**Background information**

*The following, broad definitions of data “completeness” and data “quality” are provided to clarify the scope of this questionnaire. A key aim of this questionnaire is to identify how these concepts are (or should be) treated in practice i.e. which specific issues are (or should be) considered.*

*Data “completeness” is a measure of the availability of all ‘necessary’ kinds of (meta)data / information which may include the extent of nanomaterial characterisation, both physicochemical/structural and biological, the degree to which experimental details are described as well as the availability of raw data, processed data, or derived data from the assays used for nanomaterial characterisation. N.B. [1] For the purposes of this questionnaire, data may be considered to be “complete” if they are compliant with some set of “minimum information criteria” - although definitions of “completeness” which go beyond “minimum information criteria” are also of interest. [2] The use of the term “completeness” is not meant to suggest that we completely understand all necessary independent variables - it is understood that definitions of completeness will evolve in tandem with our scientific understanding.*

*Data “quality” is a measure of the usefulness of data which encompasses both its inherent “reliability” (i.e. clarity regarding exactly what is being reported and trustworthiness/reproducibility) as well as its “relevance” (i.e. usefulness for a particular purpose). It may be considered related to data “completeness” and may encompass issues such as precision, error, sufficiency of meta-data for reproducibility etc. N.B. This concept may be considered both in a qualitative and/or quantitative sense.*

*N.B. It is recognised that the information which might be required for both “completeness” and “quality” might be use case or data type specific. Where this is the case, please indicate this insofar as possible in your responses.*

**Questions**

Section A: What is meant by data “completeness” and “quality” and why are these issues important?

1. Do you have any comments regarding the broad definitions of (a) data “completeness” and (b) data “quality” provided above? *Please note that detailed considerations (such as which physicochemical parameters to measure or checklists for assigning data quality) should be addressed in subsequent questions.*

*The definition for data completeness looks good. It may be beneficial to reference data validation in the definition of data quality. Data validation could be a quality measure for verifying both the format (e.g. conforms to an XSD or other specification) and content of the data (e.g. passes validation in specific software analysis tools).*

1. Please briefly comment on why (a) data “completeness” and (b) data “quality” is important in the context of nanomaterials. *For example, data completeness may be a pre-requisite for database interoperability?*

*Data completeness is important in support of data re-use and interoperability, reproducibility of the experiment, and to enable nanomaterial comparison.*

*Data quality is important to ensure that the data re-used and made interoperable is accurate; otherwise, the ability to reproduce the experiment and compare nanomaterials is greatly hindered.*

1. Please tell us about important, existing proposals for data “completeness” (or “minimum information criteria”) for nanomaterials that you are familiar with. *N.B. Existing proposals would include the* [*MINChar Initiative Parameters List*](http://characterizationmatters.org/parameters/)*, those implied by recent OECD reports such as* [*ENV/JM/MONO(2012)40*](http://www.oecd.org/officialdocuments/displaydocument/?cote=env/jm/mono(2012)40&doclanguage=en) *and* [*ENV/JM/MONO(2009)20/REV*](http://www.oecd.org/science/nanosafety/guidancemanualforthetestingofmanufacturednanomaterialsoecdsponsorshipprogrammefirstrevision.htm)*, as well as the proposed consolidated lists of physicochemical properties presented in* [*Stefaniak et al. 2013*](http://dx.doi.org/10.3109/17435390.2012.739664)*.*
   1. Are you familiar with the mentioned proposals?

*We are familiar with MINChar.*

* 1. Are there any others you feel are particularly important?

*caNanoLab produces a data availability matrix assessing the completeness of the nanomaterial based on both MINChar and caNanoLab identified metadata. The caNanoLab identified metadata illustrates information criterial for nanomaterial composition and specific characterizations. The Nanomaterial Registry publishes a minimal information standard (MIAN) used to calculate compliance in their repository. ISA-TAB-Nano provides a standard supporting data exchange and does provide guidance for minimal information associated with Investigations, Studies, Assays, and Materials, in collaboration with the ISA-TAB standard.*

1. Taking into account the proposals discussed above, which specific (meta)data, or “items of information”, are necessary for nanomaterial data to be considered “complete”? For each “(meta)datum” or “item of information”, please indicate
   1. why the (meta)datum is important?
   2. is the (meta)datum only important in particular contexts e.g. for particular experiments, nanomaterials or use cases? *N.B. Different use cases might be an experimentalist wishing to reproduce the results vs. a QSAR modeler etc.*
   3. is it possible to rank the (meta)datum as more or less important that any other in your list?

*N.B. 1. Please enter this information in the Excel spreadsheet you should have been provided with. 2. You may wish to refer to Tables II and III in* [*Stefaniak et al. 2013*](http://dx.doi.org/10.3109/17435390.2012.739664) *as a starting point. 3. Feel free to comment on metadata which would be valuable even if it is not currently available to you at the moment.*

*Completed via the attached spreadsheet.*

1. What additional/alternative considerations need to be accounted for when assessing nanomaterial data “quality”?

*N.B. Insofar as possible, please provide this information in the same Excel spreadsheet you should have been provided with for question A.4. However, please provide some additional, free text comments if you feel this spreadsheet does not adequately cover the points you wish to make.*

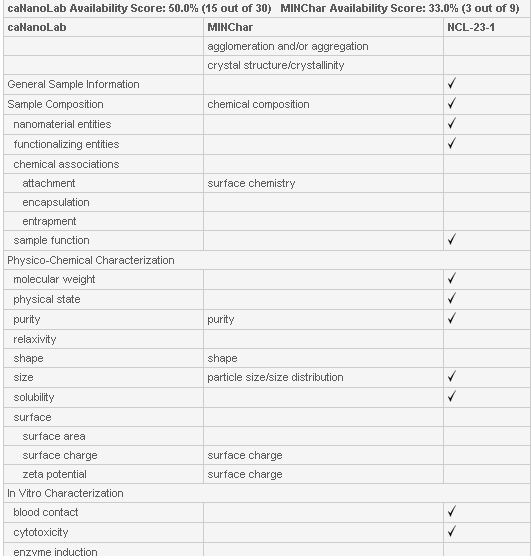
*Completed via the attached spreadsheet.*

Section B: How does the purpose of your specific nanomaterial data curation effort impact how you define/assess data “completeness” and/or data “quality”?

1. As part of your curation efforts, what formalised definition of/set of assessment criteria are you using for
   1. data “completeness”?
   2. data “quality”?

*N.B. If you do not have formalised definition(s)/set of assessment criteria, please state “None” for these questions. If the definition(s)/criteria are publicly available, please provide a link i.e. URL.*

*(a) For data completeness, caNanoLab is defining data completeness based on the mapping between submitted data against the MINchar standard and supported metadata in caNanoLab. A screen snapshot of the data availability matrix is provided below. Please note, while caNanoLab provides a calculated percentage for data completeness, it is up to the user to determine what their accepted level of completeness is when querying the data.*

**

*(b) For data quality originating from the source (e.g. publications, laboratory), the caNanoLab curator corresponds directly with the data providers to address questions or discrepancies. For data quality once submitted into caNanoLab, caNanoLab relies on feedback from the user community on any identified issues. Currently, caNanoLab does not provide any automated scripts to verify the data.*

1. If you or your organisation are using a (set of) formalised definition(s) or assessment criteria, please provide the following details. *N.B. Please provide details for both your data “completeness” and data “quality” definitions if applicable.*
   1. What is the basis for your definitions/assessment criteria? *For example, were they informed by, say, evidence in favour of the critical significance of a particular experimental condition in toxicology studies?*
   2. Are you aware of any limitations of your definitions/assessment criteria?
   3. To what extent do the specific goals of your organisation and/or purpose of your data resource/tool affect your definitions/assessment criteria?
   4. To what extent could your definitions/assessment criteria be generalised for use by other organisations involved in curating nanomaterial data?

*(a) Data completeness was informed by the NCI’s Nanotechnology Characterization Laboratory’s (NCL’s) assay cascade (*[*http://ncl.cancer.gov/working\_assay-cascade.asp*](http://ncl.cancer.gov/working_assay-cascade.asp)*) along with the MINchar standard. Data quality was primarily informed through curator analysis identifying discrepancies within the publication and clarification of uncertainties.*

*(b) The primarily limitations are the need to assess data completeness at the measured value. For example, in MINchar, agglomeration or aggregation is listed as important for the standard; however, there is no specific measurement cited in the standard for agglomeration or aggregation. This hinders nanomaterial comparison.*

*(c) caNanoLab focuses on supporting the biomedical research community. As such, definitions/assessment criteria are geared towards biomedical research.*

*(d) caNanoLab provides definitions/assessment criteria for the composition of the nanomaterial which can be used by other organizations outside of the biomedical research domain.*

Section C: What are established handling methods for addressing data “completeness” and “quality” in mature fields (e.g. biocuration)?

1. Are you familiar with specific established approaches for addressing the following issues exist in mature fields (e.g. biocuration/bioinformatics):
   1. data “completeness”?
   2. data “quality”?

*N.B. Please feel free to simply provide links to publications you feel are of particular significance.*

*(a) In genomics, the approach for data completeness can be illustrated in the TCGA data types and formats (*[*https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp*](https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp)*). Specific data types are required for diverse levels of data (Level 1, 2, and 3:* [*https://wiki.nci.nih.gov/display/TCGA/Data+level*](https://wiki.nci.nih.gov/display/TCGA/Data+level)*). In proteomics, the Protein Data Bank (PDB) is used for representing the complete structure of the protein.*

*(b) In genomics, validation tools are provided in support of data quality. For example, raw sequence files (BAMs) are validated using tools like Picard and are aligned against a reference genome using several different alignment tools (e.g. BWA-MEM). These tools help support data quality. For clinical and biospecimen data, XSDs are created and files are validated against the XSD.*

Section D: What are the key challenges related to the “completeness” and “quality” of nanomaterial data?

1. To what extent do the approaches established in mature fields (e.g. biocuration/bioinformatics) need to be modified in order to be applied in the context of nanomaterial data? Please comment on this with regards to
   1. data “completeness”
   2. data “quality”

*(a) For nanomaterial data, there are hundreds of assays. As such, standards need to be developed to establish completeness for each assay. Additionally, there needs to be established standard protocols applied to these assays. Also, for nanomaterial data as opposed to genomics, there is significant variability due to the polydisperse nature of the particle.*

*(b) For nanomaterial data, there are very few validation tools that can assess data quality. As such, the analysis of data quality is manual and subjective based on the limited information provided in publications.*

1. What are the outstanding challenges for nanomaterial data with regard to
   1. data “completeness”?
   2. data “quality”?

*See response to #1 which also provides some of the challenges.*

1. Which of these key challenges are specific to the goals of your organisation and/or data resource with regard to
   1. data “completeness”?
   2. data “quality”?

*All of the challenges outlined in response to #1 are specific to the goals of our organization and data resource.*

Section E: Are there any specific use cases to illustrate these issues and make them tangible?

1. Please provide examples of case studies and/or scenarios which illustrate the key challenges (as noted above) for nanomaterial data with regard to
   1. data “completeness”
   2. data “quality”

*(a) When the curator curates publications, there are several different types of assays with different measurement values. Additionally there are similar assays (e.g. caspase apoptosis) from different organizations that have different measurement values. One of the challenges is that without standard protocols, organizations produce different measurement values for the same types of assays. Without standard protocols, there is no reference to enforce standardized assay measurements.*

*(b) In terms of data quality, an ISA-TAB-Nano validator can be used to validate the structure of the ISA-TAB-Nano files. The use case for validating content is currently the curator reviewing the publication and submitting questions to the author to obtain clarification and additional details on figures/images.*

Section F: Recommendations. What are some practical next steps for individual stakeholders or the community as a whole?

1. To what extent do you feel there is (a) redundancy in nanomaterial data and (b) how best can this be addressed? *For example, might computational predictions be employed to substitute for missing physicochemical characteristics (PCCs) based on a subset of measured PCCs?*

*There is some redundancy due to the lack of interoperability between systems. For example, the Nanomaterial Registry exports and imports data from caNanoLab rather than directly interfacing with caNanoLab via web services. The community needs to agree on a set of common web services leveraging metadata from ISA-TAB-Nano. Additionally, the community needs to establish a common service for registering and retrieving nanomaterial identifiers.*

1. Taking into account all of the previous questions, are you able to recommend specific definitions/assessment criteria, which should be adopted by the community as a whole, or for specific scenarios, for
   1. nanomaterial data “completeness”?
   2. nanomaterial data “quality”?

*N.B. You may wish to simply propose either that (a) your own organisation’s criteria should be universally adopted (or adopted in specific contexts) or (b) that a scheme based upon compliance with the “wish list” you proposed in response to question A.4 should be applied.*

*We recommend that the community as a whole adopt the caNanoLab definitions/assessment criteria for nanomaterial composition. We recommend the biomedical community adopt the caNanoLab definitions/assessment criteria for physico-chemical, in vitro, and in vivo characterization.*

1. Are you able to recommend how to best capture the information required for “complete” data within, say, a database? *For example, might templates based upon pre-defined fields in* [*ISA-TAB-Nano*](https://wiki.nci.nih.gov/x/MwGGAg) *files be employed in some fashion?*

*We recommend that the community develop a list of data types for nanomaterial assays similar to TCGA (*[*https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp*](https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp)*) but leverage the ISA-TAB-Nano Assay File to specify the level 2 and 3 data formats. Standard ISA-TAB-Nano Assay Files with agreed upon measurement values can be created for each assay type.*

1. Are you able to make any recommendations regarding
   1. how best to implement a scoring scheme for “quality” and/or “completeness” of nanomaterial data in practice? *For example, a human expert might (a) read a nanotoxicology study and score the data using a set of predefined questions as (b) implemented in an extension of the* [*ToxRTool*](http://ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam/archive-publications/toxrtool)*.*
   2. what challenges would be associated with implementing such a scheme?
   3. existing approaches which could be extended?

*(a) For completeness, a scoring scheme can be developed leveraging the approach caNanoLab takes for developing a % complete based on the caNanoLab and MinChar standard; however, the % complete could be developed based on the agreed upon community standard leveraging the list of assays that the community could develop in recommendation question 3. For quality, scoring is complex as some quality characteristics are subjective; however, validation tools may be developed for specific files types or aggregated files and a score associated with the validation tool can be obtained. For example, in the genomics field, DNA and RNA sequences are aligned against a reference genome and scored by the tool used to perform the alignment.*

*(b) The answer for (a) addresses this. The challenges are the community agreeing on a standard for calculating % complete and the availability of validation tools for each assay type.*

*(c) The answer for (a) identifies re-use of the caNanoLab approach for data completeness and the genomic approach for validation tools.*

1. Do you have any suggestions regarding a dataset or data resource which would be suitable for a pilot study for addressing these issues?

caNanoLab

1. Do you have any recommendations from a higher-level perspective (i.e. to funding agencies, researcher associations, publishers etc) on how to move forward?

*Recommend that the community develop a list of data types for nanomaterial assays similar to TCGA (*[*https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp*](https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp)*) through the Nano WG and divide the assays amongst the different organizations to start developing example ISA-TAB-Nano Assay files. There could be an initial workshop with a goal of developing the standards for a “Nanomaterial Data Commons” leveraging the TCGA approach. We have attached an initial and incomplete draft table entitled DRAFT Data Types for Nanomaterial Assays illustrating this.*