DD-R package (Data Description in R) Inaugural Discussion

Attendees: Darren Bell, Dan Gillman, Larry Hoyle, Gregg Kellogg, Daniella Meeker, Flavio Rizzolo, Achim Wackerow, Michelle Edwards

Purpose of this group:

1. To make it easier for users to use DDI and carry metadata about the data in their own analysis environment – R will be the primary case, with Python as a second candidate.
2. Review current CSVW interoperability options in R

Background:

Current state of affairs, we have DDI4- XML schema as a preservation and exchange platform, OWL/RDF-S as a discovery layer, Java for processing, JSON-LD for web development, and R as one of the primary analysis packages used by researchers. Generally speaking, researchers are not in the habit of carrying their metadata with their data and most statistical analysis packages have very little metadata options available with no content metadata capabilities. The goal of this group is to enable the development of an R package that allows content metadata to be carried with the data which would increase the efficiency of the researcher's work. By creating a DD-R package, we would make it easier for researchers to use their metadata (in DDI) in the environment they work, R, essentially bringing DDI to their working environment.

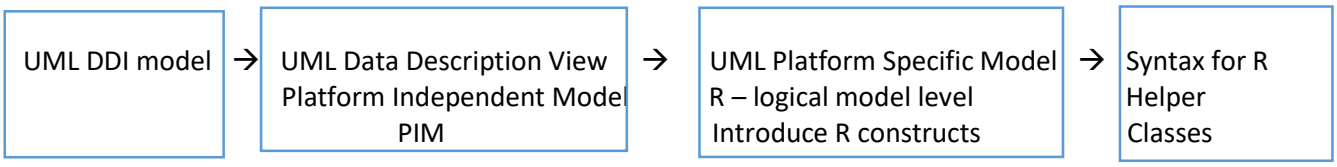
Discussion on different approaches on how this could be accomplished:

Three different approaches were discussed on how best to create the DD-R package:

1. One approach would be to create an R package independent of DDI, whilst taking advantage of the datum concept developed by the DDI-Data Description group.

   This approach would be very involved and would need a clear understanding on how the datum theory in the current DDI4 Data Description works.
   - This approach would be broader than DDI and would be broader than originally anticipated by participants initially thought.

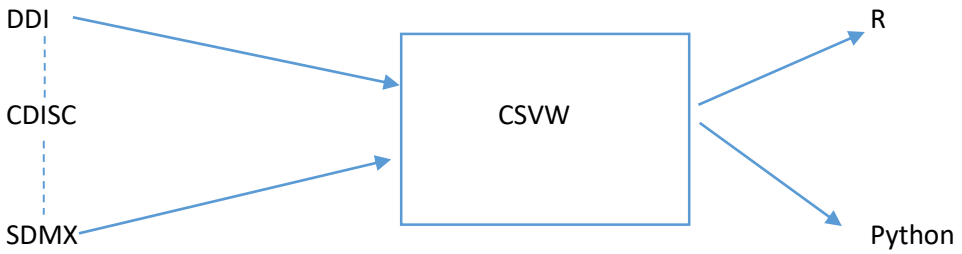
2. A second approach is to create Platform Specific Model (PSM) binding for R



   - Build R constructs that could hold content metadata
   - Only generate a structure for R – a blank data frame with variables

- Currently R cannot hold metadata such as Unit of Measure (UOM), could extend a vector in R to contain UOM

3. A third approach is to work with the CSVW (CSV on the Web) model



- All specifications use tools to transfer to CSVW
- This would allow a separation of concerns
- FHIR can currently handle many statistical analysis packages, since it sends metadata to something other than the package → should review
- CSVW
    - Common properties
    - Annotation
    - Arbitrary RDF – no conformity
    - Variables are RDF properties

A brief example:

ID	SEX	SEXRESP
1	M	

- RDF uses the predicate subject object – triple.
- We can combine the 2 columns in the dataset SEX and SEXRESP using RDF to allow us to carry metadata with the cell value
- In this case, we would have metadata about the cell value and NOT the variable per say

Action items to move this project forward:

1. Identify the requirements needed for the DD-R package.
2. There are currently 4 DDI packages in R. A thorough review of these packages should be a first step. If there is a package that already exists that we can build on, then we should consider this.
3. Gain a firm understanding of how R currently handles metadata, structural as well as content metadata. At what level? Cell, variable, dataframe?
4. Identify resources required to create the DD-R package, after reviews of items 2 and 3 are complete.
5. Identify any funding opportunities that can assist with the creation of the DD-R package.