Addendum to Agenda: For Use with Breakout Discussions

Breakout Session 1: Key Questions related to Semantics for Query and Aggregation

1. What <u>data standards, terminology standards and approaches</u> are needed to support Query and Aggregation? Considerations for discussion:
   a. Make it easy for users…
   b. What fields are likely to be collected across the Commons for effective data aggregation and analysis? Should there be vocabularies for certain domains, particularly where data standards already exist? For example: Diagnosis, Disease, Procedure
   c. What are approaches for discovery of these core elements? and how to handle extensions? Pros/cons of each
   d. What is needed to support Harmonization of data? Field level? Data value level? Common terminology/ontologies; Which terminologies give us the best coverage and what fields? What criteria should be used to determine which ones to use for specific data elements? Should we rely on mappings versus setting 'preferred standards'? What about fields where there are no de facto standards?

2. What is the appropriate balance between <u>Prescribed (Static) vs Resolved at time of Query (runtime) semantics</u>? Considerations for discussion:
   a. What are the approaches for resolving semantics at runtime? Pros/cons of each
   b. Prescribed: establish specific vocabularies for specific domains for data submission (e.g. Diagnosis)?
   c. If supporting multiple vocabularies, 'how' (and when) should we map between them? Is this something the CDA needs to provide?

3. In addition to terminology, what about underlying data model(s) to drive query, aggregation and harmonization/mapping across nodes? Data model(s) should be unseen as well as dynamic and extensible. They must also be independent of a particularly implementation. Considerations for discussion:
   a. Example: BRIDG – conceptual model harmonized with many other models, cancer projects and CDISC Data Standards. to serve as a central semantic model to support queries? BRIDG is conceptual, it does not have terminology binding for how to code the data, so many data models using different terminology binding can be mapped to the same entity e.g. Person Gender.
   b. Should there be an agreed core data model implemented by all to facilitate CDA ability to connect nodes and support interoperability?
   c. Consider semantic description of relationships and alignment.
   d. Existing model alignment; consider model landscape analysis.
   e. Best way to incentivize researchers to adopt existing common standards and models?

4. What approaches might CDA use to facilitate query across nodes? For example, should we consider adding NLP approaches to facilitate query?
    a. What approaches have been used to query across disparate data sets? What has worked and what has not? Pros/cons of each.