**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 fully appreciate all necessary independent variables which determine, say, a given result obtained from a particular biological assay - 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 their inherent “reliability” (i.e. clarity regarding exactly what is being reported and trustworthiness/reproducibility) as well as their “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 metadata 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 your responses may be considered specific to a given use case and/or type of data/study, please indicate this 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.*

*These definitions are in line with the thinking at CEINT with regard to development of the CEINT-NIKC resource. In particular, the note at the end is a critical part of our position, because the completeness and quality of data in terms of utility to our project is directly based on their ability to support the purpose/goals of our project. Based on the research question we may need to interrogate the data for relationships between specific parameters so we would need a dataset rich in data on those parameters. In this sense the data completeness from our standpoint may be most useful if defined in relation to specific queries. A dataset with a high completeness or quality score might have radically different utility for use in one query vs. another.*

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 prerequisite for database interoperability?*

*Agree with the example on interoperability. These are critical concepts because they allow us to know whether we can compare and combine datasets or not; doing so is the basis of an informatics approach, and defining what these concepts mean in the context of an ever-changing and still-emerging science is a challenge.*

*Although not specific to the context of nanomaterials, since one of our primary goals is to perform data analysis across multiple studies, we need to incorporate metadata that will allow us to group data together appropriately. We are still in the process of determining what metadata must be collected for each analytical method to allow us to appropriate group data from multiple studies together, but we expect that some parameters will be specific to the analytical method.*

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?

Yes.

* 1. Are there any others you feel are particularly important? *N.B. Feel free to simply provide a set of links to references without additional comments.*

http://www.codata.org/nanomaterials

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

From the CEINT-NIKC perspective, the most important items for the field to focus on are the environmental media preparation protocols. In this context “environmental” refers to whatever medium the measurement was made in; it could be lab based, biological We are responding with these as a lumped category of (meta)data because to date these have been often underreported or not mentioned in the data, rendering physical and chemical characterizations of nanomaterials nearly incomparable with each other and unable to be compared with other scenarios of interest.

* 1. why is the “(meta)datum” or “item of information” important?

Many of the physical chemical characteristics listed in these resources are not intrinsic to the material but rather extrinsically defined by the surrounding media. Therefore if careful and consistent preparation and reporting of the test media are not included, based on the physicochemical characteristics alone, the nanomaterials are unable be compared or integrated. Thus, the research questions on environmental and health impacts and material performance driving our mission cannot be addressed. These data are important because having them is what allows us to determine how to appropriately group and compare data.

* 1. is the “(meta)datum” or “item of information” 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 modeller etc.*

*Since our goal is to build a data analysis tool, we believe test media characterizations are important 100% of the time.*

* 1. is it possible to rank the “(meta)datum” or “item of information” as more or less important than any other in your list?

NA

* 1. whether the “(meta)datum” or “item of information” should be considered “essential” i.e. a necessary component of “minimum information criteria” as opposed to merely being required for “data completeness”?

*We believe minimum information criteria are only meaningful with respect to specific analytical purposes. Based on this definition of minimum information criteria, yes we believe test characterizations are essential to characterizing nanomaterials.*

*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) *, or the provisional NanoSafety Cluster Databases Working Group “*[*minimal standard for reporting*](http://www.nanosafetycluster.eu/working-groups/4-database-wg/tasks-2/2013-2.html)*” as a starting point with regard to physicochemical parameters and/or experimental variables (for in vitro studies) respectively. Table II of* [*Stefaniak et al. 2013*](http://dx.doi.org/10.3109/17435390.2012.739664) *might be considered a starting point for a “data completeness” proposal whilst Table III might be considered a starting point for “minimum information criteria” i.e. a specification of the highest priority “(meta)data”. [3] Feel free to comment on (meta)data or information which would be valuable even if it is not currently available to you at the moment. [4] Please provide some additional, free text comments if you feel this spreadsheet cannot adequately cover the points you wish to make. [5] Please indicate whether you feel your response has been comprehensive or not.*

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

*N.B. [1] Insofar as possible, please provide this information in thesame Excel spreadsheet you should have been provided with for question A.4. [2] Please provide some additional, free text comments if you feel this spreadsheet cannot adequately cover the points you wish to make.*

*We are focused on collecting measurements and observations made while characterizing nanomaterials and experimenting with nanomaterials. We have not focused on the manufacturing of the nanomaterial. As we move forward, it will be very important to be able to connect to a database that captures information about the manufacturing process.*

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”?

*After we have loaded data into the CEINT NIKC, we perform a series of checks to assure that each reported measurement or observation can be uniquely identified from all of the others in the database based on our required fields. We also run ad hoc queries, relevant to the specific data, that allow us to see obvious data gaps.*

* 1. data “quality”?

*Utility of data will be determined by relevance to our queries, and the completeness and quality would be defined relative to our custom structure. One of our primary goals in populating the CEINT NIKC is consistency. We have established and documented our protocols for populating the database, and are using established Web Ontology Language (OWL) descriptions to populate our fields whenever possible. We hope to develop more tools to identify anomalous data in the future.*

*N.B. [1] If you do not have formalised definition(s)/set of assessment criteria, please state “None” for these questions. [2] If the definition(s)/criteria are publicly available, please provide a link i.e. a web address.*

1. If you are using a (set of) formalised definition(s) or assessment criteria, or your organisation is, please provide the following details. *N.B. Please provide details for both your data “completeness” and data “quality” definitions if applicable.*

*Because our initial goals are to interrogate the amassed data in support of our driving research questions, our definitions exist in the form of our curation protocols; we look for relevant literature and research as a deciding factor for whether to spend resources curating the data in the first place.*

* 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?*

*The critical significance of consistent environmental media and nanomaterial characterizations drives our decision to curate data – without this the ability to compare datasets and to analyse the effects of both intrinsic and extrinsic material characteristics on behaviour and impact of the materials is lost.*

* 1. Are you aware of any limitations of your definitions/assessment criteria?

*While we have built placeholders into our database to indicate where would include the method specific metadata we think we will need to appropriately group data together when performing meaningful multi-study data analysis, we still need to develop a queryable methods compendium to be able to assess the appropriateness of grouping certain measurements together with other measurements.*

* 1. To what extent do the specific goals of your organisation and/or purpose of your data resource/tool affect your definitions/assessment criteria?

*Entirely, though we also believe that this level of fundamental description will support investigation of other mechanisms of importance within nanomaterial behaviour (e.g. product design and performance).*

* 1. To what extent could your definitions/assessment criteria be generalised for use by other organisations involved in curating nanomaterial data?

*See above answer to C.*

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, that exist in mature fields (e.g. biocuration/bioinformatics), for addressing:
   1. data “completeness”?
   2. data “quality”?

*Yes – we anticipate such formalized processes and metrics may become appropriate for CEINT-NIKC at later stages of development with a larger user community, more curators and fully established structure.*

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

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”

*Quality will always be defined relative to purpose in our view. Completeness will need to be modified based on nano-specific data categories and still-emerging measurements that describe their behaviour in systems. A mechanism for building flexibility into defining what “complete” datasets include will be necessary since measurements and descriptive parameters are still being discovered and developed.*

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

*See above answer to D.1.*

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”?

*These are field-wide challenges, highlighted by CEINT-NIKC goals but not unique to them.*

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”

*An issue that has arisen in populating the CEINT database is that, since we are not to the point where we have an established format in which we request data from our researchers, we sometimes get data in a variety of formats. For example, in one situation, researchers reported the results of spiking a soil reference material with silver as a percentage of silver recovered and as a concentration of silver found in the reference material. Since we are building an analysis tool, if the same information is included in the database in more than one format, we need to have a way of making sure that the same information is not erroneous used more than once in our computation tools. In a data repository where we are just searching and returning records for review, this is not so much of a concern, but we must be careful not to bias the quality of our analysis by not having access to the appropriate metadata.*

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?*
2. 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”?

*Our recommendations for both completeness and quality are in the realm of open communication and sharing of parameters included in resources, their definitions, data types and units and collection and reporting protocols. The NDCI is a good start toward this openness but funding for a large Center that would sort through the developing resources and their structures would help to advance communication and work toward whatever level of standardization is possible.*

*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 questions A.4 and A.5 should be applied.*

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 (and do ourselves in practice) utilizing pre-defined fields wherever appropriate within our system. Also, developing further fields and data-sharing tools to contribute to ISA-TAB-Nano, as was their project intent, would be of great assistance to the community. Perhaps the establishment of specific ISA-TAB-Nano working groups in specific areas, such as ecosystem effects measurement, would help formally advance and define data completeness and quality by narrowing to specific data type and hence, purposes.*

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 (1) read a nanotoxicology study and score the data using a set of predefined questions as (2) implemented in an extension of the* [*ToxRTool*](http://ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam/archive-publications/toxrtool) *program.*

*We strongly recommend a fit-for-purpose scoring system. Both completeness and quality could be scored relative to the utility of a dataset for a specific application. A sensitivity analysis of a specifically defined nanomaterial characteristic across multiple test media would require different data than an analysis of the effects of various characteristic values on optimal material performance, and a user of the data would be interested in how to select and access those data that could further their research, rather than simply grading who had the most entries and choosing the datasets with the most information.*

* 1. what challenges would be associated with implementing such a scheme?

*Selecting the granularity of the categories of purpose would be a challenge, as would achieving any consensus about what comprises complete data. Perhaps initial goals could include a case by case definition of completeness generated by the user generating the query (using check boxes for the various data elements that one would want to see). This type of approach would also gather data in the process about what is most often desired. Further, NanoHUB has already developed methods to gather such meta-data about user groups and their data needs; perhaps these could be leveraged as they have proven a willing and highly collaborative partner.*

* 1. existing approaches which could be extended?

*See end of F.4.b.*

1. Do you have any suggestions regarding a dataset or data resource which would be suitable for a pilot study for addressing the issues raised in this questionnaire?

*In 2016-17, CEINT-NIKC would be ready to serve as a public study, perhaps along with the NBI.*

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?

*See response to F.2.*