystematically documented
  are accessible

Accuracy:

- **Plausibility:** documented procedure about technical sources of errors and deviation/inaccuracy exists
- **Statistical Anomalies:** missing values are indicated e.g. with fill values

Level 3 Exchange Grade:

Characteristics: aspects characterized for project collaboration with systematic main data components checks conform to project requirements

Score 3 rating scales correspond to the following aspects.

Consistency:

- **Data Organisation and Data Object:**
  -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
- **Versioning and Controlled Vocabularies (CVs)**:
  -systematic versioning correspond to project requirements
  -formal CVs of main components are consistent
- **Data-Metadata Consistency:** creators/contact are correct

Completeness:

- **Existence of Data:** datasets exist, not complete and may be deleted but not overwritten unless explicitly specified
- **Existence of Core Metadata and Provenance:**
  -creators/contact exist
  -naming conventions for discovery exist
  -datasets provenance is basically documented e.g. in data header

Accessibility:

- **Technical Data Access by Identifier/Lineage:**
  -datasets are accessible by internal identifier and mapping (bijectiv) to objects are documented e.g. in data header
  -checksums are accessible

- **Core Metadata and Provenance Access by Identifier:**
  -creators/contact with naming conventions
  -datasets provenance
  are accessible by identifier

Accuracy:

- **Plausibility:**
  -documented procedure about technical sources of errors and deviation/inaccuracy exists
  -documented procedure about methodological sources of errors and deviation/inaccuracy exists

- **Statistical Anomalies:**
  -missing values are indicated e.g. with fill values
  -documented procedure about rough anomalies are available e.g. outliers concerning limits.

Level 4 Re-use Grade:

Characteristics: aspects characterized for data and metadata preservation with systematic checks conform to community standards and long-term archive requirements

Score 4 rating scales correspond to the following aspects.

Consistency:

- **Data Organisation and Data Object:**
  -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

- **Versioning and Controlled Vocabularies (CVs)**:
  -systematic versioning collection including documentation of enhancement is conform to community standards
  -old versions stored if feasible
  -formal CVs of data are conform to community standards

- **Data-Metadata Consistency:**
  -data source e.g. sensor
  -creators/contact/publisher
  -metadata for search and discovery e.g. keywords
  -quality assurance procedure for data and metadata consistency (approval and review) is documented
  -data citation metadata to a documented procedure are consistent

Completeness:

- **Existence of Data:**
  -data entities (conform to community standards) are complete (dynamic datasets - data stream is not affected)
  -number of data sets (aggregation) is consistent
  -data are persistent, as long as expiration date requires

- **Existence of Core Metadata (main components) and Provenance:**
  -data source e.g. sensor
  -creators/contact/publisher
  -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

Accessibility:

- **Technical Data Access by Identifier/Lineage:**
  -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

- **Core Metadata and Provenance Access by Identifier:**
  -main metadata components:
  -data source e.g. sensor
  -creators/contact/publisher
  -metadata for search and discovery e.g. keywords
  -quality assurance procedure for data and metadata consistency (approval and review) is documented
  -data citation metadata to a documented procedure are consistent
  with data expiration date
  -detailed description of data production steps and methods
  are accessible by identifier

Accuracy:

- **Plausibility:**
  -documented procedure about technical sources of errors and deviation/inaccuracy exists
  -documented procedure about methodological sources of errors and deviation/inaccuracy exists

- **Statistical Anomalies:**
  -missing values are indicated e.g. with fill values
  -documented procedure about rough anomalies are available e.g. outliers concerning limits.
  -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 if feasible

Level 5 Impact Grade:

Characteristics: aspects characterized for a distributed virtual research environment with checks conform to general/international standards

Score 5 rating scales correspond to the following aspects.

Consistency:

- **Data Organisation and Data Object:**
  -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

- **Versioning and Controlled Vocabularies (CVs)**:
  -systematic versioning collection including documentation of enhancement is conform to community standards
  -old versions stored if feasible
  -documentation of not included newer versions is consistent
  -CVs are general/international standardized

- **Data-Metadata Consistency:**
  -data source e.g. sensor
  -creators/contact/publisher
  -metadata for search and discovery e.g. keywords
  - Quality Assurance procedure for data and metadata consistency (approval and review) is documented
  -data citation metadata to a documented procedure are consistent
  -external metadata and data are consistent

Completeness:

- **Existence of Data:**
  -data entities (conform to general/international standards) are complete (dynamic datasets -data stream are not affected)
  -number of data sets (aggregation) is consistent
  -data are persistent, as long as expiration date requires

- **Existence of Core Metadata and Provenance:**
  -metadata is conform to general/international standards
  -data provenance chain exists including internal and external objects e.g. software, articles, method and workflow description

Accessibility:

- **Technical Data Access by Identifier/Lineage:**
  -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

- **Core Metadata and Provenance Access by Identifier:**
  -metadata with data expiration date including backup general/international standardized are accessible by global resolvable identifier
  -data provenance chain including internal and external objects e.g. software, articles, methods and workflow description
  are accessible by global resolvable identifier

Accuracy:

- **Plausibility:**
  -documented procedure about technical sources of errors and deviation/inaccuracy exists
  -documented procedure about methodological sources of errors and deviation/inaccuracy exists
  -documented procedure with validation against independent data
  -references to evaluation results (data) and methods exist

Statistical Anomalies:

-missing values are indicated e.g. with fill values

-documented procedure about rough anomalies are available e.g. outliers concerning limits.

-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 if feasible