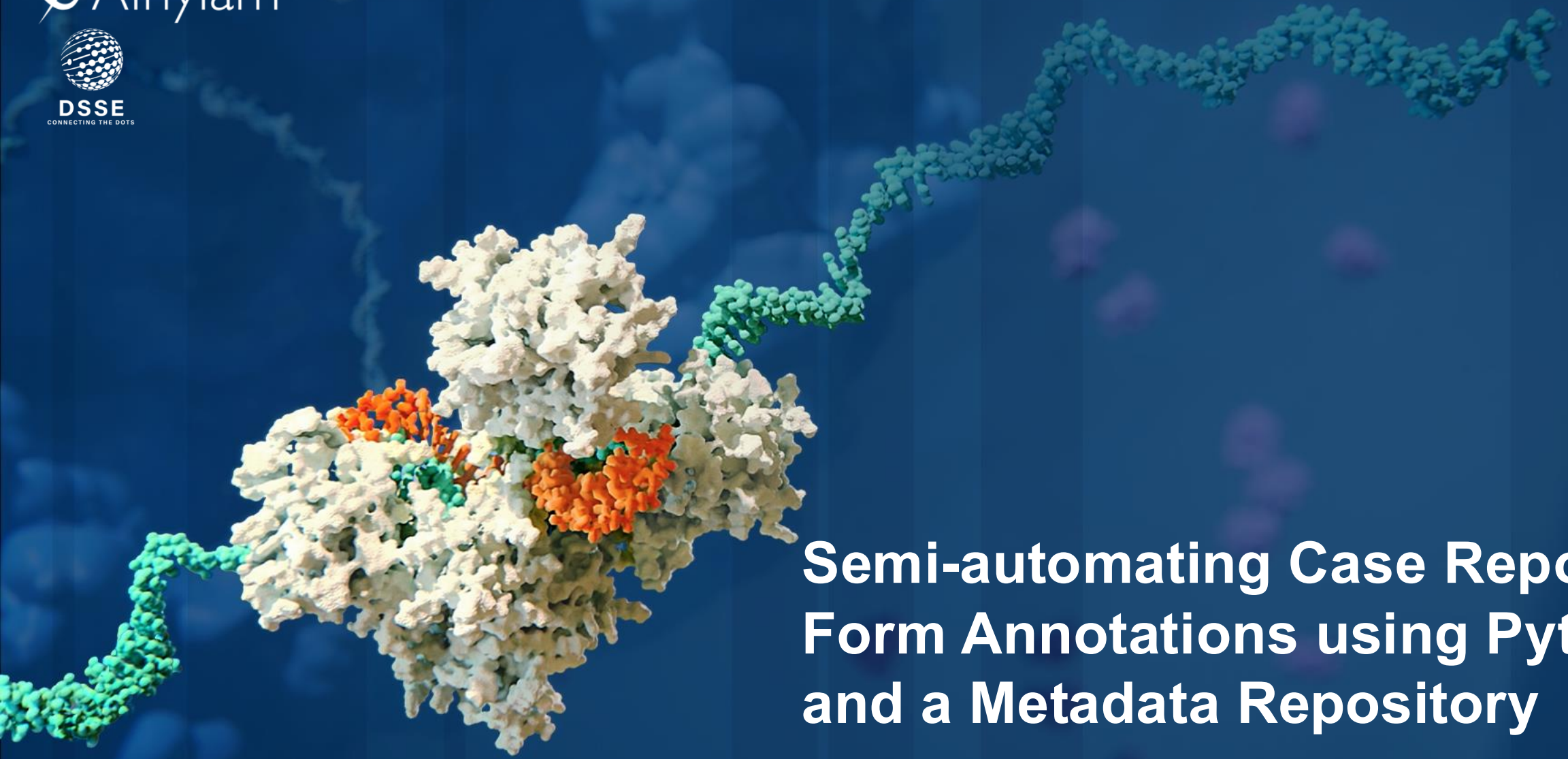




Semi-automating Case Report Form Annotations using Python and a Metadata Repository

Alvin Mathew and Ella Chee

PHUSE DH12

Agenda

Problem
CRF/aCRF
Annotations/SDTM
Current process and issues

Approach
Formatting per guidelines
Assigning standard annotations
Place annotations automatically

Solution
Extraction/Mapping
Annotation
Future updates

|| Problem

Background: CRF / aCRF

Case report forms and the annotation process

- Case Report Forms contain all pages found in EDC
- Multiple versions: blank, by visit, with EDC variable annotations
- SDTM development requires annotating fields to SDTM variables
 - Map and annotate where each page goes, any special rules, and even what is not submitted to SDTM

Year of Birth (yyyy)		
Age (years) at time of consent		Fixed Unit: years
Sex	Male	☐
	Female	☐

Annotations / SDTM

Several considerations when annotating

- Must follow Metadata Submission Guidelines v2.0, including
 - Colors
 - Font
 - Borders
 - Placement

BLUE	DM (Demographics)	191, 255, 255
YELLOW	DS (Disposition)	255, 255, 150
GREEN	SC (Subject Characteristics)	150, 255, 150
ORANGE	VS (Vital Signs)	255, 190, 155

QSSTRESC / QSSTRESN

- 7 Strongly agree
- 6 Agree
- 5 Slightly agree
- 4 Neither agree nor disagree
- 3 Slightly disagree
- 2 Disagree
- 1 Strongly disagree

QSORRES

DS (Disposition) DSCAT = PROTOCOL MILESTONE
DSTERM / DSDECOD = INFORMED CONSENT OBTAINED

INFORMED CONSENT

Informed Consent Date

DM (Demographics) DSSTDTC
RFICDTC

DEMOGRAPHICS

Birth Year BIRTHDTC

Age AGE AGEU years

Sex ☐ Female ☐ Male SEX

Race (Check all that apply)

☐ White

☐ Black or African American

☐ Asian

☐ Native Hawaiian or Other Pacific Islander

☐ American Indian or Alaskan Native

RACE

Ethnic

☐ Hispanic or Latino

☐ Not Hispanic or Latino

ETHNIC

When multiple values are selected then RACE = MULTIPLE and individual responses are RACE1, RACE2, RACE3, etc. in SUPPDM



DSSE
CONNECTING THE DOTS



Problems with Current Manual Process

Add box → add annotation → add correct color and font → adjust box border → repeat x1000

Time-consuming

- Manually slog through hundreds of pages of PDFs
- Shaping boxes to fit the text and fonts is finicky
- Apply MSG rules

Difficult to adhere to standards

- Need to reference internal and external standards
- Might be deviating from other studies
- MSG colors need to be updated if a domain is inserted

Limited updatability & collaboration

- Updating a similar annotation across multiple pages requires scrolling (or attempting to find) across all the pages
- Usually one person at a time



DSSE
CONNECTING THE DOTS



Potentially Skip MSG at Beginning

Downside: Difficult to find time to update with time crunch at start and end of study

- Tight study start-up with multiple competing responsibilities
- Teams might opt to manually write each annotation on the page and avoid MSG so they can get started with programming faster
- However:
 - Still need to adjust boxes for text and add colors
 - Complex forms may be difficult to read
 - Will eventually need to update prior to any submissions

participant a prior screen failure?

Yes ☐

No ☐

SUPPDM.QVAL when QNAM = RESCREEN

please provide the original
ant number (xxxx-xxxx)

SUPPDM.QVAL when QNAM = PSUBJIDn

ics

IRTH

BRTHDTC

s) at time of consent

Fixed Unit: years

AGE

AGEU

SEX

Male ☐

Female ☐

participant of Child Bearing
al?

RPTESTCD = CHILDPOT

RPORRES

Yes ☐

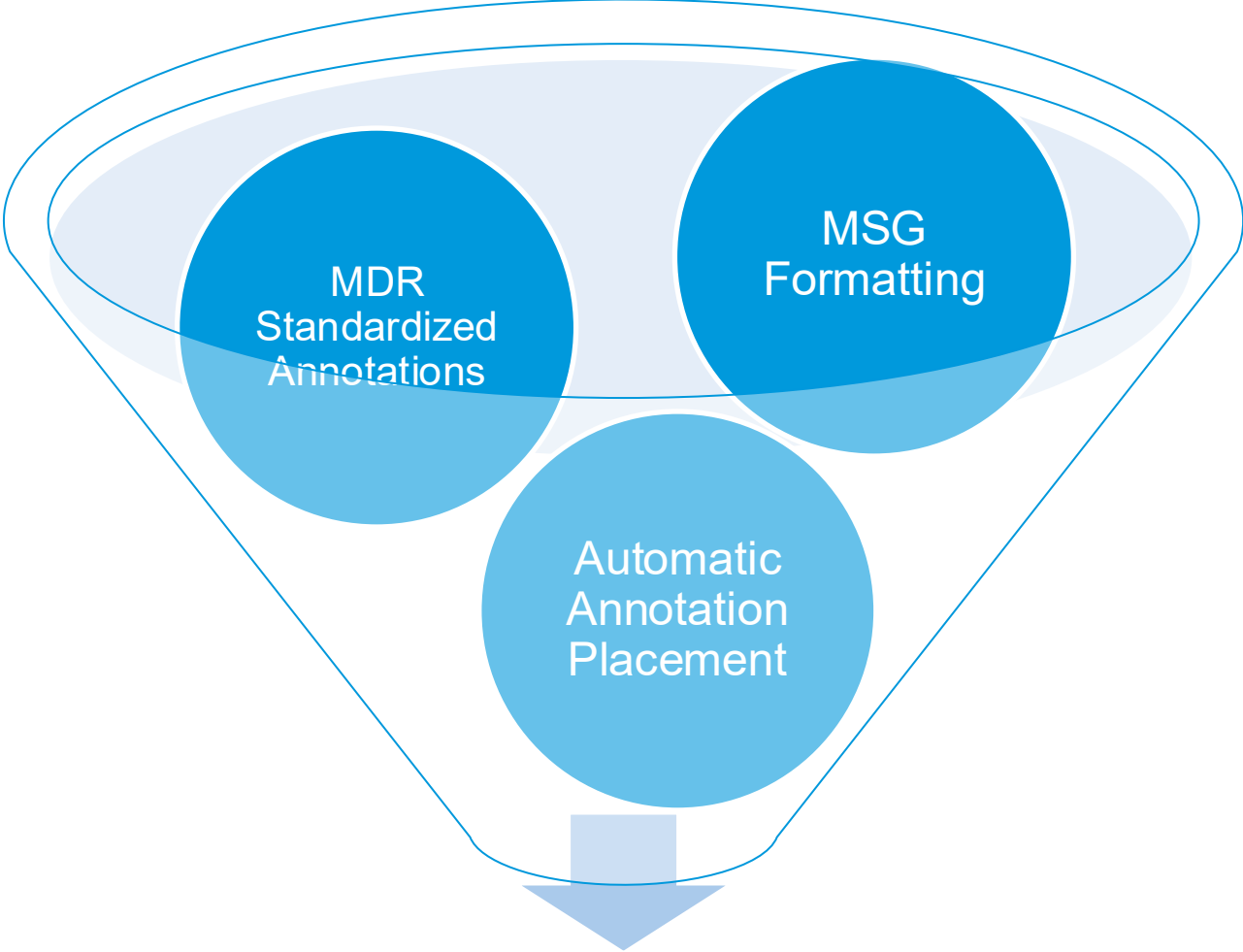
No ☐

(One or more categories may be selected)

ETHNIC = NOT HISPANIC OR LATINO

|| Approach

Utilize Standards to Automate



Semi-automated CRF Annotation Tool



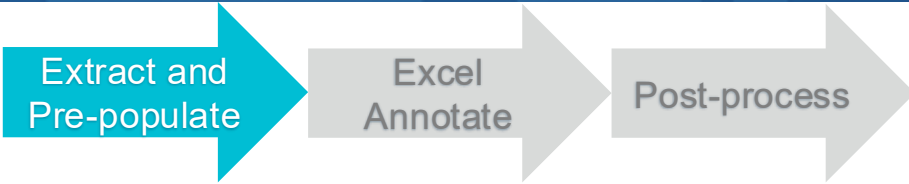
DSSE
CONNECTING THE DOTS



| | Solution

Extract PDF information into Python

PDFminer scrapes text boxes and coordinates from PDF



Form extracted from header

AC-ALNY_Library_eCRF v01.013 13JUN2025: AC-ALNY_Library eCRFs
Folder: All forms
Form: Demographics
Generated On: 27 Oct 2025 16:08:50 (GMT)

Was this participant a prior screen failure? Yes ☐ 1 No ☐

If Yes, please provide the original participant number (xxxx-xxxx) 2

Parent Study ALN-XXX-YYY 3

DEMOGRAPHICS

Year of Birth (yyyy) 5

Age (years) at time of consent Fixed Unit: years 6

Sex Male ☐ 7 Female ☐

Ethnicity Hispanic or Latino ☐ 8 Not Hispanic or Latino ☐ Not Reported ☐ Unknown ☐

Text_1 column

Text_2 column

Page	Form	Text_1	Text_2	Coord_1	Coord_2
8	Form: Demographics	AC-ALNY_Library_eCRF v01.013 13JUN2025: AC-ALNY_Library eCRFs Form: Demographics Generated On: 27 Oct 2025 16:08:50 (GMT)		90.0, 650.0, 461.3, 702.0	
8	Form: Demographics	Was this participant a prior screen failure?	Yes	90.0, 613.0, 296.4, 623.0	489.0, 613.0, 506.0, 623.0
8	Form: Demographics		No		492.7, 597.0, 506.0, 607.0
8	Form: Demographics	If Yes, please provide the original participant number (xxxx-xxxx)		95.0, 546.0, 354.9, 568.0	
8	Form: Demographics	Parent Study	ALN-XXX-YYY	90.0, 511.0, 153.0, 521.0	438.5, 511.0, 506.0, 521.0
8	Form: Demographics	DEMOGRAPHICS		90.0, 472.0, 170.9, 482.0	
8	Form: Demographics	Year of Birth (yyyy)		90.0, 437.0, 186.2, 447.0	
8	Form: Demographics	Age (years) at time of consent	Fixed Unit: years	90.0, 400.0, 239.1, 412.0	436.2, 400.0, 533.2, 412.0
8	Form: Demographics	Sex	Male	90.0, 349.0, 107.7, 359.0	483.6, 349.0, 506.0, 359.0
8	Form: Demographics		Female		470.8, 333.0, 506.0, 343.0
8	Form: Demographics	Ethnicity	Hispanic or Latino	90.0, 294.0, 131.9, 304.0	418.8, 294.0, 506.0, 304.0
8	Form: Demographics		Not Hispanic or Latino		398.3, 278.0, 506.0, 288.0
8	Form: Demographics		Not Reported		441.1, 262.0, 506.0, 272.0
8	Form: Demographics		Unknown		460.4, 246.0, 506.0, 256.0

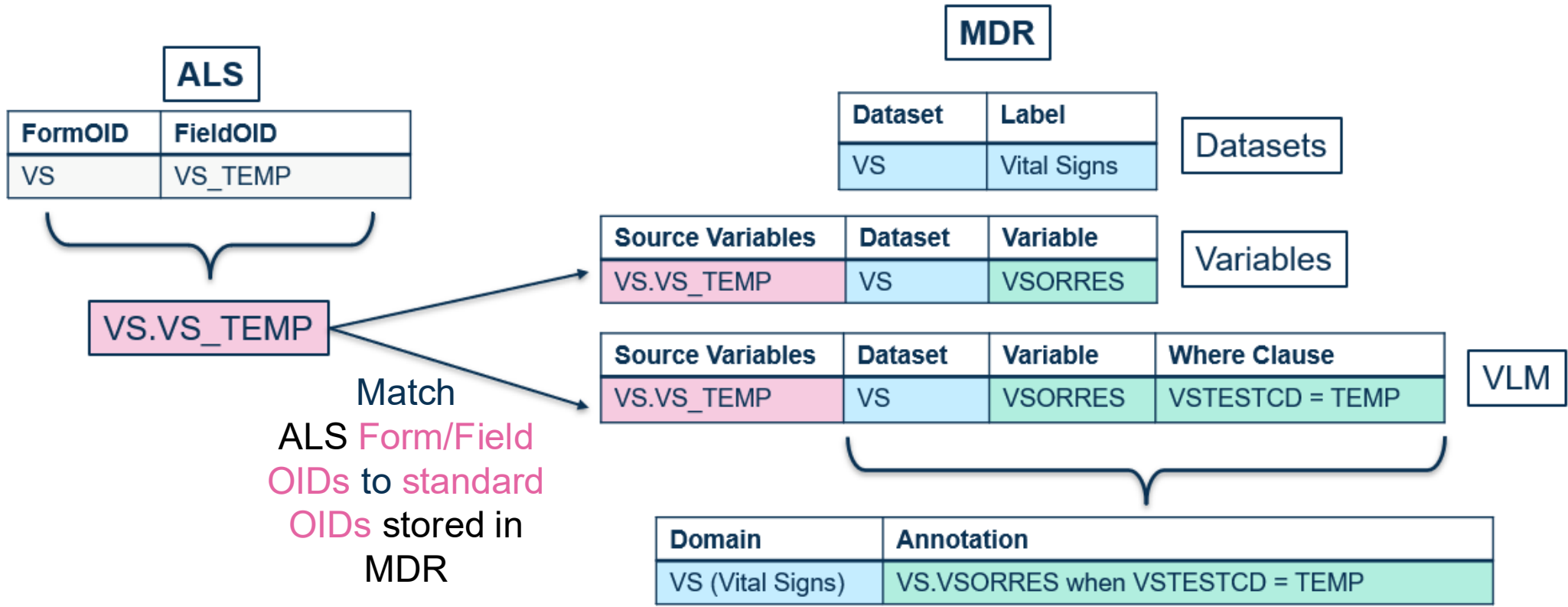
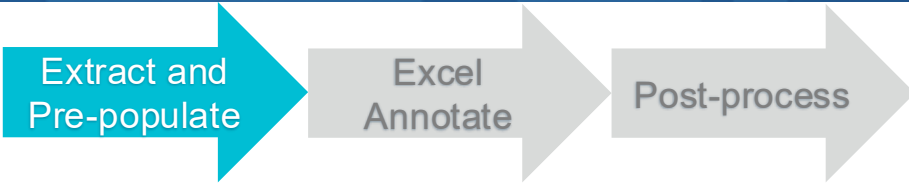


DSSE
CONNECTING THE DOTS

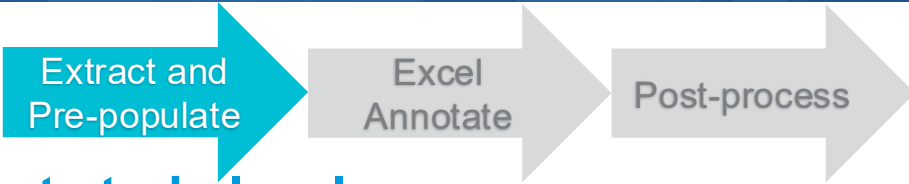


ALS linkage to MDR (OID)

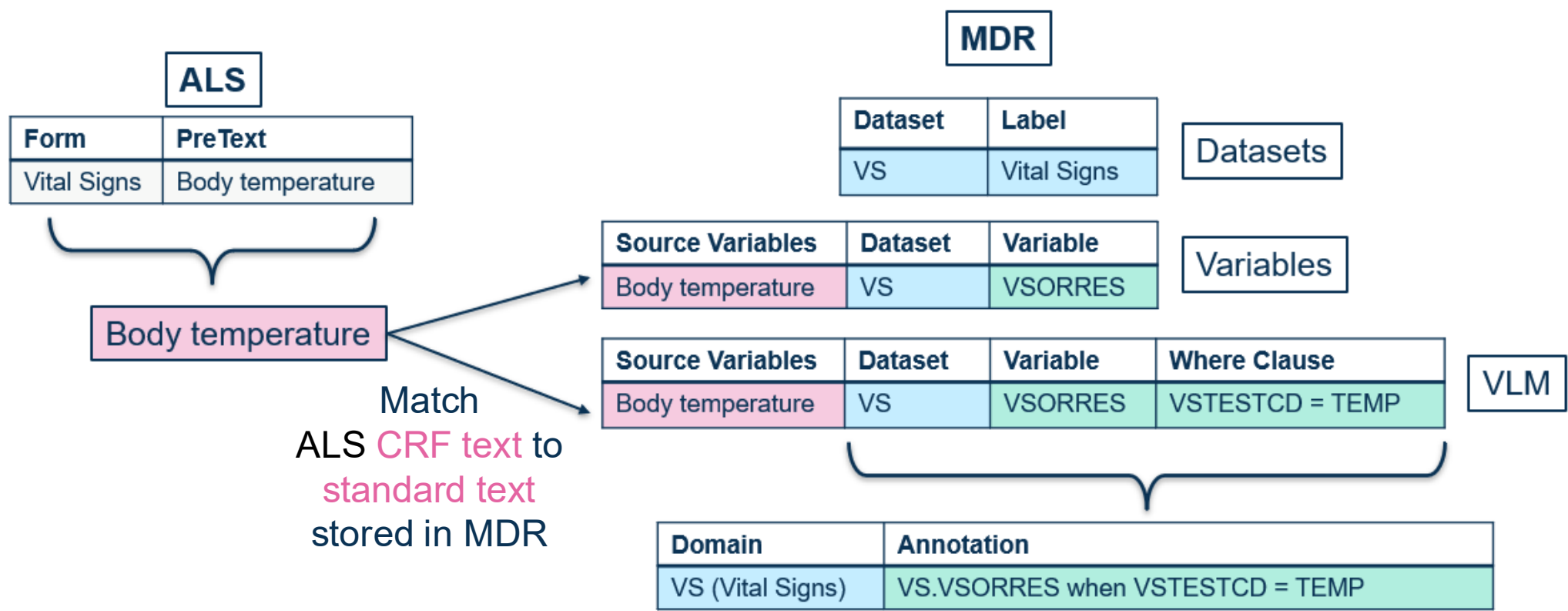
Connecting study ALS information to MDR



ALS linkage to MDR (Text)

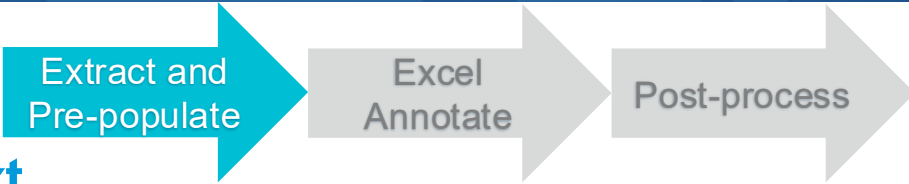


Consider CRF text itself where standard OIDs were not used at study level



CRF-ALS-MDR linkage

Overview of connection between SDTM variable and CRF text



2. ALS

FormOID	FieldOID	DataFormat	ControlType	PreText
GL_DM	XX_PREVSF	\$2	RadioButton	Was this participant a prior screen failure?
GL_DM	XX_PREVSF_NUM	\$11	Text	If Yes, please provide the original participant number (ssxxxx-xxxx)
GL_DM	LABEL1		Text	DEMOGRAPHICS
GL_DM	DM_DAT_BRTH	yyyy	DateTime	Year of Birth (yyyy)
GL_DM	DM_AGE	3	Text	Age (years) at time of consent
GL_DM	DM_AGEU	\$6	DropDownList	Age unit
GL_DM	DM_SEX	\$2	RadioButton	Sex
GL_DM	DM_CBP	\$2	RadioButton	Is the participant of childbearing potential?
GL_DM	DM_ETHNIC	\$25	RadioButton (Vertical)	Ethnicity

1. CRF

Was this participant a prior screen failure? Yes ☐ 1 No ☐

If Yes, please provide the original participant number (xxxx-xxxx) _____ 2

Parent Study ALN-XXX-YYY 3

DEMOGRAPHICS

Year of Birth (yyyy) _____ 5

Age (years) at time of consent _____ Fixed Unit: years 6

Sex Male ☐ 7 Female ☐

Ethnicity _____ 8
Hispanic or Latino
Not Hispanic or Latino

Dataset	Variable	Label	Source Variables
DM	BRTHDTC	Date/Time of Birth	DM.DM_DAT_BRTH
DM	AGE	Age	DM.DM_AGE
DM	AGEU	Age Units	DM.DM_AGEU
DM	SEX	Sex	DM.DM_SEX_STD

3. MDR



DSSE
CONNECTING THE DOTS



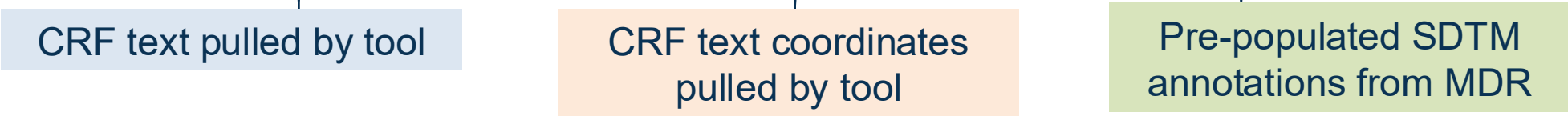
Control Sheet

Pre-populate standard field variables and export to a control sheet



page	form	text_1	text_2	coord1	coord2	domain	anno1	anno2
13	Form: Demographics	Was this participant a prior screen failure?	Yes	90.0, 627.01, 296.43, 637.01	488.95, 627.01, 506.0, 637.01		DM.RESCREEN in SUPPDM	
13	Form: Demographics		No		492.66, 611.01, 506.00000000000006, 621.01			
13	Form: Demographics	If Yes, please provide the original participant number (ssxxxx-xxxx)		95.0, 560.01, 354.85000000000001, 582.01			DM.PSUBJID in SUPPDM	
13	Form: Demographics	DEMOGRAPHICS		90.0, 525.01, 170.91, 535.01				
13	Form: Demographics	Year of Birth (yyyy)		90.0, 490.01, 186.17000000000002, 500.01			DM.BRTHDTC	
13	Form: Demographics	Age (years) at time of consent	Fixed Unit: years	90.0, 453.01, 239.07, 465.01	436.18, 453.01, 533.21, 465.01		DM.AGE	
13	Form: Demographics	Age unit	Years	95.0, 402.01, 135.66, 412.01	479.02, 402.01, 506.0, 412.01		DM.AGEU	
13	Form: Demographics	Sex	Male	90.0, 363.01, 107.66, 373.01	483.6, 363.01, 506.0, 373.01		DM.SEX	
13	Form: Demographics		Female		470.83, 347.01, 505.99999999999994, 357.01			
13	Form: Demographics	Is the participant of childbearing potential?	Yes	95.0, 308.01, 304.33000000000001, 318.01	488.95, 308.01, 506.0, 318.01		RP.RPORRES when RPTSTCD = CHILDPOT	
13	Form: Demographics		No		492.66, 292.01, 506.00000000000006, 302.01			

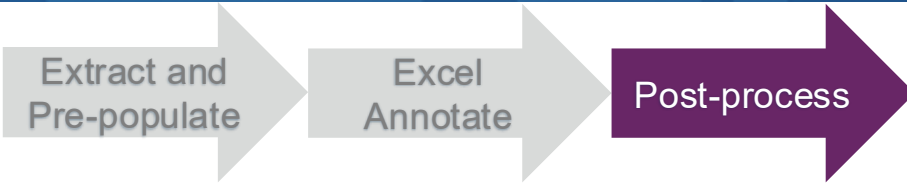
Programmer updates as needed



DSSE
CONNECTING THE DOTS

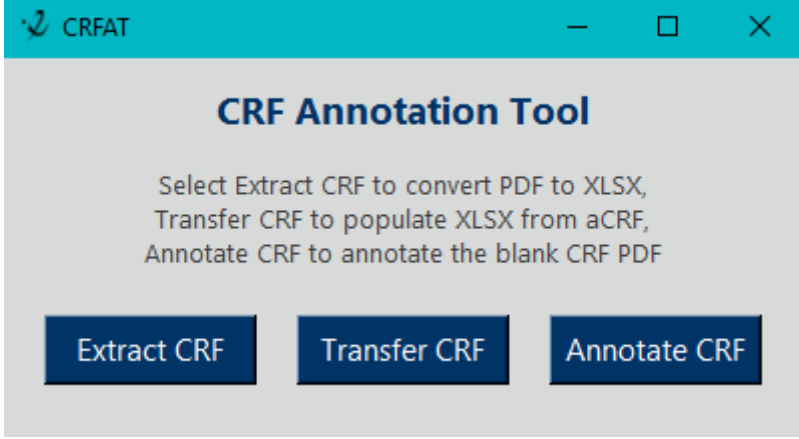


||| Annotating CRF



Ingest completed Excel control sheet

- Utilize Python Imaging Library (PIL) package to determine the width of annotation text based on font and size
- Calculate placement to avoid overlapping and truncation
- Determine coloring based on order of domains encountered per page
- Annotate directly to PDF

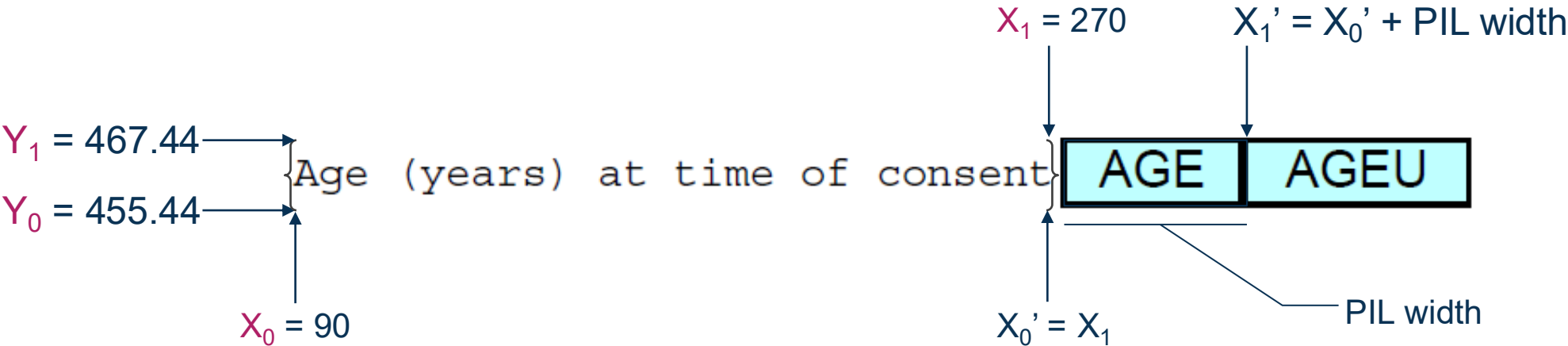


Annotating CRF

Annotation text box placement algorithm



page	text_1	coord1	domain	anno1	anno2
15	Age (years) at time of consent	90.0, 455.44, 270.0, 467.44	DM	AGE	AGEU



page	text_1	text_2	coord1	coord2	domain	anno1	anno2
15	Is Participant of childbearing potential?	Yes	90.0, 349.44, 336.0, 359.44	488.0, 349.44, 506.0, 359.44	RP	RPORRES	RPTESTCD = CHILDPOT

Is Participant of childbearing potential?

RPORRES

RPTESTCD = CHILDPOT

Yes ☒

No ☐



DSSE
CONNECTING THE DOTS



Annotating CRF

Assigning annotation text box colors



Input dataframe

Page	Form	Domain	Anno1	Anno2
1	Form: Participant		XX.[NOT SUBMITTED]	DM.SUBJID
2	Form: Demographics		DM.BRTHDTC	DC.BRTHDTC
2	Form: Demographics	DS	DSCAT = PROTOCOL MILESTONE	
2	Form: Demographics	RP	RPORRES when RPTTESTCD = CHILDPOT	
3	Form: Demographics	DS	PROTVERS in SUPPDS	



Page	Form	Anno1	Anno2	Color1	Color2
1	Form: Participant	[NOT SUBMITTED]	SUBJID	XX	DM
2	Form: Demographics	BRTHDTC	BRTHDTC	DM	DC
2	Form: Demographics	DSCAT = PROTOCOL MILESTONE		DS	
2	Form: Demographics	RPORRES when RPTTESTCD = CHILDPOT		RP	
3	Form: Demographics	PROTVERS in SUPPDS		DS	



Apply colors to each annotation

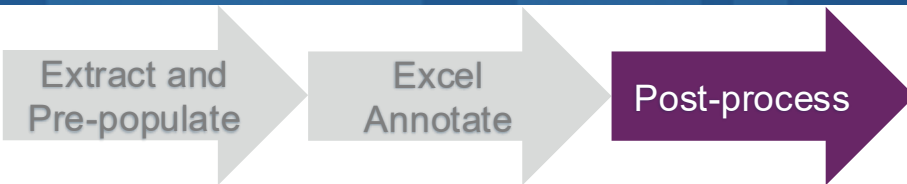


DSSE
CONNECTING THE DOTS



Annotating CRF

Applying annotations to PDF



Declare as FreeText
annotation object

Define text content
and font color

Set to viewable

```
anno_dict_1 = {  
    '/Type': '/Annot',  
    '/Subtype': '/FreeText',  
    '/Rect': [12, 768, 142, 787],  
    '/Contents': '<b>DS (Disposition) </b>',  
    '/DA': '/Arial 14 Tf 0 0 0 rg',  
    '/C': [0.75, 1, 1],  
    '/F': 4,  
}
```

DS (Disposition)

Assign coordinates

Set background color

Set dashed border

```
anno_dict_2 = {  
    '/Type': '/Annot',  
    '/Subtype': '/FreeText',  
    '/Rect': [240, 453, 255, 465],  
    '/Contents': 'When multiple screen failure  
reasons checked, the primary reason  
is mapped to DSTERM/DECOD.  
Additional reasons to be mapped to  
DSTERMx/DSDECODx in SUPPDS.',  
    '/DA': '/Arial 12 Tf 0 0 0 rg',  
    '/C': [0.75, 1, 1],  
    '/F': 4,  
    '/BS': {'/W': 1, '/S': '/D', '/D': [3]}  
}
```

When multiple screen failure
reasons checked, the
primary reason is mapped to
DSTERM/DECOD.
Additional reasons to be
mapped to
DSTERMx/DSDECODx in
SUPPDS.

Result

page	text_1	text_2	domain	anno1	anno2
15	Was this participant a prior screen failure?	Yes	DM	RESCREEN in SUPPDM	
15		No			
15	If Yes, please provide the original participant number (xxxx-xxxx)		DM	PSUBJID in SUPPDM	
15	DEMOGRAPHICS				
15	Year of Birth (yyyy)		DM	BRTHDTC	
15	Age (years) at time of consent	Fixed	DM	AGE	AGEU
15	Sex	Male	DM	SEX	
15		Female			
15	Is Participant of childbearing potential?	Yes	RP	RPORRES	RPTSTCD = CHILDPOT
15		No			
15	Ethnicity	Hispanic	DM	ETHNIC	
15		Not			
15		Not			
15		Unknown			
15	Race (check all that apply)		DM	When multiple values are selected then	
15	American Indian or Alaska Native		DM	RACE = AMERICAN INDIAN OR ALASKA NATIVE	
15	Asian		DM	RACE = ASIAN	

DM (Demographics) RP (Reproductive System Findings) DS (Disposition)

Vers 1.003 15Oct2024: xyz
Form: Demographics
Generated On: 16 Oct 2024 19:19:46 (GMT)

Was this participant a prior screen failure? Yes ☒ 1
No ☐

RESCREEN in SUPPDM

If Yes, please provide the original participant number (xxxx-xxxx) PSUBJID in SUPPDM 2

DEMOGRAPHICS

Year of Birth (yyyy) BRTHDTC 4

Age (years) at time of consent AGE AGEU Fixed Unit: years 5

Sex SEX 6
Male ☒
Female ☐

Is Participant of childbearing potential? RPORRES 7
RPTSTCD = CHILDPOT
Yes ☒
No ☐

Ethnicity ETHNIC 8
Hispanic or Latino ☒
Not Hispanic or Latino ☐
Not Reported ☐
Unknown ☐

Race (check all that apply)
When multiple values are selected then RACE = MULTIPLE and individual responses are RACE1, RACE2, RACEn in SUPPDM

American Indian or Alaska Native RACE = AMERICAN INDIAN OR ALASKA NATIVE 9

Asian RACE = ASIAN 11

Advantages of the CRF Annotation Tool

Focus on mapping and annotations

Time savings

- Easier and quicker to work in Excel
- No need to fiddle with box shapes or fonts
- Applies MSG coloring automatically

Standards adherence

- Follows external industry guidance
- Adheres to internal standards decisions
- Teams can focus on non-standard forms

Updatability & collaboration

- Can find and update annotations across multiple pages faster and more easily
- Multiple team members can edit and work across a shared Excel



DSSE
CONNECTING THE DOTS



Future Improvements

Updates

- Additional automatic formatting
- Improved mid-study update handling
- Have suggested annotations populated in grey for user review



DSSE
CONNECTING THE DOTS



|| Thanks!

