

TDF – Overview and Status of the Test Data Factory Project

Peter Schaefer (VCA-Plus, Inc.) and the TDF Team

The TDF Team

- Nancy Brucken, Syneos Health
- Jessica Dai, Vertex
- Cynthia Stroupe, UCB
- Dante DiTommaso, Omeros
- Katja Glass
- Peter Schaefer, VCA-Plus

Test Data Factory is one of the projects within PhUSE’s Standard Analyses and Code Sharing Working Group. With this project, the working group has acknowledged that testing with appropriate test data is an essential part of any process and software development. The objective of the TDF project is to provide test datasets following the CDISC standards including appropriate documentation of deviations from the standard specification that can be used by anyone for their testing effort.

At this point, the TDF project team has published two test data packages based on the SDTM and ADaM datasets of the CDISC pilot. Now the TDF team is off to tackle a bigger challenge: How to create test datasets “out of thin air”, i.e., provide test datasets based on user input about the required test datasets using an approach similar to simulation of a medical research study.

Get the Poster

Scan this



to download a copy of this poster.

Completed Phase 1:

Published SDTM and ADaM Test Data Packages

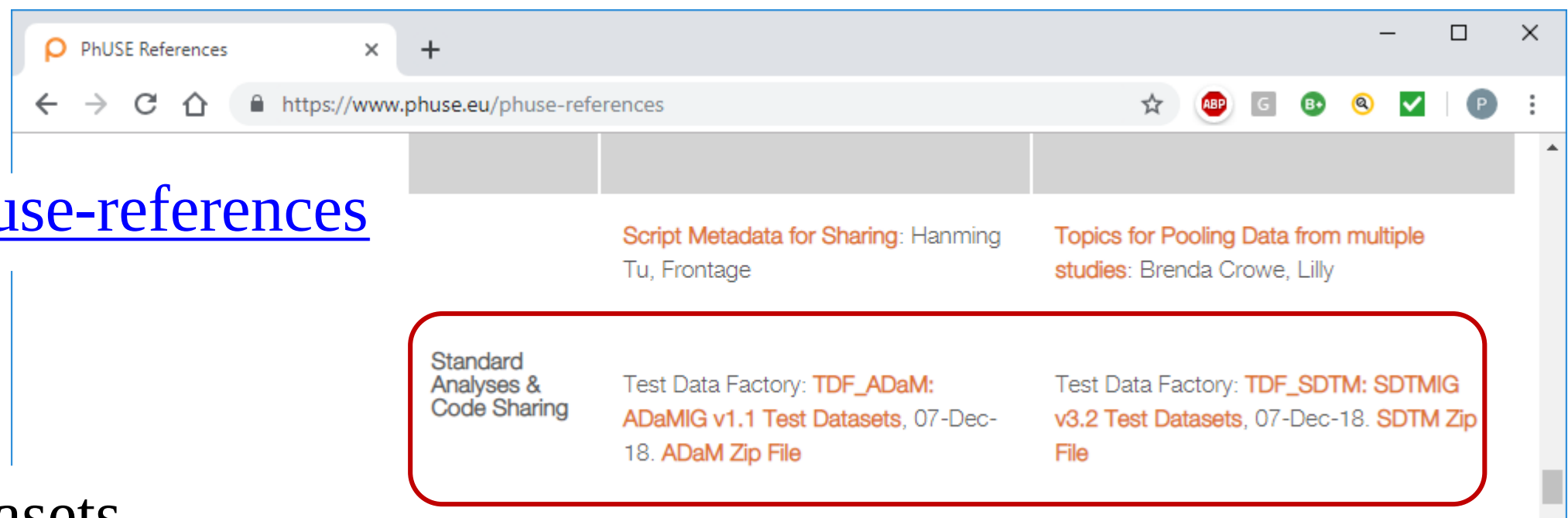
The TDF team has published two test data packages based on CDISC Pilot datasets:

- 22 SDTM datasets were updated and documented, including define.xml
- 10 ADaM datasets were updated and documented, including define.xml
- Waiting to be used, so we will get feedback

- Available at <https://www.phuse.eu/phuse-references>

Out-of-scope:

- Update of CRF, protocol
- Creation of additional datasets



Working on Phase 2:

The Vision for the TDF Project

What if we could ask a user to describe what kind of test datasets would be required? Then have some software that creates these datasets? Such as SDTM or SEND datasets that are required for testing a program or a process, with consistent data across domains and optional features like outliers or missing data? And maybe as a bonus the corresponding ADaM datasets?

Creating CDISC test datasets is the idea behind an interactive “Test Data Factory” tool, which is essentially the big-picture vision of the TDF project:

- User enter details about the required test data in some appropriate form (UI, Excel,...)
- A software program (the “TDF Tool”) generates study dataset based on this input
- User gets CDISC-compliant test datasets generated by the TDF Tool

The Roadmap towards a Test Data Factory Tool

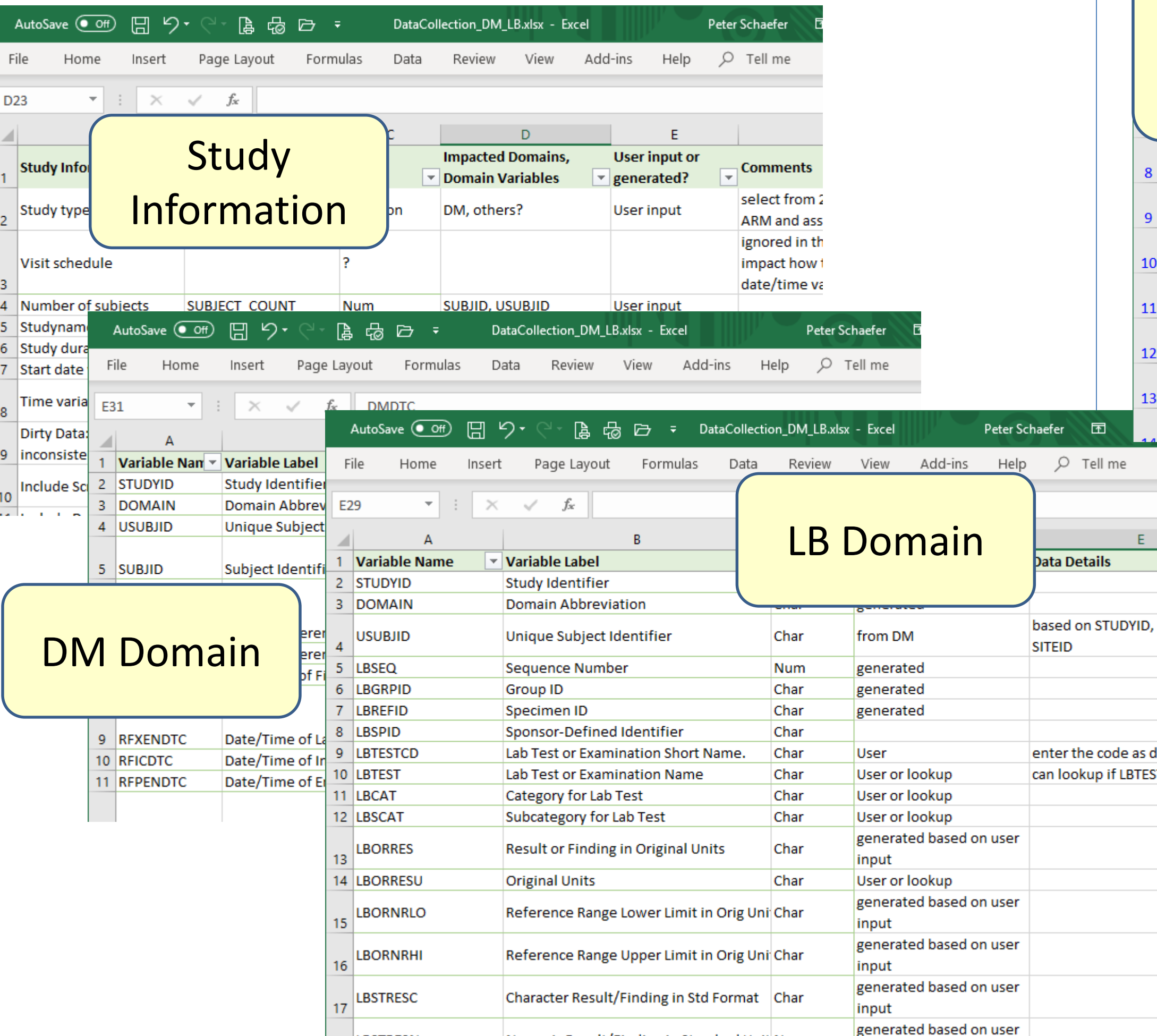
The TDF team is working towards a Test Data Factory tool in three steps:

Step 1:

Input Requirements Analysis

The basis for TDF is knowing what test datasets the user needs: Therefore the team works through each SDTM domain and each variable to understand whether the user has to provide input, which input is required, and how the user input will impact the generated test datasets.

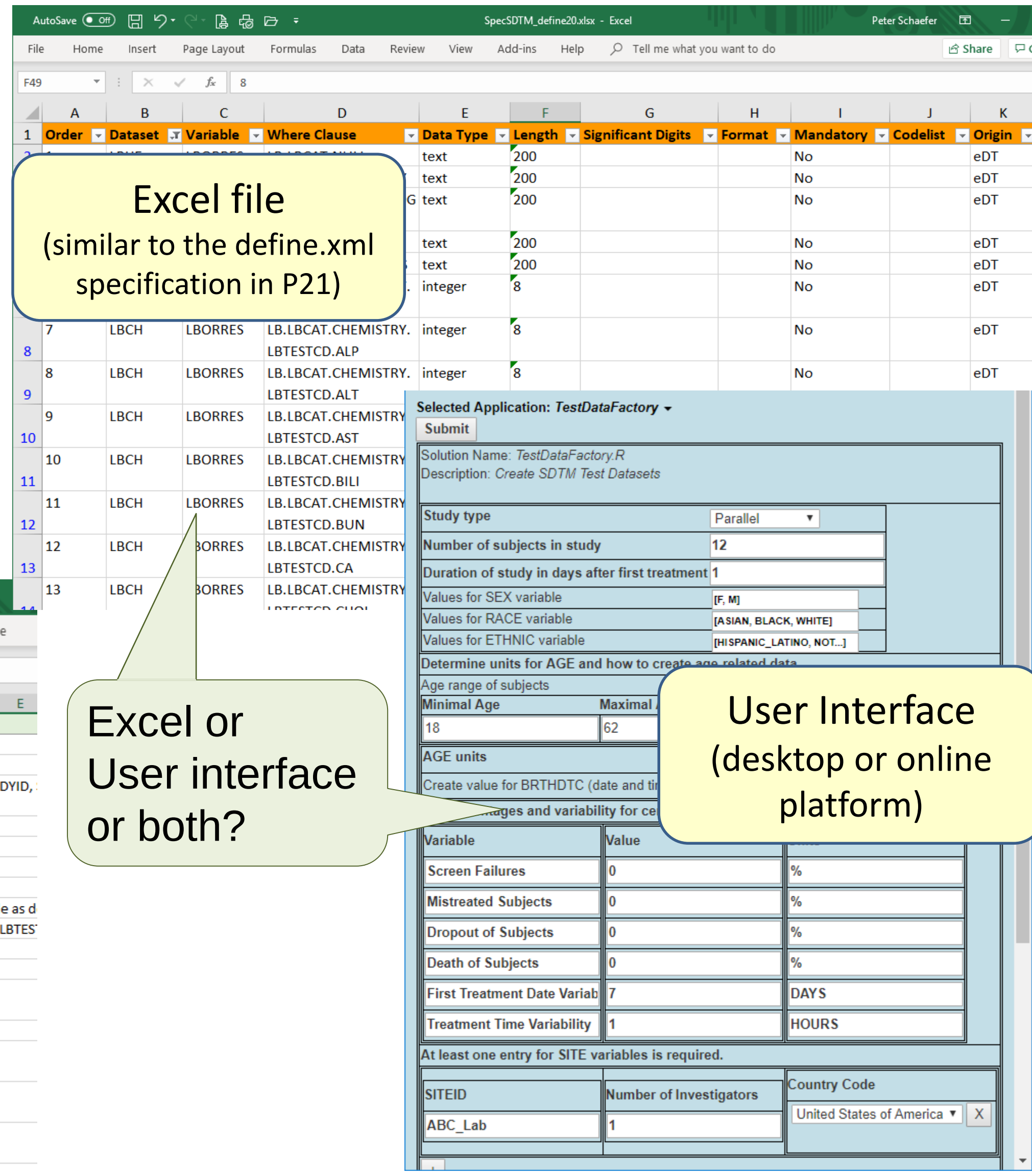
As a starting point, the TDF team is working on the DM and LB domains.



Step 2:

Formal User Input Specification

A formal specification of the required user input is the first step towards an implementation of a TDF tool. Options for formalized user input formats are Excel files or the user interface of some interactive software.



Step 3:

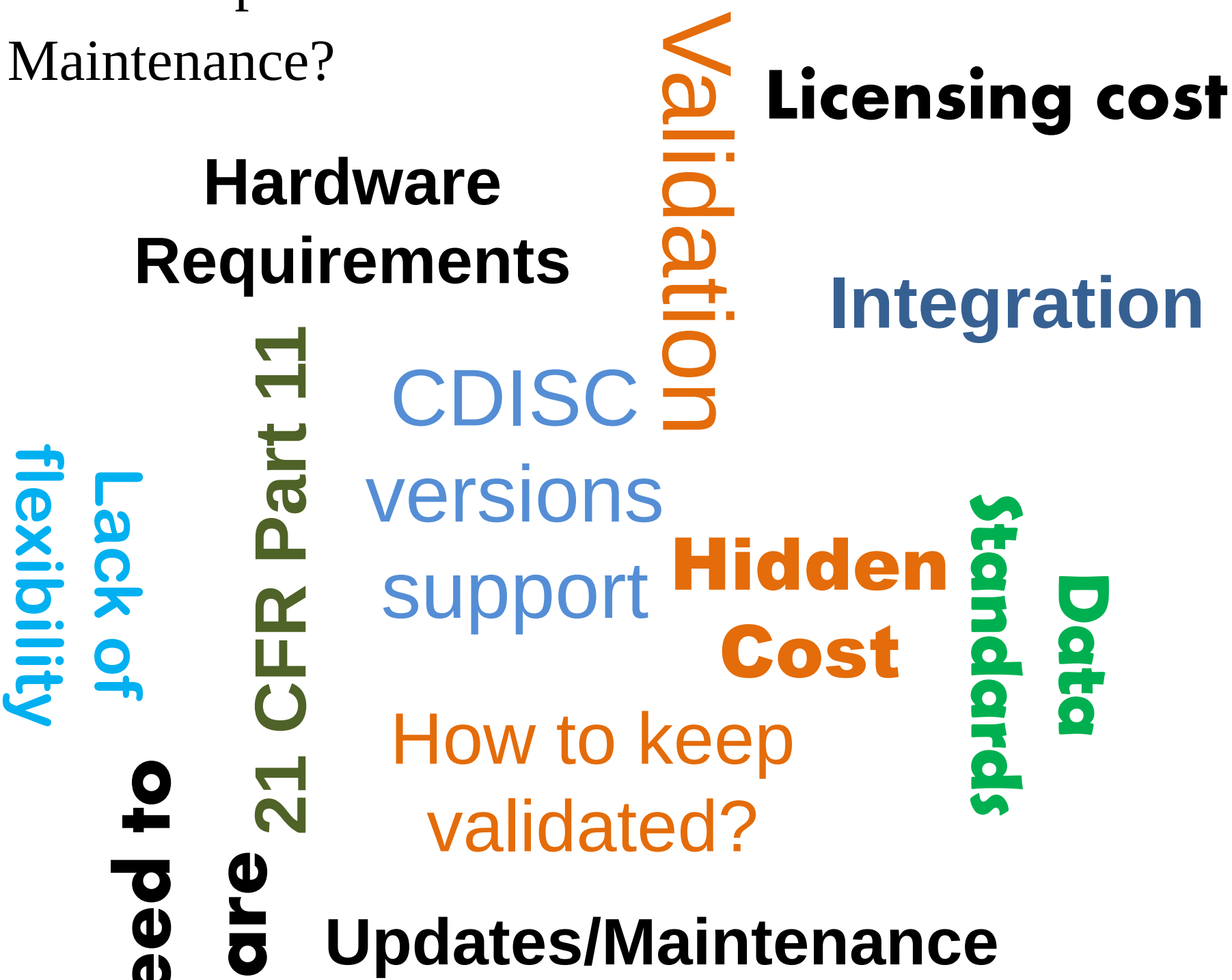
TDF Tool Implementation

The TDF team has not decided how to implement the TDF tool. We see two main alternatives:

- Scripts that each user can execute and modify as required
- A “product” that each user can use (desktop installation or SaaS) and that executes configurable scripts.

At this point, still open questions ...

- Access to CDISC specification for datasets?
- SAS, R, Python, or what?
- Ownership?
- Maintenance?



Summary

The TDF project exists to provide CDISC test datasets as an important contribution to the development and deployment of CDISC-based solutions. Publishing the first SDTM and ADaM packages based on updated CDISC Pilot datasets was only a first step.

Certainly, the idea of a software tool that enables users to ‘generate’ test datasets – a really interactive “Test Data Factory” – is somewhat of a dream. However, the technology is available and similar concepts have been used in other application areas. So, we are progressing according to plan and would welcome your contribution.

Contact: Peter Schaefer, pschaefer@vca-plus.com