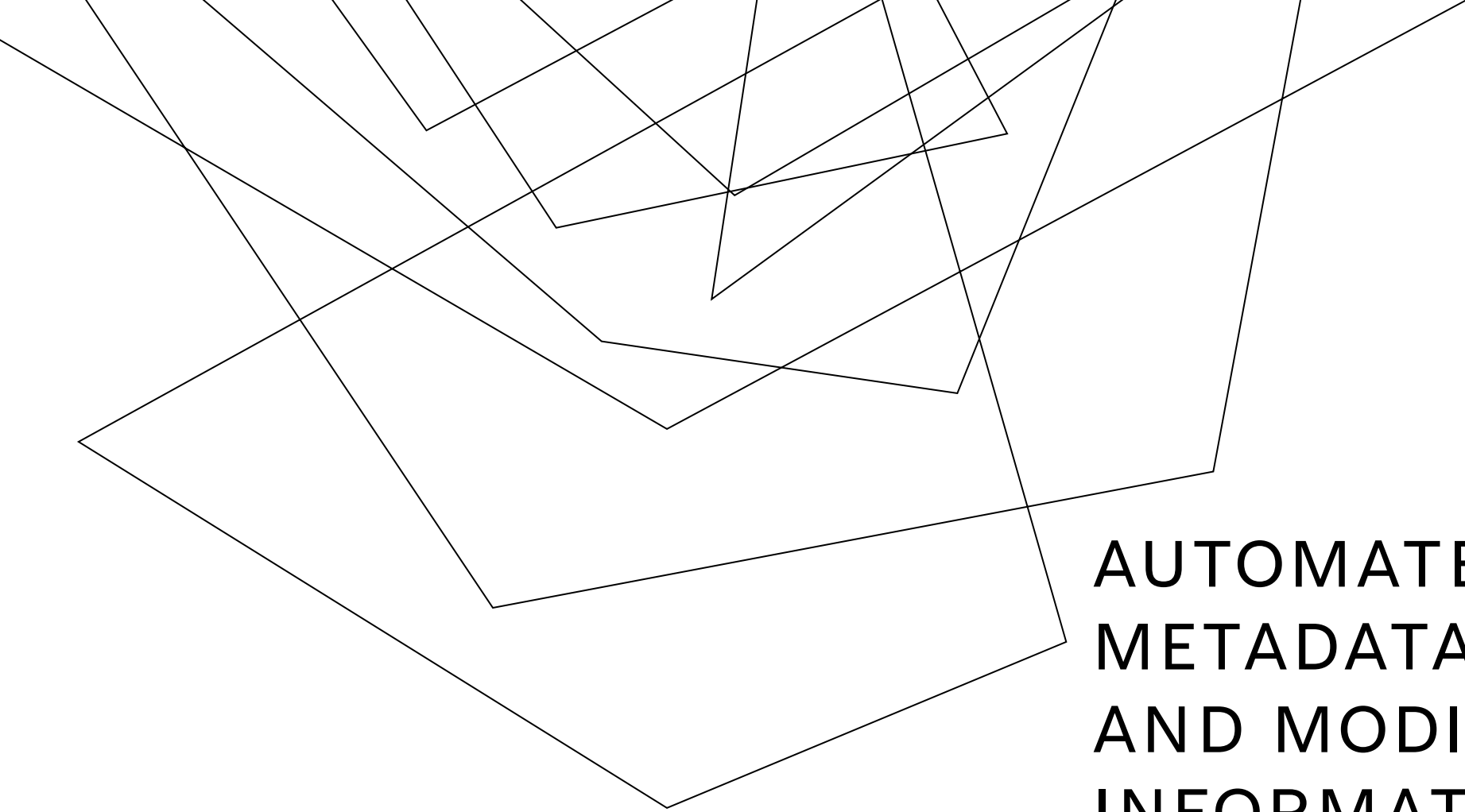




AUTOMATE YOUR METADATA: READING AND MODIFYING INFORMATION FROM PDF FILES USING PYTHON

Author: Salomo Sparna
Co-Author: Renate Scheiner-Sparna
Paper nr.: SM15



AGENDA

Introduction

Automated metadata
processing in clinical studies

Inputs

Annotation Exporter

How to proceed with the
metadata

Outlook



IMPORTANCE OF AUTOMATED METADATA PROCESSING IN CLINICAL STUDIES

IMPORTANCE OF AUTOMATED METADATA PROCESSING IN CLINICAL STUDIES

We have to manage the transition of data between different levels.



Raw data

-->



SDTM

-->



ADaM

Automated processes help in many ways.

IMPORTANCE OF AUTOMATED METADATA PROCESSING IN CLINICAL STUDIES

Reasons for automated processes:

- Less error prone – reduced manual work
- Transparent – flow can be followed
- Maintainable – structured, documented process
- Reduce amount of work – lots of hours can be saved
- Reliable metadata – you get what you had in the level before

Overall: Improved quality, less time needed

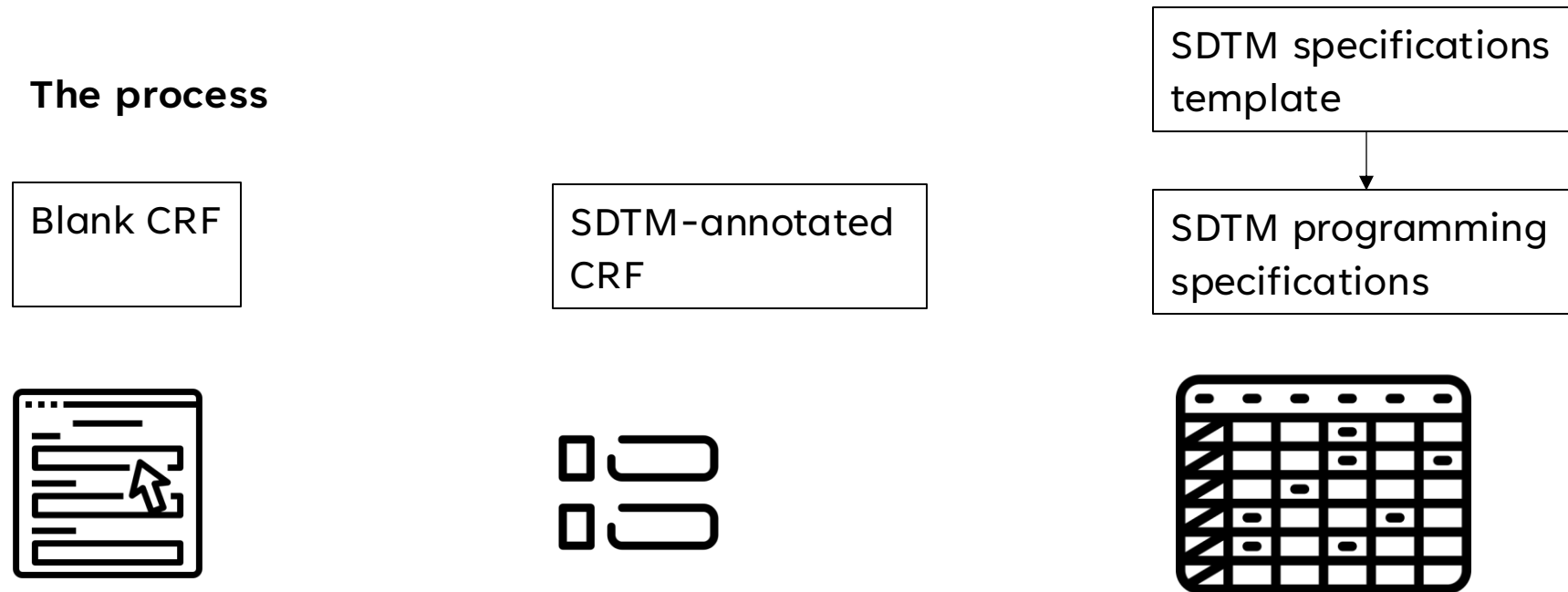
IMPORTANCE OF AUTOMATED METADATA PROCESSING IN CLINICAL STUDIES

Here: Proof on concept on SDTM level

- We come from the SDTM-annotated eCRF and want to generate the SDTM specification metadata.
- How can we use the annotations?
- How can this be integrated in the process flow?
- Which tools do we need?

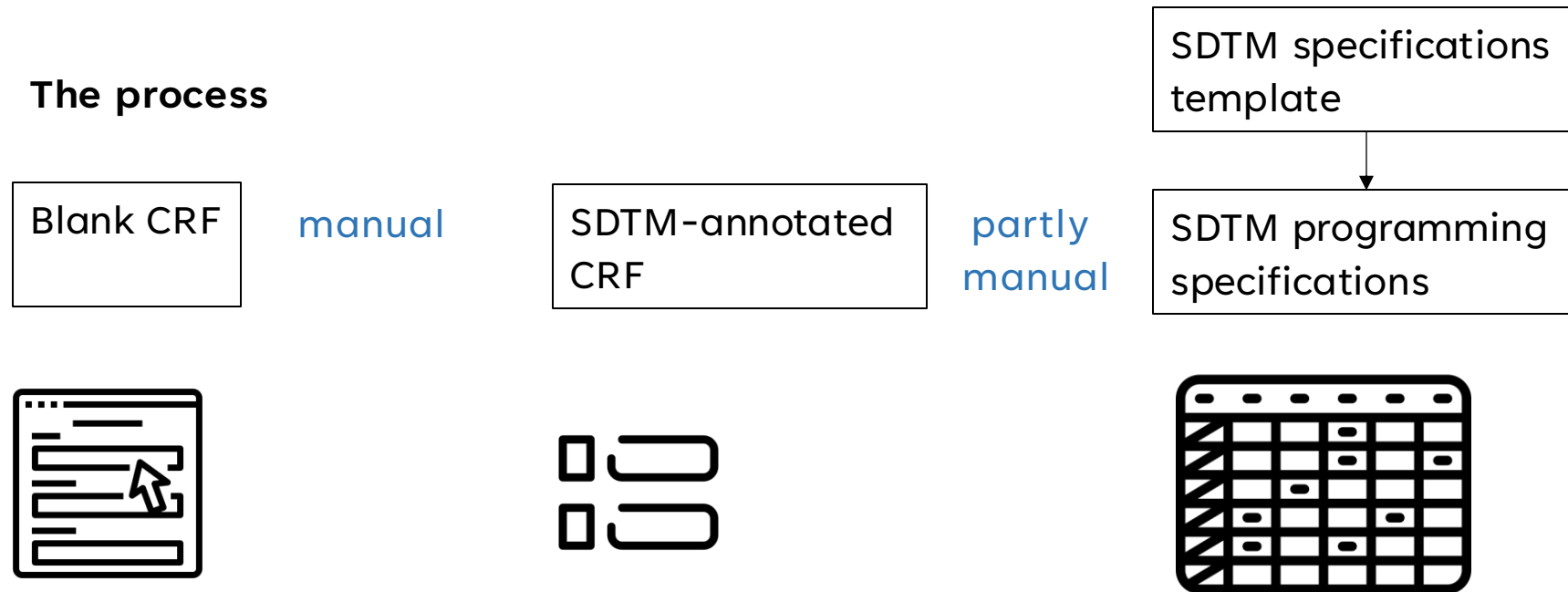
TRANSFERRING METADATA FROM SDTM ANNOTATION TO SDTM SPECIFICATION

The process



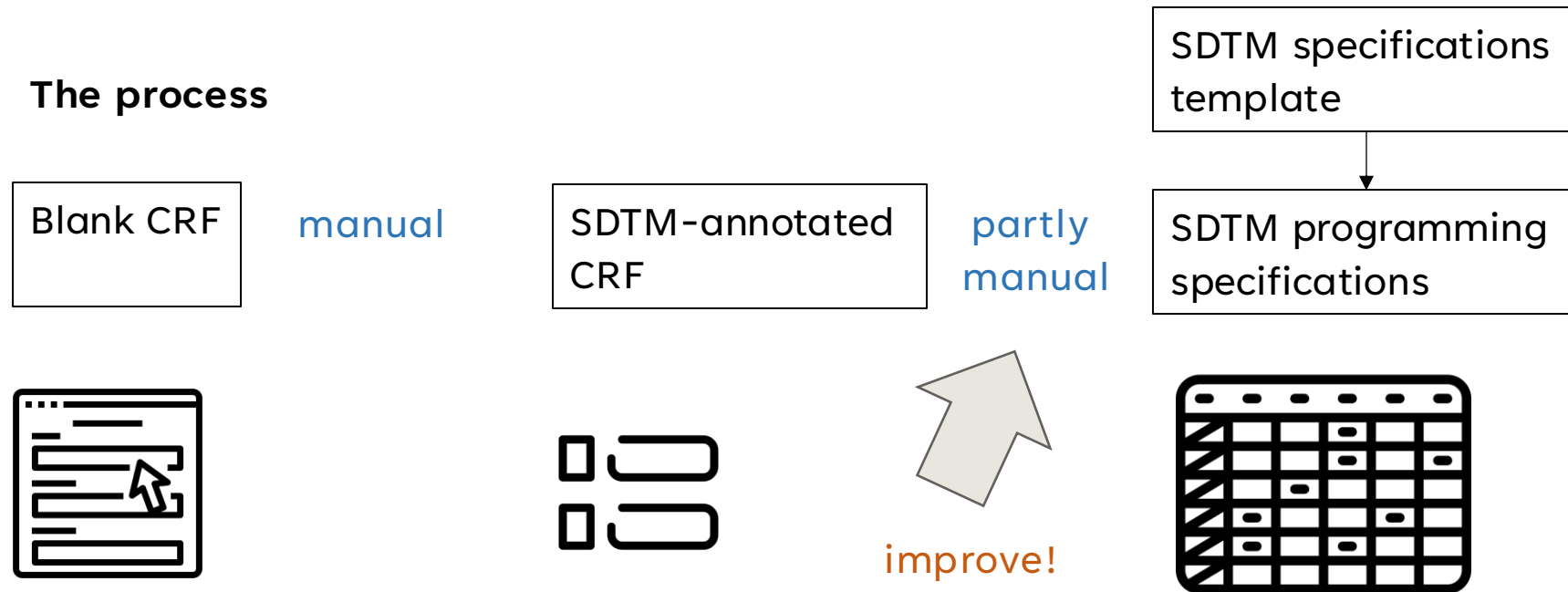
TRANSFERRING METADATA FROM SDTM ANNOTATION TO SDTM SPECIFICATION

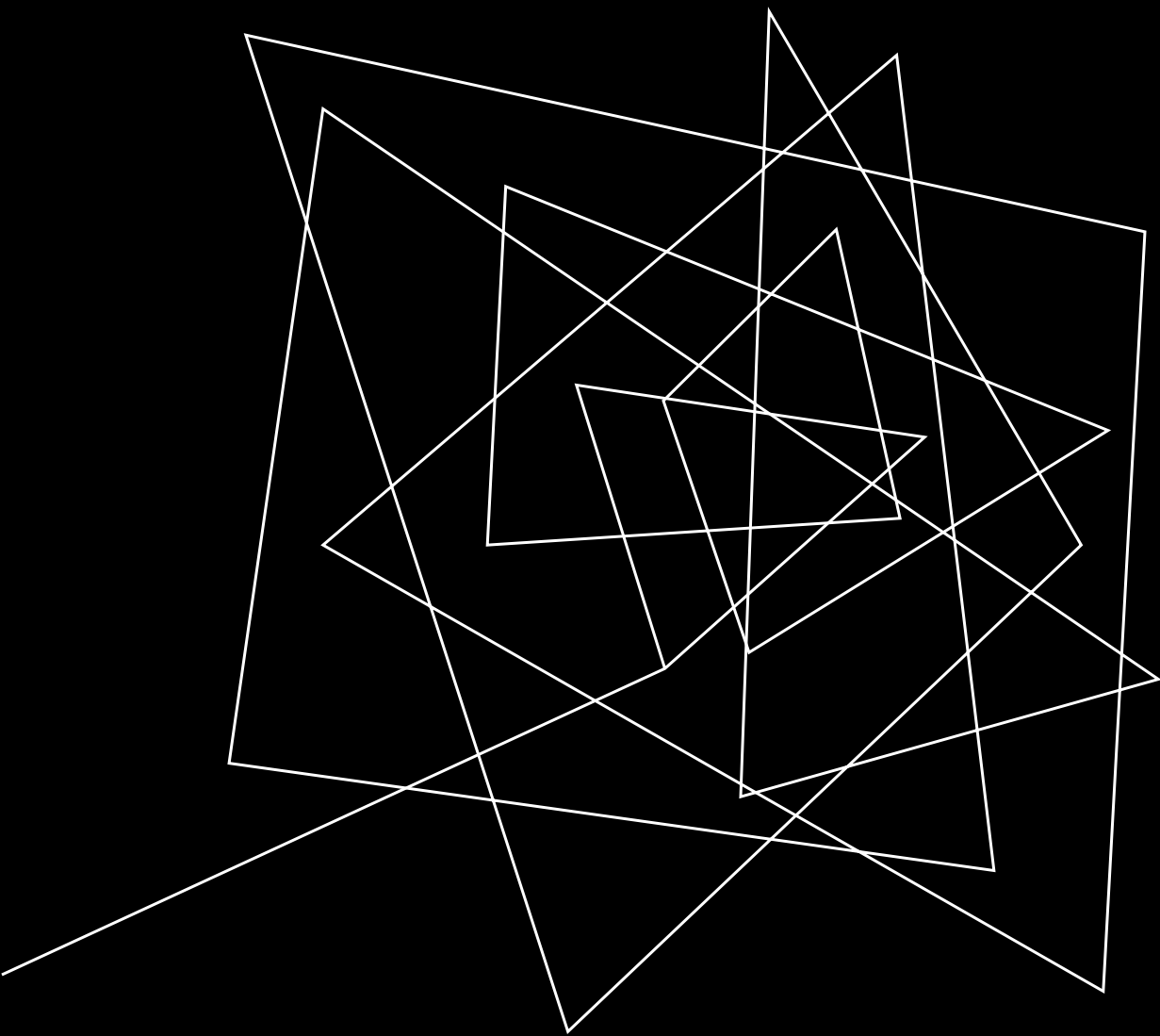
The process



TRANSFERRING METADATA FROM SDTM ANNOTATION TO SDTM SPECIFICATION

The process





INPUTS

DM (Demographics)

RP (Reproductive System Findings)

Clinical Study No.:
ABCDEF

SV (Subject Visits)

VS (Vital Signs)

Patient Identification Number

Screening day -1 or 1

VISIT = Screening

SVSTDTC

SVENDTC

Visit Date/Time (dd-MMM-yyyy HH:mm)

____-____-____

____:____

DEMOGRAPHY

Date of Birth (dd-MMM-
YYYY)

____-____-____

BRTHDTC

Please record at least the year in case the law in your country does not allow to insert the complete date of birth (i.e., UN-UNK-YYYY)

AGEU=YEARS

Age (years)

AGE

Age (years) (derived)

Sex

☐ Male

☐ Female

SEX

RPTESTCD = CHILDPOT

If female, childbearing
potential status:

☐ Childbearing potential

RPORRES=Y

☐ Postmenopausal with no menstrual bleeding for at least one year prior to study start

RPORRES=N

☐ Surgically sterilized

RPORRES=N

SUPPRP.QVAL when
SUPPRP.QNAM=NCPREAS,
RPORRES=N

Contraception method
(check all that apply)

RPTESTCD=CONMET

RPTEST=Contraception
Method

☐ None

☐ Hormonal contraception, systemic, implantable, transdermal, or injectable contraceptives from at least 2 months before the screening visit until 30 days after the last IMP dose

☐ A sterile sexual partner

☐ Abstinence

☐ Other, specify:

RPORRES
(one row for each option checked)

SUPPRP.QVAL when QNAM=OTHCM
QLABEL=Other contraception method

Self-reported race/ethnicity
(check all that apply)

☐ White

☐ Black or African American

☐ Asian

☐ Hispanic / Latino

☐ Other, specify:

(if more than 1 race selected, RACE=MULTIPLE
and add one SUPPDM.QNAM=RACEx for each
race selected)

SUPPDM.QVAL when QNAM=RACEOTH

VSTEST = Height

Height (cm)

VSORRES

VSORRESU=cm

VSORRESU=cm

VSTEST = Weight

Body weight (kg)

VSORRES

VSORRESU=kg

VSSTRESU=kg

Example CRF

- Typical CRF
- Short example
- Input for Python program

Excel Template

- Preloaded with dataset names and standard variables
- Input for Python program
- Base for STDM programming specification

	A	B	C	D	E	F	G	H	I	J
1	Order	Dataset	Variable	Label	DataType	Length	Significant	Format	Mandatory	Core
1754	1	DM	STUDYID	Study Identifier	text	200			Yes	Req
1755	2	DM	DOMAIN	Domain Abbreviation	text	200			Yes	Req
1756	3	DM	USUBJID	Unique Subject Identifier	text	200			Yes	Req
1757	4	DM	SUBJID	Subject Identifier for the Study	text	200			Yes	Req
1758	5	DM	RFSTDTC	Subject Reference Start Date/Time	datetime	19		ISO 8601	No	Exp
1759	6	DM	RFENDTC	Subject Reference End Date/Time	datetime	19		ISO 8601	No	Exp
1760	7	DM	RFXSTDTC	Date/Time of First Study Treatment	text	200			No	Exp
1761	8	DM	RFXENDTC	Date/Time of Last Study Treatment	text	200			No	Exp
1762	9	DM	RFCSTDTC	Date/Time of First Challenge Agent Admin	text	200			No	Perm
1763	10	DM	RFCENDTC	Date/Time of Last Challenge Agent Admin	text	200			No	Perm

A	B	C	F	G
Dataset	Description	Class	Purpose	KeyVariables
AE	Adverse Events	EVENTS	Tabulation	STUDYID, USUBJID, AEDECOD, AESTDTC
CM	Concomitant/Prior Medications	INTERVENTIONS	Tabulation	STUDYID, USUBJID, CMTRT, CMSTDTC
DM	Demographics	SPECIAL PURPOSE	Tabulation	STUDYID, USUBJID
DS	Disposition	EVENTS	Tabulation	STUDYID, USUBJID, DSDECOD, DSSTDTC
DV	Protocol Deviations	EVENTS	Tabulation	STUDYID, USUBJID, DVTERM, DVSTDTC
EX	Exposure	INTERVENTIONS	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC
MH	Medical History	EVENTS	Tabulation	STUDYID, USUBJID, MHDECOD
SV	Subject Visits	SPECIAL PURPOSE	Tabulation	STUDYID, USUBJID, SVTERM
VS	Vital Signs	FINDINGS	Tabulation	STUDYID, USUBJID, VSTESTCD, VISITNUM, VSTPTREF, VSTPTNUM

An abstract graphic consisting of two thin, dark gray lines that intersect diagonally on the left side of the image. One line slopes downwards from left to right, while the other slopes upwards from left to right.

ANNOTATION EXPORTER

USING THE EXPORTER

1. Run main.py
2. Enter the paths to the respective files
3. Click the “Export Annotations” Button



The screenshot shows a dark-themed application window titled "Annotation Exporter V 0.2". It contains three input fields for file paths, each with a "Browse" button to its right. The "PDF Path" field contains "PDF/example_compressed", the "Spreadsheet Path" field contains "Templates/temp", and the "Output Folder" field contains "outputs". Below these fields are two checkboxes: "Convert Old Standard" and "Create SQLite database", both of which are currently unchecked. To the right of these checkboxes is a large "Export Annotations" button.

Annotation Exporter V 0.2

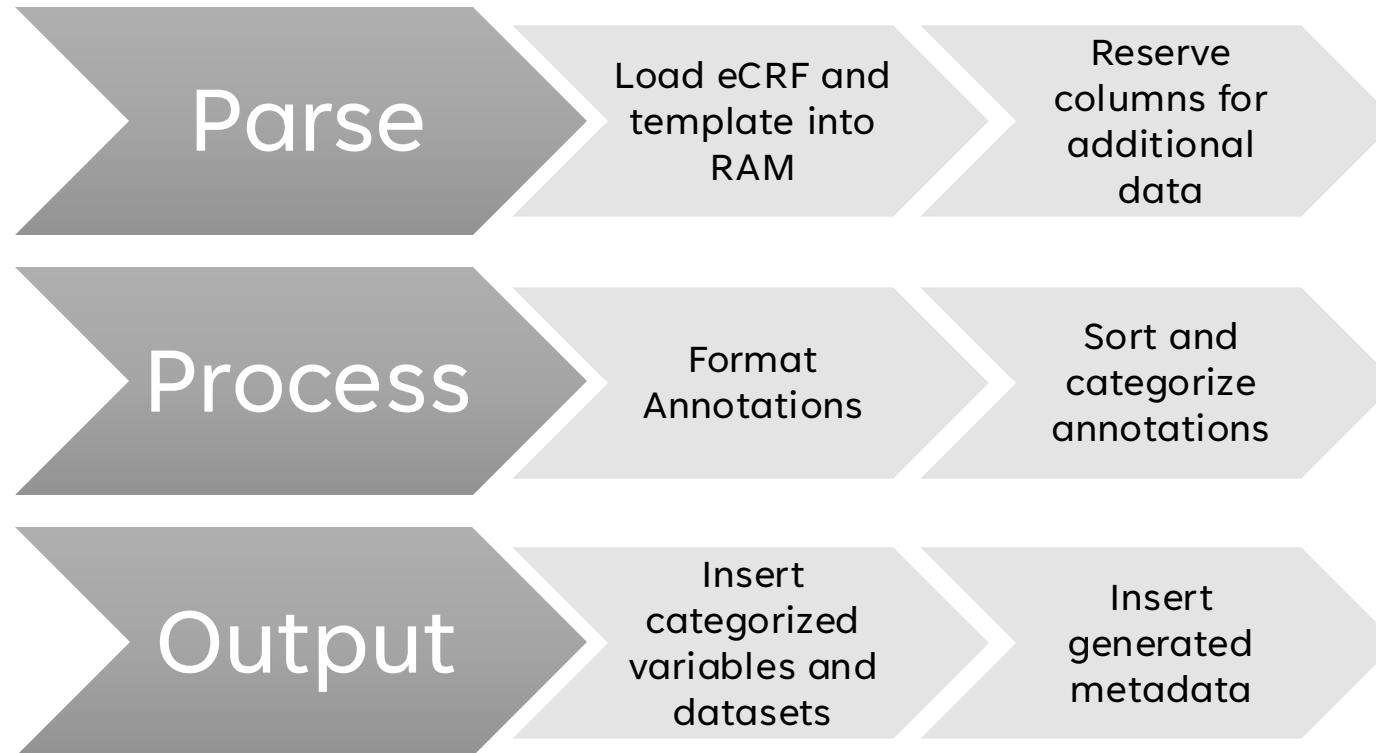
PDF Path PDF/example_compressed Browse

Spreadsheet Path Templates/temp Browse

Output Folder outputs Browse

☐ Convert Old Standard ☐ Create SQLite database Export Annotations

PROCESS STEPS



THE OUTPUT

- Modified Template
- Presence marked
- Page numbers

1	Description	Class	Structure	KeyVariables	Referer	Present in aCRF
2	Adverse Events	EVENTS	One record per adverse event per subject	STUDYID, USUBJID, AEDECOD, AESTDTC	ae.xpt	
3	Procedure Agents	INTERVENTIONS	One record per recorded intervention occurrence per subject	STUDYID, USUBJID, AGTRT, AGSTDTC	ag.xpt	
4	Biospecimen Events	EVENTS	One record per instance per biospecimen event per biospecimen identifier per subject	STUDYID, USUBJID, BREFID, BETERM, BESDTC	be.xpt	
5	Biospecimen Findings	FINDINGS	One record per measurement per biospecimen identifier per subject	STUDYID, USUBJID, BSREFID, BSTESTCD	bs.xpt	
6	Clinical Events	EVENTS	One record per event per subject	STUDYID, USUBJID, CETERM, CESTDTC	ce.xpt	
7	Concomitant/Prior Medications	INTERVENTIONS	One record per recorded intervention occurrence or constant-dosing interval per subject	STUDYID, USUBJID, CMTRT, CMSTDTC	cm.xpt	
8	Comments	SPECIAL PURPOSE	One record per comment per subject	STUDYID, USUBJID, IDVAR, COREF, CODTC	co.xpt	
9	Cell Phenotype Findings	FINDINGS	One record per test per specimen per timepoint per visit per subject	STUDYID, USUBJID, CPTSTCD, CPSPEC, VISITNUM, CPTPTREF, CPTPTNUM	cp.xpt	
10	Cardiovascular System Findings	FINDINGS	One record per finding or result per time point per visit per subject	STUDYID, USUBJID, VISITNUM, CVTESTCD, CVTPTREF, CVTPTNUM	cv.xpt	
11	Product Accountability	FINDINGS	One record per product accountability finding per subject	STUDYID, USUBJID, DATESTCD, DADTC	da.xpt	
12	Death Details	FINDINGS	One record per finding per subject	STUDYID, USUBJID, DDTESTCD, DDDTC	dd.xpt	
13	Demographics	SPECIAL PURPOSE	One record per subject	STUDYID, USUBJID	dm.xpt	Present
14	Disposition	EVENTS	One record per disposition status or protocol milestone per subject	STUDYID, USUBJID, DSDECOD, DSSTDTC	ds.xpt	Present

1	Orig	Data	Variable	Label	DataTyp	Len	Format	Man	Core	Origin	Pages	Present in aCRF
1754	1	DM	STUDYID	Study Identifier	text	200		Yes	Req			
1755	2	DM	DOMAIN	Domain Abbreviation	text	200		Yes	Req			
1756	3	DM	USUBJID	Unique Subject Identifier	text	200		Yes	Req			
1757	4	DM	SUBJID	Subject Identifier for the Study	text	200		Yes	Req	CRF	2	Present
1758	5	DM	RFSTDTC	Subject Reference Start Date/Time	datetime	19	ISO 8601	No	Exp			
1759	6	DM	RFENDTC	Subject Reference End Date/Time	datetime	19	ISO 8601	No	Exp			
1760	7	DM	RFXSTDTC	Date/Time of First Study Treatment	text	200		No	Exp			
1761	8	DM	RFXENDTC	Date/Time of Last Study Treatment	text	200		No	Exp			
1762	9	DM	RFCSTDTC	Date/Time of First Challenge Agent Admin	text	200		No	Perm			
1763	10	DM	RFCENDTC	Date/Time of Last Challenge Agent Admin	text	200		No	Perm			
1764	11	DM	RFICDTC	Date/Time of Informed Consent	text	200		No	Exp	CRF	2	Present

THE OUTPUT

- SUPP datasets
- Other „unknown“ variables (Race unknown due to wrong color in annotation).

	B	C	D	E	F	L	M	O
1	Dataset	Variable	Label	DataType	Length	Origin	Pages	Present in aCRF
1909	DS	DSSDTC			200	CRF	2	Present
1910	SUPPDS	QVAL			200	CRF	2	Present
1911	SUPPDS	QNAM			200	CRF	2	Present
1912	SUPPDS	QLABEL			200	CRF	2	Present
1913	SUPPDM	QVAL			200	CRF	3	Present
1914	SUPPDM	QNAM			200	CRF	3	Present
1915	SUPPDM	QLABEL			200	CRF	3	Present
1916	SUPPRP	QVAL			200	CRF	3	Present
1917	SUPPRP	QNAM			200	CRF	3	Present
1918	SUPPRP	QLABEL			200	CRF	3	Present
1919		RACE			200	CRF	3	Present

STANDARD MIGRATION

DM = Demographics	RP = Reproductive System Findings
Clinical Study No.: ABCDEFGH	SV = Subject Visits VS = Vital Signs

The old PDF:

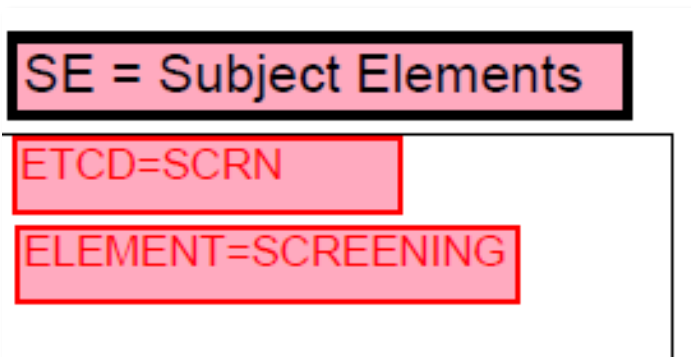
- Annotated to an older standard

DM (Demographics)	RP (Reproductive System Findings)
Clinical Study No.: ABCDEFGH	SV (Subject Visits) VS (Vital Signs)

The output PDF:

- Annotated to the newer standard
- Technical limitations

CLEAN-UP OF ANNOTATION GENERATIONS



- Unable to read some annotations
- Problem: embedded annotations!
- Solution:
 - Go through PDF
 - Find different colored text
 - Convert to FreeText annotations
- Individual solution that can be adapted for other conditions

Two thin, dark gray lines intersect diagonally on a light gray background. One line runs from the top-left towards the bottom-right, and the other runs from the top-right towards the bottom-left. They cross each other in the upper-left quadrant of the image.

WHAT DO WE
GAIN?

WHAT WE GAIN (IMMEDIATELY)

- Required variables can be manually checked.
- Non-standard variables can be checked, if permissible CDISC variables could be used. Better usage of the standard.
- Eliminate human error. Transparency.
- Time is saved, later delays are prevented.

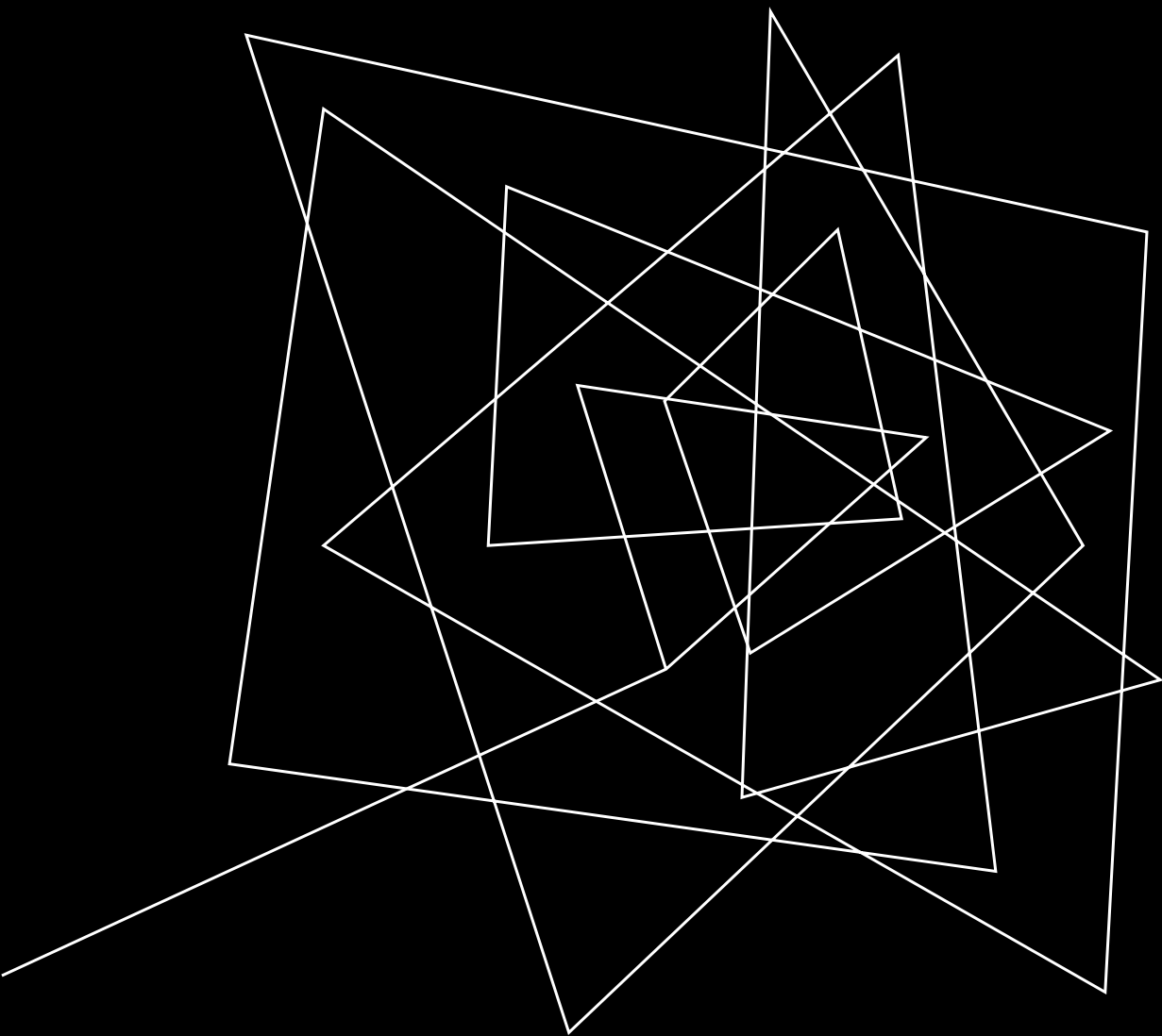
IDEAS FOR IMPROVEMENT

The program can be easily modified to:

- Verify the presence of requested variables
- Undo certain mistakes (e.g. embedded annotations, etc.)

And modified with some effort to:

- Automatically create annotations
- Verify adherence to the annotation standard (e.g. checks for font size, brackets or equal sign in dataset text boxes on top of the page)



GITHUB

<https://github.com/ssparna/Annotation-Exporter>



DOCUMENTATION

`export_annotations(template_path: str, pdf_path: str, output_folder: str) → None` [\[source\]](#)

Exports annots, this is the main function that should be called. Expects the paths to have the correct endings (.pdf, .xlsx).

Parameters:

- **template_path** (*str*) – path to the template file
- **pdf_path** (*str*) – path to the pdf file
- **output_folder** (*str*) – path to the output folder

`class generic.Annotation(annot_obj: DictionaryObject, page: Page)` [\[source\]](#)
This is the class that stores a single annotation with convenient access to frequently used properties and some useful methods

`static get_multiple_variables(annot_obj: DictionaryObject) → list[dict]` [\[source\]](#)

returns all variables from an annotation. this is used to pick up on multiple variables being in the same annotation. This has the side effect of converting the DictionaryObjects to dictionaries which loses some (hopefully irrelevant) data.

Parameters: **annot_obj** (*DictionaryObject*) – the annotation object

Returns: list of annotations as a list of dictionaries

Return type: list[dict]

 ssparna attempt to satisfy sphinx	2c8c3b9 · 17 hours ago	 64 Commits
 PDF	added support for the newest standard and tweaked readm...	last month
 Templates	Delete Templates/~\$SDTM_Specification_Template.xlsx	2 weeks ago
 annotation_exporter	attempt to satisfy sphinx	17 hours ago
 docs/source	attempt to satisfy sphinx	17 hours ago
 outputs	updated pypi metadata	last week
 .gitattributes	Initial commit	2 months ago
 .gitignore	Initial commit	2 months ago
 .readthedocs.yaml	attempt to satisfy sphinx	18 hours ago
 LICENSE	updated license	last month
 README.md	updated readme and sqlite	last week
 requirements.txt	updated readme and switched to FreeSimpleGUI	last month
 setup.py	fixed imports not working correctly	3 days ago

 **README**
 MIT license
 


Annotation-Exporter

This is an example project for the accompanying presentation and paper at PHUSE EU Connect 2024 which can be found [Here](#).

The purpose of this repo is to provide access to the code showcased in the previously mentioned presentation and paper. Feel free to use this as a baseline or inspiration to improve your workflow when dealing with the CDISC standard.



THANK YOU

Salomo Sparna

+49 1575 2998087

SalomoSparna@gmail.com