PharmaSUG 2016 - Paper HT04

Usage of Pinnacle 21 Community Toolset 2.1.1 for Clinical Programmers

Sergiy Sirichenko, Pinnacle 21, Plymouth Meeting, Pennsylvania
Michael DiGiantomasso, Pinnacle 21, Plymouth Meeting, Pennsylvania
Travis Collopy, Pinnacle21, Plymouth Meeting, Pennsylvania

ABSTRACT

All programmers have their own toolsets like a collection of macros, helpful applications, favorite books or websites. Pinnacle 21 Community is a free and easy to use toolset, which is useful for clinical programmers who work with CDISC standards. In this Hands-On Workshop (HOW) we'll provide an overview of installation, tuning, usage and automation of Pinnacle 21 Community applications including: Validator - ensure your data is CDISC compliant and FDA submission ready, Define.xml Generator - create metadata in standardized define.xml v2.0 format, Data Converter - generate Excel, CSV or Dataset-XML format from SAS® XPT, and ClinicalTrials.gov Miner - find information across all existing clinical trials.

INTRODUCTION

In 2008, the Clinical Data Interchange Standards Consortium (CDISC) had begun to make headway in its mission to develop a global set of standards. At the time, FDA had started requesting submission data in a standardized format, but software options to help ensure a submission's compliance with CDISC business rules were limited. So in October of that year, OpenCDISC was launched as an open source community dedicated to building extensible tools and frameworks for the implementation of CDISC standards. OpenCDISC Validator was the community's first product aimed at helping developers create FDA compliant SDTM datasets.

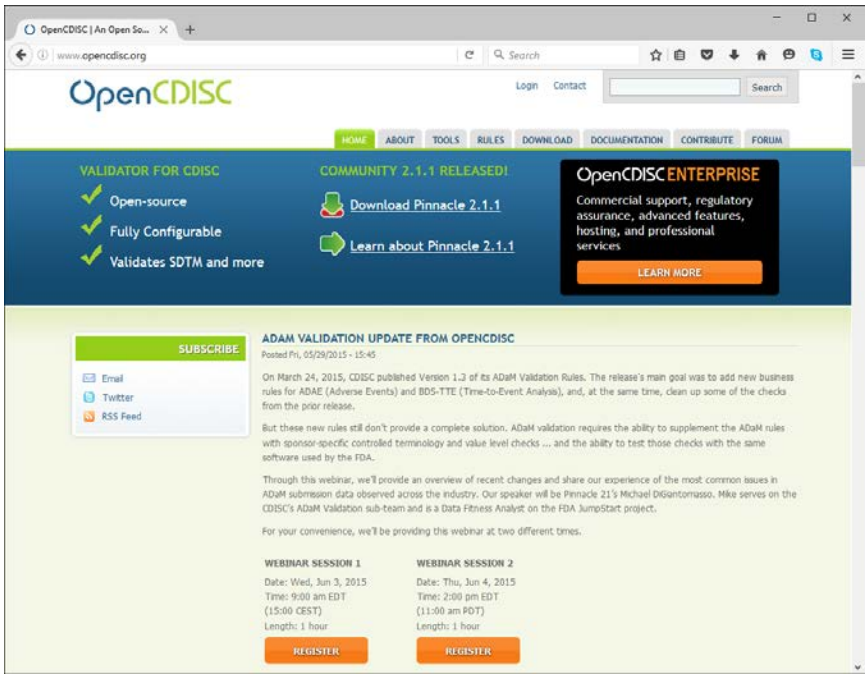
The launch of OpenCDISC Validator created an immediate buzz in the industry. It was quick and easy to download, it was absolutely free, and it worked. And because of the open, collaborative, and vendor-neutral process by which it was developed, it reached developers as a democratized solution. Word spread among developers and Validator took off, quickly expanding to support additional CDISC standards including ADaM, SEND, and Define.xml. But the project's big break occurred in 2010, when FDA evaluated and selected OpenCDISC Validator as a tool for screening all incoming submissions for compliance with CDISC business rules.

With Validator being the open source software of choice at the FDA, the momentum shifted and OpenCDISC popularity increased dramatically. However, the increased popularity also exposed some limitations of the open source project. First, there were users who simply needed the software to do more. Validator was designed as a desktop tool for individual developers or small teams to QC their work. But large companies with numerous professionals and large number of studies needed something more centralized, more robust, and with better support options. Second, a dearth of funding was holding the open source project from reaching its full potential. So, in 2011 members of OpenCDISC formed Pinnacle 21, the commercial arm of OpenCDISC. This new company created OpenCDISC Enterprise, a commercial, enterprise-wide version of the software designed to support large organizations with many users, providing all the tools, bells, and whistles advanced users needed. And by charging commercial license fees for its use, the open source project now had the financial backing it needed to continue and evolve.

In 2014, with the backing of Pinnacle 21, the open source project released OpenCDISC Community v2.0. This first major upgrade expanded the toolset to four individual tools, including the Validator to ensure your data is CDISC compliant and FDA submission ready, a Define.xml Generator to create metadata in standardized define.xml v2.0 format, a Data Converter to generate Excel, CSV or Dataset-XML format from SAS XPT datasets, and a ClinicalTrials.gov Miner to help find information across all existing clinical trials. In 2015, the tool was rebranded to Pinnacle 21 Community. This paper will describe in detail how clinical programmers can use the Pinnacle 21 Community toolkit to simplify their daily tasks and create FDA compliant deliverables.

GETTING STARTED

To get started with Pinnacle 21 Community visit www.opencdisc.org, the open source project's website (until it is completely merged into pinnacle21.net). This is where you can find the software downloads, documentation, and the latest news from the community. An active support forum is also available to ask questions and share knowledge with users from other organizations, as well as the developers of Pinnacle 21 Community. You should also subscribe to the email or twitter feed to be notified of new releases or upcoming webinars and events.



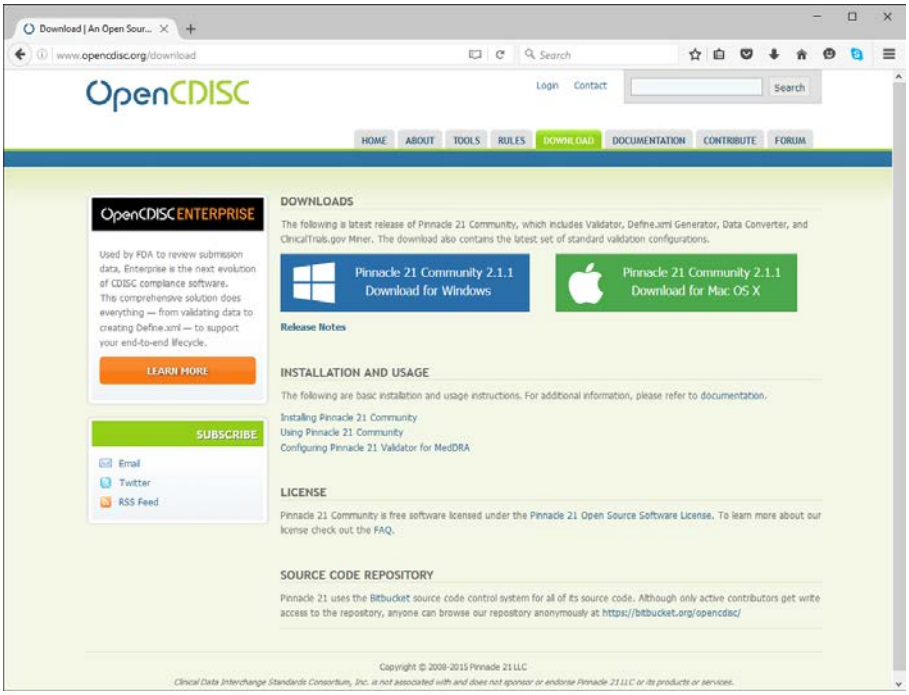
Screenshot 1. OpenCDISC.org home page

INSTALLATION

To install Pinnacle 21 Community go to the Download section on the OpenCDISC website and select and download the appropriate package for your operating system, Windows or Mac OS X. Once you have downloaded the package, unzip it to any location on your hard drive, and you are now ready to launch the application and get to work.

"Download, unzip, and run", it's that easy!

You can even unzip and run Pinnacle 21 Community from a USB flash drive if portability is important.

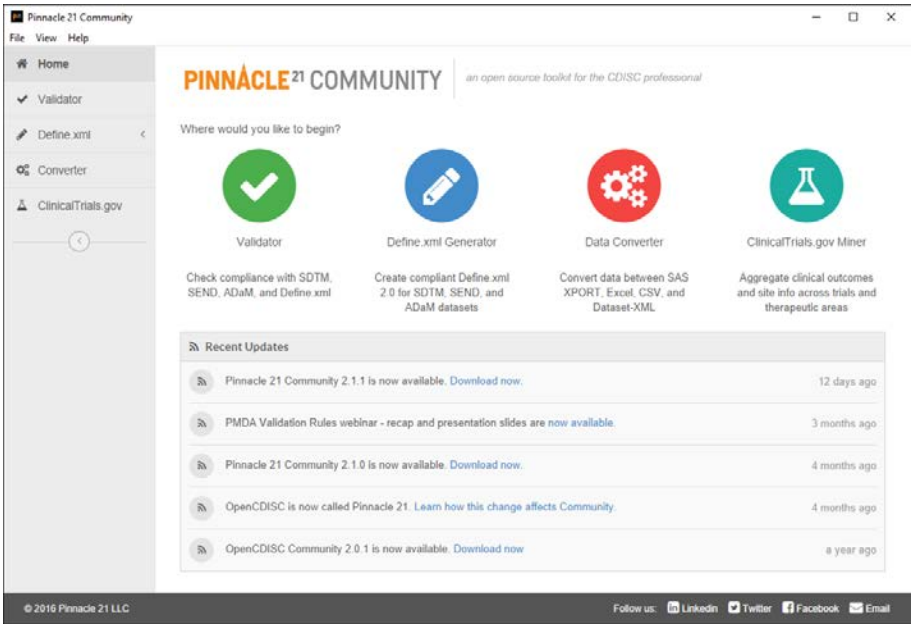


Screenshot 2. OpenCDISC.org download page

LAUNCHING THE APPLICATION

The startup wizard offers a number of enhancements built for user convenience. We recommend enabling auto-updates. This allows users to have the latest rules, controlled terminologies, and software upgrades installed automatically. By keeping up to date on the latest releases, you can ensure that your data is compliant with the published FDA validation rules for regulatory submission. Of course, in addition to utilizing auto-updates, there is always an option to re-install the software by downloading the latest package from the website.

After completing the setup wizard, the home screen is presented with options for the 4 available tools. Select the tool of your choice to begin. The home screen also provides a *Recent Updates* section, where you can stay current with the latest news from the community and industry. The menu at the top provides quick access to application preferences and help resources.



Screenshot 3. Pinnacle 21 Community home screen

TUNING AND PERFORMANCE

While Pinnacle 21 developers continuously work hard to make improvements to increase performance, there are also a few things that users can do to get the most out of running the application. Tuning the performance settings would especially help the Validator when processing large studies. To access Performance settings go to

```
Help menu → Preferences → Performance tab
```

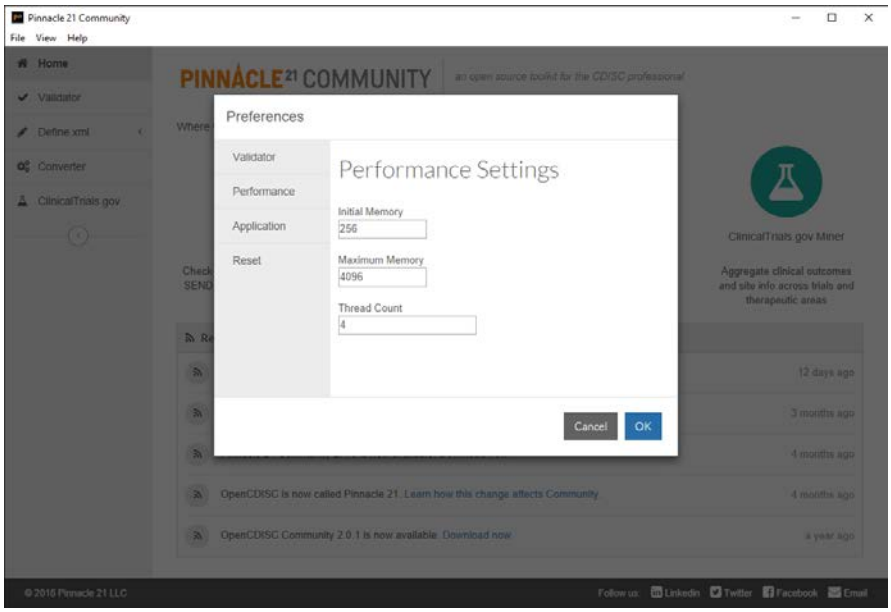
There are 3 available performance settings: Initial Memory, Maximum Memory, and Thread Count.

Given that the Validator and Data Converter do all of their processing without the help of a database or temporary files, the memory demands for very large datasets can be high. We recommend allocating up to 75% of available system RAM memory for validation. So for example, if you have a machine with 8GB of RAM you can allocate up to 6GB to the application by setting the Maximum Memory setting to 6144. Of course if you are running a 32-bit version of Windows this setting can only be increase to about 1500.

Validator also supports multicore dataset processing, where more than one dataset can be validated simultaneously on computers which have multiple processors/logical cores. So if you have a multicore machine, we recommend you increase the Thread Count to 2 or up to the number of available cores.

And don't worry if you mess up the performance settings, you can always restore default values by going to

```
Help menu → Preferences → Reset tab → Reset to Default Settings
```



Screenshot 4. Performance settings

CONFIGURING DICTIONARIES

Pinnacle 21 Validato currently uses CDISC Controlled Terminology and 4 different external dictionaries. CDISC Controlled Terminology, UNII, and NDF-RT are freely available and are included with the download package. When new versions become available they will be installed through the Auto-update feature if enabled.

MedDRA and SNOMED are proprietary dictionaries and require each company to obtain a license. Therefore, they are not included with the download package and must be installed manually. To configure MedDRA, a user needs to place the *.asc files found in the “ascii” folder of MedDRA distribution into the following Pinnacle 21 folder:

`components` → `config` → `data` → `MedDRA` → `[version number]` (for example 17.0 or 17.1)

After the MedDRA files have been correctly installed, a MedDRA drop-down box will become visible in the Validator screen, directly to the right of CDISC CT drop-down.

SNOMED configuration is a little more complicated as it requires the preprocessing of the original SNOMED files. Please refer to http://www.opencdisc.org/projects/validator/configuring-opencdisc-validator-external-dictionaries for up to date configuration instructions.

USING VALIDATOR

Pinnacle 21 Validator can be run either from a graphical user interface (GUI) or command line interface (CLI). Both options support the full set of Validator features, but are designed for different use cases. Where GUI is most commonly used for ad-hoc validation, the CLI is typically utilized for process automation. Before we show you how to run the validator, let’s first review the high-level components that enable the validation process and how they work.

VALIDATOR ARCHITECTURE

The key architectural concept behind the Validator is to decouple the definition of validation rules from application logic. This provides ultimate flexibility to create and maintain any number of validation rule definitions necessary to meet the diverse needs of sponsors, CROs, regulatory agencies, and anyone else involved in collection, storage, and exchange of clinical data. The Validator‘s architecture (Figure 1) is comprised of the following components:

- Configuration – an XML document, an extension of Define.xml 2.0 format, defines standard datasets and validation rules to be executed against each of the datasets. The rules are expressed according to the Pinnacle 21 Validation Framework, which is described at http://www.opencdisc.org/projects/validator/opencdisc-validation-framework. The configuration file also defines references between dataset variables and controlled terminology codelists that are provided as separate inputs to the validation process.

- Controlled Terminology and External Dictionaries – CDISC CT, MedDRA, UNII, NDF-RT, and SNOMED files can be provided to the validation process to enable validation checks that compare values of controlled variables to the content of the referenced codelists as defined in the configuration files.
- Validation Engine – the core component of the architecture is developed in Java and houses the application logic, which reads and parses input datasets, interprets and executes validation rules described in the configuration file, and outputs a validation report. The current Validator release supports SAS XPORT, delimited text files, and Dataset-XML as input.
- Validation Report – the results of validation are rendered in Excel or CSV format, based on user preferences, and contains issue messages, descriptions, and details of data records that failed validation.

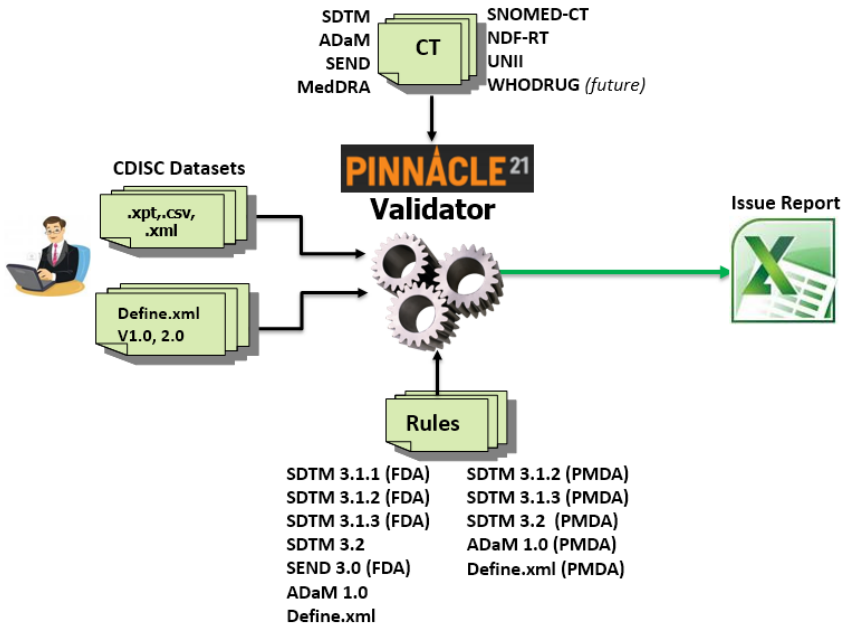
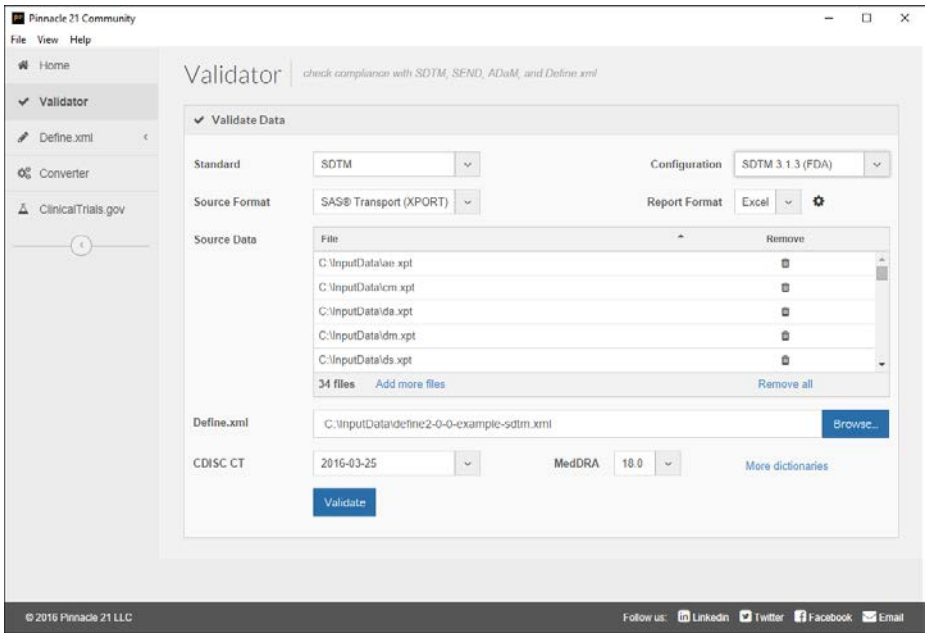


Figure 1. Pinnacle 21 Validator Architecture

RUNNING VALIDATOR FROM GUI



Screenshot 5. Validator screen

The Validator can be used to check compliance with CDISC SDTM, SEND, ADaM, and Define.xml standards. It also executes the FDA published business rules for submission data. The following is a list of provided validation configurations. These are located in `components → config` directory:

- **SDTM**
  - SDTM 3.1.1 (FDA)
  - SDTM 3.1.2 (FDA)
  - SDTM 3.1.3 (FDA)
  - SDTM 3.1.2 (PMDA)
  - SDTM 3.1.3 (PMDA)
  - SDTM 3.2 (PMDA)
  - SDTM 3.2
- **SEND**
  - SEND 3.0 (FDA)
- **ADaM**
  - ADaM 1.0 (PMDA)
  - ADaM 1.0
- **Define.xml**
  - Define.xml (PMDA)
  - Define.xml

Validation configurations designated with "(FDA)" represent executable versions of FDA published business rules for submission data. Currently FDA only publishes rules for SDTM and SEND, but ADaM and Define.xml are coming in the future. PMDA is similar, but instead does not recognize 3.1.1 and does publish ADaM version which overrides some rule severities (i.e. ERROR to REJECT or WARNING)

To run the Validator

- Click on Validator icon on the home screen or select Validator from navigation menu
- Select the Standard and version by picking the desired Configuration
- Select Source Format, with SAS XPORT, Delimited text, or Dataset-XML as supported options
- The default and recommended Report Format is Excel, but user can change to CSV if expecting more than about 1 million issues thus exceed the Excel record limit. To control the amount of issues generated you can also modify the reporting options (gear to the right of Report Format) and reduce the Excel Message Limit that controls the number of detail records that will be created for each issue. A user can also modify the validation report file name by changing the File Name Format setting.
- There are 2 options to select the datasets for validation. You can either drag and drop them into the Source Data area or browse and select them using the Browse button.
- If Define.xml is available, make sure to select it when validating SDTM, SEND, or ADaM datasets to ensure consistency between metadata definition and actual datasets. While this is optional for XPT files, it's required when validating data in Dataset-XML format.
- Finally, ensure that correct CDISC CT and external dictionary versions are selected. If a version is not selected, Validator will automatically use the latest available version.
- Now click Validate to start the validation

Once validation has completed a report is generated. It can be opened from the validation summary page or it can be found in the following folder:

`components → reports`

The Excel validation report includes 4 worksheets:

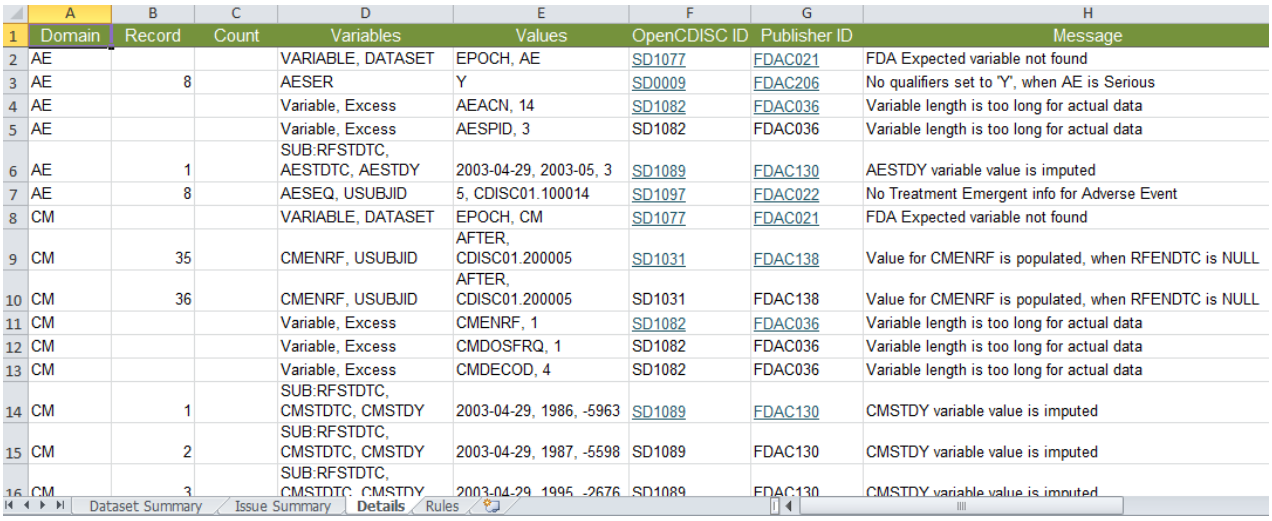
- Dataset Summary – a listing of validated datasets, number of records, errors, warnings, and notices for each. The header section shows select validation options, such as the version of dictionaries, configuration, and Validator version.
- Issue Summary – a listing of issues grouped by dataset, with both Pinnacle 21 and FDA IDs (if available), severity, and the number of reported instances for each issue
- Details – information for each reported issue instance, including the dataset name, records number, and values of the effected variables
- Rules – a listing of validation rules in executed validation configuration, including the detailed descriptions. Pinnacle 21 and FDA ID values on other tabs are hyperlinked to the Rules worksheet for quick reference.

	A	B	C	D	E	F	G	H	I
1	Pinnacle 21 Validator Report								
2									
3	Configuration: ...\P21C\2.1.1\components\config\SDTM 3.1.3 (FDA).xml								
4	Define.xml: C:\InputData\define2-0-0-example-sdtm.xml								
5	Generated: 2016-04-08T19:48:09								
6	CDISC CT Version: 2016-03-25								
7	MedDRA Version: 18.0								
8	UNII Version: 2016-01-21								
9	NDF-RT Version: 2016-03-07								
10	Software Version: 2.1.1								
11									
12	Processed Sources								
13	Domain	Label	Class	Source	Records	Rejects	Errors	Warnings	Notices
14	GLOBAL	Global Metadata	--	--	--	0	0	0	0
15	AE	Adverse Events	EVENTS	ae.xpt	16	0	4	12	0
16	CM	Concomitant Medications	INTERVENTIONS	cm.xpt	36	0	26	3	0
17	DA	Drug Accountability	FINDINGS	da.xpt	16	0	3	1	0
18	DM	Demographics	SPECIAL PURPOSE	dm.xpt	5	0	4	8	0
19	DS	Disposition	EVENTS	ds.xpt	14	0	4	1	0
20	EG	ECG Test Results	FINDINGS	eg.xpt	56	0	4	113	0
21	EX	Exposure	INTERVENTIONS	ex.xpt	17	0	2	1	0
22	IE	Inclusion/Exclusion Criteria Not Met	FINDINGS	ie.xpt	1	0	0	1	0
23	LB	Laboratory Tests Results	FINDINGS	lb.xpt	83	0	5	52	0

Screenshot 6. Dataset Summary worksheet

	A	B	C	D	E	F
1	Pinnacle 21 Validator Report					
2						
3	Configuration: ...\P21C\2.1.1\components\config\SDTM 3.1.3 (FDA).xml					
4	Define.xml: C:\InputData\define2-0-0-example-sdtm.xml					
5	Generated: 2016-04-08T19:48:09					
6	CDISC CT Version: 2016-03-25					
7	MedDRA Version: 18.0					
8	UNII Version: 2016-01-21					
9	NDF-RT Version: 2016-03-07					
10	Software Version: 2.1.1					
11						
12	Issue Summary					
13	Source	Pinnacle 21 ID	Publisher ID	Message	Severity	Found
14	AE					
15		SD0009	FDAC206	No qualifiers set to 'Y', when AE is Serious	Error	1
16		SD1082	FDAC036	Variable length is too long for actual data	Error	2
17		SD1089	FDAC130	AESTDY variable value is imputed	Error	1
18		SD0057	FDAC020	SDTM Expected variable AEBDSYCD not found	Warning	1
19		SD0057	FDAC020	SDTM Expected variable AEHLGT not found	Warning	1
20		SD0057	FDAC020	SDTM Expected variable AEHLGTCD not found	Warning	1
21		SD0057	FDAC020	SDTM Expected variable AEHLT not found	Warning	1
22		SD0057	FDAC020	SDTM Expected variable AEHLTCD not found	Warning	1
23		SD0057	FDAC020	SDTM Expected variable AELLT not found	Warning	1
24		SD0057	FDAC020	SDTM Expected variable AELLTCD not found	Warning	1
25		SD0057	FDAC020	SDTM Expected variable AEPTCD not found	Warning	1
26		SD0057	FDAC020	SDTM Expected variable AESOC not found	Warning	1
27		SD0057	FDAC020	SDTM Expected variable AESOCCD not found	Warning	1
28		SD1077	FDAC021	FDA Expected variable EPOCH not found	Warning	1
29		SD1097	FDAC022	No Treatment Emergent info for Adverse Event	Warning	1
30	CM					
31		SD1082	FDAC036	Variable length is too long for actual data	Error	3
32		SD1089	FDAC130	CMSTDY variable value is imputed	Error	22
33		SD1093	FDAC135	CMENDY variable value is imputed	Error	1
34		SD1031	FDAC138	Value for CMENRF is populated, when RFENDTC is NULL	Warning	2
35		SD1077	FDAC021	FDA Expected variable EPOCH not found	Warning	1

Screenshot 7. Issue Summary worksheet



	A	B	C	D	E	F	G	H
1	Domain	Record	Count	Variables	Values	OpenCDISC ID	Publisher ID	Message
2	AE			VARIABLE, DATASET	EPOCH, AE	SD1077	FDAC021	FDA Expected variable not found
3	AE	8		AESER	Y	SD0009	FDAC206	No qualifiers set to 'Y', when AE is Serious
4	AE			Variable, Excess	AEACN, 14	SD1082	FDAC036	Variable length is too long for actual data
5	AE			Variable, Excess	AESPID, 3	SD1082	FDAC036	Variable length is too long for actual data
6	AE	1		SUB:RFSTDTC, AESTDTC, AESTDY	2003-04-29, 2003-05, 3	SD1089	FDAC130	AESTDY variable value is imputed
7	AE	8		AESEQ, USUBJID	5, CDISC01.100014	SD1097	FDAC022	No Treatment Emergent info for Adverse Event
8	CM			VARIABLE, DATASET	EPOCH, CM	SD1077	FDAC021	FDA Expected variable not found
9	CM	35		CMENRF, USUBJID	AFTER, CDISC01.200005	SD1031	FDAC138	Value for CMENRF is populated, when RFENDTC is NULL
10	CM	36		CMENRF, USUBJID	AFTER, CDISC01.200005	SD1031	FDAC138	Value for CMENRF is populated, when RFENDTC is NULL
11	CM			Variable, Excess	CMENRF, 1	SD1082	FDAC036	Variable length is too long for actual data
12	CM			Variable, Excess	CMDOSFRQ, 1	SD1082	FDAC036	Variable length is too long for actual data
13	CM			Variable, Excess	CMDECOD, 4	SD1082	FDAC036	Variable length is too long for actual data
14	CM	1		SUB:RFSTDTC, CMSTDTC, CMSTDY	2003-04-29, 1986, -5963	SD1089	FDAC130	CMSTDY variable value is imputed
15	CM	2		SUB:RFSTDTC, CMSTDTC, CMSTDY	2003-04-29, 1987, -5598	SD1089	FDAC130	CMSTDY variable value is imputed
16	CM	3		SUB:RFSTDTC, CMSTDTC, CMSTDY	2003-04-29, 1995, -2676	SD1089	FDAC130	CMSTDY variable value is imputed

Screenshot 8. Details worksheet

The validation report is also available in CSV format, which includes only the information from the Details worksheet above. The CSV format is useful when the number of validation issues exceed 1 million records, an Excel limit. This format is also helpful when the results are used in SAS programs or loaded into a database.

Helpful tips:

- Always use the most recent version of Pinnacle 21 Community
- Don't forget to configure MedDRA and SNOMED
- When validating ADaM, include SDTM DM, AE and EX domains for cross-reference validation
- Validate Define.xml file first before using it in SDTM, SEND, and ADaM validation
- If found a bug, report it to Pinnacle 21 so that it gets fixed promptly (support@opencdisc.org). Please include a message of what you were doing along with environment details and any supporting files and screenshots. The application should automate the environment details

RUNNING VALIDATOR FROM CLI AND SAS

Using the Command Line Interface (CLI) can enable automation of organization's specific workflows. The CLI can be kicked off periodically or when new data has been placed in a specific location. Another usage example is to run Validator at the end of your SAS program as a QC step.

Here is an example of how to run the Validator using CLI from SAS x command:

```
x
  java
  -jar "C:\pinnacle21-community-2.1.1\components\lib\validator-cli-2.1.1.jar"
  -type=sdtm
  -source:type=sas
  -source="C:\InputData\*.xpt"
  -config="C:\pinnacle21-community-2.1.1\components\config\SDTM 3.1.3 (FDA).xml"
  –config:cdisc=2014-03-28
  -config:meddra=8.0
  -report="C:\pinnacle21-community-2.1.1\components\reports\ValidationReport.xls"
  -report:type=excel
  –report:overwrite=yes
;
```

For Validator command line syntax please refer to online documentation available at
http://www.opencdisc.org/using-opencdisc-validator-cli

USING DEFINE.XML GENERATOR

Creating a high quality Define.xml has in the past required a solid knowledge of the standard and mastery of XML. Pinnacle 21's goal in creating Define.xml Generator was to eliminate the need for the latter and lower the barrier to learning and becoming proficient with the standard. Define.xml Generator is based around Excel, which allows you to focus on the metadata content instead of the complex XML syntax.

There are two basic approaches to creating Define.xml, both of which are supported by Define.xml Generator:

- **Descriptive approach** – aims to create a Define.xml from completed study datasets. This typically occurs after all data has been collected and the study has been closed. The datasets are scanned and all possible metadata is extracted into a specification. The developer then fills out the missing components of the specification and then generates the Define.xml. This is currently the most popular approach, because it's perceived to be the cheapest and only needs to be performed if study data will be included in an FDA submission. This approach however has many disadvantages. Since Define.xml is only available at the end, it's not possible to utilize it to drive data mapping or validation during study conduct to ensure the metadata matches the actual data. Define.xml files created using the descriptive approach also seem to be the least useful as sponsors seem to populate only the minimally required content and leave out important information such as the Value Level metadata.
- **Prescriptive approach** – aims to create a Define.xml during study setup. In this scenario, Define.xml is used as the study specification for data collection, data mapping, and validation. A developer typically starts with the company's standards metadata specification or with a specification from a similar prior study. The goal is to define study specific metadata including expected datasets, variables, codelists, and value level metadata before any data has been collected. This specification is then converted into Define.xml and used by the company to verify that incoming study data matches the study specification. The prescriptive approach is becoming popular with sponsors who outsource study conduct to CROs including the creation of CDISC datasets. The sponsor and CRO use the Define.xml to communicate requirements and to ensure compliance of created datasets, which results in higher quality data and improves overall sponsor/CRO relationship.

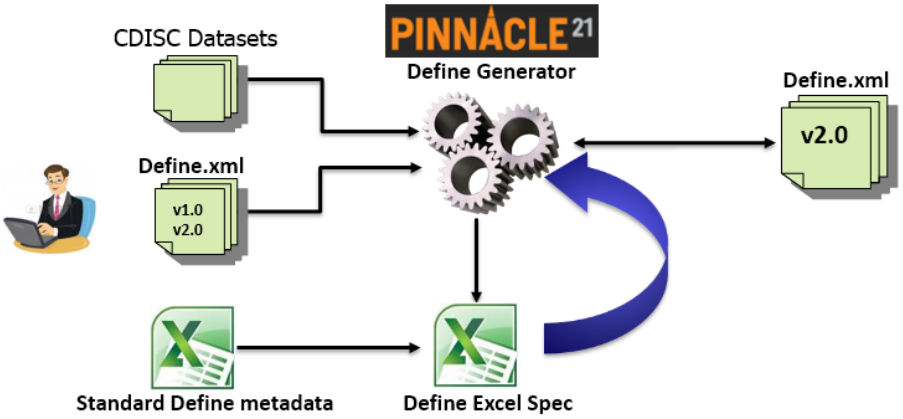


Figure 2. Approaches to creating Define.xml

The following sections describe how to use Define.xml Generator for both approaches.

CREATING DEFINE.XML USING DESCRIPTIVE APPROACH

Using the descriptive approach, Define.xml Generator starts with a completed set of SAS XPORT datasets and scans them to extract dataset and variable metadata. This metadata is then used to create and populate an Excel specification.

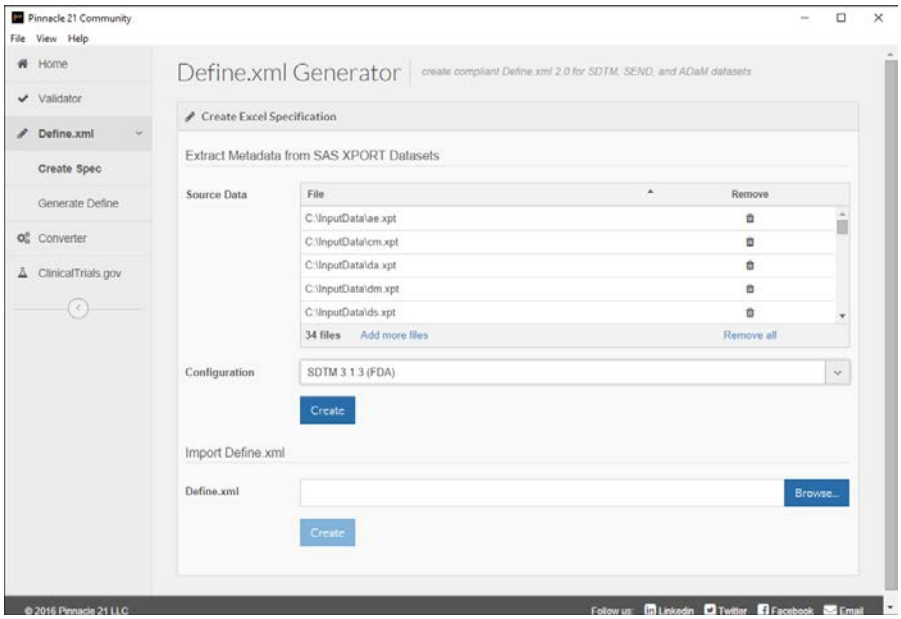
Creating Excel Specification

To run the Define.xml Generator to create an Excel specification

- Click on Define.xml Generator icon on the home screen or select Define.xml → Create Spec from navigation menu
- Drag and drop or browse to select the Source Data
- Select the standard and version used by the datasets by picking the desired Configuration. The Configuration will be used to supplement metadata extracted from the datasets with additional metadata stored in the configuration spec.

- Now click Create to start the metadata extraction process

An alternative option is to begin with an existing Define.xml to create and populate the Excel specification. This method could be used to migrate a Define.xml v1.0 into Define.xml v2.0 or to finish an incomplete Define.xml created outside of Pinnacle 21. It's also a great way for beginners to learn the tool and familiarize themselves with the template of the specification. Just take an existing high quality Define.xml and import it to generate a completed Excel specification.



Screenshot 9. Create Define.xml Spec screen

Once the Define.xml specification is created, open it in Excel and take a few minutes to review the 10 worksheet tabs that comprise the specification:

- Study – specifies basic information about the study including name, description, protocol, and standard
- Datasets – list of datasets and their corresponding metadata. The Dataset name and Description was extracted from the datasets while the remaining information was merged from the standard configuration. You need to review and update the information as necessary, especially the Structure and Key Variables that are study specific.
- Variables – a list of variables found by the scanning process. Just like datasets, much of the information was extracted directly from SAS XPORT files including Variable name, Label, Data Type, and Length. The remaining columns will need to be completed by the user.
- ValueLevel, WhereClauses, Codelists, Dictionaries, Methods, Comments, and Documents – are unpopulated but provide a clear template for users to follow to complete the specification.

	Order	Dataset	Variable	Label	Data Type	Length	Codelist	Origin	Pages	Method	Comm
2	1	AE	STUDYID	Study Identifier	text	7		Protocol			
3	2	AE	DOMAIN	Domain Abbreviation	text	2	AE.DOMAIN	Assigned			
4	3	AE	USUBJID	Unique Subject Identifier	text	14		Derived		USUBJID	
5	4	AE	AESEQ	Sequence Number	integer	1		Derived		SEQ	
6	5	AE	AESPID	Sponsor-Defined Identifier	text	4		CRF	21		
7	6	AE	AETERM	Reported Term for the Adve	text	25		CRF	21		
8	7	AE	AEMODIFY	Modified Reported Term	text	9		Assigned			
9	8	AE	AEDECOD	Dictionary-Derived Term	text	18	AEDICT_F	Assigned			
10	9	AE	AEBODSYS	Body System or Organ Class	text	52	AEDICT_F	Assigned			
11	10	AE	AESEV	Severity/Intensity	text	8	AESEV	CRF	21		
12	11	AE	AESER	Serious Event	text	1	NY	CRF	21		
13	12	AE	AEACN	Action Taken with Study Tre	text	30	ACN	CRF	21		
14	13	AE	AEREL	Causality	text	16	AEREL	CRF	21		
15	14	AE	AESTDTC	Start Date/Time of Adverse	date			CRF	21		
16	15	AE	AEENDTC	End Date/Time of Adverse E	date			CRF	21		

Study Datasets Variables ValueLevel WhereClauses Codelists Dictionaries Methods Comm

Screenshot 10. Define.xml Excel specification

Completing Excel Specification

At this point a user has many options to complete the specification. One option is to just follow the template and manually fill out the rest of the specification. Another option is to use VLOOKUP Excel function to merge with external metadata contained in mappings specifications, controlled terminology listings, etc.

Whatever options you choose, use the following helpful tips to overcome Excel data entry limitations and other common issues that could prevent you from generating a valid Define.xml file:

- Be careful of Excel auto-correction, like “ACN” → “CAN”
- When copying and pasting from Word, make sure to use Paste Special to avoid introducing special characters that are not allowed in XML and could lead to an unreadable Define.xml file.
- Define.xml is case sensitive, where “COUNTRY” is not the same as “Country”. So please use consistent case, especially for ID columns
- Pay special attention to ID columns and how they are referenced from other tabs
    - Remove trailing space characters. They are difficult to notice, so just use the Excel TRIM function methodically to remove them.
    - A Codelist assigned to a Variable or ValueLevel item must match an ID value defined on the Codelist tab
    - A Where Clause on the ValueLevel item must match an ID value defined on the WhereClauses tab
    - A Comment on the Dataset, Variable or ValueLevel tabs must match an ID value defined on the Comments tab
    - A Document on the Comments or Methods tab must match an ID value defined on the Documents tab
- When Origin=CRF, then Pages column should be populated
- When populating Codelists tab, make sure all codelist items available to the investigator should be included, not just the ones collected. This means the additional codelist items on the annotated CRF should be added / appended to the terms that are found in the data.
- Value level metadata should be populated for all SUPPQUAL and FINDINGS datasets
- When creating a complex Where Clause with multiple Variable/Value conditions, populate each condition on a separate row on the WhereClauses tab, but give them the same ID.
- Include at least the Reviewers Guide and Annotated Case Report Form (acrf.pdf) in Documents tab, but additional documents such as Complex Algorithm are recommended.
- When Comment or Method descriptions become too long, it is recommended to include them as a separate document defined on the Documents tab and reference on the Methods and Comments.
- The Href column on the Documents tab should contain a relative path to the document with the exact file name of the document. This will ensure that the links in Define.xml are generated correctly.

Get Additional Help

For additional information on how to populate the various tabs in Define.xml Excel specification, refer to the YouTube recording of an Pinnacle 21 webinar at https://www.youtube.com/watch?v=8PzYO0YlO0I. This webinar shows how to creating a Define.xml with Pinnacle 21 Enterprise, which addresses many of the issues described above by automating additional tasks and by safeguarding users from common data entry mistakes. Pinnacle 21 Enterprise can help you with:

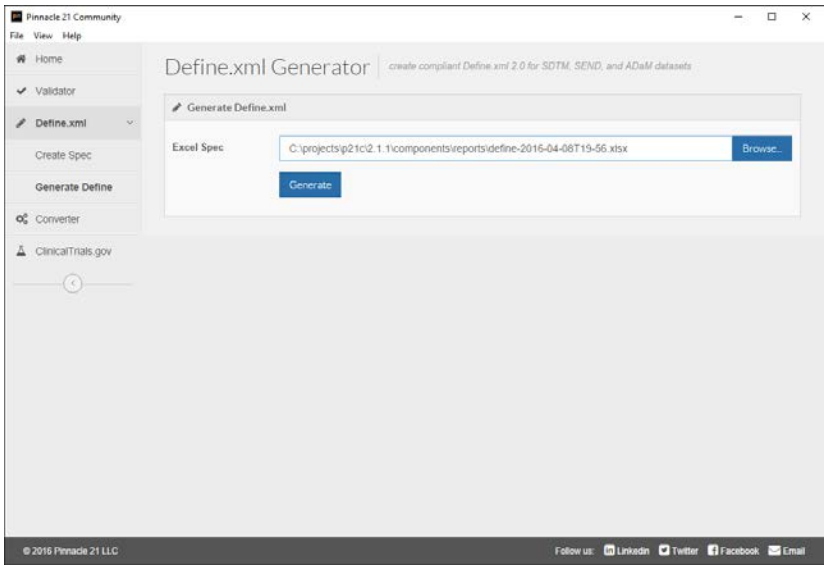
- Automatically populating Codelists and ValueLevel tabs with metadata extracted by scanning the datasets
- Automatically populating Page numbers on Variables and ValueLevel tabs by scanning annotated CRF’s
- Provides real-time on screen validation that highlights errors in the Define.xml metadata helping users avoid common data entry issues and help identify the root cause when a problem occurs
- If your organization utilizes global or therapeutic area standards for Methods and Comments, these too can be automatically populated to ensure consistently across studies and projects.
- Keeps versions and provides comparison tools to help track changes in your specification over time

Generating Define.xml

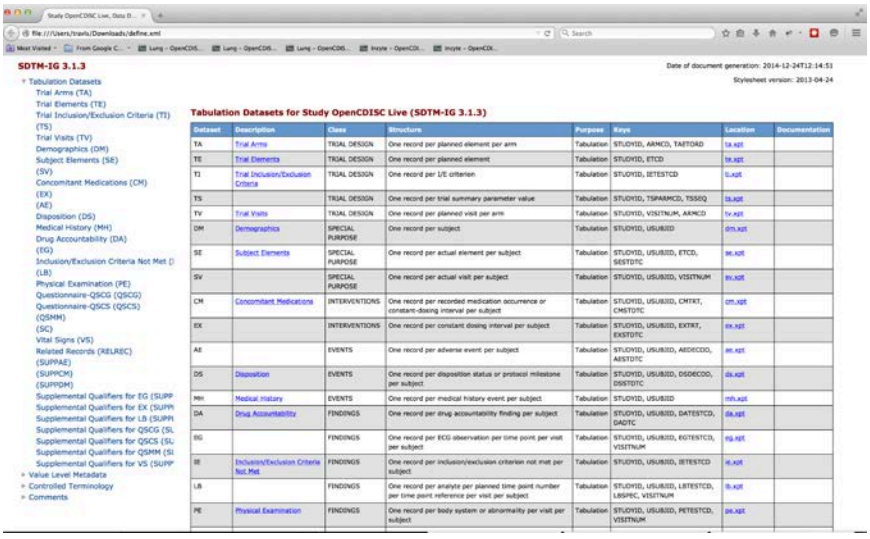
Once your Define.xml specification is complete, the final step is to generate the Define.xml.

To generate Define.xml

- Select Define.xml → Generate Define from navigation menu
- Browse to your Excel specification
- Now click Generate to start the generation process
- Once complete, click Open Define.xml button to open and view the completed Define.xml in a web browser



Screenshot 11. Generate Define.xml screen



Screenshot 12. Sample Define.xml

CREATING DEFINE.XML USING PRESCRIPTION APPROACH

Using the prescriptive approach, Define.xml can be created by starting from your organization's existing metadata to populate the study specific Excel specification. The Excel specification can be populated from metadata that is kept in a metadata repository, or from metadata that was created and maintained in a standard Excel specification and used as a template. Standards can be managed at different levels, such as global or therapeutic areas. Regardless how the standards are managed by an organization, this standard metadata can then be copied or applied to each new study during study setup.

Once study's Excel specification is copied from the standard, the metadata that is not utilized by the study, such as datasets in the standard that are not being used by the study, are removed. Additional study specific items can then be added to the Excel specification. To get started, not all study specific items need to be added immediately. First focus on items such as Variable definitions, Codelists, and Value Level metadata, since these help facilitate study execution and data validation.

Study information such as the set of all terms that should be available in the Electronic Data Capture system can be provided in the excel spec for these to be implemented by the Clinical Research Organization. Alternatively, if there are other processes in place to create the Electronic Data Capture system already, then these terms may be extracted from the EDC system and placed in the Codelists tab. Extracting these from the source specification or electronic system is better approach than scanning the data to find the terms, because inevitably there are additional terms which need to be appended to the scanned data in order to reflect the complete set of terms that the investigator has available to select.

Once the metadata is defined in the Excel specification, or instantiated as Define.xml generated from the specification, it can be shared with Clinical Research Organizations, or mapping programmers so they can develop the data collection system and data mapping programs. By including value level metadata that details when certain codelists, data types, and mandatory properties should be applied to results for specific test codes, the Define.xml serves not only as the spec for development and validation, but it is ready to be submitted to agencies to aid their analysis.

As study startup activities proceed and additional deliverables are developed, they can be used to more fully complete the Excel specification. For example, when annotated CRFs become available the page numbers can be applied to the Variable and Value Level items. This information is generally not needed early in study startup so they can be deferred.

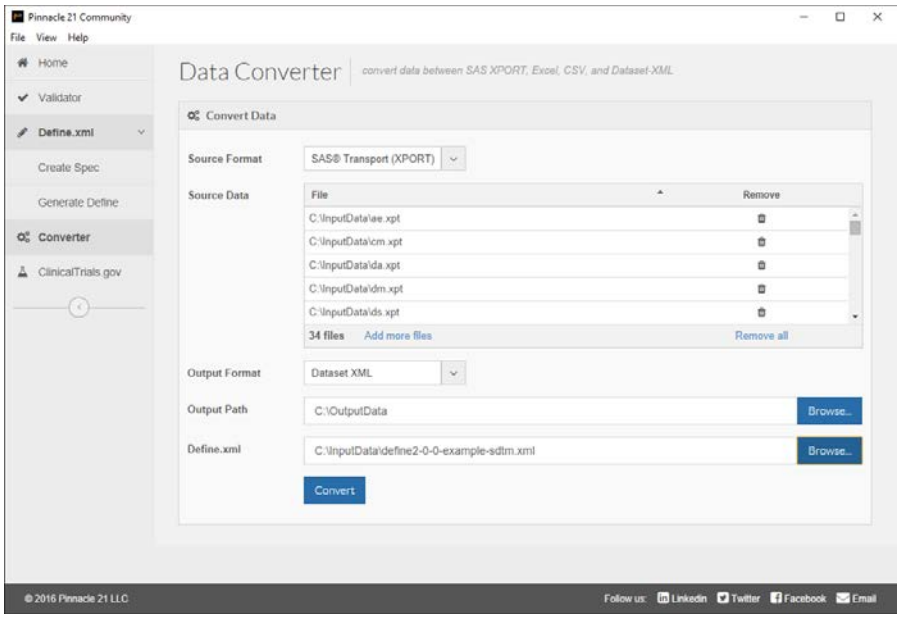
Later, Methods and Comments can be assigned if they are not already implemented by copying from the standard. Methods should be applied to all derived Variables and Value Level items, and Comments should be applied to Datasets, Variables and Value Level metadata as appropriate for the study. As these components are added, Define.xml can be regenerated using Define.xml Generator and then check for consistency with the datasets using the Validator.

USING DATA CONVERTER

The Data Converter can be used to covert SAS XPT files to various formats, MS Excel, CSV, and Dataset-XML. Clinical or data management personnel, or anyone who does not frequently use XPT file types, may find this useful. Mac OS users may also find this useful because SAS viewer is not available for Mac.

When converting to MS Excel, three tabs are created 1) dataset metadata 2) variable metadata and 3) data records.

When creating Dataset-XML, a define.xml must be specified. To create the Define.xml, a pre-existing Define.xml can be used, or one can be created using the Pinnacle 21 define.xml generator. Dataset-XML format does not store metadata needed to work with data files. Therefore variable names, data type, and other vital metadata can be provided only through a supplemental define.xml file. Pinnacle 21 Define.xml Generator enables source XPT files to be scanned and generate a define.xml file with all expected basic metadata to work with data to convert it into Dataset-XML format.



Screenshot 13. Data Converter screen

USING DATA CONVERTER FROM COMMAND LINE

From a command prompt, a user can run the following to get access to the help menu. This will guide them in how to actually use the program on the command line: `java -jar data-converter-1.0.1.jar –-help`

```
usage: java -jar data-converter-1.0.1.jar  -s <source> -i <sourceType> -o <output> -e
<outputType> -d <define> -c <config>
```

Parameter	Meaning	Description
-s	source	Full path to source data files.
-i	sourceType	Data type of source files. Currently supported file types are [xpt] (default: xpt)
-o	output	Full path to place output data files
-e	outputType	Desired data type of output files. Currently supported file types are [xlsx, csv, xml] (default: xlsx)
-d	define	For use in converting to Dataset-XML. Path to the define.xml for your datasets. If you do not have a define.xml you must specify the path to a configuration file so one can be generated.
-c	config	For use in converting to Dataset-XML. Path to the configuration xml file for generating a define.xml. An incomplete define.xml will be generated based off of this configuration file. The configuration files are packaged with Pinnacle 21 Community.

Table 1. CLI parameters for Converter

Here is an example of how to run the Converter using CLI from SAS x command:

```
x
   java -jar C:\pinnacle21-community-2.1.1\components\lib\ data-converter-1.0.1.jar
   -s=C:\InputData\*.xpt
   -i=xpt
   -o=C:\OutputData
   -e=xml
   -d=C:\InputData\define.xml
;
```

It converts all SAS XPT files in “C:\InputData” folder into Dataset-XML format and put new files into “C:\OutputData” folder. Existing define.xml file is utilized during data conversion and should be included into a new Dataset-XML data package.

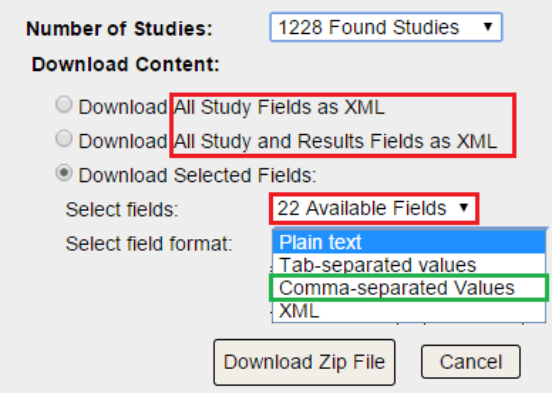
USING CLINICALTRIALS.GOV MINER

ClinicalTrials.gov miner is designed for clinical people, or anyone who prefers to look at an aggregated report of clinicaltrials.gov results. It generates results of queries it passes to Clintrials.gov as a summary table in MS Excel format focused on either outcome or site information.

Clinicaltrials.gov website provides patients, their family members, health care professionals, researchers, and the public with access to timely clinical study information across all 50 states and 188 countries. The scope includes over 187,000 interventional, observational and expanded access studies across a wide range of diseases and conditions. Users of the site work with data in the form of NCT study records which contain data such as disease/condition, intervention, study design, endpoints, eligibly criteria, site locations and contact information.

Query results are provided in 5 formats. The obvious HTML pages are available for browsing 1 study at a time. You may also download a zip file containing 1 xml file per study. The advantage is that all available fields are displayed or downloaded, albeit in a less than human readable format. The other three formats provide 1 file containing all studies, although severely limited to 20+ data fields. These formats are .xml, plain text and delimited (.csv, tab).

If you do not want the data limited to 20+ fields, it is quite challenging to deal with several xml files, 1228 in the example below. ClinicalTrials.gov contains well over 100 unique pieces of information which are listed in their public schema: https://clinicaltrials.gov/ct2/html/images/info/public.xsd. To date, only the CTTI AACT DB provides the entire clinicaltrials.gov with all its fields as a relational database requiring a heavy IT footprint to download, install and analyze. More problematic is that it hasn't been updated since September 2014.



Screenshot 14. Clinicaltrial.gov download results dialog

Originally developed to support FDA, NCI and CPATH research for priority therapeutic area endpoints analysis and TA standards development, Pinnacle 21 bridges the gaps by meeting several uses cases.

- Data from clinicaltrials.gov in real time
- Aggregation of many studies into one tabular file without being limited to certain fields
- Data in analysis ready excel format
- Tailored reports with respect to record structure, field selection and order which focus on specific data mining opportunities, i.e.
  - *Outcomes* – 1 record per study per outcome
  - *Sites* – 1 record per study per site
- Enrichments to automatically categorize and group records with user defined tags

MINING FOR AGGREGATED OUTCOMES / ENDPOINTS

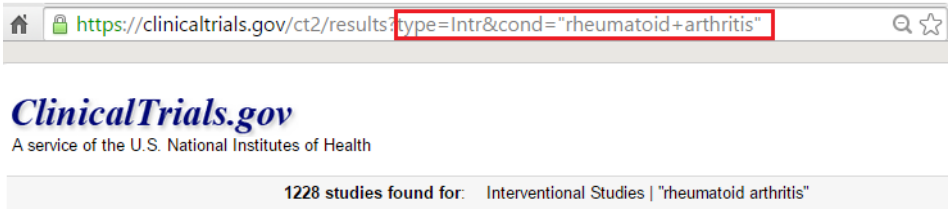
In clinical trials, an outcome or endpoint is an event that can be measured objectively to determine whether the intervention being studied is beneficial. The endpoints of a clinical trial are usually included in the study objectives. Some examples of endpoints are survival, improvements in quality of life, relief of symptoms, and disappearance of the tumor.

From Pinnacle 21 Community, select ClinicalTrials.gov and select the following 3 options

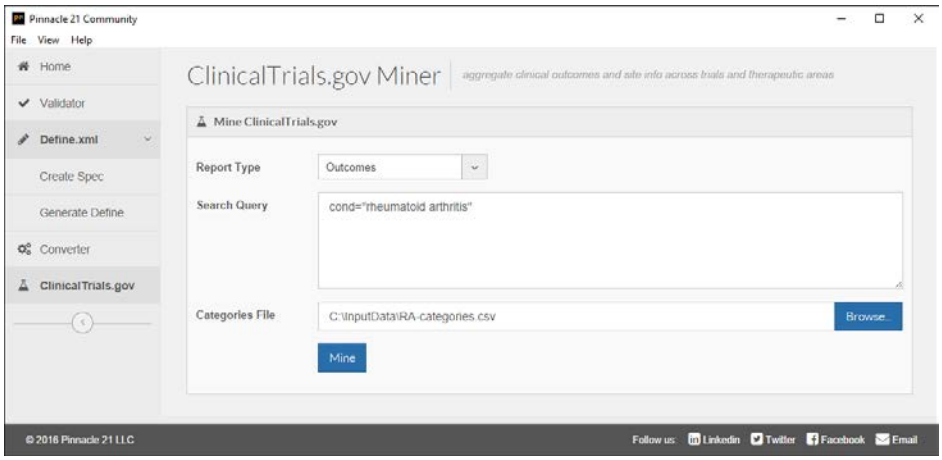
- **Report Type**: Outcomes
- **Search Query**: text string which conforms to a valid clinicaltrials.gov query. It can be beneficial to build and test it in clinicaltrials.gov using their basic or advanced search. Plug in the actual piece of the URL which follows the question mark as shown below. https://clinicaltrials.gov/ct2/results? If your query is malformed, clinicaltrial.gov may do one of two things. Reject it or attempt to download the all 187k studies to your machine. This is true of the actual website.
- **Categories File**: Choose path to any .csv file that contains 2 columns named ***SEARCH_TERM*** and ***DISPLAY_CATEGORY***. The SEARCH_TERM text is case insensitive. If the search term is found as part of the outcome fields, those records will be categorized with the DISPLAY_CATEGORY.

SEARCH_TERM	SEARCH_TERM
ACR	CLINICAL RESPONSE (ACR/EULAR)
AMERICAN COLLEGE OF RHEUMATOLOGY	CLINICAL RESPONSE (ACR/EULAR)
EULAR	CLINICAL RESPONSE (ACR/EULAR)
EUROPEAN LEAGUE AGAINST RHEUMATISM	CLINICAL RESPONSE (ACR/EULAR)
RESPONSE INDICATOR	CLINICAL RESPONSE (ACR/EULAR)
DAS	DAS(28)
DISEASE ACTIVITY	DAS(28)
CTLA4	ELISA CTLA4 RESPONSE
ELISA	ELISA CTLA4 RESPONSE
Enzyme-Linked Immunosorbant Assay	ELISA CTLA4 RESPONSE
RADIOGRAPHIC	Radiology Findings
RADIOLOGICAL	Radiology Findings
X-RAY	Radiology Findings

Table 2. Sample categories for Rheumatoid Arthritis



Screenshot 15. Browser address showing clinicaltrials.gov search URL



Screenshot 16. Miner screenshot for Outcomes

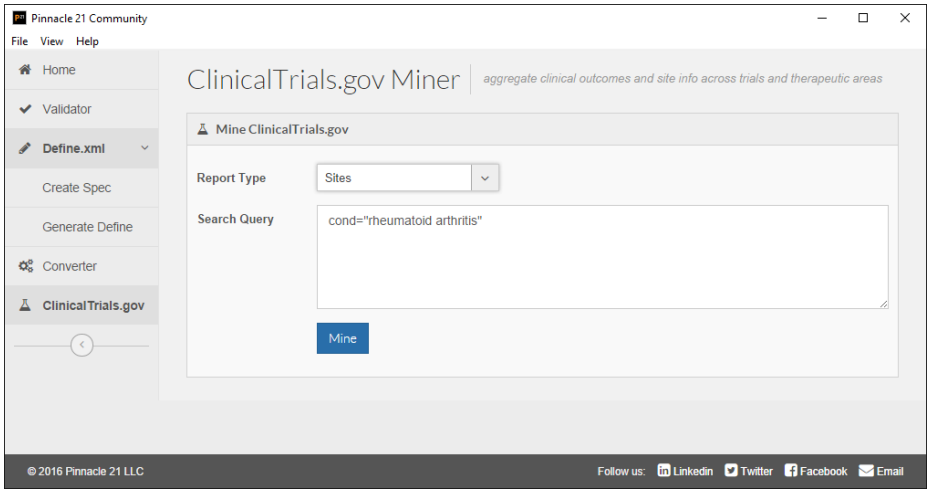
Study_ID	Categorization	End_Point_Type	End_Point_Title	End_Point_Details	Conditions	Intervention
NCT00000395		Primary	Determine the effect of Folic		Rheumatoid Arthritis|Adju	Methotrexate|Folinic acid|Folic acid
NCT00000395		Secondary	Determine the effect of folic		Rheumatoid Arthritis|Adju	Methotrexate|Folinic acid|Folic acid
NCT00000395	DAS(28)	Secondary	Correlate disease activity wi		Rheumatoid Arthritis|Adju	Methotrexate|Folinic acid|Folic acid
NCT00000401		Primary	Repeated measures analysis		Rheumatoid Arthritis	Oral bovine type II collagen
NCT00000401		Secondary	A Pearson correlation coeffic		Rheumatoid Arthritis	Oral bovine type II collagen
NCT00000416		Primary	job losses	Periods of work cessation	Rheumatoid Arthritis|Syst	Rehabilitation counseling
NCT00000416		Secondary	limitation in ability to work	Extent of on the job limita	Rheumatoid Arthritis|Syst	Rehabilitation counseling
NCT00000435	CLINICAL RESPONSE (ACR/EULAR)	Primary	Area under the curve or 'AUC		Rheumatoid Arthritis	dnaJ peptide|None-placebo
NCT00000435	CLINICAL RESPONSE (ACR/EULAR)	Secondary	Day 112 ACR 20 score		Rheumatoid Arthritis	dnaJ peptide|None-placebo
NCT00010335	MORTALITY/DEATH	Primary	Mortality		Systemic Sclerosis|System	Stem Cell Transplantation|CD34 select
NCT00010335		Secondary	Immune reconstitution, engr		Systemic Sclerosis|System	Stem Cell Transplantation|CD34 select
NCT00023205		Primary	Adherence to treatment	self report of medication (	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater
NCT00023205		Primary	Self efficacy	self report on self efficacy	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater
NCT00023205		Primary	Satisfaction with medical car	self report of satisfaction	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater
NCT00023205		Secondary	Heath Status	Health Assessment questi	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater
NCT00023205		Secondary	Mental Health	SF-36 (5 items)	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater
NCT00023205		Secondary	Understanding of medicatior	open ended questionnaire	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater
NCT00023205		Secondary	Perceived usefulness of mat	Interview	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater
NCT00023205		Secondary	Appointment keeping	Clinical record	Rheumatoid Arthritis|Psor	11th grade reading level arthritis mater

Screenshot 17. Miner screenshot for Outcomes Report. One record per Study outcome with categories if applicable

MINING FOR AGGREGATED SITE AND CONTECT INFORMATION

A patient or family member may want to query clinicaltrials.gov to determine what trials are taking place related to a specific disease condition in a particular geographical area.

Though other websites like Centerwatch aim to provide such a capability, they are not "real-time" because they need to first ingest the clinicaltrials.gov data, collate it and publish it. With Pinnacle 21 you can also choose where to do your filtering. It may be easier to just search all sites in clinical trials.gov first for say Rheumatoid Arthritis, then filter by recruiting status and City in excel. In addition to study id, sponsor, facility name and location, you will often get investigator, sub investigator, primary and backup contact names in addition to their phone numbers and email addresses shown in below screenshot.



Screenshot 18. Miner screenshot for Sites

FACILITY_STATUS	FACILITY_NAME	FACILITY_CITY	STATE	ZIP	INVESTIGATOR	CONTACT_NAME	CONTACT_PHONE	CONTACT_EMAIL
Recruiting	Pinnacle Research Group	Anniston	Alabama	36207	Vishala Chindalore, MD	April Bolt	256-236-0055	abolt@pinnacletrials.com
Recruiting	Pinnacle Research Group, LLC	Anniston	Alabama	36207	Vishala Chindalore, MD	Lorie Campbell	256-236-0055	lcampbell@pinnacletrials.
Recruiting	University of Alabama at Birmingham	Birmingham	Alabama	35294	Jeffrey R Curtis, MD	Jeffrey R. Curtis, MD	205-934-7727	
Recruiting	Investigator Site	Birmingham	Alabama	35216		Ablynx NV Belgium	329-262-0000	clinicaltrials@ablynx.com
Recruiting	Arthrocare Arthritis Care and Research PC	Gilbert	Arizona	85234	Michael Fairfax, DO	Chanel Tate	480-245-7663	research3@arthrocarepc.c
Recruiting	Arizona Arthritis & Rheumatology Research	Glendale	Arizona	85304	Eric Peters		480-626-6654	
Recruiting	Sun Valley Arthritis Center, LTD	Peoria	Arizona	85381	Joy Schechtman		623-566-3550	
Recruiting	Valley Arthritis Care, LLC	Phoenix	Arizona	85023	Nehad Soloman		613-851-2690	
Recruiting	University of Arizona College of Medicine	Tucson	Arizona	85724	Jeffrey Lisse		520-626-2655	
Recruiting	University of Arizona	Tucson	Arizona	85724	Charles Raison, MD	Kimberly Kelly, MPA	520-621-0181	kkelly@psychiatry.arizona
Recruiting	Arkansas Specialty Orthopaedics	Little Rock	Arkansas	72205	C. Lowry Barnes, M.D.	Cara Petrus	501-246-4439	cpetrus@hipkneearkansas
Recruiting	Valley Endocrine, Fresno	Fresno	California	93720	Paul Norwood		559-261-0992	
Recruiting	UCLA	Los Angeles	California	90095	Daniel E Furst, MD	Harold E Paulus, MD	310-825-6439	hpaulus@mednet.ucla.ed
Recruiting	UCLA, 1000 Veteran Avenue, Rehab Building	Los Angeles	California	90095	Dinesh Khanna, M.D., MS	Amber Bechtel	310-825-0425	abechtel@mednet.ucla.ed
Recruiting	UCLA Pediatric Pain Program Research Offic	Los Angeles	California	90064	Lonnie K Zeltzer, M.D.	Lonnie K Zeltzer, M.D.	310-825-0731	LZeltzer@mednet.ucla.ed
Recruiting	Sanguine Biosciences	Los Angeles	California	91403	Carlyn Crisostomo, MBA	Desiree Roman	877-864-3053	study@sanguinebio.com

Screenshot 19. Miner screenshot for Site Report. One record per Study site

CONCLUSION

Pinnacle 21 Community is a must have toolkit for every clinical programmer who works with CDISC standards. It's completely free and easy to use and install. It can be used as a desktop application or as part of an automated process within SAS. Regardless of your use case, you can take advantage of the available tools to ensure your data is CDISC compliant and FDA submission ready, create your study metadata in Define.xml v2.0 format, generate Excel, CSV or Dataset-XML files from SAS XPORT, and mine ClinicalTrials.gov to find information across all existing clinical trials.

This paper described in detail how clinical programmers could utilize the Pinnacle 21 Community toolkit, including Validator, Define.xml Generator, Data Converter, and ClinicalTrials.gov Miner, to simplify their daily tasks and create FDA compliant deliverables. It also provided helpful tips, best practices, and recommendations. The most important recommendation is to make sure you always use the latest version of the toolkit. This will ensure you are using the latest rules and controlled terminologies to stay compliant with the constantly evolving FDA regulatory requirements.

Pinnacle 21 has grown over the last 8 years from a small grassroots project to one adopted by most regulatory agencies, pharmaceutical and biotech companies, and clinical research organizations around the world. This would have not been possible without you, the Pinnacle 21 community. We would like to thanks all contributors for your continued support and hope to bring you additional value as we strive toward a common baseline for data quality and standards compliance.

USEFUL LINKS

- http://www.opencdisc.org/
- http://www.pinnacle21.net/
- http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm
- http://cdisc.org/standards-and-implementations
- http://clinicaltrials.gov

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Sergiy Sirichenko
Company: Pinnacle 21 LLC
Work Phone: 908-781-2342
E-mail: ssirichenko@pinnacle21.net

Name: Michael DiGiantomasso
Company: Pinnacle 21 LLC
Work Phone: 267-331-4433
E-mail: mike.digian@pinnacle21.net

Name: Travis Collopy
Company: Pinnacle 21 LLC
Work Phone: 267-331-4432
E-mail: tcollopy@pinnacle21.net