



INTEGO
GROUP

Making data mapping process easier and smarter with SAS and R

Andrii Artemchuk, Kharkiv, Ukraine



Amsterdam 2019



Main idea

To automate the process of the raw data transformation into SDTM CDISC compliant data without using additional software*

* Apart from SAS and R





Results overview

1. Transformation of the raw CRF values to the SDTM CDISC compliant values automatically
2. Checking if the SDTM data was mapped correctly





Required software

- SAS version 9.2 or higher
- R version 3.0.1 or higher
- R studio version 1.2.5001 or higher for convenient work with the R code





Part 1

Transformation of the raw CRF values to
SDTM CDISC compliant values automatically





CRF limitations

1. The PDF version of CRF must have text that can be highlighted or found
2. Only CRF fields that contain tick-boxes will be analyzed
3. The tick-boxes and their corresponding text must be left or right aligned





1. A CRF must have a PDF version

The text in the PDF document must be highlighted or found on a page

Epoch	STUDY DISCONTINUATION	
Date subject discontinued from study	  	
Select the primary reason the subject discontinued from the study	Adverse Event Pregnancy Death Lost to follow-up Protocol violation Non-compliance with study drug	





2. Only fields that contain pre-defined text values or tick-boxes are processed

Tick-boxes and associated text values are the easiest to be found among the other CRF fields

Epoch	STUDY DISCONTINUATION
Date subject discontinued from study	
Select the primary reason the subject discontinued from the study	Adverse Event Pregnancy Death Lost to follow-up Protocol violation Non-compliance with study drug Non-compliance





3. The tick-boxes must be right or left aligned

That makes them easy to find and locate. The example of a page with **not** aligned tick-boxes:

End of treatment	
Primary reason for end of treatment:	☐ Adverse event(s) <i>(also specify on the Adverse events page)</i>
	☐ Abnormal laboratory value(s) <i>(please specify below)</i>
	☐ Abnormal test procedure result(s) <i>(please specify below)</i>
	☐ Subject withdrew consent <i>(please specify below)</i>
	☐ Lost to follow-up
	☐ Administrative problems <i>(please specify below)</i>
	☐ Death <i>(please complete Serious adverse events pages and record adverse events leading to death on the Adverse events page)</i>
Study evaluation completion	
Primary reason for study evaluation completion:	☐ Subject withdrew consent <i>(please specify below)</i>
	☐ Lost to follow-up
	☐ Administrative problems <i>(please specify below)</i>
	☐ Death <i>(please complete Serious adverse events pages and record adverse events leading to death on the Adverse events page)</i>





Extract the text from the PDF

PDFTOOLS package in R

Function to extract text and put into R data frame:
pdf_data(<pdf_file>)





R code to parse a PDF version of the CRF and output each page to a separate CSV file

```
install.packages("callr");
install.packages("pdftools");

library(pdftools);
library(callr);

pdf_path      <- "<Path to CRF in PDF format>"; # USER INPUT
total_pages   <- <number of CRF pages>;         # USER INPUT
csv_path      <- "<Path to output CSV files>";  # USER INPUT

for (i in 2:total_pages) {
  pdfdata <- r(function(x) pdftools::pdf_data(x), args = list(pdf_path))[[i]];
  path    <- paste(csv_path, "page", i-1, ".csv", sep = "");

  print(paste("Page", i, "of", total_pages, "was output into file:", path, sep = " "));

  write.csv(pdfdata, path, row.names = FALSE);
}
```





Result of R code execution after being imported into SAS dataset using PROC IMPORT:

	width	height	x	y	text
21	16	9	90	169	Age
22	16	9	(1) 90	184	Age
23	15	9	109	184	unit (2)
24	15	(3) 9	90	199	Sex
25	32	9	489	184	YEARS
26	20	9	485	199	Male
27	29	9	476	215	Female
28	(4) 36	9	90	234	Ethnicity
29	35	9	431	234	Hispanic
30	8	9	469	234	or
31	26	9	479	234	Latino
32	15	9	413	250	Not
33	35	9	431	250	Hispanic
34	8	9	469	250	or
35	26	9	479	250	Latino
36	15	9	455	266	Not
37	33	9	472	266	reported
38	39	9	466	282	Unknown

Day month and year of birth

Age (x;y) word start coordinates (1)

Age unit text (2)

Sex height (3)

Ethnicity

width (4)

YEARS

Male ☐

Female ☐

Hispanic or Latino ☐

Not Hispanic or Latino ☐

Not reported ☐

Unknown ☐





SAS macro that rules the process



```
%auto_mapping(func_outlib_name      = <area.ds_name.user_name for function>,  
              sdtm_term_xlsfile_path = "<SDTM Controlled Terminology file path>",  
              sdtm_term_sheet_name   = "<SDTM Controlled Terminology Excel sheet>",  
              crf_page_path           = "<Path to CRF pages in CSV>",  
              print_test_pagenum      = <enter any number to print test page>,  
              keep_or_drop_text       = "<word pattern for keep or drop>",  
              keep_or_drop_mode       = <KEEP | DROP: mode for a pattern>,  
              start_pagenum           = <start page number>,  
              end_pagenum              = <end page number>,  
              tbox_line_end            = <tick-box word end position>,  
              corr_tbox_end            = <tick-box word end pos. correction>,  
              del_prev_page_fl         = <Y | N: delete prev. file with code?>,  
              output_code_path         = "<Path to output files with code>",  
              show_temp_datasets       = <Y | N: show/hide temp. datasets?> );
```





Select only tick-boxes out of all the CRF data



User input required: find out which words are **the last** within every tick-box. Determine the common end position value for the last words found





Example

Purpose: to look at tick-box text and find the end horizontal coordinate for the last word within each tick-box text





CRF text value	Horizontal end position of text	Horizontal start position of text
Large	468	445
Small	468	445
Gall	471	454
Urinary	471	440
Abdominal	475	431
Soft	476	460
62	493	483
of	504	496
Cavity	505	479
Gland	505	481
Bone	505	484
Brain	505	483
Breast	505	480
Bronchus	505	467
Cervix	505	478
Colon	505	481
Duodenum	505	461
Esophagus	505	462
Bladder	505	474



Site code	Abdominal Cavity	☐
	Adrenal Gland	☐
	Bone	☐
	Brain	☐
	Breast	☐
	Bronchus	☐
	Cervix	☐
	Colon	☐
	Duodenum	☐
	Esophagus	☐
	Gall Bladder	☐
	Large Intestine	☐
	Kidney	☐
	Liver	☐
	Lung	☐
	Lymph Nodes	☐



The coordinate
is equal to **505**

CRF text value	Horizontal end position of text	Horizontal start position of text
Large	468	445
Small	468	445
Gall	471	454
Urinary	471	440
Abdominal	475	431
Soft	476	460
62	493	483
of	504	496
Cavity	505	479
Gland	505	481
Bone	505	484
Brain	505	483
Breast	505	480
Bronchus	505	467
Cervix	505	478
Colon	505	481
Duodenum	505	461
Esophagus	505	462
Bladder	505	474



Site code	Abdominal Cavity	☐
	Adrenal Gland	☐
	Bone	☐
	Brain	☐
	Breast	☐
	Bronchus	☐
	Cervix	☐
	Colon	☐
	Duodenum	☐
	Esophagus	☐
	Gall Bladder	☐
	Large Intestine	☐
	Kidney	☐
	Liver	☐
	Lung	☐
	Lymph Nodes	☐



Result of the tick-box selection

	section	linenum	crf_header	crf_term
1	1	1	Surgical Procedure	Resection
2	1	2	Surgical Procedure	Biopsy
3	2	1	Intent of surgery	Diagnostic
4	2	2	Intent of surgery	Palliative
5	2	3	Intent of surgery	Curative

Site	LIVER
Surgical Procedure	
Section 1	Resection ☐
	Biopsy ☐
Date of surgery	
Intent of surgery	
Section 2	Diagnostic ☐
	Palliative ☐
	Curative ☐





Match an SDTM term to each CRF section

Dose Frequency		Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)
As needed	☐	C71113		Yes	Frequency	FREQ	Frequency
Continuous	☐	C64496	C71113		Frequency	BID	BD; Twice per day
Intermittent	☐	C71129	C71113		Frequency	BIM	Twice per month
Unknown	☐	C53279	C71113		Frequency	CONTINUOUS	Continuous
Other	☐	C64954	C71113		Frequency	OCCASIONAL	Occasional
Every hour	☐	C64499	C71113		Frequency	PRN	As needed
Every 2 hours	☐	C64516	C71113		Frequency	Q2H	Every 2 hours
Every 3 hours	☐	C64533	C71113		Frequency	Q3D	Every 3 days
Every 4 hours	☐	C64517	C71113		Frequency	Q3H	Every 3 hours
Every 5 hours	☐	C25473	C71113		Frequency	QD	Daily
Every other day	☐	C64510	C71113		Frequency	QH	Every hour
2 times per day	☐	C64525	C71113		Frequency	QOD	Every other day





Problems that were solved

- A CRF term may be split into several lines

Outcome of AE	Fatal	☐
	Not recovered/Not resolved	☐
	Recovered/Resolved	☐
	Recovered/Resolved with sequelae	☐

- A CRF section may be split into 2 pages
- CRF tick-boxes may lie not exactly one under another

Outcome of AE	Fatal	☐
	Not recovered/Not resolved	☐
	Recovered/Resolved	☐
	Recovered/Resolved with sequelae	☐





Result of macro execution

An RTF file for each CRF page where tick-boxes were found.
Each file contains the SAS code – a simple PROC FORMAT call

```
/*Code generated from page 122*/  
proc format;  
  value <insert_user_value_name>; /* Anatomical location of biopsy */  
    "Adrenal gland" = "ADRENAL GLAND"  
    "Ascites" = "ASCITES"/*Warning: corresponding value was not found in SDTM!*/  
    "Bone" = "BONE"  
    "Bronchus" = "BRONCHUS"  
    "Cervix" = "CERVIX"/*Warning: corresponding value was not found in SDTM!*/  
    "Colon" = "COLON"  
    "Duodenum" = "DUODENUM"  
    "Esophagus" = "ESOPHAGUS"  
    "Gall bladder" = "BLADDER"/*Warning: check this line in CRF and SDTM for consistency*/;  
run;
```





Special case – a CRF section and its text fields were not found in the SDTM Terminology

It happens to all study-specific sections. Text fields
will be mapped to themselves in the upper case:

```
/*Code generated from page 26*/  
proc format;  
    value <insert_user_value_name>; /* If not done, please indicate reason not done*/  
  
        "Subject refused to complete" = "SUBJECT REFUSED TO COMPLETE" /*Warning: corresponding  
value was not found in SDTM!*/  
        "Subject unable to complete (physically or mentally)" = "SUBJECT UNABLE TO COMPLETE  
(PHYSICALLY OR MENTALLY)" /*Warning: corresponding value was not found in SDTM!*/  
        "Other" = "OTHER" /*Warning: corresponding value was not found in SDTM!*/  
  
run;
```





Part 2

Check if CRF data was mapped to SDTM
Terminology data correctly





A SAS macro that performs the checking

```
%sdm_check(func_outlib_name      = <lib.ds_name.user_name for function>,  
           dsin_sdm             = <SDTM dataset lib.name>,  
           variable_list        = <space-delimited list of variables>,  
           sdm_term_xlsfile_path = "<SDTM Controlled Terminology file path>",  
           sdm_term_sheet_name  = "<SDTM Controlled Terminology Excel sheet>",  
           report_file_path     = "<Path to output files with report>");
```





Principle

The macro compares each of the variable values to the SDTM CDISC Submission values in the SDTM CT. The report files are created separately for each variable in the RTF format.





Checking the mapping can produce three types of results

- a match
- no match while the list of SDTM CDISC Submission values is extensible
- no match while the list of SDTM CDISC Submission values is **not** extensible;





Example of the RTF report file for SDTM.AEOUT variable

Results of checking for variable AEOUT (Outcome of Adverse Event) mapped
values against CDISC SDTM

Value:	Check result:
FATAL	OK - mapping is correct
NOT RECOVERED/NOT RESOLVED	OK - mapping is correct
RECOVERED/RESOLVED	OK - mapping is correct
RECOVERED/RESOLVED WITH SEQUELAE	OK - mapping is correct
RECOVERING/RESOLVING	OK - mapping is correct
UNKNOWN	OK - mapping is correct
MULTIPLE	Not found in the SDTM Terminology!





What is the benefit?

Time: it takes ~1 hour to generate RTF files with the SAS code starting from the reading in the CRF of 200 pages using the R code



Accuracy: Test runs show that no data is lost.
If a CRF text value does not have a pair in the SDTM CT it will still be printed in the code with the warning message





So what? or Conclusion

The described macros show that reading and processing PDF data is quite possible by the combined efforts of SAS and R.

The most important thing is to show by a particular example that it is possible to significantly speed up the SDTM mapping process.





Macros and codes are available at GitHub

https://github.com/aartemchuk/SDTM_MAPPING





INTEGO
GROUP

Thank you for your attention!

Andrii Artemchuk

andrii.artemchuk@intego-group.com

Kharkiv, Ukraine

www.intego-group.com

