Building SQL Scripts Using Google Sheets, R and R Shiny

Huiming Tu, FHL Consulting
Hanming Tu, TuCai Consulting

PhUSE Connect US 2019: Clinical Data Science Conference
Feb 24~27, 2019
Baltimore Marriott Waterfront Hotel
700 Aliceanna St, Baltimore, MD
Agenda

• Data model and data modeling
• Simple data model for data modeling
• Google sheets as data store
• R and R shiny as tool
• PL/SQL scripts for Oracle database
• JSON scripts for MongoDB
• Visualization of data models
• Conclusion
Data Model

- Data model is an abstract framework that organizes elements of data and standardizes how they relate to one another and to the properties of the real world entities.

  - CDISC Study Data Tabulation Model (SDTM) is a data model (standard) for organizing and formatting data to streamline processes in collection, management, analysis and reporting.
  
  - CDISC Operational Data Model (ODM) is designed to facilitate the regulatory-compliant acquisition, archive and interchange of metadata and data for clinical research studies.
Data modeling is an important skill for data scientists involved with data analysis and is the process of documenting a complex software system design and creating entity and relationship (ER) diagram.

- There are some very sophisticated systems developed just to create the ER diagram.
- We are going to explore a simple way to document the data entity and relationship using some free tools.
## The Simple Data Model

<table>
<thead>
<tr>
<th>variable</th>
<th>seq</th>
<th>label</th>
<th>type</th>
<th>req</th>
<th>length</th>
<th>kt</th>
<th>fk</th>
<th>desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>table</td>
<td>0</td>
<td>collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Collection of table variables</td>
</tr>
<tr>
<td>id</td>
<td>1</td>
<td>Variable ID</td>
<td>numeric</td>
<td>T</td>
<td></td>
<td></td>
<td>pk</td>
<td>Unique identifier for table variable</td>
</tr>
<tr>
<td>db</td>
<td>2</td>
<td>Database Name</td>
<td>string</td>
<td>T</td>
<td>50</td>
<td></td>
<td></td>
<td>Database name</td>
</tr>
<tr>
<td>collection</td>
<td>3</td>
<td>Collection Name</td>
<td>string</td>
<td>T</td>
<td>50</td>
<td></td>
<td></td>
<td>Collection / Table name</td>
</tr>
<tr>
<td>variable</td>
<td>4</td>
<td>Variable Name</td>
<td>string</td>
<td>T</td>
<td>50</td>
<td></td>
<td></td>
<td>Variable name</td>
</tr>
<tr>
<td>seq</td>
<td>5</td>
<td>Variable Sequence</td>
<td>numeric</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td>Variable sequence</td>
</tr>
<tr>
<td>label</td>
<td>6</td>
<td>Variable Label</td>
<td>string</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
<td>Variable label</td>
</tr>
<tr>
<td>type</td>
<td>8</td>
<td>Variable Type</td>
<td>string</td>
<td>T</td>
<td>20</td>
<td></td>
<td></td>
<td>Variable type</td>
</tr>
<tr>
<td>req</td>
<td>9</td>
<td>Required?</td>
<td>string</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td>Whether the variable is required(T), permissible (P) or optional (O, blank)</td>
</tr>
<tr>
<td>length</td>
<td>16</td>
<td>Variable Length</td>
<td>numeric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Variable Length</td>
</tr>
<tr>
<td>kt</td>
<td>17</td>
<td>Key Type</td>
<td>string</td>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td>Key type such primary key (pk), foreign key (fk), etc.</td>
</tr>
<tr>
<td>fk</td>
<td>18</td>
<td>Foreign Key</td>
<td>string</td>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>Foreign key in the format of collection.variable</td>
</tr>
<tr>
<td>desc</td>
<td>23</td>
<td>Description</td>
<td>string</td>
<td></td>
<td>1000</td>
<td></td>
<td></td>
<td>Description of the definition</td>
</tr>
</tbody>
</table>
Modelling **SDTM metadata model**

- Standard Version (m_std_versions)
- Standard Domain (m_std_domains)
- Standard Variable (m_std_variables)
Google Sheets as Data Store

Steps: 1) Share the Google Sheet; 2) Publish the Google Sheet; 3) Capture the published link

<table>
<thead>
<tr>
<th>CDISC_Data_Models_pub</th>
<th>s_definitions</th>
<th>s_schemas</th>
<th>s_cfgvars</th>
<th>m_std_versions</th>
<th>m_std_domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 cdisc</td>
<td>m_std_versions</td>
<td></td>
<td>table</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>61 cdisc</td>
<td>m_std_versions</td>
<td></td>
<td>v_id</td>
<td>1</td>
<td>Version ID</td>
</tr>
<tr>
<td>62 cdisc</td>
<td>m_std_versions</td>
<td></td>
<td>p_vid</td>
<td>2</td>
<td>Parent Version</td>
</tr>
<tr>
<td>63 cdisc</td>
<td>m_std_versions</td>
<td></td>
<td>sdo</td>
<td>3</td>
<td>Standard Dev</td>
</tr>
<tr>
<td>64 cdisc</td>
<td>m_std_versions</td>
<td></td>
<td>class</td>
<td>4</td>
<td>Class Name</td>
</tr>
<tr>
<td>65 cdisc</td>
<td>m_std_versions</td>
<td></td>
<td>name</td>
<td>5</td>
<td>Model Name</td>
</tr>
</tbody>
</table>
R and R Shiny as a Tool

- **R** is an open source programming language for statistical computing and graphics.
- **Shiny** is an R package for building interactive web apps straight from R.
- **R Package** is the fundamental unit of shareable code bundled with data, tests, examples, and docs.
- Use R and R Shiny built a tool called Database Script Builder to use the data model metadata to generate SQL and JSON codes.

---

Computational Science Working Groups
Database Script Builder

Select a Model:
- GOOGLE SHEET: GS02
- GOOGLE SHEET: GS01
- GOOGLE SHEET: GS02
- ORACLE DB: DB01
- EXCEL: EXL01
- GOOGLE SHEET: AnyGS
- ORACLE DB: AnyDB
- EXCEL: AnyEX

Table:

<table>
<thead>
<tr>
<th>Spreadsheet title: CDISC Data Models pub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spreadsheet author: hamming.tu</td>
</tr>
<tr>
<td>Date of google sheet registration: 2018-11-14 02:16:12 GMT</td>
</tr>
<tr>
<td>Date of last spreadsheet update: 2018-11-14 02:09:57 GMT</td>
</tr>
<tr>
<td>visibility: public</td>
</tr>
<tr>
<td>permissions: rw</td>
</tr>
<tr>
<td>version: new</td>
</tr>
</tbody>
</table>

Contains 5 worksheets:
- (Title): (Nominal worksheet extent as rows x columns)
  - s_definitions: 34 x 33
  - s_schemas: 94 x 36
  - s_cfrgvars: 4 x 26
  - m_std_versions: 1009 x 26
  - m_std_domains: 1010 x 27

Key: 1ZTycKUaMhh4gbqNg814zBqszsrqN6kHbvAvjCwYwm20
Browser URL: https://docs.google.com/spreadsheets/d/1ZTycKUaMhh4gbqNg814zBqszsrqN6kHbvAvjCwYwm20/
Two ways to provide metadata models
- Configure the YML file
- Provide through the UI

Three sources
- Google sheet (GS)
- MS Excel (EX)
- Oracle database (DB)
Functional Tabs on the UI

- **Info**: displays info about the model source.
- **WS**: worksheet tab to view the content of each worksheet in GS, MS Excel or Oracle DB.
- **DB**: database tab to view table/collection definitions.
- **Script**: to generate Oracle SQL or MongoDB JSON scripts.
- **Load**: to load metadata into target DD or insert the data into Google Sheets or MS sheets.
Use Oracle SQL Developer

Visualization of Data Models
Demo
The premier community for people working in the biometric area

phuse.eu
@PhUSETwitta
/PhUSE
phusewiki.org