

C2Metadata Project Meeting Notes March 3, 2017

Agenda

1. Announcements (All)
2. WG1 Target Transformations for Year 1
 - A. Missing values (George, Jie, Sanda)
3. Software Developers Group
 - A. Modelling RECODE (Ornulf, Jeremy)
 - B. Variable Comparison tool (J)
4. WG4 Outreach (Jared)
 - A. NADDI presentations
 - B. IASSIST presentations

Related documents:

Recording of the meeting:

<https://drive.google.com/open?id=oB5stM2UCJoPYVjVReGZrNkxTZDQ>

Ornulf's slides:

<https://drive.google.com/open?id=oB5stM2UCJoPYUo1TVFdEQ3F3SFhEYWhTXXZsUFMzSG8yTC1Z>

1. Announcements (All)

2. WG1 Target Transformations for Year 1

A. Missing values (George, Jie, Sanda)

George asked if it would help the developers to have some reference examples of commands translated into VTL. This question was tabled for discussion later.

Sanda and Jie are working on a list of functions in the 4 languages and VTL, so that we can find similarities, differences, and gaps. Jie has found that VTL lacks many functions, and we need to prepare a strategy for dealing with this.

3. Software Developers Group

A. Modelling RECODE (Ornulf, Jeremy)

Ornulf and Jeremy demonstrated their applications for parsing RECODE statements in Stata and SPSS. Information from the original statement is captured in a JSON file. Ornulf explained that we are developing a JSON schema for each type of statement, which captures the structure and rules in the statement. The JSON files are outputs from the AST (Abstract Syntax Tree) approach that was described at the last meeting. JSON captures information about each statement in a structured format that is easy to work with. This will make moving to the next step (e.g. VTL) much easier.

NSD and Colectica are building a library of JSON schemas for various commands in Github. The Stata and SPSS RECODE statements are very similar, and they may be able to use a single JSON schema for both packages. NSD and Colectica will coordinate, but sharing JSON schemas across stat packages is not a requirement.

Ornulf gave an example showing two ways to get the same result with a Stata RECODE statement. Both methods could be represented in the same way in JSON.

Consequently, moving from the JSON representation to Stata may result in different a Stata statement that has the same effect.

Ornulf also raised the question of whether VTL can capture everything that will appear in our JSON schemas. Jeremy pointed out that our proposal originally referred to an SDTL (Standard Data Transformation Language). If VTL does not develop quickly enough, we could create our own SDTL or an extended version of VTL.

The RECODE discussion also raised issues about how we will incorporate information into DDI and EML. For example, on RECODE statement can be used to transform many variables. Should that information be copied into each variable that is affected, or can it be represented one time in DDI? George proposed focusing on this question in our April meeting.

B.Variable Comparison tool (J)

J reported on progress on the XML Comparison Tool. MTNA has created a Google doc for describing the tool. They expect it to be available as a free web-service

4.WG4 Outreach (Jared)

Jared reported that Outreach WG has a meeting scheduled on March 27.

Presentations have been accepted for both NADDI and IASSIST. Jared and Jeremy are coordinating preparations.

George reported that the National Academies of Sciences Workshop on Transparency in Federal Statistics seems to be on track for the end of June. Jag will present on this project, and others are encouraged to attend.

Action Items

Mirror JSON schemas for stat commands in the C2Metadata GitLab repository	Ornulf, Jeremy
Arrange conference call with VTL team	George