**NCIP Nanotechnology Working Group**

**Consensus Response to Data Readiness Levels**

The individual DRL categories and the document itself are a useful starting point for a more extensive dialog across disciplines. Our questions and suggestions are offered in that context.

***Questions***

* How do these readiness levels apply to the traditional data levels (1-raw, 2-processed, 3-interpreted, 4-summary)
  + Because for example if it is Level 1 data on the traditional scale, by definition you cannot reach DRL 6
  + Were these categories known at the time of the DRL drafting and intended to map into them or are they independently conceived
  + These traditional levels have been established within genomics/next-gen sequencing, established bioinformatics data levels) <https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp>
* Formats for data must be specific to levels and information within specific assays and constraints have to be defined at this level in order to allow inter-comparison
* Interesting that meta-data are pulled out separately from the data considerations within the DRLs.
  + If a dataset is scaled, interpreted, and is becoming a standard, it would seem necessary to have all of the raw data associated with that as well to move ahead in using the dataset comparably
* A number of the questions arising for our group are definitional in nature
  + What is raw data? – the actual measure, the raw data dump from an analytical equipment
  + What is meant by “meta-data” vs. part of “data”– equipment, timepoints, environmental system characterization
  + What does it mean to fit into a larger scientific evidence? To be translatable and comparable with disciplines outside of the one generating the dataset?
  + Precision
  + Noise
  + Uncertainty
* It appears that the data and data readiness document seem to reflect an analytical chemist’s view, which is inline with its scope statement, or purpose, which is “improving analytical methods and validating or calibrating models.”  However, in the NKI presentation of the DRL’s to the NCIP NanoWG (Lisa Friedersdorf, 4/10/2014) and in the follow-up discussion meeting held by the NCIP NanoWG to develop DRL feedback (5/8/2014), we assumed it was applicable to all data and all purposes. From a biologists perspective there may be some critical aspects missing that could preclude utility and comparability of datasets without being addressed (animal models, etc.)
* Societal perspective question – should/do the DRLs include any aspect of data relevance in addition to data completeness, maturity? (e.g. relevant to informing consumer risk information)? This feeds into questions of data purpose.
* Should there be a separate set of criteria for the scalability, the scientific relevance, the reduction of uncertainty so that we don’t try to “normalize” across all and lose the granularity.
  + You could then delineate the DRLs for these various aspects separately
  + Is there some analogy between this effort and current efforts to assess relative data quality for toxicological data quality
* Perhaps workflow for assigning DRLs might be developed? There seems to be an analogy between assignment of toxicological data “reliability”. Klimisch [http://dx.doi.org/10.1006/rtph.1996.1076], ToxRTool [http://ihcp.jrc.ec.europa.eu/our\_labs/eurl-ecvam/archive-publications/toxrtool] and fuzzification of ToxRTool [http://dx.doi.org/10.1002/minf.201200082]
* Is there a tension between the maturity of the data and the purpose for which it is being utilized, which may have different quality/maturity requirements? Invalid data, for example, is an awkward phrase as there may be a great deal of data-in-waiting that contain information, but are difficult to interpret.  What may be invalid to one group (database) may be valid to another.  Who determines “of little practical value.”
* The aspect of relative establishment, or readiness levels and granularity of specific assays that are used to generate the data also feeds into the resulting datasets 🡪 do particular methods with more detailed protocols associated automatically confer higher DRLs
* Some fields of science are so protocol driven that by nature of utilizing the established protocols they have generated data that will earn higher DRL s (e.g. in physical chemistry you would report pH but in biology you may include the protocol for having assessed pH).
* The nanoWG has grappled with issues of having a single database or a system of federated databases sharing a file transfer protocol. The role of the curator is localized in the case of federated databases, but the value added is the ability to tailor metadata to the field (or focus or purpose) specific to that database. Characterizing metadata as “poor” based on “failure to include” or “ambiguity in interpreting the data” may be best decided at the local level.

***Suggestions***

The NanoWG feedback is that the DRLs, in order to be truly useful in practice, either instead of or in addition to having a single aggregate DRL scale, they need to also be set up either in a matrix or a decision tree format so people can incorporate their purposes and perhaps their more specifically defined needs for detail in assessing how ready a dataset is for what they need.

1. Context dependency is important to consider – one might say that how “ready” the data are depends entirely on what the data must be ready for.

* Generally speaking – does this mean comparability to other data sets (perhaps the current lens)
  + Currently this seems to stress analytical chemistry and perhaps doesn’t address the parameter needs
    - OR is Analytical Chemistry always a 6 and biology is always a 3 by nature?
  + This gap in data readiness, between biology and analytical chemistry, was the stark realization of the ASTM’s inter-laboratory study.  It also exists in the Nanomaterial Registry, which has a compliance score for physico-chemical data, but not for the biological or environmental studies.
* DRLs across multiple decision purposes (or groupings of purposes that share a similar comparability requirements, granularity of raw data definition, uncertainties, etc.)
* DRLs across multiple fields with different comparability requirements

1. Perhaps a decision tree would be helpful in this context.

* What is your purpose for using these data? (e.g. ranking for research prioritization vs. QSAR calculations)
* Once the purpose is chosen, perhaps the discipline and associated
* Decision tree plus distinct levels for separate concepts of quality, usefulness, accessibility (licensing)?

1. Could also draw on the existing DRLs in other *scientific* fields (in addition to the influence of the DRLs from technology readiness), including genomics, in order to preserve the concept of an aggregated single scale to assess the overall readiness of a dataset in a single number.
   * These traditional levels have been established within genomics/next-gen sequencing, established bioinformatics data levels) <https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp>
2. Publish a glossary along with the DRL table to have a simple way for users to understand the precise meanings of terms used (see questions section above).

***DRLs and associated definitions, explanations will be most useful if they address:***

* How transportable are these data to other purposes from the ones for which they were intended? How easy is it for users to understand and apply these DRLs in utilizing and sharing data?
* Education and communication for data providers about what is the added value of the data to the wider/broader community
* Expressing the difference between disciplines, recognizing that different disciplines are at different levels of transportability or aggregation in data sharing
  + Not quite “fair” to compare the readiness directly between vastly different fields
  + Communication between your discipline might be important, communication beyond your discipline might be another level (this is captured in the concept of “related to the larger bod of scientific knowledge) addressed in DRLs 4 and 5.
* DRLs could provide a platform for controlled data sharing, a framework for describing datasets and the willingness to share. Perhaps once established this could facilitate increased data sharing.
* Would be good to include multiple aspects of defining availability of the data; not simply addressing whether the data could be shared if someone gave you permission, but also whether it’s licensed so that you know whether you can utilize it for whatever you want sans permission, etc.