Embracing change:
A journey in SDTM creation

Dennis Jørgensen & Aik Hoe Seah 2018

Agenda

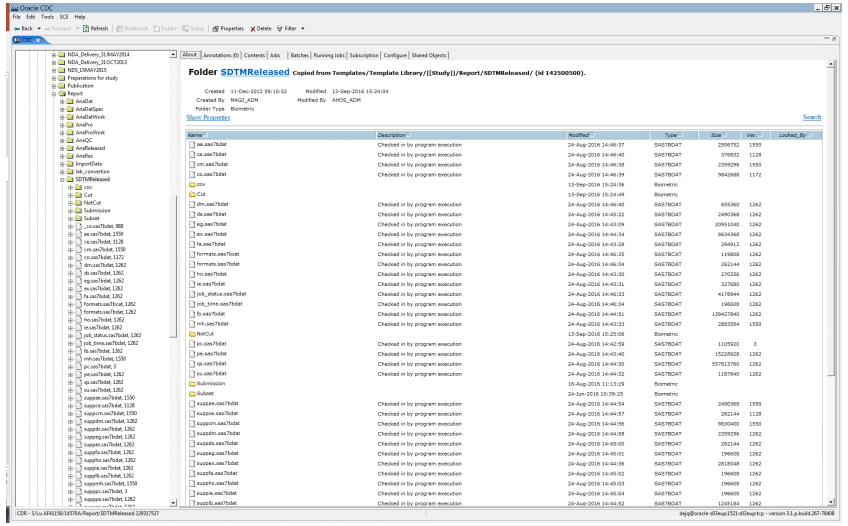
- Background
- Migrating to a new CDR
- From raw data to final output
- The old process
- The new process

Background

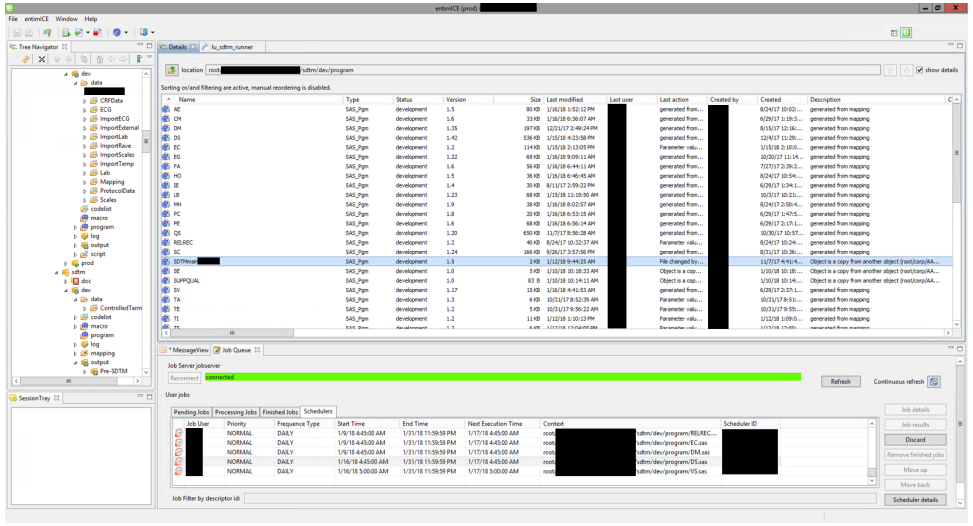
- For several years, Lundbeck had maintained an in-house SDTM creation tool, built solely using SAS and Excel, that handled data from multiple sources, variable mapping, controlled terminology, and advanced transformations. This resulted in tightly coupled system components, great complexity, a steep learning curve and on occasion, problems debugging code.
- The implementation of a new data repository, which includes a built-in ETL tool, offers new possibilities to see if a simpler, less coupled system was possible, and if some of the decoupled tasks could now be performed without SAS and expert system knowledge.

Migrating to a new CDR

- Was previously using a Clinical Data Repository (CDR) solution called Waban / Oracle CDC (Phase Forward / Oracle).
- Migrated to a new CDR solution called entimICE (entimo AG)

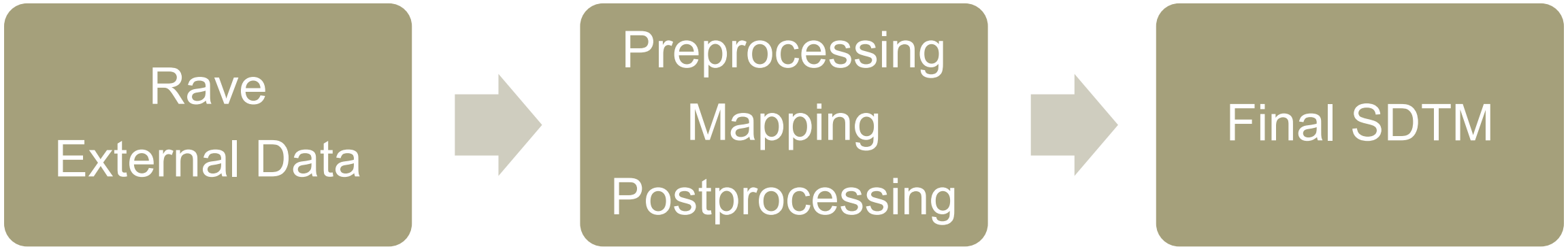


Waban CDR



entimICE CDR

From raw data to final output



The old way: Pre processing

- Prepare data for the mapper
  - Make it look like Rave data
  - Merge key massage (AENO = 2 vs. SAEID = 2 - Headache – 01 Jan 2018)
  - Simple derivations (Screen Fail at V1 and V2? First dose at V3 and WD)
  - Evaluation date = Visit Date?
  - Hardcodings (e.g. lab name)
- Actions registered in Excel (Set & Merge)
- Integrated with Mapping process
  - Debugging

Mapping and controlled terminology

- Mapping of names
  - Map_Log, MappingDictionary_Log, MappingDictionary_Std
  - Excel
- Controlled terminology
  - Events domains: Cat_term_decod
    - Excel
  - Findings domains: CTD_Log, CTDictionary_Log, CTDictionary_Std
    - Excel (20 – 30 columns)
  - SasFormats (codelists)
    - Excel

Postprocessing

- Visit numbers, sequence numbers, Supp domains, format dates…
- Post processing macros
  - By domain (10 - 20 calls)
  - Excel

Where to go from here

Pros:

- Extremely powerful
- High level of standardization possible

Cons:

- Peak Excel (Keep in sync, or else!)
  - Study local: 6 files, 39 sheets
  - Change control: 6 files, 8 sheets
- Steep learning curve, and SAS experience needed – despite Excel
- Hard to parallelize (can I borrow map_log.xls?)
- Cracks in standardization forming

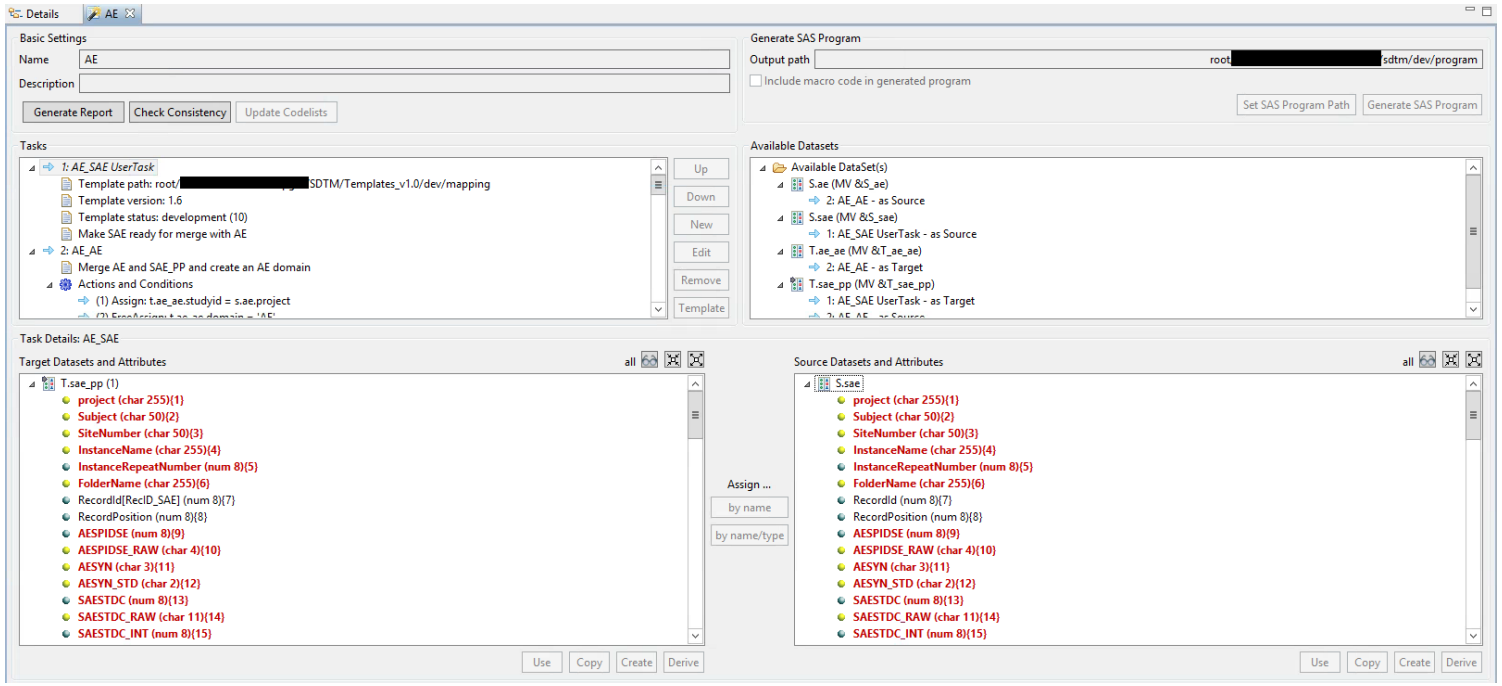
The new way: Pre processing

- Pre processing: anything that needs to be done programmatically before performing the mapping task

- Examples:
  1) Create Trial domains from Protocol Data like Trial Design and SDS
  2) Create SAS Formats and Study Terminology datasets by appending both Lundbeck and Study Terminology
  3) Fixing inconsistent Rave or external data across domains
  4) Deriving First Dose/Last Dose Date for both DM and DS
  5) Converting exceptional non-ISO8601 dates
  6) Transforming data for study-specific sponsor defined SDTM domains

Mapping

- Now a decoupled task which could be performed without SAS and expert system knowledge



Mapping

- Map raw data to PreSDTM dataset using the Mapping Tool
- Standard GLIB forms → Use standard templates for mapping in Entimice
- Study specific forms → Modify the mappings from standard template,
  → refer to CDISC's SDTM Guide
- New data types → refer to CDISC's SDTM Guide, CDISC terminology
  → consult with CDM/SP/BS for best solution
- Make use of CDISC standard terminology for testcd, test, cat, scat
- Add other fields for SUPP domains

Post processing

- Post processing: a set list of tasks done programmatically after mapping

	PROGRAMS	PROGRAM TASK
1	Postprocessing	This is the driver post processing program which calls the individual macros. This runs on all the main domains found in pre-SDTM folder.
2	genAge	This macro derives AGE by subtracting year of birth from year of the reference date. Currently, it uses either of *RFSTDTC*, *RFXSTDTC* or *RFICDTC* as the reference date in the calculation. This runs only for DM domain.
3	genDosTxt	This populates *xxDOSTXT* or *xxDOSE*. If the dose value is in text, it puts it in *xxDOSTXT*. If it is numeric, it puts it in *xxDOSE*. This runs for domains which have both *xxDOSTXT* and *xxDOSE* variables.
4	genReasnd	This sets the *xxREASND* variable to "NOT COLLECTED" if *xxREASND* is missing and the *xxSTAT* is "NOT DONE".
5	genEpoch	This macro derives the *EPOCH*. It gets the distinct list of EPOCH from SE. It then assigns the *EPOCH* by checking *xxSTDTC* (or *xxDTC*) against the *SESTART* of the EPOCH. The EPOCH having the max(*SESTART*) <= *xxSTDTC* (or *xxDTC*) will be the EPOCH that will be assigned to that record. For MH and ZD, if the date variable is not missing, the *EPOCH* variable is set to the first EPOCH. Otherwise, if date variable is missing, the *EPOCH* is also left blank.
6	genCaptureCo	This macro captures the comments variables from all the pre-SDTM datasets. The comments variables are those having _CO at the end of the variable name. This also assigns *RDOMAIN*, *COREF*, *IDVARVAL* and *IDVAR*. It transposes the dataset that will result to all the comments placed in 1 column (*COVAL*) and outputs the result to *&work..COMMENTS_&domain*, which will then later be used by genCO macro.

	PROGRAMS	PROGRAM TASK
7	genCO	This macro outputs the final CO dataset. It appends all the *&work..COMMENTS_&domain*, which are produced by genCaptureCO. It also checks if pre-SDTM CO exists and if it does, it also appends it to the *&work..COMMENTS_&domain* datasets. This also splits the *COVAL* into chunks of 200. It also derives *COSEQ* and outputs final dataset in SDTM data model structure. Note that the final CO dataset is created only when there are comments variables and/or pre-SDTM CO exists. Otherwise, CO dataset is not created.
8	genVisit	This populates *VISITNUM* and *VISITDY* based on TV dataset. *VISIT* should be populated from the mapping for the *VISITNUM* and *VISITDY* to be derived during the postprocessing. If the *VISIT* value is not found in TV, the *VISIT* is set to "Unscheduled". *VISITNUM* for unscheduled visits is also derived in this macro.
9	genUnschVisit	This runs only for SV domain. It takes all the unscheduled visits from the other domains and appends it to SV.
10	genDY	This populates the *xxDY*, *xxSTDY* and/or *xxENDY* variables. *xxDY* value is derived as *xxDTC* - <reference date> + 1; *xxSTDY* and *xxENDY* are derived the same way but uses *xxSTDTC* and *xxENDTC* instead. This currently uses coalescec(*RFSTDTC*, *RFXSTDTC*) as the reference date.
11	genBLFL	This derives the baseline flag (*xxBLFL*). *BLFL* is set to Y for the maximum *xxDTC* that is less than or equal to the *RFXSTDTC*. The *xxDTC* and *RFXSTDTC* values are also converted to datetime, ie., for those dates that do not have the time, the time is imputed to 00:00:00. Currently, this macro does not include checks on the format of the date variables, it assumes that the date variables used have, at the minimum, year, month and day. *xxBLFL* will not be populated if the *xxDTC* values for that subject, for all the records are missing. Alternatively, *xxBLFL* can also be assigned based on the Baseline visit though it must be noted that the baseline visit may not be the same across all domains.

Post processing

	PROGRAMS	PROGRAM TASK
12	genUnitConversion	This macro converts the results from original unit to standard unit. It runs only for domains that have Unit Conversion entries in the Conversion dataset (Conversion tab in Study Terminology).
13	genStdConversion	This converts the original results to standard results. It runs only for domains that have Standard Conversion entries in the Conversion dataset (Conversion tab in Study Terminology). The tests that are having _PARSEVAL_ in the STRESC in the conversion table are expected to have character values only. Therefore, *xxSTRESN* is not derived for these.
14	genSeq	This populates the SEQ variable. The dataset is first sorted by STUDYID, DOMAIN, USUBJID, xxSPID, VISITNUM, VISIT, xxDTC, xxCAT, xxSCAT, XXTESTCD/xxTRT/xxTERM before the SEQ is derived.
15	genEndDomain	This outputs the final SDTM datasets by following the SDTM data model structure found in */root/standards/lundbeck/data_models/SDTM/SDTMIG_v3.2/dev/data*
16	genTDomain	This directly copies the pre-SDTM datasets that do not need post-processing from the pre-SDTM folder to the SDTM folder. Those pre-SDTM datasets are the trial domains (TA, TE, TI, TS, TV) and SE.
17	genSUPP	This program creates the supplemental qualifier domains. This refers to the SUPPQUAL tab in Study Terminology for the list of variables that should be placed in SUPP domains.

Pre-SDTM dataset (LB)

Library LB

Freeze Hide Show... Format Filter... Font... Find

Table View

	DOMAIN	EPOCH	LBASSYN	LBBLFL	LBCAT	LBDRVFL	LBDTC	LBDY	LBELTM	LBENDTC	LBFAST	LBGRPID	LBLOINC	LBMETHOD	LBNAM	LBNRIND	LBORNRHI	LBORNRLO	LBORRES	LBORRESU	LBREASND	LBREFID
1	LB		N		URINALYSIS		2017-07-25					LB_URN:							POSITIVE			
2	LB		N		URINALYSIS		2017-11-01					LB_URN:							POSITIVE			
3	LB		N		URINALYSIS		2017-10-06					LB_URN:							POSITIVE			
4	LB		N		URINALYSIS		2017-10-06					LB_URN:							POSITIVE			
5	LB				URINALYSIS		2017-11-09					LB_URN:							POSITIVE			
6	LB				URINALYSIS		2017-11-09					LB_URN:							POSITIVE			
7	LB		N		URINALYSIS		2017-07-25					S_URIN:							NEGATIVE			

Library LB

Freeze Hide Show... Format Filter... Font... Find

Table View

	LBSTRESC	LBSTRESN	LBSTRESU	LBTEST	LBTESTCD	LBTOX	LBTOXGR	LBTPT	LBTPTNUM	LBTPTREF	STUDYID	USUBJID	VISIT	VISITDY
1				Benzodiazepine	BNZDZPN								Visit 1 (Screening)	
2				Cannabinoids	CANNAB								Visit 1 (Screening)	
3				Benzodiazepine	BNZDZPN								Visit 1 (Screening)	
4				Cannabinoids	CANNAB								Visit 1 (Screening)	
5				Benzodiazepine	BNZDZPN								Withdrawal Visit (1)	

Final SDTM dataset (LB)

Library | LB

Freeze | Hide | Show... | Format | Filter... | Font... | Find

Table | View

	STUDYID	DOMAIN	USUBJID	LBSEQ	LBGRPID	LBREFID	LBSPID	LBTESTCD	LBTEST	LBCAT	LBSCAT	LBORRES	LBORRESU	LBORNRLO	LBORNRHI	LBSTRESC
1	[redacted]	LB	[redacted]	1	S_URIN[redacted]		0	GLUC	Glucose	URINALYSIS		NEGATIVE				NEGATIVE
2		LB		2	S_URIN[redacted]		0	KETONES	Ketones	URINALYSIS		NEGATIVE				NEGATIVE
3		LB		3	S_URIN[redacted]		0	OCCBLD	Occult Blood	URINALYSIS		NEGATIVE				NEGATIVE
4		LB		4	S_URIN[redacted]		0	PROT	Protein	URINALYSIS		NEGATIVE				NEGATIVE
5		LB		1	S_URIN[redacted]		0	GLUC	Glucose	URINALYSIS		NEGATIVE				NEGATIVE
6		LB		2	S_URIN[redacted]		0	KETONES	Ketones	URINALYSIS		NEGATIVE				NEGATIVE
7		LB		3	S_URIN[redacted]		0	OCCBLD	Occult Blood	URINALYSIS		NEGATIVE				NEGATIVE

LBBLFL	LBFAST	LBDRVFL	LBTOX	LBTOXGR	VISITNUM	VISIT	VISITDY	EPOCH	LBDTC	LBENDTC	LBDY
Y					1	Visit 1 (Screening)	-21	SCREENING	2017-12-20		-15
Y					1	Visit 1 (Screening)	-21	SCREENING	2017-12-20		-15
Y					1	Visit 1 (Screening)	-21	SCREENING	2017-12-20		-15
Y					1	Visit 1 (Screening)	-21	SCREENING	2017-12-20		-15
					4	Visit 4 (Week 2)	14	BASELINE 1	2018-01-16		13
					4	Visit 4 (Week 2)	14	BASELINE 1	2018-01-16		13
					4	Visit 4 (Week 2)	14	BASELINE 1	2018-01-16		13
					4	Visit 4 (Week 2)	14	BASELINE 1	2018-01-16		13

Final SDTM dataset (SUPPLB)

Library | SUPPLB

Freeze | Hide | Show... | Format | Filter... | Font... | Find

Table | View

	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
1	[redacted]	LB	[redacted]	LBSEQ	1	LBASSYN	Clincally sgnificant and potential risk?	N	CRF	
2		LB		LBSEQ	2	LBASSYN	Clincally sgnificant and potential risk?	N	CRF	
3		LB		LBSEQ	3	LBASSYN	Clincally sgnificant and potential risk?	N	CRF	
4		LB		LBSEQ	4	LBASSYN	Clincally sgnificant and potential risk?	N	CRF	
5		LB		LBSEQ	1	LBASSYN	Clincally sgnificant and potential risk?	N	CRF	

Lessons learnt

1) Strengthen cooperation between Statistical Programmers & Clinical Data Managers
2) CDM have given feedback that they have gained a better understanding of SDTM and entimICE ☺
3) Templates made mapping tasks easier across studies/compounds
4) Folder structure is clear and simplified compared to before
5) Less Excel files compared to before
6) Flexibility to update post-processing macros as needed by study requirements (previous macros were not easy to modify)

Q&A

Thank you!

Special thanks goes to KGE, TLY, PHJK, PRSN for management support.
SDTM post processing taskforce: AMI, PRYA, VERD, CJON, NODA
CDMs who contributed greatly: JTJ, EVSA, YUEN, KABN, EDSE
SDTM experts no longer with Lundbeck: JSTS, RUKA
The list is long, it was truly a team effort, apologies if we missed anyone!

Name: Dennis Jørgensen
& Aik Hoe Seah
Organization: H. Lundbeck A/S
E-mail: dejq@lundbeck.com
aikh@lundbeck.com