

Automated creation of a mapping specification for a pooled database

Thomas Wollseifen

Date 15th-19th November 2021



15th-19th November

EU Connect 2021

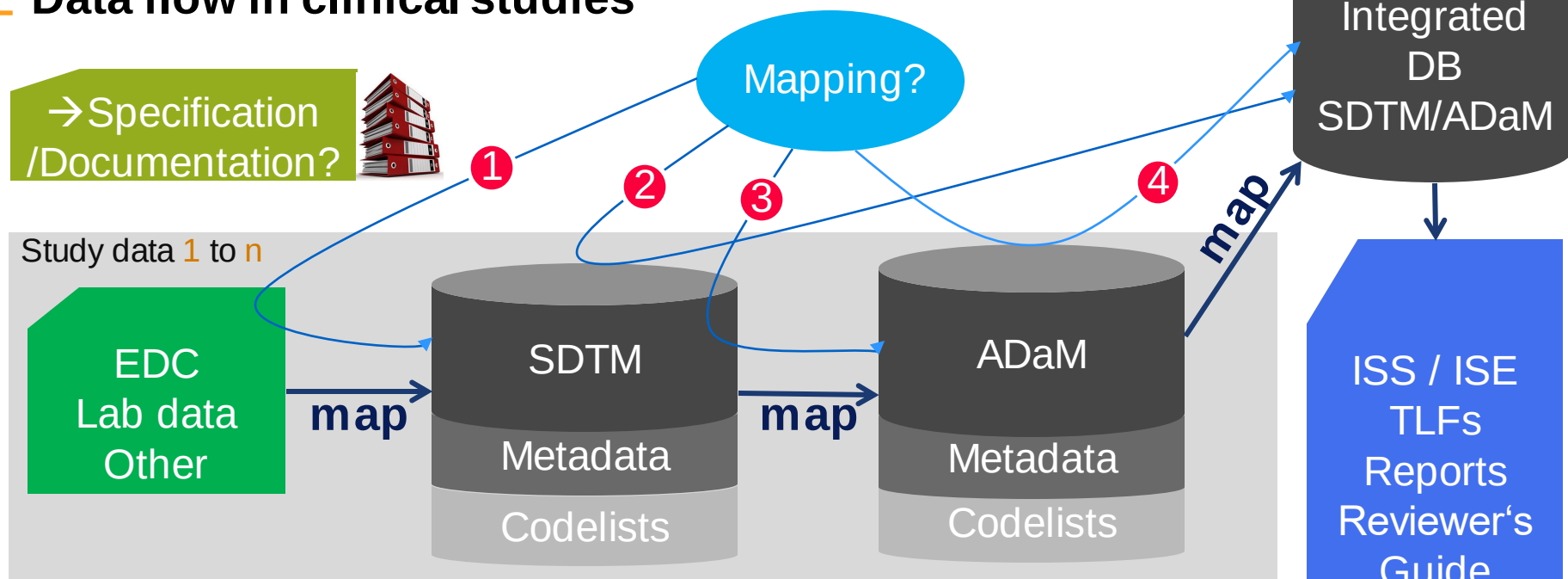
The Clinical Data
Science Conference

AGENDA

- 01 Introduction - Data flow in clinical studies
- 02 Macro %m_create_metadata_docu_xls()
- 03 Mapping suggestions – Fuzzy matching
- 04 Add previous mapping description
- 05 Summary

01 Introduction - Data flow in clinical studies

Data flow in clinical studies



Manual mapping programming

Semi-automatic mapping

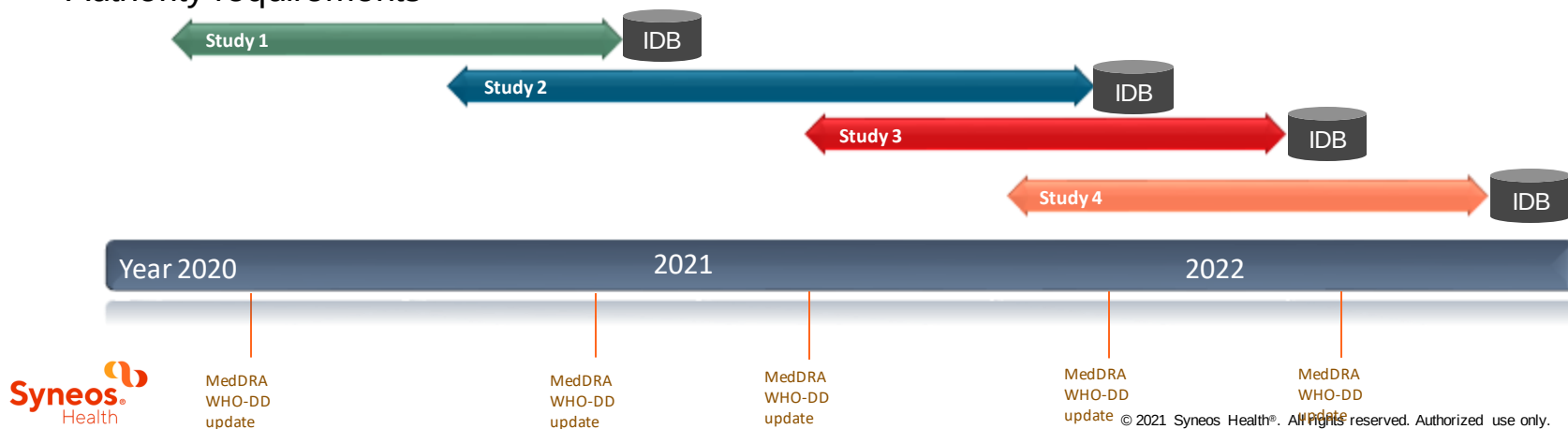
Automatic mapping

→ Meta data repository (MDR)

Features of Study flow/IDB that influence the documentation

- Study data is added at different timepoints
- Mapping might be changed due to requests
- New variables could be added / New domains
- Structure of the IDB might change
- SDTM or ADaM IDB
- Maintenance of the IDB (MedDRA updates, WHO-DD updates, corrections)
- Authority requirements

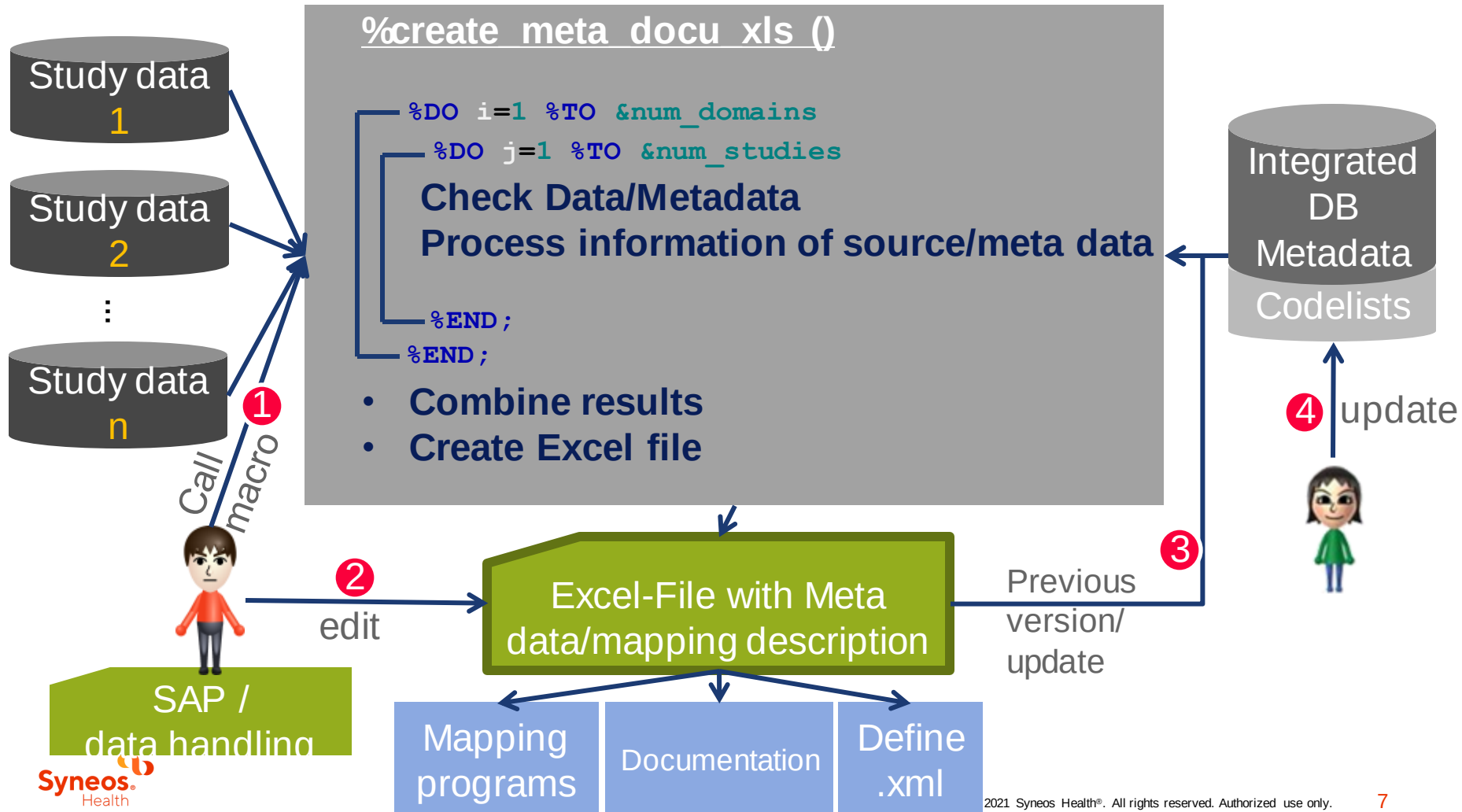
→ How do we document this?



02 Macro %m_create_metadata_docu_xls()

→ Documentation

A		B		C		D								SN_1000	SN_1001	SN_1002
Domains Pool		SN_1000		SN_1001		SN_1002		1	2	3	4	5	6	7	8	9
2	AE	AE	AE	AE	AE	AE	AE	TV	TRIALNO					SN_1000	SN_1001	SN_1002
3	CM	COMED	CM	CM	CM	CM	CM	TV	VISITSP					SN_1000	SN_1001	SN_1002
4	DM	not available or not assigned	DM	DM	DM	DM	DM	TV	VIS_SDR					SN_1000	SN_1001	SN_1002
5	DS	PROTOCOL ENDPOINT DV	DS	DS	DS	DS	DS	TV	VIS_SDRS					SN_1000	SN_1001	SN_1002
6	EX	DOSE	EX	EX	EX	EX	EX	TV	STUDYID	1	Study Identifier	C	10	ADS SOTM	TV STUDYID	TV STUDYID
7	LB	LAB	LB	LB	LB	LB	LB	TV	DOMAIN	2	Domain Abbreviation	C	2	DOMAIN	TV DOMAIN	TV DOMAIN
8	MH	MHIS	MH	MH	MH	MH	MH	TV	VISITNUM	3	Visit Number	N	9.4	P_VISIT	ADS SOTM	TV VISITNUM
9	SE	not available or not assigned	SE	SE	SE	SE	SE	TV	VISIT	4	Visit Name	C	40	ADS SOTM	TV VISIT	TV VISIT
10	SV	VISIT	SV	SV	SV	SV	SV	TV	VISITN	5	Visit Name (N)	N	14.6	ZVISIT	ADS	TV VISITN
11	TA	not available or not assigned	TA	TA	TA	TA	TA	TV	VISITDY	6	Planned Study Day of Visit	N	8	ADS SOTM	TV VISITDY	TV VISITDY
12	TE	not available or not assigned	TE	TE	TE	TE	TE	TV	ARMCD	7	Planned Arm Code	C	20	ZARMCD	ADS SOTM	TV ARMCD
13	TS	PROTOCOL	TS	TS	TS	TS	TS	TV	ARM	8	Description of Planned Arm	C	50	ADS SOTM	TV ARM	TV ARM
14	TV	VISDES	TV	TV	TV	TV	TV	TV	TVSTR	9	Visit Start Rule	C	100	ADS SOTM	TV TVSTR	TV TVSTR
15	VS	not available or not assigned	VS	VS	VS	VS	VS	TV	TVENRL	10	Visit End Rule	C	100	ADS SOTM	TV TVENRL	TV TVENRL
16																
17																
Info		AE	CM	DM	DS	EX	LB	MH	SE	SV	VS	TA	TE	TS	TV	



Mapping documentation – Example

Info – with mapping overview

Sheet for every domain

A		B		C		D		A		B		C		D		E		F		G		H		I		J		K	
1	Domains Pool	SN 1000	SN 1001	SN 1002	1	DOMAIN	VARIABLE	VARNUM	LABEL	TYPE	LENGT	CODLIST	BSP_TARGET	SN 1000	SN 1001	SN 1002													
2	AE	AE	AE	SUPPAE AE																									
3	CM	COMED	CM	CM																									
4	DM	not available or not assigned	DM	SUPPDM DM	24	AE	AESHOSPP							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESHOSPP													
5	DS	PROTOCOL ENDPOINT DV	DS	DS																									
6	EX	DOSE	EX	SUPPEX EX	25	AE	AESHOSPR							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESHOSPR													
7	LB	LAB	LB	SUPPLB LB	26	AE	AESOSP							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESOSP													
8	MH	MHS	MH	MH																									
9	SE	not available or not assigned	SE	SE	35	AE	AETRLPRC							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AETRLPRC													
10	SV	VISIT	SV	SV																									
11	TA	not available or not assigned	TA	TA	71	AE	STUDYID	1	Study Identifier	C	10		OAD ADS SDTM		AE STUDYID	AE STUDYID.Length=meta data definition													
12	TE	not available or not assigned	TE	TE	72	AE	DOMAIN	2	Domain Abbreviation	C	2	DOMAIN	OAD SDTM		AE DOMAIN	AE DOMAIN.Codelist is different. \$													
13	TS	PROTOCOL	TS	TS	73	AE	USUBID	3	Unique Subject Identifier	C	27		OAD ADS SDTM		AE USUBID	AE USUBID													
14	TV	VIDES	TV	TV																									
15	VS	not available or not assigned	VS	VS	74	AE	SUBJID	4	Study (N)	N	11		OAD ADS		AE SUBJID														
16					75	AE	AESEQ	5	Sequence Number	N	8.1		OAD ADS SDTM		AE AESEQ	AE AESEQ													
17					76	AE	AENYN	6	Any AEs? (N)	N	3	NY	OAD		AE AENYN														
					77	AE	AEGRPID	7	Group ID	C	20		OAD ADS SDTM		AE AEGRPID														
					78	AE	AERFID	8	Reference ID	C	30		OAD ADS SUPP		AE AERFID														

%create_meta_docu_xls (

```

project = Drugname
,libmeta = meta
,studyid = 1000#1001#1002
,pathstudydata = &source1 #&source2 #&source3
,inbase = patinfo#dm#dm#dm
,domains = ae cm dm ds ex lb mh se sv vs ta te ts tv vs
,pathoutfile = &outpath./docum
,outfilename = drugname_mapping_documentation_&date._test.xlsx
,standard = SDTM
,mapto = 1000: COMED->CM# (PROTOCOL ENDPOINT DV) ->DS#DOSE->EX#LAB->LB#MHS->MH#VISIT->SV#
        PROTOCOL->TS#VIDES->TV
        /1001:
        /1002: (SUPPAE AE) ->AE# (SUPPDM DM) ->DM# (SUPPEX EX) ->EX# (SUPPLB LB) ->LB
    
```

Mapping documentation – Domains and Mapping assignment

	A	B	C	D
1	Domains Pool	SN_1000	SN_1001	SN_1002
2	AE	AE	AE	SUPPAE AE
3	CM	COMED	CM	CM
4	DM	not available or not assigned	DM	SUPPDM DM
5	DS	PROTOCOL ENDPOINT DV	DS	DS
6	EX	DOSE	EX	SUPPEX EX
7	LB	LAB	LB	SUPPLB LB
8	MH	MHIS	MH	MH
9	SE	not available or not assigned	SE	SE
10	SV	VISIT	SV	SV
11	TA	not available or not assigned	TA	TA
12	TE	not available or not assigned	TE	TE
13	TS	PROTOCOL	TS	TS
14	TV	VISDES	TV	TV
15	VS	not available or not assigned	VS	VS

```
%create_meta_docu_xls(
    ,domains      = AE CM DM DS EX LB MH SE SV VS TA TE TS TV VS
    ,studyid      = 1000#1001#1002
    ,mapto        = 1000: COMED->CM#(PROTOCOL ENDPOINT DV)->DS#DOSE->EX#LAB->LB#MHS->MH#VISIT->SV# PROTOCOL->TS#VISDES->TV
                    /1001:
                    /1002: (SUPPAE AE)->AE#(SUPPDM DM)->DM#(SUPPEX EX)->EX#(SUPPLB LB)->LB
```

- 1:1 mapping / Multiple source domains / SUPP-Domains
- Information if source domain is not available or not assigned
- Mapto = **STUDYID: (domain1...domain n | SUPPdomain) -> TARGET**

Variable matching in different domains

	A	B	C	D	E	F	G	H	I	J	K
1	DOMAIN	VARIABLE	VARNUM	LABEL	TYPE	LENGT	CODELIST	BSP_TARGET	SN_1000	SN_1001	SN_1002
16	AE	AEMROBT							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AEMROBT
17	AE	AESBDTC							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESBDTC
18	AE	AESBYN							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESBYN
24	AE	AESHOSPP							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESHOSPP
25	AE	AESHOSPR							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESHOSPR
26	AE	AESOSP							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AESOSP
35	AE	AETRLPRC							Variable does not exist	Variable does not exist	SUPPAE QVAL where SUPPAE QNAM eq. AETRLPRC
1	AE	STUDYID	1	Study Identifier	C	10		OAD ADS SDTM		AE STUDYID	AE STUDYID, Length=meta data definition
72	AE	DOMAIN	2	Domain Abbreviation	C	2	DOMAIN	OAD SDTM		AE DOMAIN	AE DOMAIN, Codelist is different; \$
73	AE	USUBJID	3	Unique Subject Identifier	C	27		OAD ADS SDTM		AE USUBJID	AE USUBJID
74	AE	SUBJIDN	4	Study (N)	N	11		OAD ADS		AE SUBJIDN	
75	AE	AESEQ	5	Sequence Number	N	8,1		OAD ADS SDTM		AE AESEQ	AE AESEQ
76	AE	AENYN	6	Any AEs? (N)	N	3	NY	OAD		AE AENYN	
77	AE	AEGRPID	7	Group ID	C	20		OAD ADS SDTM			
78	AE	AEREFID	8	Reference ID	C	30		OAD ADS SUPP		AE AEREFID	
79	AE	AEREFID2	9	Reference ID 2	C	30		OAD ADS SUPP			
80	AE	AESPID	10	Sponsor-Defined Identifier	C	60		OAD ADS SDTM		AE AESPID	AE AESPID
81	AE	AELNKID	11	Link ID	C	65		OAD ADS SDTM			
82	AE	AETERM	12	Reported Term for the Adverse Event	C	200		OAD ADS SDTM		AE AETERM	AE AETERM

Additional variables which could not be mapped

Target /matched/unmatched variables

- Colors indicate if variable is **found** / **not found** / **additionally** available
- Mapping suggestions could be added for **additionally** found variables
- Information about content and checks against metadata

03 Mapping suggestions – Fuzzy matching

Levenshtein distance

F	O	R	M
F	O	R	K

substitution

M to K

Costs=1

Mapping suggestions – Fuzzy matching techniques

```
data suggestion&_std.&target.;
  set suggestion&_std.&target.;
  dist1=spedis(sasname, variable);
  dist2=complev(sasname, variable);
  sound_dist=complev(soundex(sasname), soundex(variable));
  dist3_label=complev(label,label&_std);
  dist4_label=spedis(label, label&_std);
  if (type='C' and type&_std.=2) or
    (type='N' and type&_std.=1)
  then type_cost=0; else type_cost=3;
  if soundex(sasname)=soundex(variable) then sound_cost=0;
  else sound_cost=2;
  sum_costs=sum(dist1,dist2,comp_,type_cost, sound_dist,
                dist3_label,dist4_label);
  if sasname=variable then match=1;
run;
```

- **SOUNDEX** – phonetic matching
- **SPEDIS** – spelling distance
- **COMPLEV** – Levenshtein Distance
- Check variable names, labels, types
- Summing costs – chose lowest costs as match

Mapping suggestions – Example 1

DOMAIN	VARIABLE	LABEL	TYP	LENG	CODEL	BSP_TARGET	SN_1000	SN_1001	SN_1002
DM	BIRTH_D						PATIENT.BIRTH_D , mapping suggestion: BRTH_D	Variable does not exist	Variable does not exist
DM	BIRTH_M						PATIENT.BIRTH_M , mapping suggestion: BRTH_M	Variable does not exist	Variable does not exist
DM	BIRTH_Y						PATIENT.BIRTH_Y , mapping suggestion: BRTH_Y	Variable does not exist	Variable does not exist
DM	BRTHDTC	Birth	C	19		SDTM		DM.BRTHDTC	DM.BRTHDTC
DM	BRTH_D	Day of Birth	N	2		ADS		DM.BRTH_D	
DM	BRTH_M	Month of Birth	N	2		ADS		DM.BRTH_M	
DM	BRTH_Y	Year of Birth	N	4		ADS		DM.BRTH_Y	

variable Variable Name	label1 Variable Label	SASNAME	suggest1	dist	dist2	dist3	sound1	sound2	sound_dist	type_cost	sound_cost	sum_costs	match	cost_min
BIRTH_D	Day of Birth	BRTH_D	BRTH_D	8	1	100	B633	B633	0	0	0	11	.	11
BIRTH_M	Month of Birth	BRTH_M	BRTH_M	8	1	100	B635	B635	0	0	0	11	.	11
BIRTH_Y	Year of Birth	BRTH_Y	BRTH_Y	8	1	100	B63	B63	0	0	0	11	.	11
RACEC	Race	RACE	RACE	12	1	10	R2	R22	1	0	2	19	.	19

Suggestions only when costs < threshold (user defined =75)

Suggestions only for variables not yet mapped and found in study

Suggestion could be turned off

Mapping suggestions – Example 2

DOMAIN	VARIABLE	TY	LI	SN 1000	SN 1001	SN 1002
MH	MHNY			MHIS.MHNY, Variable not in metadata , mapping suggestion: MHNYN	Variable does not exist	Variable does not exist
MH	MHNYN	N	NY	MHIS.MHNY mapped to MHNYN according to NY codelist	MH.MHNYN	Set to 1 when medical history is available

variable Variable Name	SASNAME	suggest1	dist	dist2	sound_dist	comp_	dist4_label	type_cost	dist5_label	sum_costs
MHNY	MHNYN	MHNYN	7	1	1	5	8	0	20	42

→ Levenshtein distance=1

→ SPEDIS distance=7 (probability of two phrases matching)

Suggestions only when costs < threshold (user defined =75)

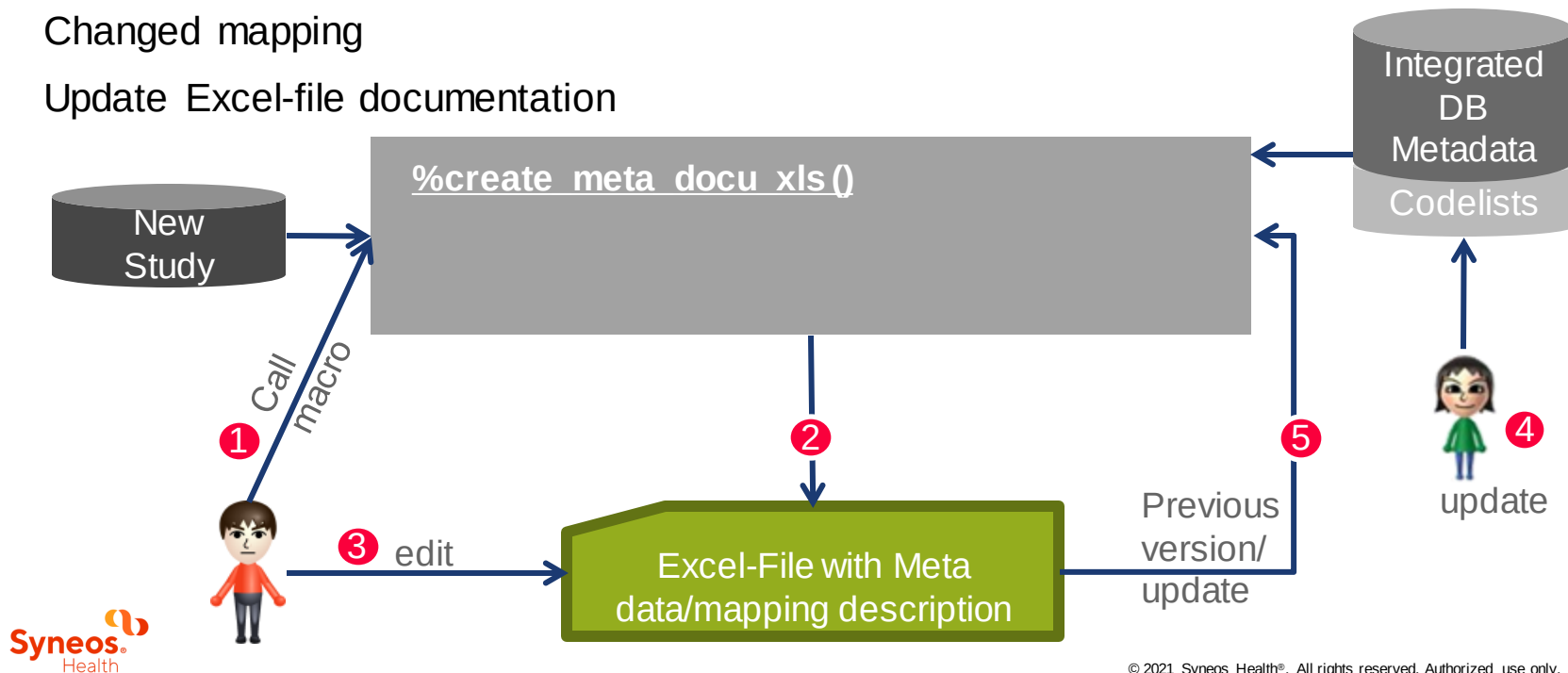
Suggestions only for variables not yet mapped and found in study

Suggestions could be turned off

04 **Add previous description**

Integration of previous Excel documents

- Update of meta data of the IDB
- New study
- Changed mapping
- Update Excel-file documentation



Example: adding new study with %create_meta_docu_xls

Only few things need to be changed!

1	Domains Pool	SN_1000	SN_1001	SN_1002
2	AE	AE	AE	SUPPAE AE
3	CM	COMED	CM	CM
4	DM	assigned	DM	SUPPDM DM
5	DS	DV	DS	DS
6	EX	DOSE	EX	SUPPEX EX
7	LB	LAB	LB	SUPPLB LB
8	MH	MHS	MH	MH
9	SE	assigned	SE	SE
10	SV	VISIT	SV	SV
11	TA	assigned	TA	TA
12	TE	assigned	TE	TE
13	TS	PROTOCOL	TS	TS
14	TV	VISDES	TV	TV
15	VS	assigned	VS	VS

Study
1003

1	Domains Pool	SN_1000	SN_1001	SN_1002	SN_1003
2	AE	AE	AE	SUPPAE AE	AE
3	CM	COMED	CM	CM	CM
4	DM	assigned	DM	SUPPDM DM	DM
5	DS	DV	DS	DS	DS
6	EX	DOSE	EX	SUPPEX EX	EX
7	LB	LAB	LB	SUPPLB LB	LB
8	MH	MHS	MH	MH	MH
9	SE	assigned	SE	SE	SE
10	SV	VISIT	SV	SV	SV
11	TA	assigned	TA	TA	TA
12	TE	assigned	TE	TE	TE
13	TS	PROTOCOL	TS	TS	TS
14	TV	VISDES	TV	TV	TV
15	VS	assigned	VS	VS	VS

```
%create_meta_docu_xls(
    project          = Drugname
    ,libmeta         = meta
    ,studyid         = 1000#1001#1002#1003
    ,pathstudydata   = &source1 #&source2 #&source3#&source4
    ,inbase          = patinfo#dm#dm#dm
    ,domains         = ae cm dm ds ex lb mh se sv vs ta te ts tv vs
    ,pathoutfile     = &outpath./docum
    ,outfilename     = drugname_mapping_documentation_&date._test.xlsx
    ,standard        = SDTM
    ,previousfile    = &path_to_excel_documentation/project drugname_2021-06-10.xlsx
    ,mapto           = 1000: COMED->CM#(PROTOCOL ENDPOINT DV)->DS#DOSE->EX#LAB->LB#MHS->MH#VISIT->SV# PROTOCOL->TS#VISDES->TV
                      /1001:
                      /1002: (SUPPAE AE)->AE#(SUPPDM DM)->DM#(SUPPEX EX)->EX#(SUPPLB LB)->LB
                      /1003: <add mapping if necessary>
);
```

Example: Use manually changed Excel file as input

DOMAIN	VARIABLE	VARNUM	LABEL	TYPE	LENGTH	CODLIST	SN_1000	SN_1001	SN_1002
MH	STUDYID	1	Study Identifier	C	10			MH.STUDYID	MH.STUDYID,Length>meta data definition
MH	DOMAIN	2	Domain Abbreviation	C	2	DOMAIN		MH.DOMAIN	MH.DOMAIN,Codelist is different;\$
MH	USUBJID	3	Unique Subject Identifier	C	27			MH.USUBJID	MH.USUBJID
MH	SUBJIDN	4	Subject Identifier for the Study (N)	N	11			MH.SUBJIDN	
MH	MHSEQ	5	Sequence Number	N	8			MH.MHSEQ	MH.MHSEQ
MH	MHNYN	6	Any Medical History findings (N)	N	3	NY		MH.MHNYN	
MH	MHGRPID	7	Group ID	C	20				

DOMAIN	VARIABLE	VARNUM	LABEL	TYPE	LENGTH	CODLIST	SN_1000	SN_1001	SN_1002
MH	STUDYID	1	Study Identifier	C	10		MHIS.TRIALNO in character format	MH.STUDYID	"1002"
MH	DOMAIN	2	Domain Abbreviation	C	2	DOMAIN	"MH"	MH.DOMAIN	MH.DOMAIN
MH	USUBJID	3	Unique Subject Identifier	C	27		Concatenation of MHIS.TRIALNO and MHIS.PATNO in character format	MH.USUBJID	MH.USUBJID
MH	SUBJIDN	4	Subject Identifier for the Study (N)	N	11		MHIS.PATNO	MH.SUBJIDN	Substring of USUBJID in numeric format (=input((substr(USUBJID,19)),best))
MH	MHSEQ	5	Sequence Number	N	8		MHIS.MHSEQ	MH.MHSEQ	MH.MHSEQ
MH	MHNYN	6	Any Medical History findings (N)	N	3	NY	MHIS.MHNY mapped to NY codelist	MH.MHNYN	Set to 1 when medical history is available
MH	MHGRPID	7	Group ID	C	20		Variable is not populated	Variable is not populated	Variable is not populated

%create_meta_docu_xls

Excel-File with manually changed mapping description



- Macro merges entries in manually entered Excel file with information processed during macro run
- Darker green color indicates if original found content is changed with Excel file

05 Summary

Features of %create_meta_docu_xls

- Processes any source, SDTM or ADaM
- Multiple studies
- Different source domains to one target domain (source1...source N) → target
- Adds information from SUPP-domains
- Adds information about data / variables
- Adds mapping suggestions, variables found
- Processes a previously created Excel document
- Handles updated source / target metadata
- Adds colours to support users
- Could be used as mapping specification (draft) / define.xml / documentation



Thomas Wollseifen
thomas.wollseifen@syneoshealth.com

Shortening the Distance from Lab to Life[®].

