Nanotechnology Signature Initiative
Nanotechnology Knowledge Infrastructure (NKI): Enabling National Leadership in Sustainable Design

Draft Discussion Document: May 9, 2013

Data Readiness Levels

A critical aspect of sharing data is an understanding of the maturity or quality of the data. Representatives from the collaborating agencies of the NKI Signature Initiative have developed a nomenclature for communicating the maturity of data. Analogous to Technology Readiness Levels, the Data Readiness Levels provide a shorthand method for conveying coarse assessments of data from experiments or model predictions for use in improving analytical methods and validating or calibrating models, and for comparisons with legacy datasets. Data Readiness Levels (DRLs) are seven graded definitions (0-6) of data quality and data maturity. DRLs provide common, simple descriptors of data quality and maturity. Unlike Technology Readiness Levels (TRLs), DRLs are augmented with metadata qualifiers that enable further assessment, reproduction, or use of the data by others. Metadata vary by discipline, as well measurement or computational considerations. The use of both DRL levels and metadata qualifiers provide a common basis for a peer-reviewed "literature" to support informed data sharing, to augment data citation in print publications, and to accelerate the translation of research to design and manufacture.

This discussion document is intended to engage the broader community regarding key issues critical to achieving the goals set forth in the NKI. For more information, please contact info@nnco.nano.gov.

The ideas presented within are of the authors alone and do not necessarily represent the views or positions of the United States Government.

Nanotechnology Signature Initiative
Nanotechnology Knowledge Infrastructure (NKI):
Enabling National Leadership in Sustainable Design

Data Readiness Levels

To use data effectively, the data's maturity or quality must be known. To aid in data sharing and use, representatives from the Government collaborating agencies of the NKI Signature Initiative have developed seven graded definitions (0-6) of Data Readiness Levels (DRLs). DRLs are similar to DoD/NASA Technology Readiness Levels. DRLs provide common, simple descriptors of data quality and maturity. They are a shorthand for conveying coarse assessments of data for use in validating models or assessing other data.

Unlike TRLs, DRLs are augmented with metadata qualifiers that aid data assessment, reproduction, and use. Metadata vary by discipline and measurement or computational considerations. Metadata provide information on data curation and provenance.

Table 1. Data Readiness Level Definitions

Data Readiness Level (DRL)	**Description**
0. Invalid data	Data that have been assessed and found to be invalid or so inaccurate or inadequately documented as to be of little practical value.
1. Raw or unscaled data	Data from sensors or calculations not converted to final (appropriate) physical units. *An example is a recording of the electrical output from a speedometer not scaled to units of distance per unit time, or archived unscaled output from a numerical simulation.*
2. Scaled data	Data converted to the intended final physical units. Noise levels are undefined. Data precision is undefined. Duplicate measurements are not cited or are not available. Data uncertainty estimates are not cited, or are not accepted. *Example are speedometer data appropriately scaled to units of miles/hour or km/hour, but without adequate information on calibration and background noise.*
3. Scaled data with defined precision **or** noise level	Data precision or data noise levels defined by accepted duplicate measurements or accepted noise measurements. Data not confirmed by independent observations. Models fit to data are not yet related to the larger body of accepted scientific knowledge. Data uncertainty estimates are not cited, or are not accepted. *Examples are scaled speedometer data, adjoined by data defining the noise levels of the instrument and recording system or adjoined by multiple speed measurements of the same event.*
4. Scaled data with defined precision **and** noise levels, but **not related** to the larger body of scientific knowledge	Data precision and data noise levels defined by accepted duplicate measurements and accepted noise measurements. Data confirmed by independent observers using similar methods. Models fit to data are not yet related to the larger body of accepted scientific knowledge but are speculative and inspire scientific debate. Data uncertainty estimates are speculative. DRL 4 data often lead to fundamental scientific advances. *An example is the 1887 Michelson-Morley data of the speed of light before Einstein's development of the Theory of Special Relativity, or the 1965 Penzias and Wilson radio telescope data before incorporation into the Big Bang Theory.*
5. DRL 4 data **related** to the larger body of scientific knowledge, but with measurement uncertainty too large for data standards	Data precision and data noise levels defined by accepted duplicate measurements and accepted noise measurements. Data confirmed by independent observers using similar methods. Models fit to data relate to the larger body of accepted scientific knowledge and can be used for coarse validation of existing models or development of refined models. Data uncertainty is larger than current standards for like data. *An example is the Michelson-Morley data after the development and acceptance of the Theory of Special Relativity.*
6(X). Standards-quality data of X % measurement uncertainty	Data precision and data noise levels defined by accepted multiple duplicate measurements and accepted noise measurements. Data confirmed by independent observers using independent methods. Data uncertainties are accepted by the scientific community to be X % or less. Models fit to data relate to the larger body of accepted scientific knowledge and can be used to validate existing models to X % accuracy, or to advance higher-fidelity scientific hypotheses. Data are used as a standard to validate other data, providing the bases for data traceability. *An example is the light travel time interval specified by U.S. NIST to define the meter.*

Table 2. Summary of DRLs Versus Data Attributes

Attribute	DRL 0	DRL 1	DRL 2	DRL 3	DRL 4	DRL 5	DRL 6
Units		maybe	yes	yes	yes	yes	yes
Precision and Noise				either	both	both	both
Independent Confirmation				possibly	yes	yes	yes
Related to Larger Body of Scientific Knowledge					no	yes	yes
Measurement Uncertainty					speculative	high	low
Example or use	little to none	unscaled sensor data	scaled sensor data	scaled data; noise levels defined	major scientific advances	coarse validation of theory	theory refinement and methods validation

Metadata

The three metadata qualifiers: poor, acceptable, and excellent, are used in conjunction with the DRLs to indicate the state of the data's curation and provenance. Examples are: "DRL 2 with poor metadata" or "DRL 5 with excellent metadata."

Poor Metadata: Failure to include critical information so that the data cannot be reproduced by others, or so that there is ambiguity in interpreting the data, or so that acceptable measurement uncertainty estimates cannot be made; for example, failure to adequately describe the measurement/computational methods used, or failure to provide complete and unambiguous descriptions of data format.

Acceptable Metadata: Inclusion of all key parameters so that others can reproduce the data, and so that there is little/no ambiguity in interpreting the data, and so that acceptable measurement uncertainty estimates can be made, including adequate descriptions of measurement/computational methods used, boundary and initial conditions, and complete and unambiguous descriptions of data format, curation, and provenance (history).

Excellent Metadata: Inclusion of all key parameters so that others can reproduce the data and unambiguously interpret the data, and so that acceptable measurement uncertainty estimates can be made, as well as other information to judge the data such as names and pedigree of the data creators; data format, curation, and provenance; and validation of the measurement/computational methods. DRL 6(X), by its nature, can only be data that have excellent metadata.

Additional Definitions

The following definitions apply and, as appropriate, were developed to be compliant with terms used in JCGM 100 and 200: 2008 (http://www.iso.org/sites/JCGM/). The definition of data includes both calculated and measured quantities.

Data: Information obtained from investigations including measurements, calculations, computations, simulations, estimates, and statistics.

Data Accuracy: The degree of agreement between measured or calculated values to the "true" value. In archery, accuracy refers to the distance from the average location of the arrows to the bullseye.

Data Curation: The active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains its quality, adds value, and provides for re-use over time (from University of Illinois Graduate School of Library and Information Science).

Data Precision: The closeness of agreement between measured/calculated values obtained by replicate measurements/calculations of the same or similar phenomena under the same conditions. In archery, precision refers to the aerial extent of the arrow pattern on the target. An arrow pattern could have good accuracy and poor precision, or good precision and poor accuracy, or be good or poor in both attributes.

Data Provenance: Provenance (also called lineage) captures where data came from, how it was derived, manipulated, and combined, and how it has been updated over time (from Stanford University Infolab).

Data Scaling: A mathematical transformation applied to raw data to produce data with appropriate physical units. Thus an output quantity Y (scaled data), is obtained from measured/computed quantities X_1,X_2, ..., X_N (unscaled raw data) by a functional relation: $Y = f\ X_1, X_2, \cdots, X_N$, where the function $f\ X_1, X_2, \cdots, X_N$, depends on the "scaling model" adopted to express the dependence of Y on the unscaled raw data $X_1, X_2, \cdots, X_N$.

Data Uncertainty: An expression of the "level of confidence" of the validity of the result of a measurement or calculation, quantifying the likelihood of producing the best value consistent with presently available knowledge related to the quantity being measured or calculated. In the case of theoretical computations involving no random variables, Data Uncertainty is understood to be the error of the result as compared to an accepted best value.

Metadata: Information about data such as format, means of creation, purpose, time and date of creation, author/creator of data, and standards used in creating data.

Noise: Spurious data or signal superimposed on intended measured or calculated data values caused by phenomena other than that which the data is intended to describe, such as wind noise superimposed on an intended recording of a musical instrument. Noise can be random and/or systematic, with systematic noise often resulting in data bias.

Readiness: As described in this document, a proposed systematic measure of reliability.

Relevance: The extent to which particular data can contribute to the resolution of a problem.

Reliability: The extent to which the provenance of data and their associated metadata are grounded in the methods and procedures of science.