Addendum to Agenda: For Use with Breakout Discussions

Breakout Session 2: Key Questions related to supporting Data Submission

1. Data submitters know more about their data than the repository owner. The information required to be submitted for large scale data sets, can miss some of the data originator's knowledge.
    a. How can a semantic infrastructure compliment what the researcher is already doing? How can it capture what the researcher wants to say about their data?
    b. Should we be shifting from an approach based on "these are the columns you must provide" to "what are the columns that define your study" or "what are the columns of data in your study". Is this approach feasible in terms of supporting query and aggregation? If so what are the approaches that could work? Pros/cons?

2. What support and services are required to minimize the effort for data owners to prepare data for submission?
    a. Should data owners be expected to split apart and recode their own data for submission? What would help facilitate this task?
    b. What approaches have people used, what works, lessons learned?

3. Other considerations for approaches to support metadata annotation for the CRDC:
    a. How to reduce the burden for producing the various types of metadata annotations that are needed (experiment, field, data values (1s, 2s)? i.e. PDC is planning to use constrained templates to make it easier and more consistent when submitting data.
    b. What types of semantic metadata are highest priority and who and how should it be created (e.g. Data owner, Repository Owner)?
    c. Should researchers re-using data for analytics be able to update the source data's metadata?
    d. What are some semi-automated approaches to facilitate the different types of metadata annotation with standards or common metadata? How might they be evaluated?
    e. Should support for metadata annotation and validation be available on demand? Versus prepackaged as part of a predefined process (e.g. in advance of data submission or only during submission)? What is the best approach? Prons/cons of each
    f. What kinds of metadata annotation support should CDA provide? (experiment metadata, field metadata, meaning of data values, etc)

4. What approaches are needed to support metadata validation?
    a. Who is responsible for validation?
    b. When should validation occur? For example, for GDC we are providing "early metadata validation" for terminology / metadata, supporting users in finding the

appropriate concepts and data elements to represent their data. The TCIA team is considering offering templates for users to insert their data that contain the data constraints up front, vs the templates GDC offers which are just column headings.

5. What approaches are needed to support transformation and mapping, and when is it best to use?
    a. If data are not pre-harmonized by using common standards for data capture, who is responsible for mapping and transformation in preparation for submission? Can metadata annotations help? If so, how?
    b. Should mapping and transformation (of submitted data) happen at time of submission or runtime? e.g. 1, 2 vs M, F, or if annotated with diff vocabularies such as ICD-9 or 10)
        c. Should CRDC provide tools for researchers to use to map and transfomr their existing data into format required by a particular CRDC node vs individual data repository/CRDC owners)?

6. Should the CRDC provide human curation support/services for data submission?
    a. If so, at what point in data submission process and who should provide it?
    b. Are there pros/cons of centralized support?
    c. What are the pros/cons of each node providing this support?