

A quickly way to translate SDTM/ADaM using SAS

Zhixuan Zhao, Yilong Xie, Xiaodong Shi, SIMCERE Zaiming Pharmaceutical Co. Ltd

ABSTRACT

Clinical trials are usually conducted in multiple countries or regions, and new drug applications will be submitted to more than one country. In general, the submission package in accordance with country official language is needed. Chinese is needed for NMPA CDE. Clinical data in the format of XPT, define and corresponding review guide are needed to translated into local official language to meet requirement. Therefore, the paper explored a faster language-translation workflow base on a bilingual control table to submit NDA together with resources saving. The method presented in this paper is in the UTF-8 encoding SAS environment.

INTRODUCTION

Regarding the enormous amount of clinical trial data, translation is a big challenge. The whole workflow is divided into three parts. 1) translation of SDTM and ADaM; 2) translation of statistical analysis reports; 3) translation of define.xml. When translation of SDTM is complete, the same method could be used to translate ADaM and TFLs due to same logic. Therefore, SDTM translation would be split into following sections. 1) remapping MedDRA and WHODrug; 2) translation of dataset descriptions and variable labels; 3) translation of other text. All translation methods and techniques discussed in this article were implemented in the UTF-8 encoding environment.

APPROACH

REMAPPING OF STANDARD MEDICAL DICTIONARY

We know that multilingual MedDRA or WHODrug dictionaries share the same code in the same version. In this part, we will use this character to replace it according to code (such as ptcodes in MedDRA, atccodes in WHODrug), to realize the function of translation. Before performing the substitution, we can integrate the dictionaries of different language versions into a dataset (as shown in Figure 1) to facilitate fast translation between different language versions.

观测	ATCCODE	LEVEL	ATC_EN	ATC_CN
1	A	1	ALIMENTARY TRACT AND METABOLISM	消化和代谢系统药物
2	A01	2	STOMATOLOGICAL PREPARATIONS	口腔病用药
3	A01A	3	STOMATOLOGICAL PREPARATIONS	口腔病用药
4	A01AA	4	CARIES PROPHYLACTIC AGENTS	预防龋齿药物
5	A01AB	4	ANTIINFECTIVES AND ANTISEPTICS FOR LOCAL ORAL TREATMENT	口腔局部用抗感染和抗菌药物
6	A01AC	4	CORTICOSTEROIDS FOR LOCAL ORAL TREATMENT	口腔局部用皮质激素类药物
7	A01AD	4	OTHER AGENTS FOR LOCAL ORAL TREATMENT	其他口腔局部用药物
8	A01W	3	HERBAL STOMATOLOGICAL REMEDIES	治疗口腔疾病的草药
9	A01WA	4	HERBAL STOMATOLOGICAL REMEDIES CONTAINING TANNINS	含有单宁治疗口腔疾病的草药
10	A01WB	4	HERBAL STOMATOLOGICAL REMEDIES CONTAINING MUCILAGE	含有粘液治疗口腔疾病的草药

Figure 1. WHODrug Dictionary Chinese and English Comparison Table

TRANSLATION OF DATASET DESCRIPTIONS AND VARIABLE LABELS

From a technical perspective, we can use "proc datasets" for replacement (see the specific syntax in Figure 2). This is a straightforward method for the individual variable within a single dataset. However, it is worth considering how to perform batch replacements for labels across dozens of datasets and thousands of variables.

The use of loop statements may solve this issue. Before implementing these, we need to prepare a control table on the dataset level and the variable level (such as in Table 1 and Table 2). It contains a comparison of Chinese and English labels for all datasets and variables that require replacement.

This file should cover all necessary information for replacing labels across datasets and variables.

```
1 □ proc datasets noprint;
2     modify adsl(label='受试者水平分析数据集');
3     label SITEID="研究中心标识符";
4 run;
```

Figure 2. Code About Rename SAS Data Description and Variable Label

DOMAIN	LABEL_EN	LABEL_CN
AE	Adverse Events	不良事件
CE	Clinical Events	临床事件
CM	Concomitant/Prior Medications	既往与合并用药
DD	Death Details	死亡详情
DM	Demographics	人口学
DS	Disposition	受试者分布

Table 1. Dataset Description Sheet

DOMAIN	NAME	LABEL_EN	LABEL_CN
AE	STUDYID	Study Identifier	研究标识符
AE	DOMAIN	Domain Abbreviation	域名缩写
AE	USUBJID	Unique Subject Identifier	受试者唯一标识符
AE	AESEQ	Sequence Number	序号
AE	AEGRPID	Group ID	组合 ID
AE	AESPID	Sponsor-Defined Identifier	自定义标识符
AE	AETERM	Reported Term for the Adverse Event	不良事件报告名称

Table 2. Variable Label Sheet

TRANSLATION OF OTHER TEXTS

There is a huge amount of work to translate text except for the parts of standard medical dictionaries and labels. It is consisting of controlled terminology (CT) and free texts. The data management department needs support for the translation of CT from CRF, because CRF has been translated before this process. Free-text translation requires more specialized support from professional departments, such as Medical department.

Here's how we approach it step by step(The flowchart is shown in Figure 3), translating from English to Chinese as an example:

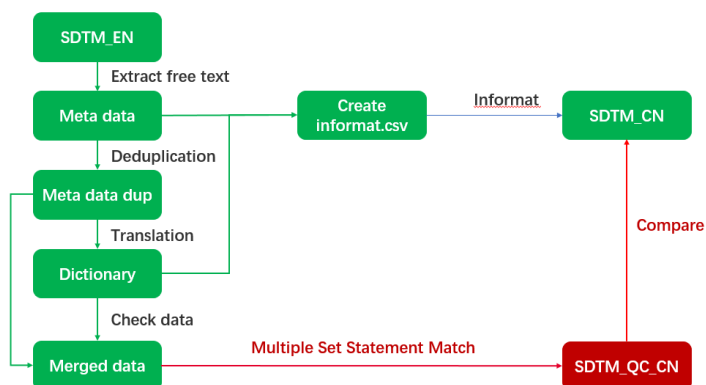


Figure 3. Free Text Translation Flowchart

1. Get text translation metadata. Once completing the SDTM datasets in English version, we start to do SDTM translation. We need to find out the free contexts as well as location which be translated. Hence, when we pull out the metadata information, we also record the “position” in datasets. “position” includes domain name, variable name, variable label, and record row in datasets. Get variable information from SasHELP.VCOLUMN, put them into a loop, rename the variable to “_CONTENT” and drop the other, and use “Proc append” to combine them. The following table is an example of metadata file. We assume that one English sentence has one meaning of Chinese expression, deduplication by “_content” is a good way to make work easier.

_DOMAIN	_NAME	_LABEL	_ORDER	_CONTENT
AE	AETERM	Reported Term for the Adverse Event	169	nausea
...	...	...	...	...

Table 3. Text Translation Metadata

2. Establish a Translation Dictionary. We send metadata to our medical colleagues for translation. They will check the correctness of CT and translate free texts. Once we receive the translation results, we organize them into a table called “Translation Dictionary”. It’s the main core of language switching in SDTM. Hence, we need to make sure the mapping relationship between two versions in different languages is correct. It is better to make a version control on this dictionary. There is one advantage of a translation dictionary. When we go to another project, we could use this dictionary to make a pre-translation to improve the efficiency of translation.

EN_VER	CH_VER	SOURCE
nausea	[Here is the Chinese expression of “nausea”]	[Project number]
...	...	...

Table 4. Translation Dictionary

3. Check no sentence is missed. We merge the translation dictionary to the text translation metadata or the deduplicated one. (we call this merged data “Merged Data” here for convenience) If the Chinese expression (“CH_VER”) is missing after merging. It represents the mapping is failed and you need to do additional work to fix it. Usually, some special characters and newline characters will cause this problem. So far, all the preparations are done. We can go into the last step of SDTM translation. We perform two different methods, one for datasets production and another for quality control, to translate the language.
4. Use “informat” to translate STDM. We create an “informat” from a CNTLIN= data set. Then, we perform an “informat” processing on each variable. By following steps (see Figure 4 and Figure 5), we can translate other texts. However, how to effectively and efficiently perform batch translations

remains a challenging question.

DOMAIN	FMNAME	VARNAME	START	END	LABEL
AE	AEACN	AEACN	DOSE NOT CHANGED	DOSE NOT CHANGED	无变化
AE	AEACN	AEACN	DRUG WITHDRAWN	DRUG WITHDRAWN	停药
AE	AEACN	AEACN	NOT APPLICABLE	NOT APPLICABLE	不适用
AE	AEACNTH	AEACNTH	CONCOMITANT MEDICATION	CONCOMITANT MEDICATION	合并用药
AE	AEACNTH	AEACNTH	MULTIPLE	MULTIPLE	多种
AE	AEACNTH	AEACNTH	NONE	NONE	无
AE	AEACNTH	AEACNTH	OTHER: L'Occitane moisturizing cream on top	OTHER: L'Occitane moisturizing cream on top	外涂欧舒丹润肤凝霜
AE	AEACNTH	AEACNTH	OTHER: Oxygen therapy	OTHER: Oxygen therapy	其他: 氧疗
AE	AEACNTH	AEACNTH	OTHER: hot compress	OTHER: hot compress	其他: 热敷

Figure 4. Informat Excel Example Sheet

Name	Size	Type
 Ae_aeacn	0.4KB	Infmtc
 Ae_aeacnoth	0.6KB	Infmtc
 Ae_aeenrtpt	0.3KB	Infmtc
 Ae_aeout	0.5KB	Infmtc

Figure 5. Result Of Informat

- Independent programming to make quality control. Another approach is that we can assign values using the "Tranwrdr" function or using an equal sign. We use multiple set statements to match data from "merged data" to SDTM and save the corresponding "CH_VER" when the sentence we need to translate is the same as the "EN_VER" variable in "Merged data" in the subset statement. A loop will be executed for domain and variable one by one, and when the domain and variable exist in "merged data", the steps mentioned in above will be activated.

Overall, free texts translation requires more comprehensive and complicated process, it can be accomplished effectively through well designed planning and execution.

TRANSLATION OF DEFINE.XML

Define-XML transmits metadata that describes any tabular dataset structure. When used with the CDISC Foundational Standards, it provides the metadata for human and animal model datasets using the SDTM and/or SEND standards and analysis datasets using ADaM.

Define-XML is required by the NMPA, FDA, and PMDA for every study in each electronic submission to inform the regulators which datasets, variables, controlled terms, and other specified metadata were used.

Pinnacle21 is currently a free, open-source software for the production of define files in the world. One of its main functions is that it can convert Excel and define.xml files to and from each other, which provides us with ideas for translation. We can translate the Excel Specification (see Table 5) that created the define.xml, thereby indirectly implementing the translation of define.xml.

DATASET	VARIABLE	LABEL	DATA TYPE
DI	STUDYID	Study Identifier	text
DI	DOMAIN	Domain Abbreviation	text
DI	SPDEVID	Sponsor Device Identifier	text
DI	DISEQ	Sequence Number	integer
DI	DIPARMCD	Device Identifier Element Short Name	text
DI	DIPARM	Device Identifier Element Name	text
DI	DIVAL	Device Identifier Element Value	text

Table 5. Excel Specification-Variables sheet

NOTE

The focus and difficulties of translation work are not limited to technical implementation, but also the following points require special attention:

1. Preparation of Chinese-English comparative documents (Excel and others)。
2. How to ensure that the translated version of the CRF is consistent with the translation of the dataset。
3. The translated define.xml corresponds to the preparation of the companion file XSL.

CONCLUSION

The translation of the database not only saves the economic cost for the company but also saves more time cost for the promotion of the project. This work will not be done only once, and it may occur in many cases, such as license in/out, pooling, and NDA submission in multiple countries or regions. When the translation framework and workflow are established, we are more relaxed when face with language-switching task.

REFERENCES

1. SAS Institute Inc. 2023. "Creating an Informat from a CNTLIN= Data Set." May 1, 2023. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/proc/n09a6xpr9nj0brn1089gge2hdcun.htm
2. CDISC 2019. "Define-XML v2.1." 8 May 2019. <https://www.cdisc.org/standards/data-exchange/define-xml/define-xml-v2-1>
3. pinnacle21. "DEFINE.XML." <https://www.pinnacle21.com/products/define-xml>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Zhixuan Zhao
SIMCERE Zaiming Pharmaceutical Co. Ltd
+86-18821202053
Zhaozhixuan@zaiming.com
xieyilong@zaiming.com
xiaodong.shi@zaiming.com