

An aerial photograph of a city, likely New York City, with a red overlay. The overlay is a semi-transparent red rectangle that covers the central part of the image, behind the text.

SM09: Using Python to Automate the Reviewer's Guide

Matt Methereell & Simon Purkess



Acknowledgements

P21 Enterprise Edition™

Much of this functionality is already provided by P21 Enterprise Edition™. If you're using this to validate your study, then use the Reviewer's Guides that they generate.

<https://www.pinnacle21.com>

Python-docx package

A really useful python library that allows us to edit Word documents at a metadata level. This paper would not be possible without this package.

<https://python-docx.readthedocs.io>

openpyxl package

Another useful python library that allows us to read data from Excel documents into Python.

<https://openpyxl.readthedocs.io>

| Agenda

- Reviewer's Guides
- Python-docx
- Trial Design Model (TDM) datasets
- Supplemental variables
- Conformance issues (Pinnacle 21)

Reviewer's Guides

These are documents to help an independent reviewer from the regulatory agencies to navigate the study submission. It summarises key points from the protocol, key statistical analysis, and anything else deemed relevant.

<https://advance.phuse.global/>

An important feature is that they have a clearly-defined structure that requires specific information in specific sections, so we are able to automate some of these.

| Reviewer's Guides

- In section 3.4.x, all of the supplemental variables are listed:

3.4.2 DS – Disposition

The following variables are mapped to SUPPDS:

QNAM	Description
SUPPVAR1	Supplemental Variable 1
SUPPVAR2	Supplemental Variable 2
SUPPVAR3	Supplemental Variable 3
SUPPVAR4	Supplemental Variable 4

Reviewer's Guides

- In section 4.2, all of the outstanding conformance issues are listed with associated explanations:

4.2 Issues Summary

Dataset	Diagnostic Message	Severity	Count	Explanation
AE	No Treatment Emergent info for Adverse Event	Warning	10	This issue is due to partial dates, so AETRTEM is correct as NULL. A TEAE analysis flag will be derived in ADAE based on the imputation rules for partial dates.
CM	CMROUTE value not found in 'Route of Administration Response' extensible <u>codelist</u>	Warning	16	New terms added to extensible <u>codelist</u> : 'OTHER'. The "Other Specify" is stored in SUPPCM.CMRTOTH.
GLOBAL	Missing PP dataset	Warning	1	PP is not needed/relevant for this study

Reviewer's Guides

- In Appendix I, all of the inclusion and exclusion criteria are listed:

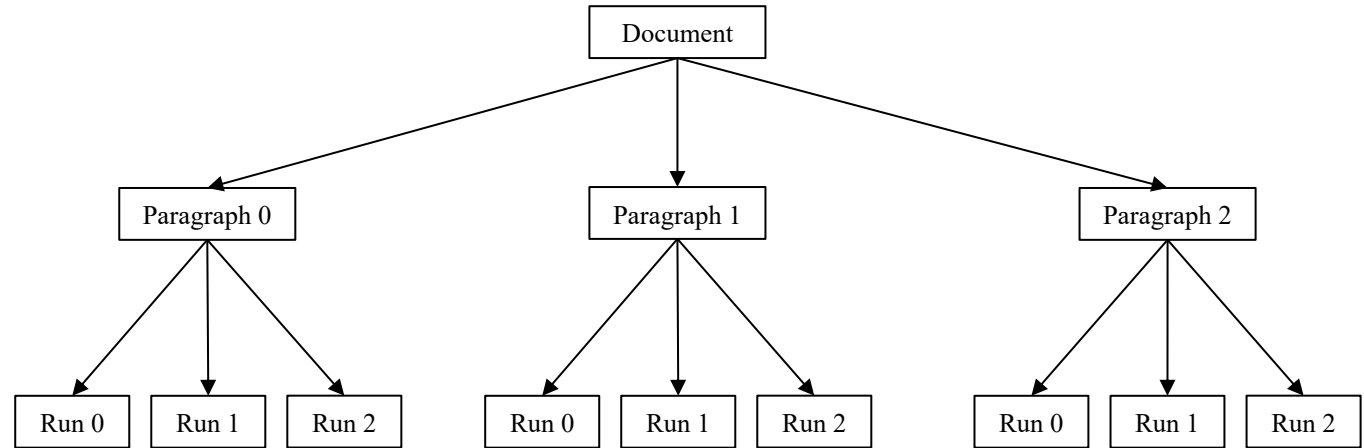
Appendix I: Inclusion/Exclusion Criteria

Protocol/ Amendment Version	Category	IETESTCD	Full Text of Criterion
V1.0	INCLUSION	IN01	Subject is aged ≥ 18 years of age at the screening visit.
V1.0	EXCLUSION	EX01	Treatment with any investigational product within the last 30 days.

Python-docx

This is a python package that reads and saves a Word document purely as a metadata object.

It sees each document as a collection of paragraphs, and each of those as a collection of runs.



Python-docx: Documents

- Top of the python-docx hierarchy is a document

```
#Import original CDISC template  
document = Document('Word/csdrp-template.docx')
```

Clinical Study Data
Reviewer's Guide

<Sponsor Name>

Study <Protocol Number>

CDISC Template Version 0000-mm-dd

Python-docx: Paragraphs

- A document is a collection of paragraphs

```
#Select paragraph  
paragraph = document.paragraphs[5]
```

```
#Display paragraph text  
print(paragraph.text)
```

<Sponsor Name>

0	
1	
2	Clinical Study Data
3	Reviewer's Guide
4	
5	<Sponsor Name>

Study <Protocol Number>

CSORG Template Version 0000-mm-dd

Python-docx: Runs of text

0 1 2 3

• A **paragraph** of text is made up of runs

- A run is a piece of text that has consistent attributes
- So anytime the font or size changes, it's a new run.
- Any time underlining, emboldening or any other stylization starts or ends, it's a new run.
- So how many runs are in the top bullet here?

Python-docx: Runs of text

- In order to update the text in a paragraph, we usually have to update the runs.

```
#Select run to update  
run = paragraph.runs[0]
```

```
#Extract attributes  
font = run.font
```

```
#Update text  
run.text = "Metherell-Purkess Industries inc."
```

```
#Save document  
document.save("Word/csdrp-updated.docx")
```

Clinical Study Data Reviewer's Guide

<Sponsor Name>

Study <Protocol Number>

Python-docx: Tables

- Tables exist directly under the document in the metadata hierarchy.

Appendix I: Inclusion/Exclusion Criteria

Protocol/ Amendment Version	Category	IETESTCD	Full Text of Criterion

```
#Identify inclusion/exclusion table
```

```
ti_table = document.tables[7]
```

```
#Check contents to confirm correct table
```

```
ti_table.cell(0,1).text
```

```
'Category'
```

Python-docx: Tables

- We don't need to worry about runs in formatted tables, as the text styling is already present in the table in the template.

```
#Identify inclusion/exclusion table
```

```
ti_table = document.tables[7]
```

```
#Identify row
```

```
row = ti_table.rows[1]
```

```
#Identify the cells in the row
```

```
row_cells = row.cells
```

```
#Update specific cell(s) in row
```

```
row_cells[0].text = "Final"
```

```
#Save document
```

```
document.save("Word/csdrp-updated.docx")
```

Protocol/ Amendment Version	Category	IETESTCD	Full Text of Criterion

Trial Design Model (TDM) datasets

There are typically five of these, and they are described in sections 2.3.x of the CSDRG.

- 2.3.1: TA – Trial Arms
- 2.3.2: TE – Trial Elements
- 2.3.3: TV – Trial Visits
- 2.3.4: TI – Trial Inclusion/Exclusion Criteria
- 2.3.5: TS – Trial Summary

At the minimum, the contents of TI will be placed in Appendix I, but it is not uncommon for a table of Visits to be included as an appendix, which will be a tabulation of the TV dataset.

I TI – Trial Inclusion/Exclusion Criteria

- Extending the table example earlier, we can populate it with the contents of a SAS dataset:

```
#Read in TI SAS dataset
ti_df = pd.read_sas('SAS/ti.sas7bdat', encoding="utf-8")

#Convert to array
ti_arr = ti_df.to_numpy()

#Enter each row of TI contents into new row of table
for studyid, domain, ietestedcd, ietest, iecat, tivers in ti_arr:
    row_cells = ti_table.add_row().cells
    row_cells[0].text = tivers
    row_cells[1].text = iecat
    row_cells[2].text = ietestedcd
    row_cells[3].text = ietest

#Save document
document.save("Word/csdrgr-updated.docx")
```

I TI – Trial Inclusion/Exclusion Criteria

- Extending the table example earlier, we can populate it with the contents of a SAS dataset:

Protocol/ Amendment Version	Category	IETESTCD	Full Text of Criterion
V1.0	INCLUSION	IN01	Subject is aged ≥ 18 years of age at the screening visit.
V1.0	EXCLUSION	EX01	Treatment with any investigational product within the last 30 days.

Supplemental variables

As alluded to earlier, every other domain has its own section in the Reviewer's Guide as 3.4.x.

This contains info about the type of data, and links to other datasets, and a list of all the variables in its supplemental dataset.

This list of supplemental variables can be easily extracted and placed into the CSDRG.

Supplemental Variables

- Rather than import one dataset, this time we want to import all datasets, and extract the distinct values of QNAM and QVAL.
- Therefore, we make a function that we can call multiple times:

```
#Function to extract variable names from SUPPxx SAS dataset
```

```
def supp_extract(fname):
```

```
    #Read in SUPP dataset
```

```
    file = 'SAS/'+fname
```

```
    df = pd.read_sas(file, encoding="utf-8")
```

```
    #Find unique values
```

```
    return pd.unique(df[['QNAM', 'QLABEL']].values.ravel())
```

Supplemental Variables

- Rather than import one dataset, this time we want to import all datasets, and extract the distinct values of QNAM and QVAL.
- Therefore, we make a function that we can call multiple times:

```
#Call macro to extract
```

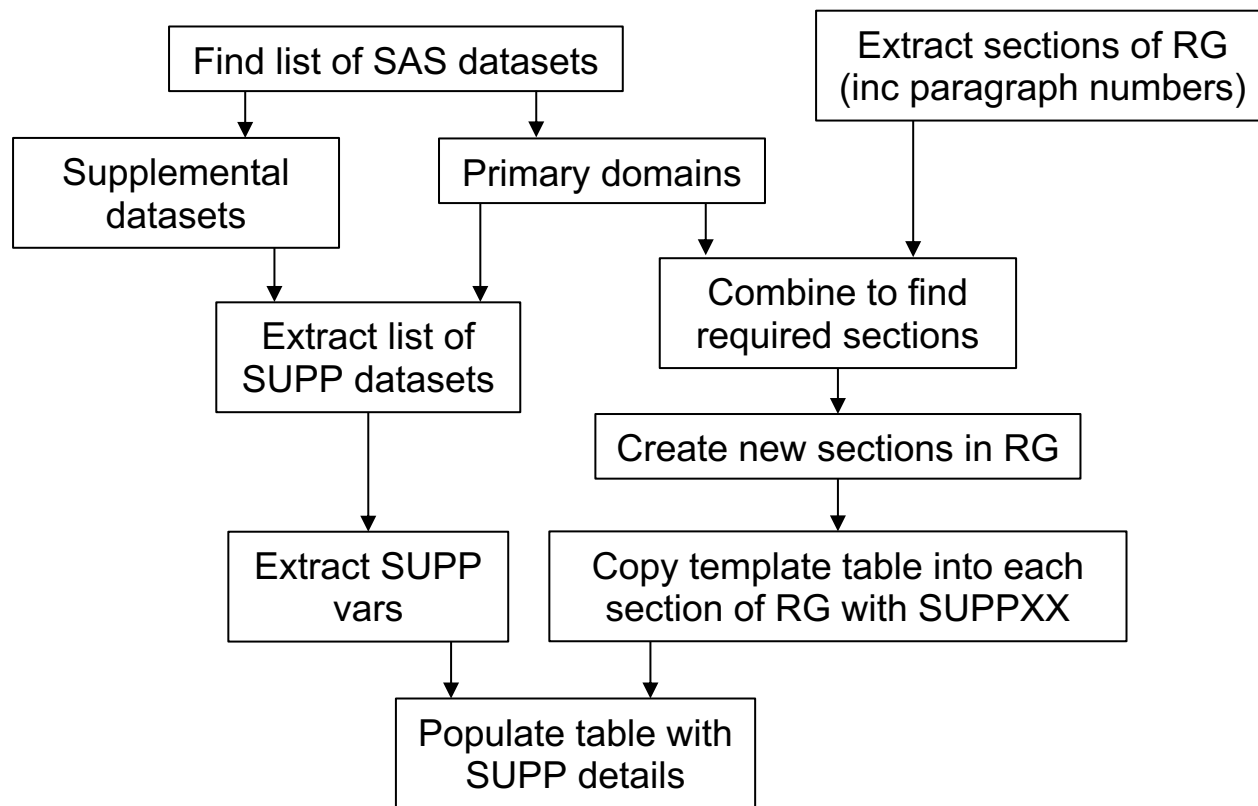
```
supp_extract('suppae.sas7bdat')
```

```
array(['SUPPVAR1', 'Supplemental Variable 1',  
      'SUPPVAR2', 'Supplemental Variable 2',  
      'SUPPVAR3', 'Supplemental Variable 3'], dtype=object)
```

QNAM	Description
SUPPVAR1	Supplemental Variable 1
SUPPVAR2	Supplemental Variable 2
SUPPVAR3	Supplemental Variable 3

Supplemental Variables

- We can't go through every function we create, but this is an overview:



Supplemental Variables

2.3.4 TI – Trial Inclusion/Exclusion Criteria
2.3.5 TS – Trial Summary
▲ 3. Subject Data Description
3.1 Overview
3.2 Traceability Flow Diagram
3.3 Annotated CRFs
▲ 3.4 SDTM Subject Domains
3.4.1 AE – Adverse Events
3.4.2 DS – Disposition
3.4.3 EX – Exposure
3.4.4 Dataset – Dataset Label
▲ 4. Data Conformance Summary
4.1 Conformance Inputs

3.4.1 AE – Adverse Events

(Text and/or supplemental qualifier inventory here)

3.4.2 DS – Disposition

(Text and/or supplemental qualifier inventory here)

3.4.3 EX – Exposure

(Text and/or supplemental qualifier inventory here)

3.4.4 Dataset – Dataset Label

(Text here)

QNAM	Description

Supplemental Variables

Headings	Pages	Results
2.3.5 TS – Trial Summary		
3. Subject Data Description		
3.1 Overview		
3.2 Traceability Flow Diagram		
3.3 Annotated CRFs		
3.4 SDTM Subject Domains		
3.4.1 AE – Adverse Events		
3.4.2 DS – Disposition		
3.4.3 EX – Exposure		
3.4.4 DA – Drug Accountability		
3.4.5 DM – Demographics		
3.4.6 EG – ECG Test Results		
3.4.7 FA – Findings About Events or Interventions		
3.4.8 IE – Inclusion/Exclusion Criteria Not Met		
3.4.9 LB – Laboratory Test Results		
3.4.10 SE – Subject Elements		
3.4.11 SV – Subject Visits		
3.4.12 VS – Vital Signs		
4. Data Conformance Summary		
4.1 Conformance Inputs		
4.2 Issues Summary		
4.3 Additional Conformance Details		
Appendix I: Inclusion/Exclusion Criteria		
Appendix II: Conformance Issues Details		

3.4.6 EG – ECG Test Results

(Text and/or supplemental qualifier inventory here)

QNAM	Description
EGVAR1	Supplemental EG Variable 1
EGVAR2	Supplemental EG Variable 2
EGVAR3	Supplemental EG Variable 3

3.4.7 FA – Findings About Events or Interventions

(Text and/or supplemental qualifier inventory here)

3.4.8 IE – Inclusion/Exclusion Criteria Not Met

(Text and/or supplemental qualifier inventory here)

3.4.9 LB – Laboratory Test Results

(Text and/or supplemental qualifier inventory here)

QNAM	Description
LBVAR1	Supplemental LB Variable 1
LBVAR2	Supplemental LB Variable 2
LBVAR3	Supplemental LB Variable 3

3.4.10 SE – Subject Elements

(Text and/or supplemental qualifier inventory here)

3.4.11 SV – Subject Visits

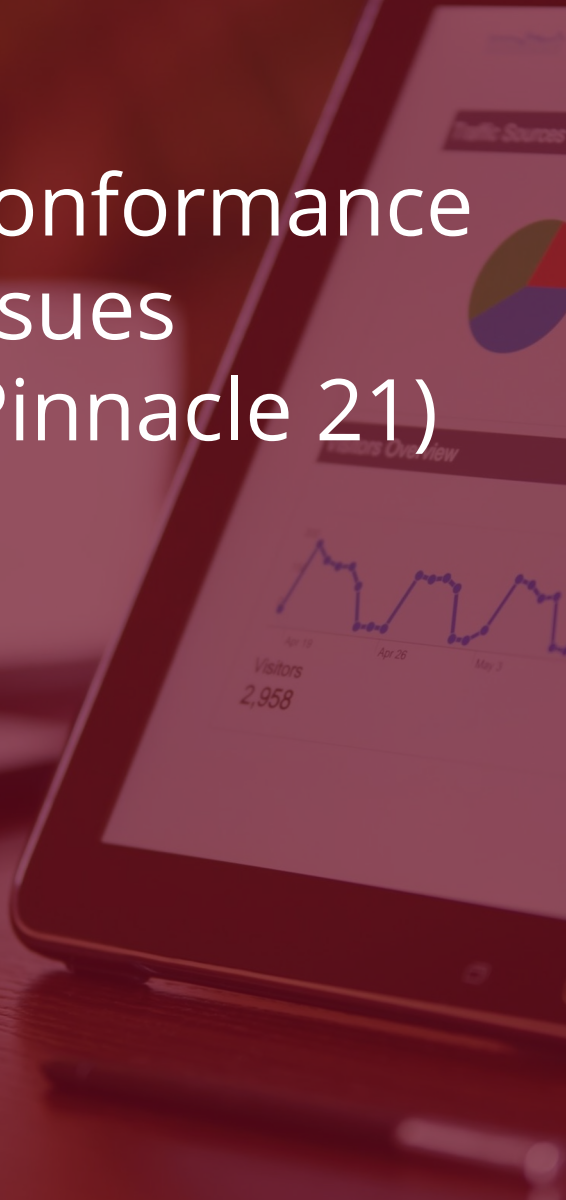
(Text and/or supplemental qualifier inventory here)

3.4.12 VS – Vital Signs

(Text and/or supplemental qualifier inventory here)

QNAM	Description
VSVAR1	Supplemental VS Variable 1
VSVAR2	Supplemental VS Variable 2
VSVAR3	Supplemental VS Variable 3

Conformance issues (Pinnacle 21)

A tablet is shown at an angle, displaying a dashboard. The dashboard includes a pie chart at the top right and a line graph below it. The line graph is labeled 'Visitors' and shows a fluctuating trend over time, with specific dates like 'Apr 11', 'Apr 26', and 'May 7' marked on the x-axis. Below the line graph, the text 'Visitors 2,958' is visible. The background of the slide is a dark red gradient.

- cSDRG Section 4 and ADRG Section 6 include a list of dataset conformance issues
- P21 Community provides these issues in an Excel spreadsheet
- We can use Python to copy the issues from the spreadsheet to the Reviewer's Guide

	A	B	C	D	E	G	H
1	From this... Pinnacle 21 Validator Report						
2							
3	Issue Summary						
4	Source	Pinnacle 21 ID	Message	Severity	Found	Action	Detail
5	AE						
6		CT2002	EPOCH value not found in 'Epoch' extensible codelist		236	None	Extended EPOCH values required for this study
7		SD1097	No Treatment Emergent info for Adverse Event		251	None	This study has a complex definition of Treatment Emergent across the two phases of the study, so this derived at ADaM level
8		SD1143	No Details info for AESMIE Adverse Event in SUPPAE domain		2	None	AESOSP is not part of sponsor standards - additional AE details are stored in AEDESC01/AEDESC02
9	CE						
10		SD0001	No records in data source		1	None	No relevant data collected
11		SD1076	Model permissible variable added into standard domain		11	None	No relevant data collected
12		SD1078	Permissible variable with missing value for all records		19	None	No relevant data collected
13		SD1149	Expected variable with missing value for all records		1	None	No relevant data collected
14	CM						
15		CT2001	CMSTRF value not found in 'Relation to Reference Period' non-extensible codelist		391	None	Synonyms from CRF used instead of NCI terms

Columns added
by P21

Columns added
by user

...to this

Dataset	Diagnostic Message	Count	Explanation
AE	CT2002: EPOCH value not found in 'Epoch' extensible codelist	236	Extended EPOCH values required for this study
AE	SD1097: No Treatment Emergent info for Adverse Event	251	This study has a complex definition of Treatment Emergent across the two phases of the study, so this derived at ADaM level
AE	SD1143: No Details info for AESMIE Adverse Event in SUPPAE domain	2	AESOSP is not part of sponsor standards - additional AE details are stored in AEDESC01/AEDESC02
CE	SD0001: No records in data source	1	No relevant data collected
CE	SD1076: Model permissible variable added into standard domain	11	No relevant data collected
CE	SD1078: Permissible variable with missing value for all records	19	No relevant data collected
CE	SD1149: Expected variable with missing value for all records	1	No relevant data collected
CM	CT2001: CMSTRF value not found in 'Relation to Reference Period' non-extensible codelist	391	Synonyms from CRF used instead of NCI terms

Importing from Excel to Python

- The openpyxl package can be used to read data from an xlsx document into Python
- The load_workbook function can be used to load the xlsx file into a Python data structure, which we can then manipulate into our desired format

```
#Import openpyxl library for XLSX, and docx for DOCX  
from openpyxl import load_workbook
```

```
#Create Workbook data structure  
wb = load_workbook(filename = 'Excel\p21 report.xlsx')
```

```
#Select sheet of interest - Issue Summary  
sheet = wb['Issue Summary']
```

Referencing Cells

- There are two ways of referencing a cell within a Worksheet:

`sheet['A1'].value` *#Letter/number coordinate reference*

`sheet.cell(1,1).value` *#Row/column coordinate reference*

- Method 1 corresponds closer to Excel notation, but Method 2 is easier to loop over
- We've created a shorthand function to improve readability – `c(row,col)` is equivalent to `sheet.cell(row,col).value`

```
def c(r,c,worksheet=sheet): #Shorthand function for cell references  
    return(worksheet.cell(r,c).value)
```

Creating a Table - Header

```
#Create table object  
p21table = document.add_table(rows=1, cols=4)
```

```
#Add header row  
header=p21table.rows[0].cells  
header[0].text="Dataset"  
header[1].text="Diagnostic Message"  
header[2].text="Count"  
header[3].text="Explanation"
```

Dataset	Diagnostic Message	Count	Explanation
AE	CT2002: EPOCH value not found in 'Epoch' extensible codelist	236	Extended EPOCH values required for this study
AE	SD1097: No Treatment Emergent info for Adverse Event	251	This study has a complex definition of Treatment Emergent across the two phases of the study, so

Creating a Table - Body

Problem: Dataset name is on a different row to the dataset's issue rows...

	A	B	C	D	E	F	G	H
1			Pinnacle 21 Validator Report					
2								
3			Issue Summary					
4	Source	Pinnacle 21 ID	Message	Severity	Found	Action	Detail	
5	AE							
6		CT2002	EPOCH value not found in 'Epoch' extensible codelist		236	None	Extended EPOCH values required for this study	
7		SD1097	No Treatment Emergent info for Adverse Event		251	None	This study has a complex definition of Treatment Emergent across	
8		SD1143	No Details info for AESMIE Adverse Event in SUPPAE domain		2	None	AESOSP is not part of sponsor standards - additional AE details a	
9	CE							
10		SD0001	No records in data source		1	None	No relevant data collected	
11		SD1076	Model permissible variable added into standard domain		11	None	No relevant data collected	
12		SD1078	Permissible variable with missing value for all records		19	None	No relevant data collected	
13		SD1149	Expected variable with missing value for all records		1	None	No relevant data collected	
14	CM							
15		CT2001	CMSTRF value not found in 'Relation to Reference Period' non-extensible codelist		391	None	Synonyms from CRF used instead of NCI terms	

Creating a Table - Body

Solution: Create dset variable to store the most recent variable name seen

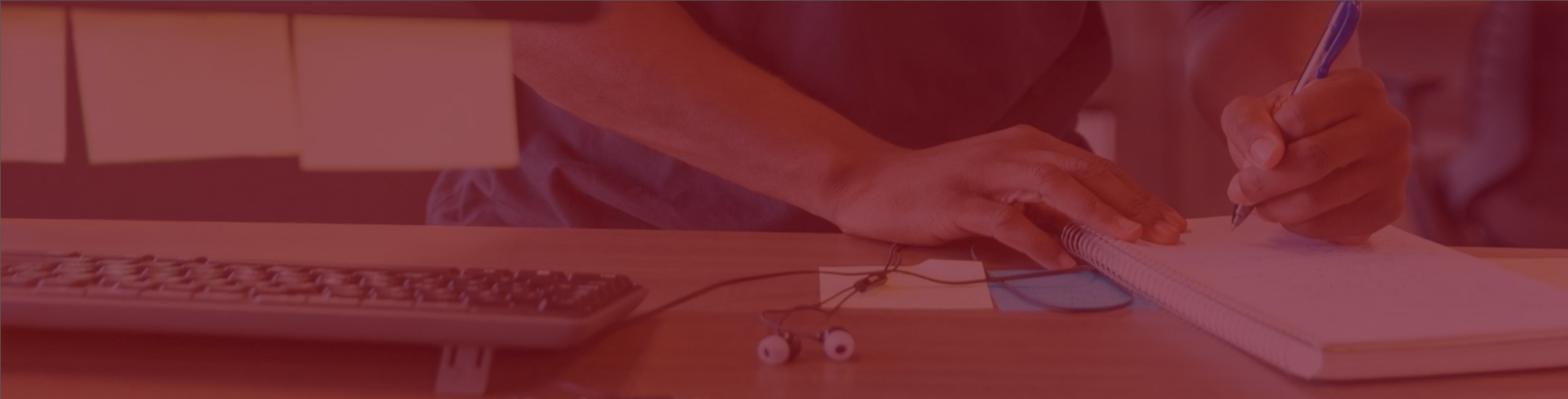
```
#Populate rows of P21 messages
dset=''
for i in range(5,sheet.max_row+1):
    #Dataset name is only given once per section so it must be retained
    if c(i,1) is not None:
        dset=c(i,1)
    else: #P21 message row
        row_cells=p21table.add_row().cells
        row_cells[0].text=dset
        row_cells[1].text=f'{c(i,2)}: {c(i,3)}'
        row_cells[2].text=str(c(i,5))
        row_cells[3].text=str(c(i,8))
```

Creating a Table – Final Result

Dataset	Diagnostic Message	Count	Explanation
AE	CT2002: EPOCH value not found in 'Epoch' extensible codelist	236	Extended EPOCH values required for this study
AE	SD1097: No Treatment Emergent info for Adverse Event	251	This study has a complex definition of Treatment Emergent across the two phases of the study, so this derived at ADaM level
AE	SD1143: No Details info for AESMIE Adverse Event in SUPPAE domain	2	AESOSP is not part of sponsor standards - additional AE details are stored in AEDESC01/AEDESC02
CE	SD0001: No records in data source	1	No relevant data collected
CE	SD1076: Model permissible variable added into standard domain	11	No relevant data collected
CE	SD1078: Permissible variable with missing value for all records	19	No relevant data collected
CE	SD1149: Expected variable with missing value for all records	1	No relevant data collected
CM	CT2001: CMSTRF value not found in 'Relation to Reference Period' non-extensible codelist	391	Synonyms from CRF used instead of NCI terms

Conclusion

- Python-docx allows us to easily manipulate Word documents programmatically.
- Many aspects of the Reviewer's Guides require data we already have stored in a structured way.
- Therefore, a lot of effort can be saved!



Any questions?

- matthew.metherell@phastar.com
- simon.purkess@phastar.com

Other Phastar presentations

- TT10: TFL Automation the CDISC 360 Way, Monday 5pm, Hall 6a
- PP03: When you're helping create life-saving treatments, who's looking after you? Tuesday 5:30pm