

# Single Day Event

Leveraging Open Source Technologies for  
Data Extraction from SDRG/ADRG  
Unstructured PDF files

Akihiro Ichikawa,  
Janssen Pharmaceutical K.K.  
(Johnson & Johnson Innovative Medicine)



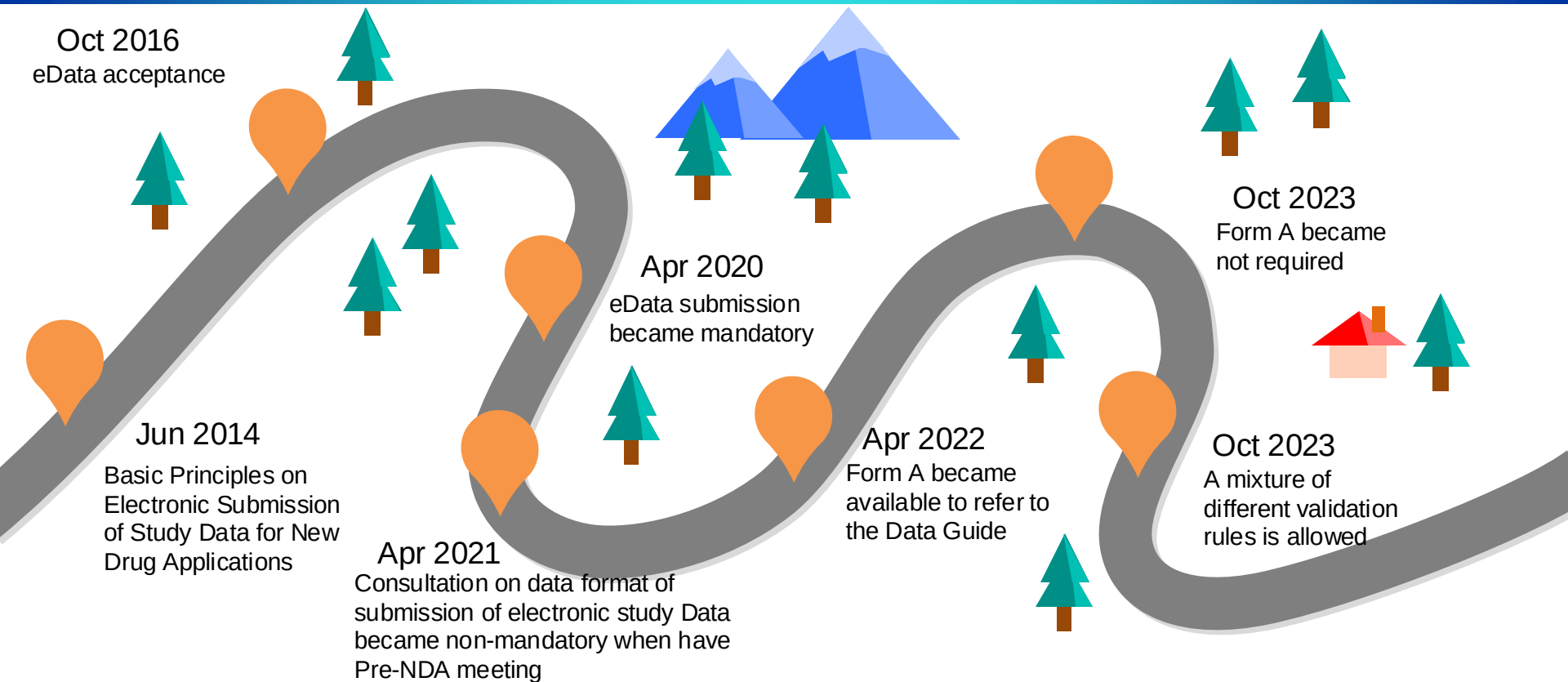
# Disclaimer

The views and opinions expressed in the following PowerPoint slides are those of the individual presenter and should not be attributed to Phuse and Johnson & Johnson.

This presentation is incomplete without accompanying verbal commentary.



# The Journey of eData Submission





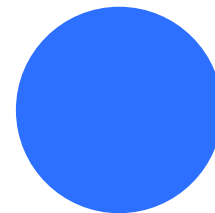
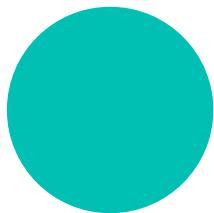
# Quality of eData Is Still Important



Accumulated Learning and Trust



Efficient Process



Quality is still important  
even if the regulation is relaxed



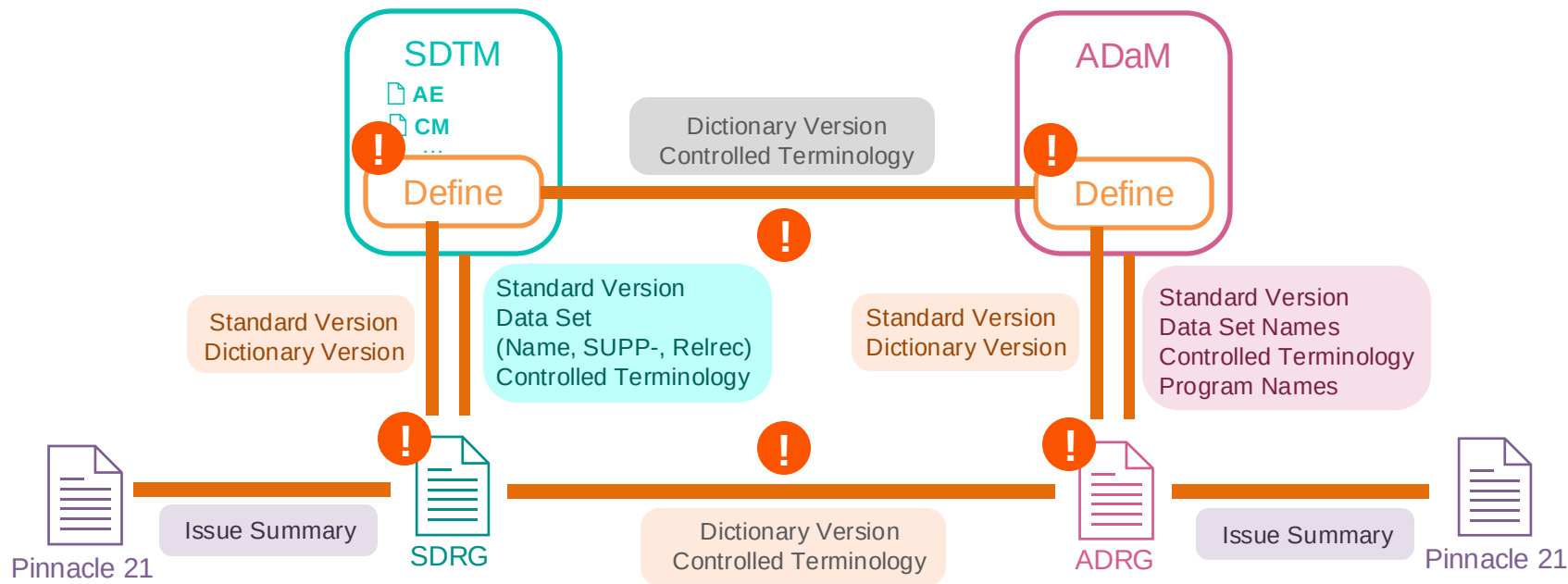


# Gaps in Quality



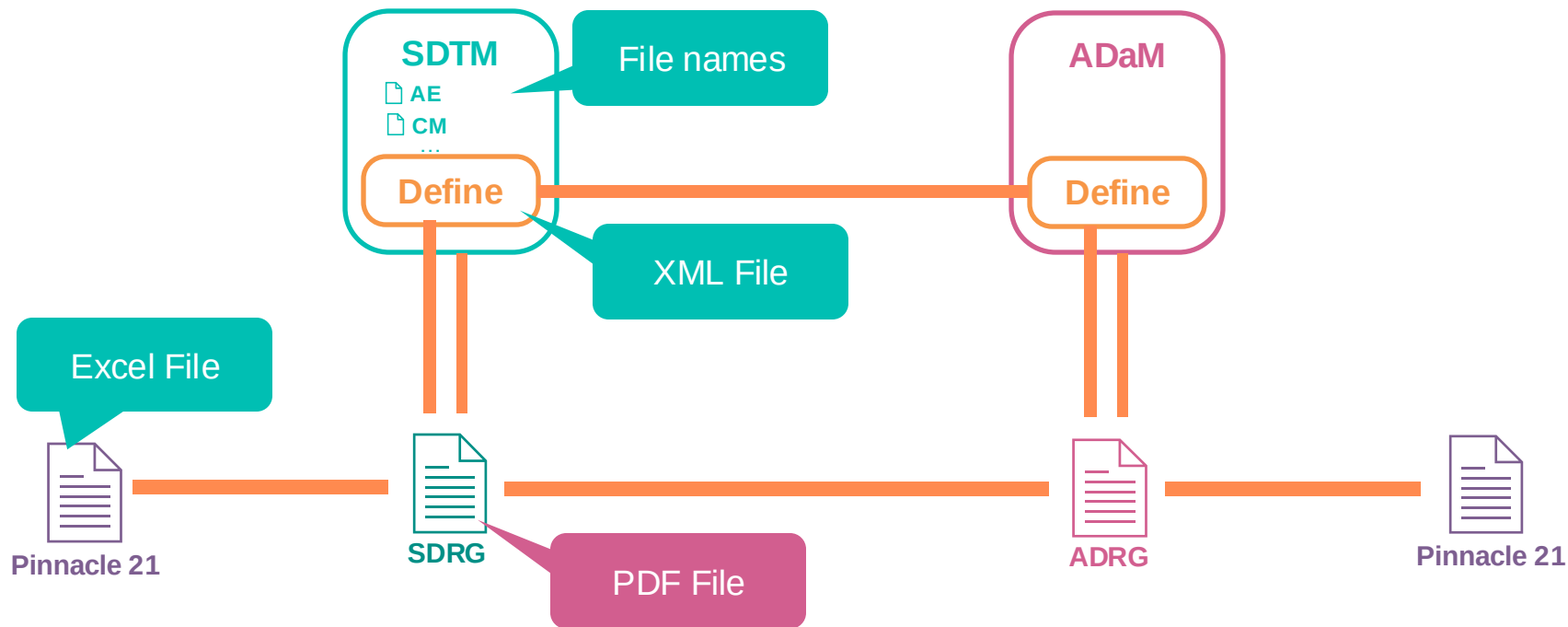
There are some gaps in the approach to package quality, and personnel for PMDA submission are under pressure to bridge those gaps.

# Potentially Inconsistency





# Check Target





# eData Checker

eData Checker 0.14

Submission: PMDA

Study:

**SDTM**

Folder:

P21 Report:

**ADaM**

Folder:

P21 Report:

**Ready**

Folder path of SDTM package

Path of P21 SDTM Report

Folder path of ADaM package

Path of P21 ADaM Report





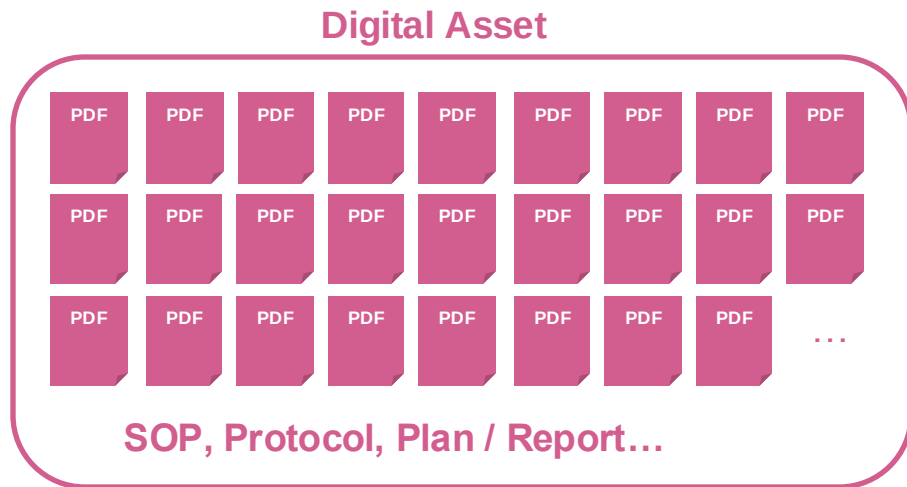
# Check Report

| SDTM                                          |               |                   |           |                 |               | ADAM                                          |            |                   |           |                 |            |
|-----------------------------------------------|---------------|-------------------|-----------|-----------------|---------------|-----------------------------------------------|------------|-------------------|-----------|-----------------|------------|
| SDRG                                          | Completed     | Define.xml        | Completed | P21             | NotChecked    | ADRG                                          | Completed  | Define.xml        | Completed | P21             | NotChecked |
| Study Data Standards and Dictionary Inventory | Completed     | MetaDataVersion   |           |                 |               | Study Data Standards and Dictionary Inventory | Completed  | MetaDataVersion   |           |                 |            |
| Define                                        | 2.0.0         | Define            | 2.0.0     |                 |               | Define                                        | 2.0        | Define            | 2.0.0     |                 |            |
| SDTM                                          | 1.4           | Standard          | SDTM-IG   |                 |               | SDTM                                          | 1.4        | Standard          | ADaM-IG   |                 |            |
| SDTM IG                                       | 3.2           | Standard Ver      | 3.2       |                 |               | SDTM IG                                       | 3.2        | Standard Ver      | 1.1       |                 |            |
| SDTM CT                                       | 2021-03-26    | SDTM CT           |           |                 |               | SDTM CT                                       | 2021-03-26 | SDTM CT           |           |                 |            |
|                                               |               | .ExternalCodeList |           |                 |               | ADaM                                          | 2.1        | .ExternalCodeList |           |                 |            |
|                                               |               |                   |           |                 |               | ADaM IG                                       | 1.1        |                   |           |                 |            |
| MedDRA                                        | 25.1          | MedDRA            | 25.1      | MedDRA          | 25.1          | ADaM CT                                       | 2023-06-30 | ADaM CT           |           | MedDRA          | 25.1       |
| WHODD                                         | September2023 | WHODD             | 2023-09   | WHODD           | September2023 | MedDRA                                        | 25.1       | MedDRA            | 25.1      | WHODD           | 2023-09    |
| Conformance Inputs                            | Completed     |                   |           |                 |               | Conformance Inputs                            | Completed  |                   |           |                 |            |
| P21 Version                                   | 5.4.0         |                   |           | P21 Version     | 5.4.0         | P21 Version                                   | 5.4.0      |                   |           | P21 Version     | 5.4.0      |
| P21 SDTM IG                                   | 3.2           |                   |           | P21 SDTM IG     | 3.2           | P21 FDA Engine for ADaM                       | 2304.3     |                   |           | P21 ADaM IG     | 2304.3     |
| P21 SDTM CT                                   | 2021-03-26    |                   |           | P21 SDTM CT     | 2021-03-26    | P21 FDA Engine for Define                     |            |                   |           | P21 SDTM CT     |            |
|                                               |               |                   |           |                 |               |                                               |            |                   |           | P21 ADaM CT     |            |
| P21 FDA Engine                                |               |                   |           | P21 FDA Engine  |               | P21 PMDA Engine for ADaM                      | 2211.1     |                   |           | P21 FDA Engine  | 2211.1     |
| P21 PMDA Engine                               | 2211.1        |                   |           | P21 PMDA Engine | 2211.1        | P21 PMDA Engine for Define                    |            |                   |           | P21 PMDA Engine |            |
| Issues Summary                                | Completed     |                   |           |                 |               | Issues Summary                                | Completed  |                   |           |                 |            |
| FDA                                           |               |                   |           | FDA             |               | FDA                                           |            |                   |           | FDA             |            |
| Reject                                        | 0             |                   |           | Reject          | 0             | Reject                                        | 0          |                   |           | Reject          | 0          |
| High                                          | 0             |                   |           | High            | 0             | High                                          | 1          |                   |           | High            | 1          |
| Midium                                        | 0             |                   |           | Midium          | 0             | Midium                                        | 0          |                   |           | Midium          | 0          |
| Low                                           | 0             |                   |           | Low             | 0             | Low                                           | 1          |                   |           | Low             | 1          |
| NA                                            | 235           |                   |           |                 |               | NA                                            | 22         |                   |           |                 |            |
|                                               |               |                   |           |                 |               |                                               |            |                   |           |                 |            |
| PMDA                                          |               |                   |           | PMDA            |               | PMDA                                          |            |                   |           | PMDA            |            |
| Reject                                        | 0             |                   |           | Reject          | 0             | Reject                                        | 0          |                   |           | Reject          | 0          |
| Error                                         | 33            |                   |           | Error           | 33            | Error                                         | 7          |                   |           | Error           | 7          |
| Warning                                       | 202           |                   |           | Warning         | 202           | Warning                                       | 16         |                   |           | Warning         | 16         |
| NA                                            | 0             |                   |           |                 |               | NA                                            | 1          |                   |           |                 |            |



# PDF or Word

Word is by far the easiest way to analyze documents, but...



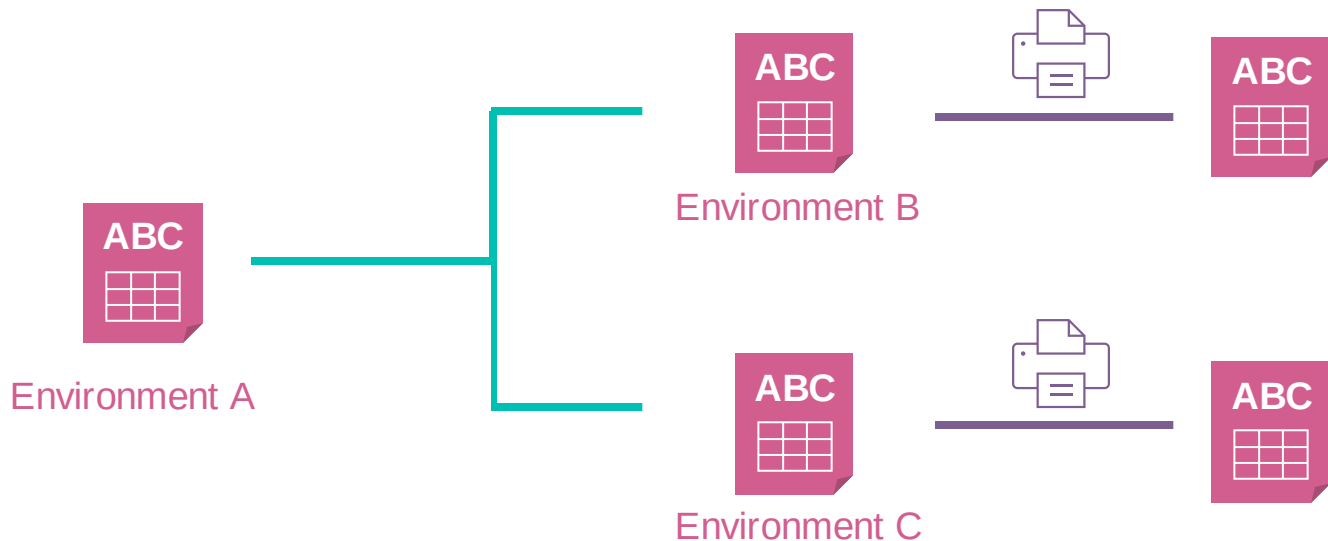
**Process Automation**  
**Big Data Analysis**  
**AI / Machine Learning**

Decided to take on the challenge of reading PDFs, which is a highly difficult task



# What is PDF?

PDF is a data format that is intended to be able to be displayed and printed with the same look and feel in any environment.



It has been standardized by ISO and is widely used due to its excellent visual readability.  
On the other hand, machine reading is not taken into account.



# Opensource Library

## Python 3.12

### PyPDF 4.2.0

For reading text from PDF files

BSD License

<https://pypi.org/project/pypdf/>

### camelot-py 0.11

For extracting tables from PDF files

MIT License

<https://pypi.org/project/camelot-py/>



# Items to Be Checked

| Section                                           | Type  |                                                                                             |
|---------------------------------------------------|-------|---------------------------------------------------------------------------------------------|
| 1.3 Study Data Standards and Dictionary Inventory | Table | SDTM ver, SDTM IG ver, CT ver, Define ver, Medication Dictionary, Medical Events Dictionary |
| 2.3 Trial Design Datasets                         | Text  | Trial Design Datasets                                                                       |
| 3.4 SDTM Subject Domains                          | Table | Dataset, SUPP-, RELREC                                                                      |
| 4.1 Conformance Inputs                            | Text  | P21 ver, Validation rules, Engine                                                           |
| 4.2 Issues Summary                                | Table | Dataset, P21 Rule ID, Message, Severity, Count                                              |

**PDFs do not have section-specific information.  
First, you must identify the section you want to retrieve.**



# Mysterious Whitespace

PDF

## 2.3.1. TA – Trial Arms

The Trial Arms (TA) domain contains the planned arms of the study.



Extracted by PyPDF

Extract

## 2.3.1. TA - Trial Arms

The Tri

al Arms

(TA) domain contains the planned arms of the study.

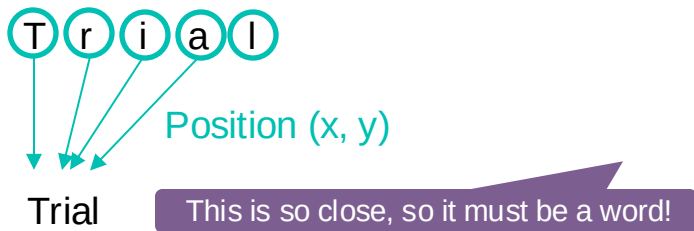
Unexpected line break

Extra blank



# PDF Doesn't Have Any Context

PDF only have information of letters and their position



The information of word, sentence and context are lost when the PDF is created. PyPDF reconstructs the text by inferring from the position of the letters whether they are connected or not.

**Extra whitespaces and line breaks are inevitable in extracting text from PDF.**  
**We need to account for the extra whitespaces and line breaks that seem to occur at random.**



# Find Headers in Sea of Text - 1

2.3 Trial Design Datasets

n.n XXXXXXXXXXXXX

Regular expression

`([1-9]|[1-2][0-9])\.([1-9]|[1-2][0-9])\s.+\\n`

## Issue #1

Text extracted from PDF contains extra whitespace and line breaks.

Then just take out the whitespace and line breaks!

e.g.) 2.3TrialDesignDatasets

## Issue #2

Excluding line breaks, it is indistinguishable from a reference.

e.g.) see [2.3TrialDesignDatasets](#)

## Issue #3

Mistaking a mere number with a decimal point for a section

e.g.) 2.5 mg of the investigational drug was administered





# Find Headers in Sea of Text - 2

- 1. *Introduction*.....
  - 1.1 Purpose .....
  - 1.2 Acronyms .....
  - 1.3 Study Data Standards and Dictionary Inventory .....
- 2. *Protocol Description* .....
- 2.1 Protocol Number and Title .....
- 2.2 Protocol Design.....
- 2.3 Trial Design Datasets .....

1. Step.1 Get headers from contents list
2. Step.2 Remove whitespace and line break
3. Step.3 Seek the header from text which is removed whitespace and line break
4. Step.4 Check the order of the found header



# The Table Is Not a Table

In PDF, a table is not a table, but only a combination of lines and text.  
There is no information of how many rows and columns exist in the table.

Camelot-py detects tables and get values as a pandas (DataFrame) object.

| Dataset | P21 Rule ID | Diagnostic Message                                                                   | Severity | Count | Explanation                                                                                     |
|---------|-------------|--------------------------------------------------------------------------------------|----------|-------|-------------------------------------------------------------------------------------------------|
| AE      | CT2001      | AEACN value not found in 'Action Taken with Study Treatment' non-extensible codelist | Error    | 6     | MULTIPLE has been added as an extended value into CODELIST 'Action Taken with Study Treatment'. |
| AE      | TS0012      | Analysis Required variable AESEV not found                                           | Error    | 1     | AESEV is not collected for this study, AETOXGR is collected instead of AESEV per CTCAE 5.0.     |
| AE      | SD0021      | Missing End Time-Point value                                                         | Warning  | 570   | Not an issue, the outcome of the AE records are 'Not recovered/Not resolved'.                   |



Just load the PDF and it will automatically retrieve the table. It is very easy.  
However, this is where things got tough.



# Points in Need of Care

| Dataset(s) | P21 Rule ID | Diagnostic Message    | Severity (FDA) | Severity (PMDA) | Count/ Issue Rate | Explanation                                                                                                                                   |
|------------|-------------|-----------------------|----------------|-----------------|-------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| DEFINE     | DD0068      | Invalid use of Length | Error          | Error           | 19                | According to our interpretation of the Case Report Tabulation Data Definition Specifications(define.xml), the length attribute is allowed for |

Need to consider extra whitespace and line break when you check the table.

Page 19 of 28

Analysis Data Reviewer's Guide

| Dataset(s) | P21 Rule ID | Diagnostic Message | Severity (FDA) | Severity (PMDA) | Count/ Issue Rate | Explanation                                                                               |
|------------|-------------|--------------------|----------------|-----------------|-------------------|-------------------------------------------------------------------------------------------|
|            |             |                    |                |                 |                   | datetime values. The length represents the maximum length of ISO8601 formatted dates(19). |

Camelot-py treats continuing rows as separate rows



# Not Detected as It Appears

| Dataset | P21 Rule ID | Diagnostic Message                                                                   | Severity (FDA) | Severity (PMDA) | Count | Explanation                                                                                                                                                                                                                                                     |
|---------|-------------|--------------------------------------------------------------------------------------|----------------|-----------------|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AE      | CT2001      | AEACN value not found in 'Action Taken with Study Treatment' non-extensible codelist | NA             | Error           | 8047  | According to the CDISC guidance, 'MULTIPLE' is used in case multiple responses are available. This should not be considered as an error. As there are multiple study treatments, AEACN = MULTIPLE and the action taken per study treatment is mapped to SUPPAE. |

| Column             | Value                                                                                       |
|--------------------|---------------------------------------------------------------------------------------------|
| Dataset            | AE                                                                                          |
| P21 Rule ID        |                                                                                             |
| Diagnostic Message | CT2001 AEACN value not found in 'Action Taken with Study Treatment' non-extensible codelist |

▼ Sometimes values are detected as those of the next cell

This problem cannot be predicted where it will occur, therefore all columns need to be considered



# 0 Inquiries from Authority



Record of 0 consecutive inquiries from PMDA due to incorrect descriptions is being renewed!!



# Conclusion

**Although PDF is a structured format, it is equivalent to dealing with unstructured data in the secondary use of text.**

1. PDFs don't always keep context for secondary use as long as they look the same.
2. There are excellent open-source libraries that extract data from PDFs, but they have limitations.
3. It is best to give up on extracting complete data and use it to focus on key points.

**The more accuracy of the extracted text is required,  
the much greater the volume and difficulty of the code.**

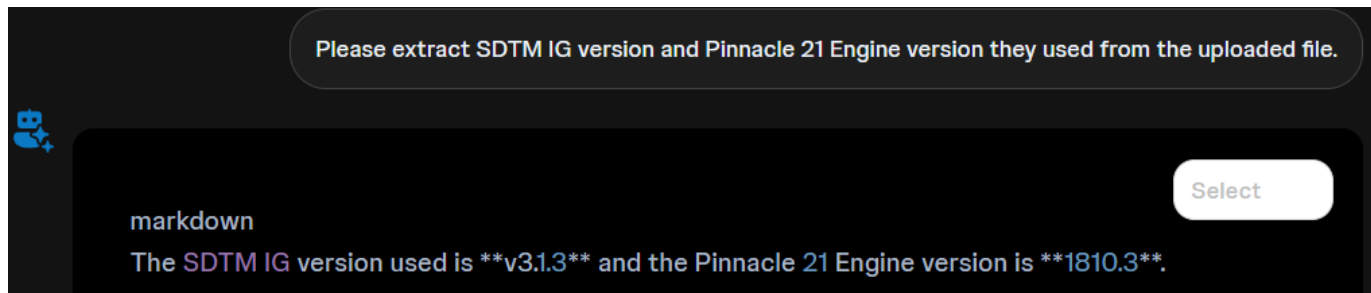


# Best Practice...?



Extract text from word files using python-docx

Let AI extract text

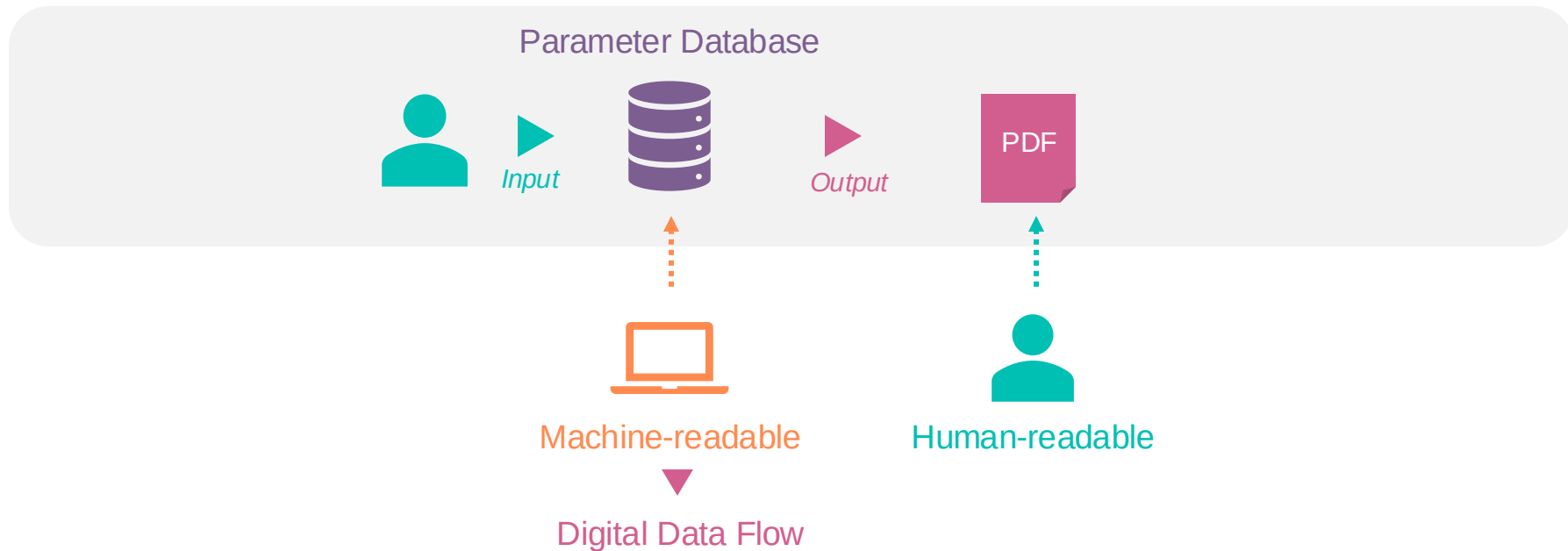


Chat GPT extracted version information! But the engine for FDA that was written with it was ignored.



# To Digital Data Flow

Documentation is moving into an era of both human readability and machine readability







## The Global Healthcare Data Science Community

### Contact Channels

📞 UK +44 1843 609600  
✉ [office@phuse.global](mailto:office@phuse.global)  
🌐 [phuse.global](http://phuse.global)

### Social Media

🐦 [@phusetwitta](https://twitter.com/phusetwitta)  
📘 [/phusebook](https://facebook.com/phusebook)  
▶ [/phusetube](https://youtube.com/phusetube)  
🌐 [/company/phuse](https://company.phuse.com)