

A Contribution to Better Systems – The Test Data Factory (TDF) Project

Peter Schaefer, VCA-Plus; Nancy Brucken, inVentiv Health; Cynthia Stroupe, UCB

The “Standard Analyses and Code Sharing” Working Group has started the new Test Data Factory (TDF) project. With this project, PhUSE and the working group acknowledge that the use of appropriate test data is an essential part during any solution development. Several PhUSE projects describe medical research methods, features, or processes, and some even create scripts that handle SDTM or ADaM datasets. These activities require a variety of test data. The typical fallback position is to use the outdated CDISC pilot project and/or anonymized study data provided by team members. The TDF project strives to support the testing effort by providing test datasets for a more systematic approach.

The project team has started with updating the CDISC Pilot datasets to comply with more recent versions of the CDISC standards. This poster will provide an update on TDF and encourage conference participants to contribute to this effort. In addition, the poster will outline a vision for provisioning test datasets based on simulation of research studies and the creation of datasets based on user input and according to system requirements.

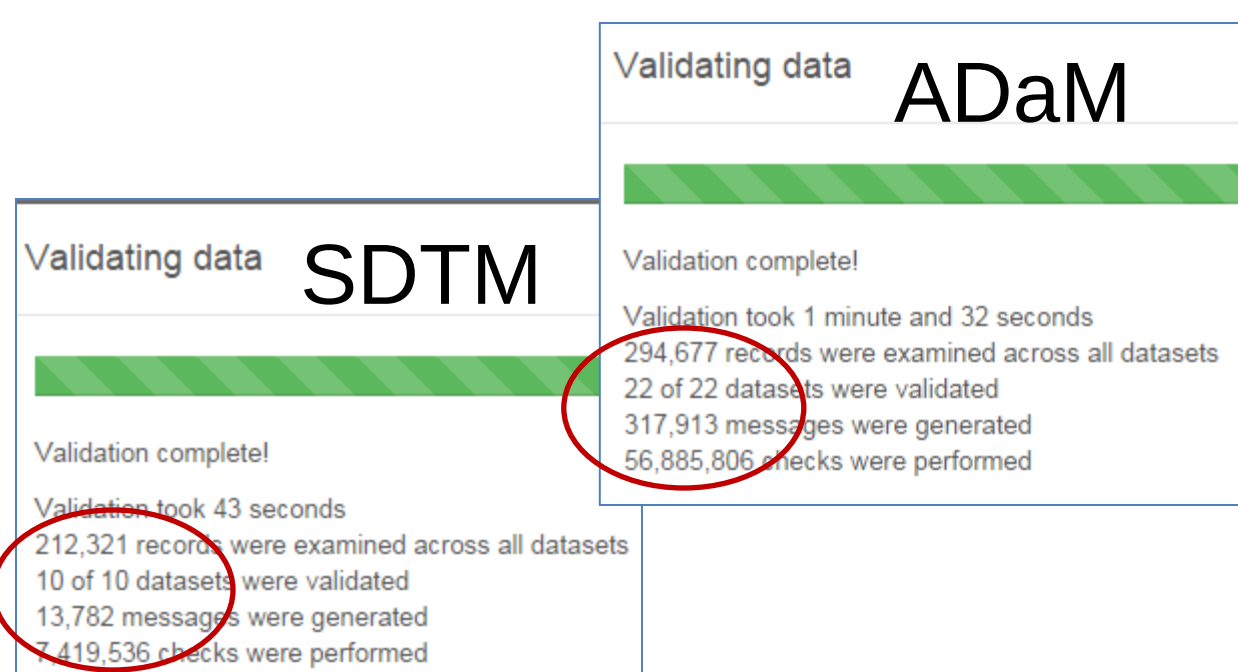
The CDISC Pilot

More than 10 years ago the CDISCPILOT01 study was created – a de-identified study provided by a large pharmaceutical company to test the then current versions of SDTM and ADaM for their usefulness in submitting and reviewing study data. Later the pilot study was updated to reflect the feedback from the community. And until now, in many cases the CDISC Pilot02 study is still a reference and starting point when CDISC-based processes and systems are developed and tested.

But are the datasets in the CDISC Pilot really good for testing?

BEFORE

The CDISC Pilot study datasets are outdated and incomplete.

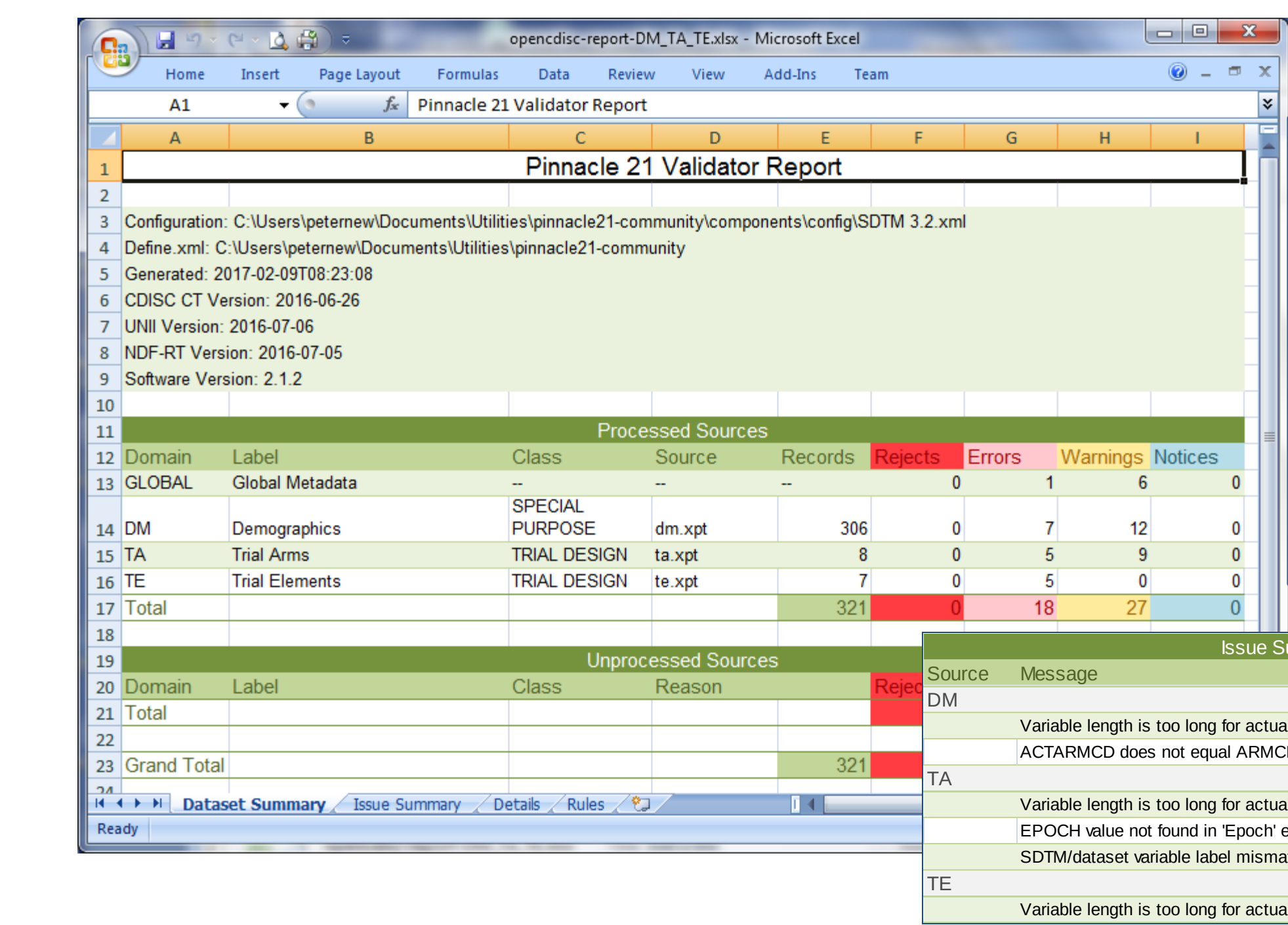


- These are just ‘formal’ validation errors
- Many are rather easy to fix
- We also found ‘content’ errors; these will require us to generate additional data

AFTER

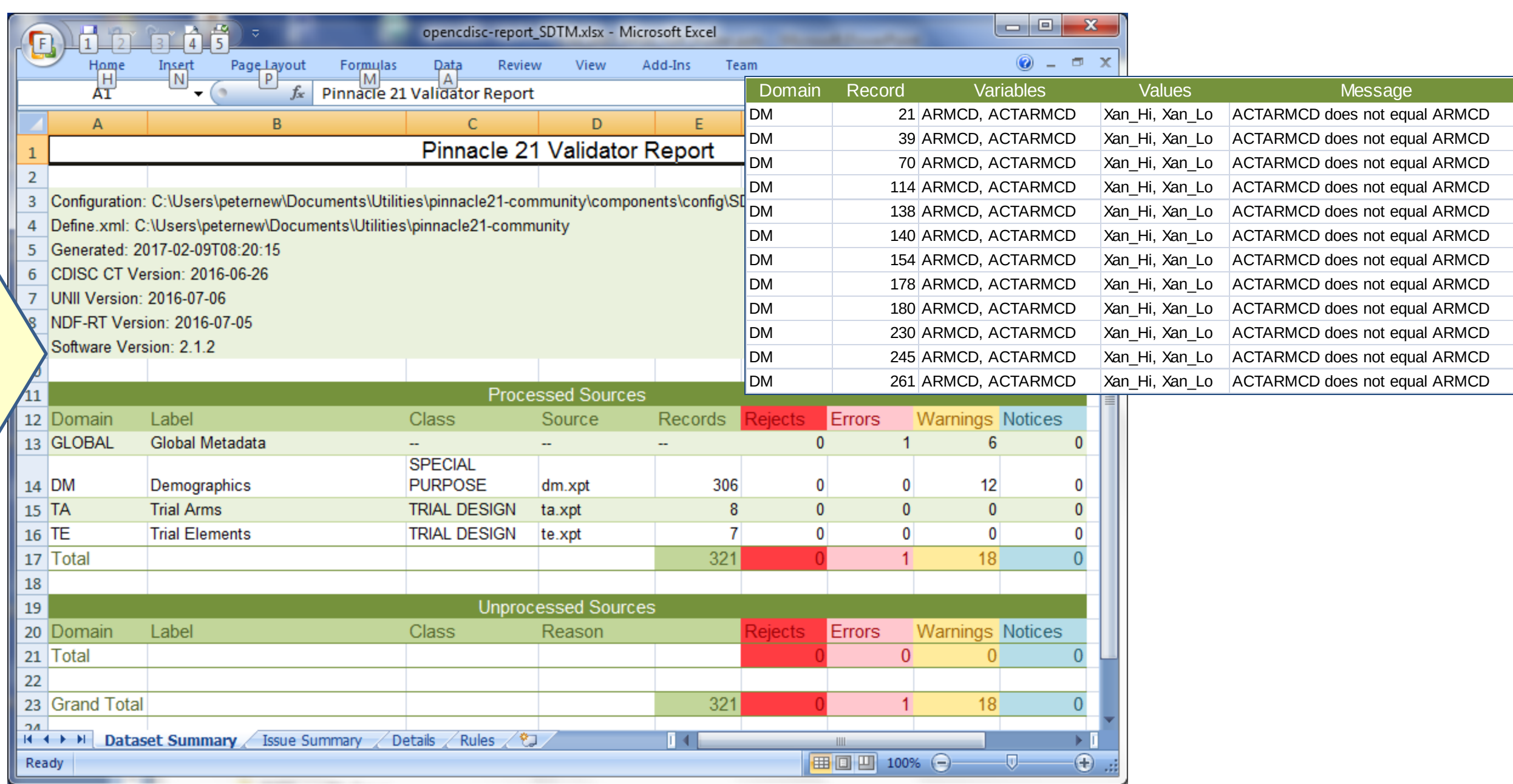
The initial goal:

- Updated datasets that pass validation using Pinnacle 21 Community Edition.
- Documentation of remaining validation warnings/errors using SDRG template
- Publish datasets and documentation in Github



These were easy:

- Trim variable length
- Leave incorrectly treated subjects in dataset
- Add, update, or correct Controlled Terminology
- Update variable label



There is way more to do for the CDISC Pilot update ...

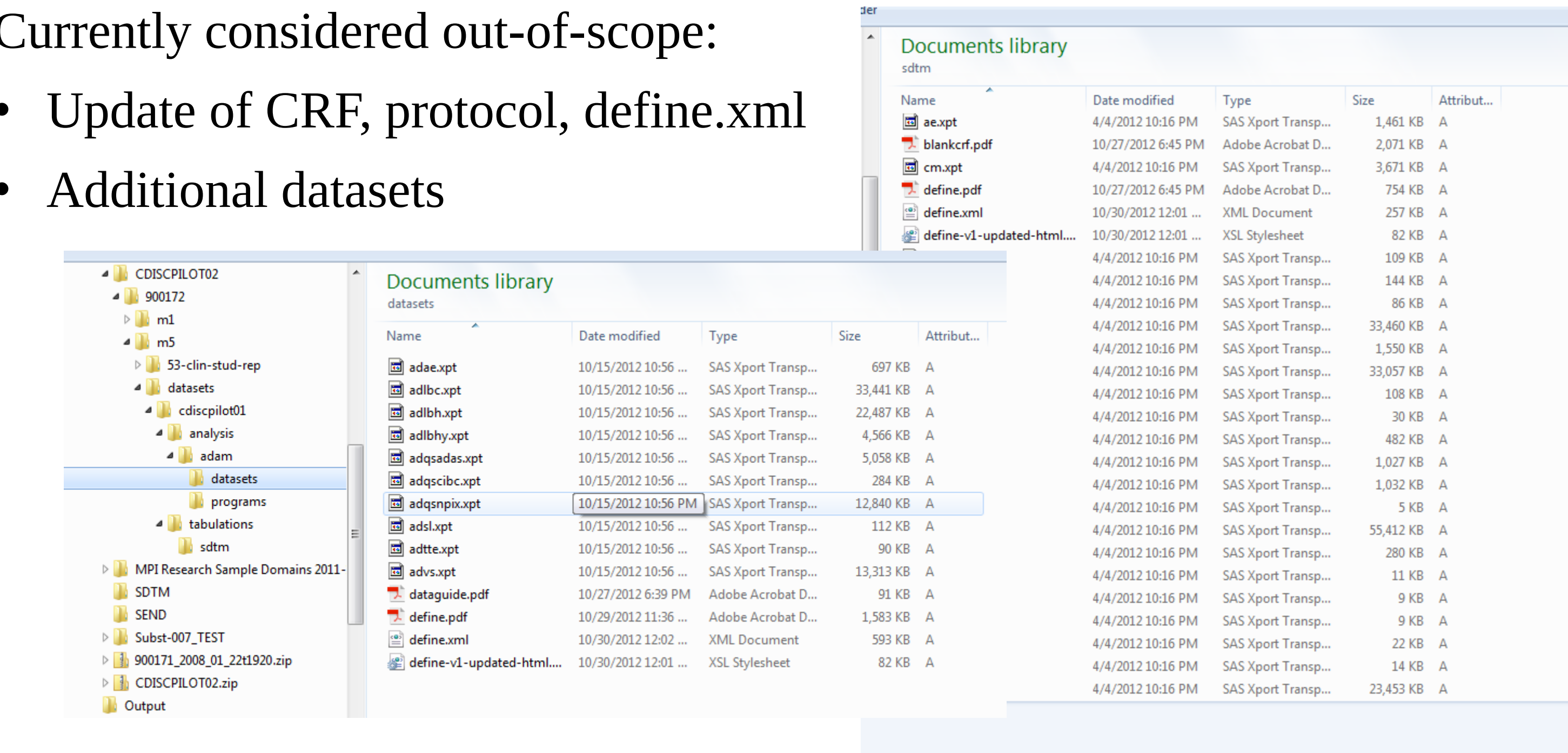
22 SDTM and 10 ADaM datasets

In-scope:

- Update all datasets
- Go beyond ‘formal’ compliance and identify other data issues

Currently considered out-of-scope:

- Update of CRF, protocol, define.xml
- Additional datasets

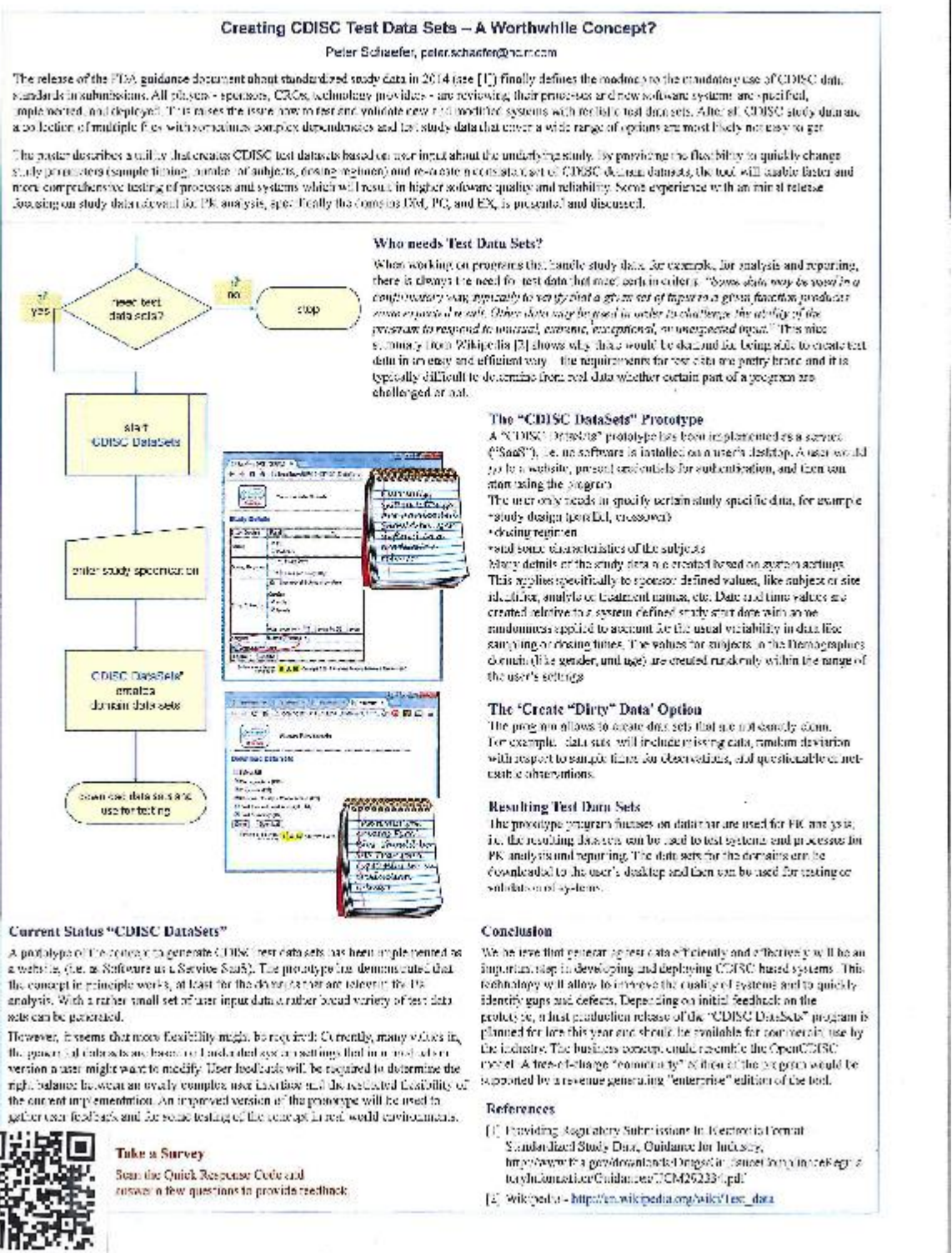


Sometimes you have to believe in flying cars

What if we could create tools that allow a user to describe what kind of test study date are wanted and have the tools create the datasets? SDTM domains that are requested, with consistent observations and optional features like ‘dirty’ data? And even the resulting ADaM datasets?

Well, already two years ago, a CSS poster presented the idea of “Creating CDISC Test Data Sets”, which is essentially the vision of the Test Data Factory project:

- User enters details of the study specification in a web browser UI
- Cloud-based R scripts are executed to generate the datasets based on this input
- CDISC datasets are created as xpt files and can be downloaded to the user’s desktop



Maybe it’s time to pick up this idea again!!!

Summary

The TDF team believes that providing CDISC test datasets will be an important contribution to the development and deployment CDISC-based solutions. Updating the CDISC Pilot datasets might not seem like a big step, but we want to walk before we run and contribute useful results quickly.

Certainly, the idea of a software tool that enables users to ‘generate’ test data sets – a real interactive “Test Data Factory” – is somewhat of a dream. However, the technology is available, similar concepts have been used in other application areas, and maybe there is enough interest in a tool like this.

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