



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

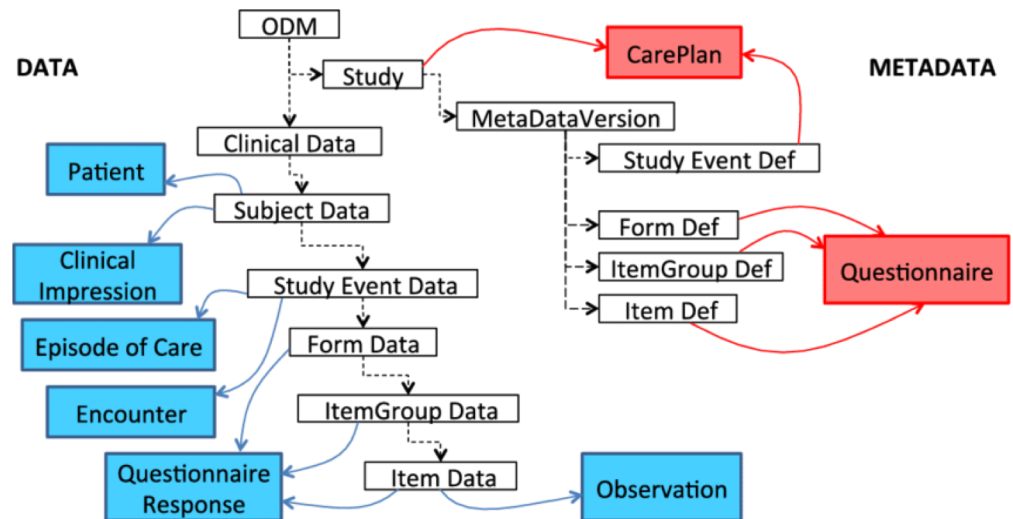


**Computational Science
Working Groups**



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.



**Computational Science
Working Groups**



Data Modeling

- 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.



**Computational Science
Working Groups**



The Simple Data Model

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



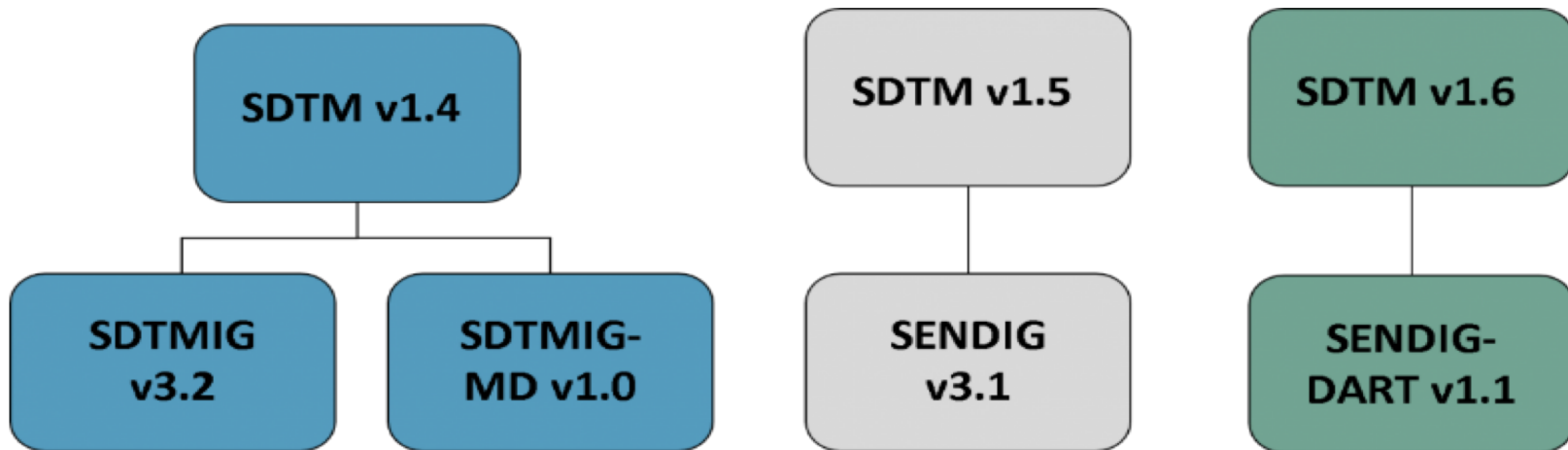
**Computational Science
Working Groups**



SDTM Metadata Model

➤ Modelling SDTM metadata model

- Standard Version (m_std_versions)
- Standard Domain (m_std_domains)
- Standard Variable (m_std_variables)



**Computational Science
Working Groups**



Google Sheets as Data Store

- **Steps:** 1) Share the Google Sheet; 2) Publish the Google Sheet;
3) Capture the [published link](#)

CDISC_Data_Models_pub

	s_definitions	s_schemas	s_cfgvars	m_std_versions	m_std_domains
60	cdisc	m_std_versions	table	0	
61	cdisc	m_std_versions	v_id	1	Version ID
62	cdisc	m_std_versions	p_vid	2	Parent Versio
63	cdisc	m_std_versions	sdo	3	Standard Dev
64	cdisc	m_std_versions	class	4	Class Name
65	cdisc	m_std_versions	name	5	Model Name



**Computational Science
Working Groups**



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:

GOOGLESHEET : GS02 ▲

GOOGLESHEET : GS01

GOOGLESHEET : GS02

ORACLEDDB : DB01

EXCEL : EXL01

GOOGLESHEET : AnyGS

ORACLEDDB : AnyDB

EXCEL : AnyEX

[[R](#) | [R Shiny](#) | [RStudio](#) | [R Packages](#) | [CRAN](#) | [Submit](#)]

Info WS DB Script Load

```
Spreadsheet title: CDISC_Data_Models_pub
Spreadsheet author: hanming.tu
Date of googlesheets registration: 2018-11-14 02:16:12 GMT
Date of last spreadsheet update: 2018-11-14 02:09:57 GMT
visibility: public
permissions: rw
version: new

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

Key: 12TycKUaMhh4gbgNg814zBqszsrqN6kHbvAvjCwYwm20
Browser URL: https://docs.google.com/spreadsheets/d/12TycKUaMhh4gbgNg814zBqszsrqN6kHbvAvjCwYwm20/
```



Computational Science
Working Groups



Metadata Sources

- 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)



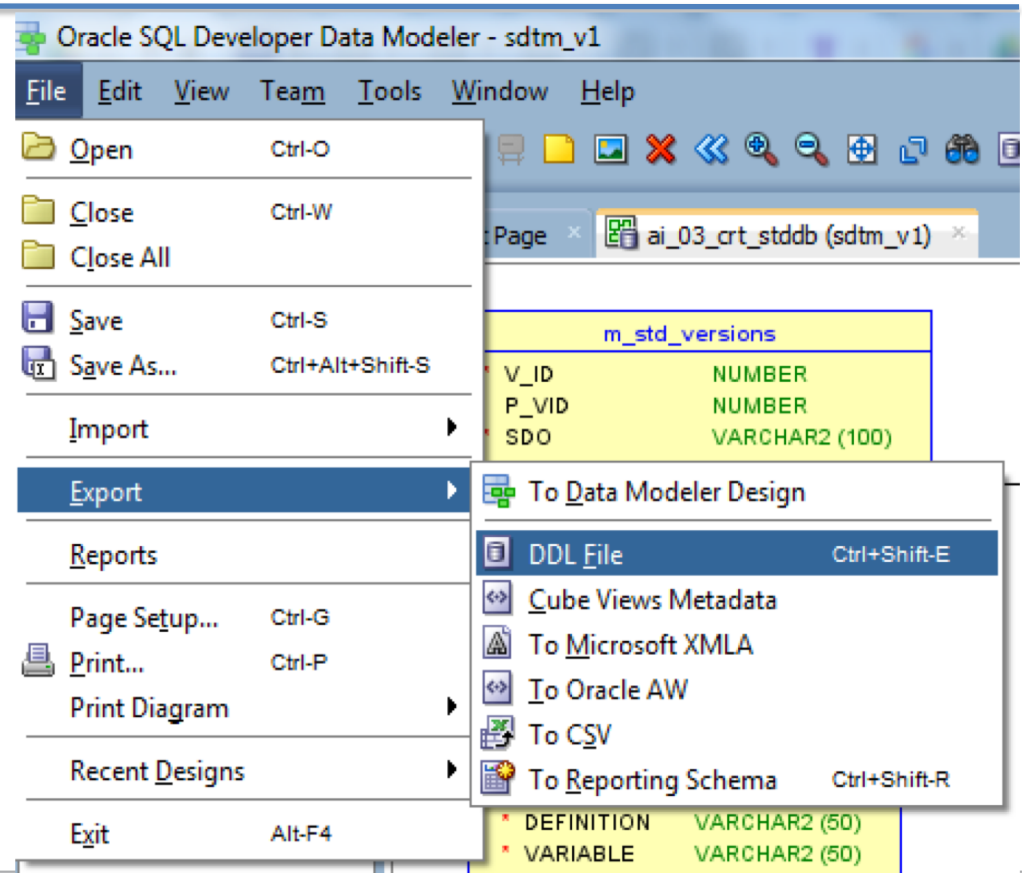
Functional Tabs on the UI

- **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.



Visualization of Data Models

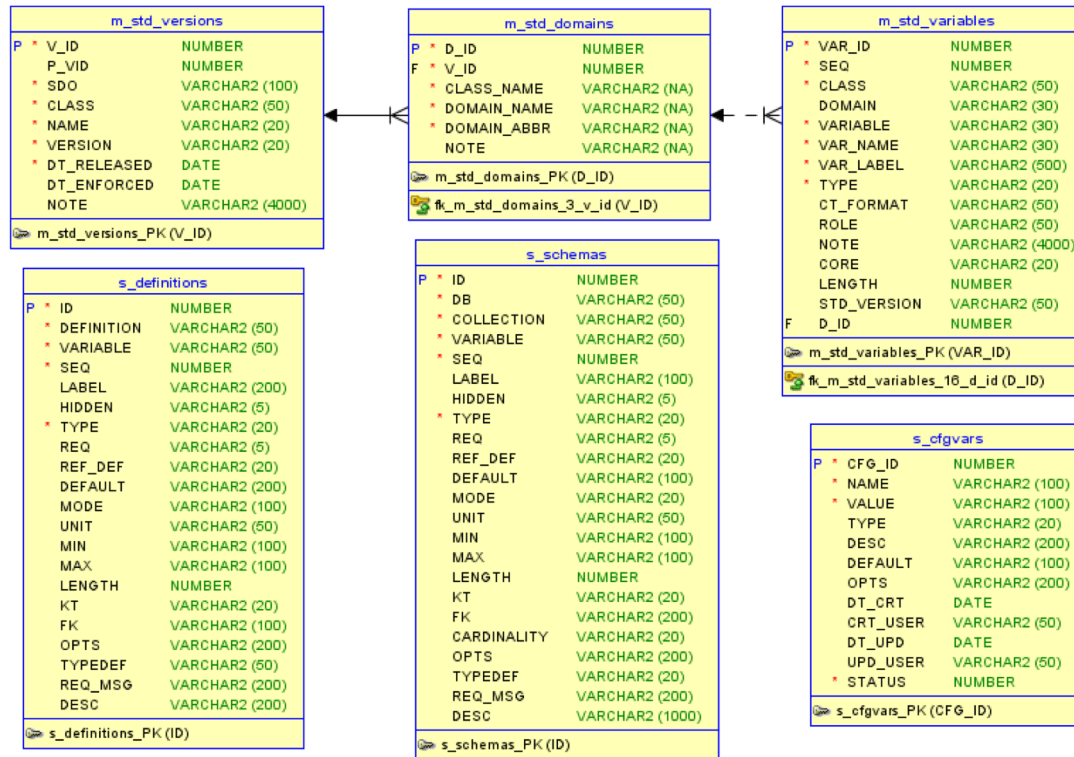
➤ Use Oracle SQL Developer



**Computational Science
Working Groups**



Example of Data Model



Computational Science
Working Groups



Demo



**Computational Science
Working Groups**



Q & A



**Computational Science
Working Groups**





**The premier community for people
working in the biometric area**

 ***phuse.eu***

 ***@PhUSETwitta***

 ***/PhUSE***

 ***phusewiki.org***