



Filtering non-ASCII characters for Submission Packages using CHECII

Lihao Wang

Jiangsu Hengrui Pharmaceuticals Co., Ltd.

Background

Internationalization:

- As we can see from the bar chart on the right, the pharmaceutical manufacturing market will become larger and larger. Internationalization is the best way to expand the scope of markets and gain more resources. And it's also the best way to improve brand influence.
- On the other hand, in “The 14th Five-Year Plan”, it clearly states six objectives for the pharmaceutical industry. Among them, the sixth says we should accelerate the overall process of the internationalization. As one of the biggest local pharmaceutical companies in China, Hengrui has already started to do so.
- But, what's the relationship between internationalization and CHECII?



Reason for CHECII

- Internationalization means submission in different areas such as EU, JPN, US. Each medical agency has its own submission requirements.
- Let's take FDA submission as an example:
 - In the FDA “STUDY DATA TECHNICAL CONFORMANCE GUIDE”, it states that dataset using UTF-8 encoding is not guaranteed to have no error code. Therefore, under their environment, submission packages with no non-ASCII character would be the safest. And it should be the same for other type of files.

Contains Nonbinding Recommendations



3.3.5 Special Characters: Variables and Datasets

Variable names, as well as variable and dataset labels should include American Standard Code for Information Interchange (ASCII) text codes only. Variable values are the most broadly compatible with software and operating systems when they are restricted to ASCII text codes (printable values below 128). Use UTF-8 for extending character sets; however, the use of extended mappings is not recommended. Transcoding errors, variable length errors, and lack of software support for multi byte UTF-8 encodings can result in incorrect character display and variable value truncations. Ensure that LBSTRESC and controlled terminology extensions in LBTEST do not contain byte values 160-191 as some character mappings in that range may interfere with agency processes.

Rule ID	Publisher ID	Message	Description	Domains	FDA
SD1029	TCG 3.3.5	Non-ASCII or non-printable characters in variable	Variables value must not include non-ASCII or non-printable characters (outside of 32-126 ASCII code range), limited to variables which values may be converted into new variable name or label (~TEST, ~TESTCD, ~PARM, ~PARMOD, QLABEL, QNAM).	FINDINGS, FINDINGS ABOUT, RELREC, SUPPQUAL, TS	X

Reason for CHECII

ASCII control characters			ASCII printable characters					
00	NULL	(Null character)	32	space	64	@	96	`
01	SOH	(Start of Header)	33	!	65	A	97	a
02	STX	(Start of Text)	34	"	66	B	98	b
03	ETX	(End of Text)	35	#	67	C	99	c
04	EOT	(End of Trans.)	36	\$	68	D	100	d
05	ENQ	(Enquiry)	37	%	69	E	101	e
06	ACK	(Acknowledgement)	38	&	70	F	102	f
07	BEL	(Bell)	39	'	71	G	103	g
08	BS	(Backspace)	40	(	72	H	104	h
09	HT	(Horizontal Tab)	41	)	73	I	105	i
10	LF	(Line feed)	42	*	74	J	106	j
11	VT	(Vertical Tab)	43	+	75	K	107	k
12	FF	(Form feed)	44	,	76	L	108	l
13	CR	(Carriage return)	45	-	77	M	109	m
14	SO	(Shift Out)	46	.	78	N	110	n
15	SI	(Shift In)	47	/	79	O	111	o
16	DLE	(Data link escape)	48	0	80	P	112	p
17	DC1	(Device control 1)	49	1	81	Q	113	q
18	DC2	(Device control 2)	50	2	82	R	114	r
19	DC3	(Device control 3)	51	3	83	S	115	s
20	DC4	(Device control 4)	52	4	84	T	116	t
21	NAK	(Negative acknowl.)	53	5	85	U	117	u
22	SYN	(Synchronous idle)	54	6	86	V	118	v
23	ETB	(End of trans. block)	55	7	87	W	119	w
24	CAN	(Cancel)	56	8	88	X	120	x
25	EM	(End of medium)	57	9	89	Y	121	y
26	SUB	(Substitute)	58	:	90	Z	122	z
27	ESC	(Escape)	59	;	91	[	123	{
28	FS	(File separator)	60	<	92	\	124	
29	GS	(Group separator)	61	=	93	]	125	}
30	RS	(Record separator)	62	>	94	^	126	~
31	US	(Unit separator)	63	?	95	_		
127	DEL	(Delete)						

- What is ASCII characters? It stands for the "American Standard Code for Information Interchange".

This is the table for a total number of 128 ASCII characters. Usually, we may only include #32-#126 characters (printable values). Any character other than that should be excluded. And this is what CHECII is doing.

	A	B
1	Single-Byte	Double-Byte
2	a b	a b
3	a;	a;
4	a, b	a, b

- Dummy-Study-15.xlsx

	E	F	G	H
5996				+++++

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Sheet	Row	Column	nonASCII1	nonASCII2	nonASCII3	nonASCII4	nonASCII5	nonASCII6	nonASCII7	nonASCII8	nonASCII9	Num
2	Study-6	5996	H	升	白	、	升	中	性	粒	治	疗	9
	...	Dummy-Study-2.xlsx		Dummy-Study-3.xlsx		Dummy-Study-4.xlsx		Dummy-Study-5.xlsx		Dummy-Study-6.xlsx		Dumr	

Reason for CHECII

- Generally speaking, there would be three parts of files related to the submission package. CHECII can filter any number of following files simultaneously by just one click, which can save a lot labor.



Text files (including .sas, .xml, .txt) and SAS datasets (including .sas7bdat and .xpt).



Microsoft files (including .xlsx, .xls, .rtf and .docx). Many files are stored in this type, e.g. SDTM/ADaM specification, codelist, translation files, TFL outputs, ...



PDF files. Some documents are required in pdf format.

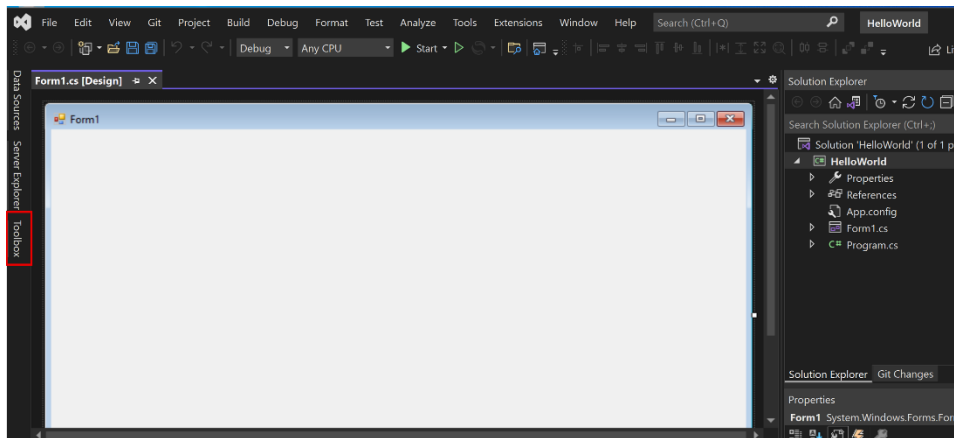
Although we want our final files prepared with no non-ASCII character, its not realistic to examine that after everything is done. Because it may cause chain effects, which would increase the overall workload. Therefore, quality control is necessary, and CEHCII will be frequently used.

CHECII

- Now let's get back to CHECII itself:
 - This small tool is designed using C#. It is a modern, object-oriented, and type-safe programming language. C# enables developers to build many types of secure and robust applications that run in .NET. C# has its roots in the C family of languages and will be immediately familiar to C, C++, Java, and JavaScript programmers.
 - Windows Forms is a Graphical User Interface(GUI) class library which is bundled in .NET Framework. Its main purpose is to provide an easier interface to develop the applications for desktop, tablet, PCs. It is also termed as the WinForms.

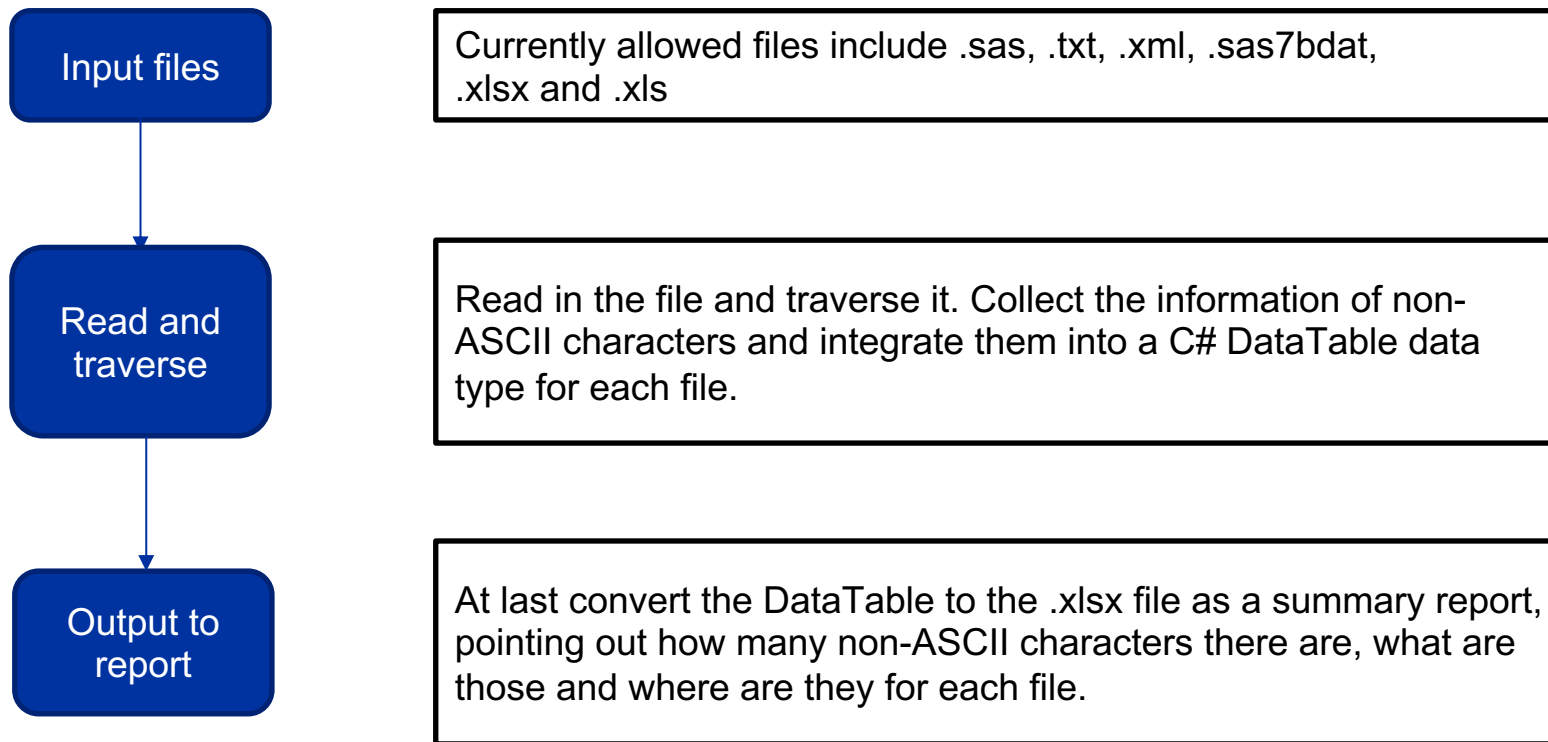


.NET Framework 4.8



Structure of CHECII

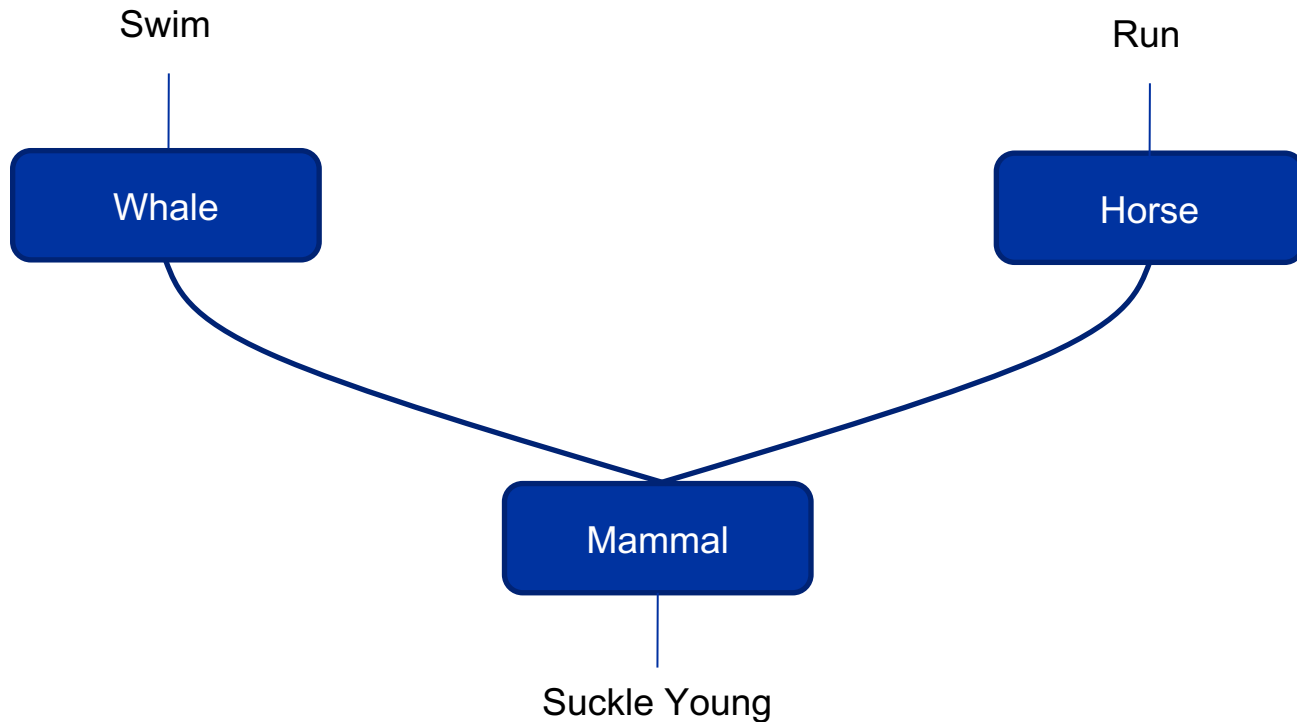
- The overall logic is not that complicated:



Structure of CHECII

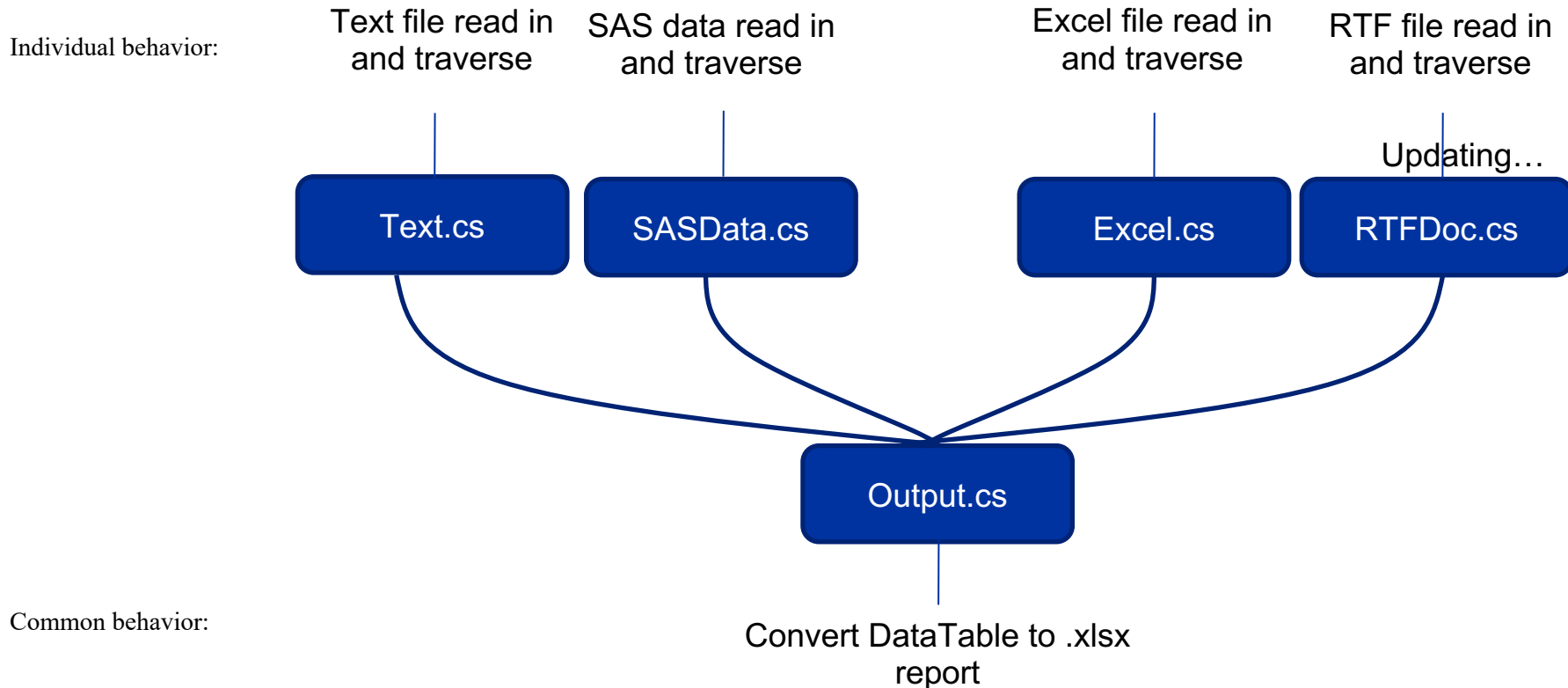
- Inheritance structure:

Individual behavior:



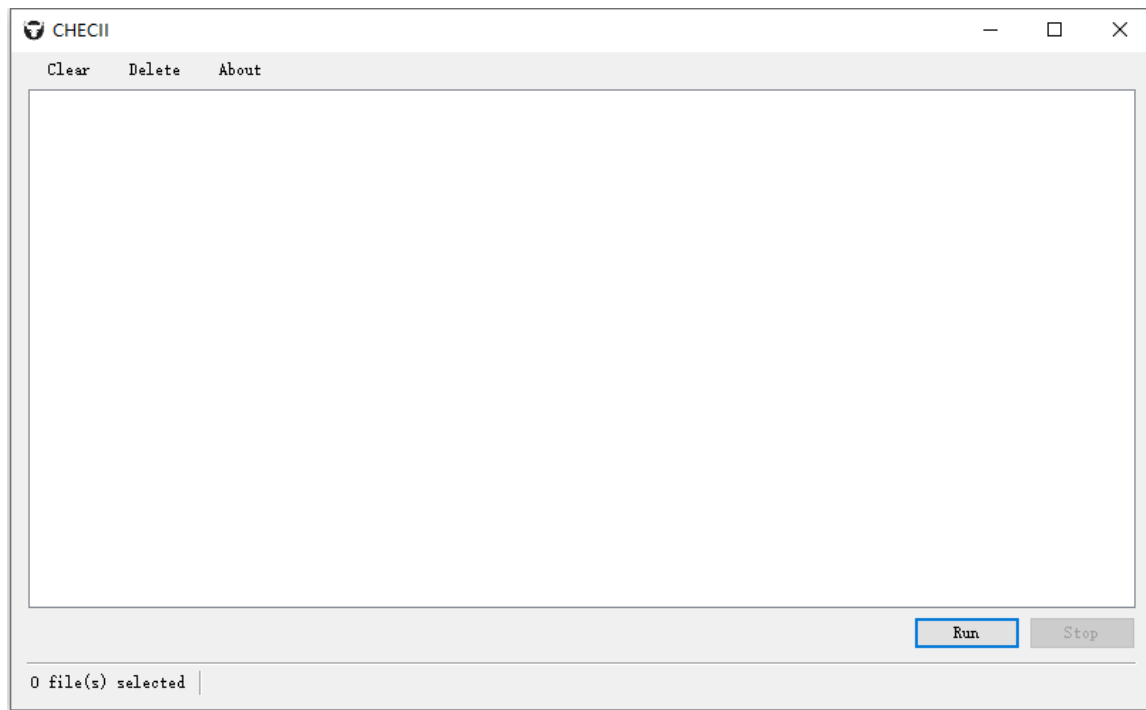
Common behavior:

Structure of CHECII



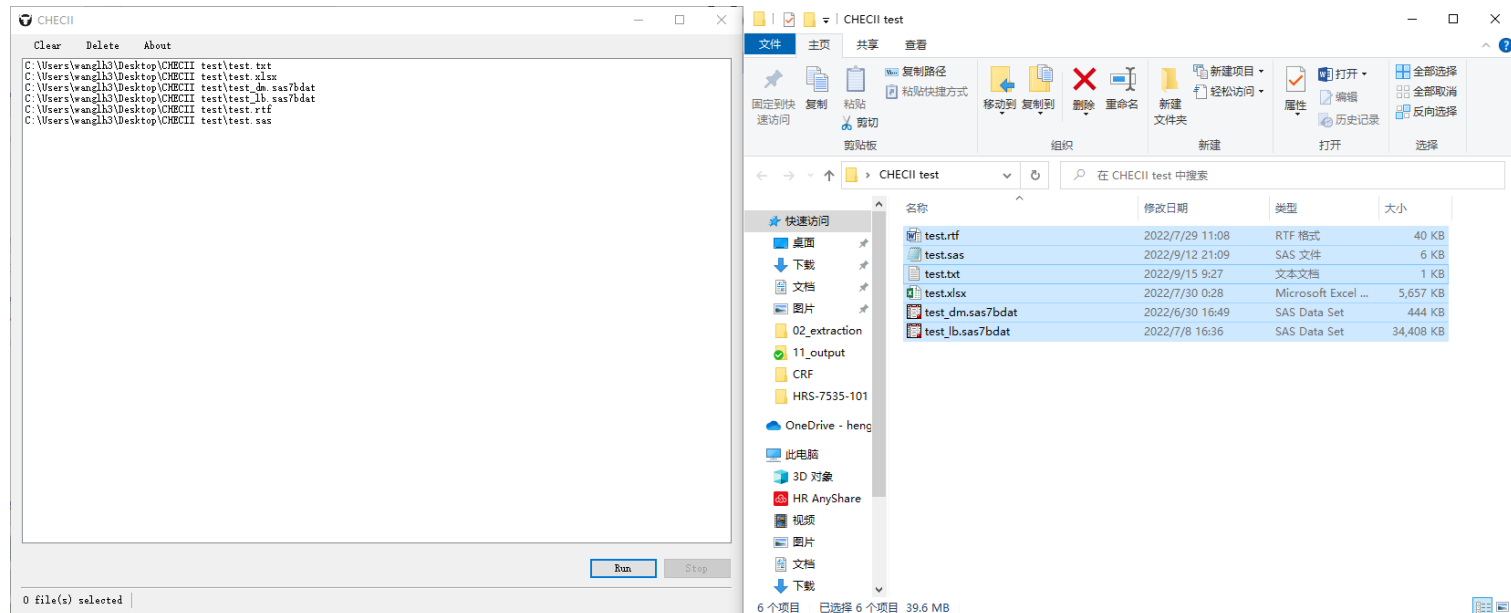
CHECII

- This is the UI of CHECII. Next, I will show you how to do batch checking for different types of files.



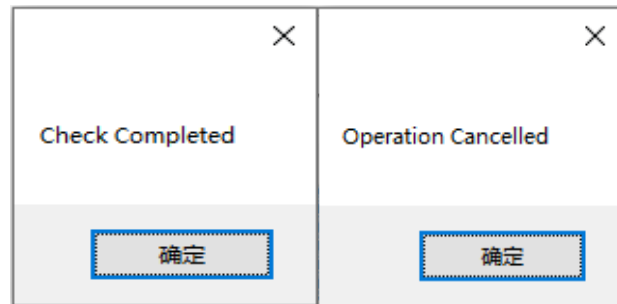
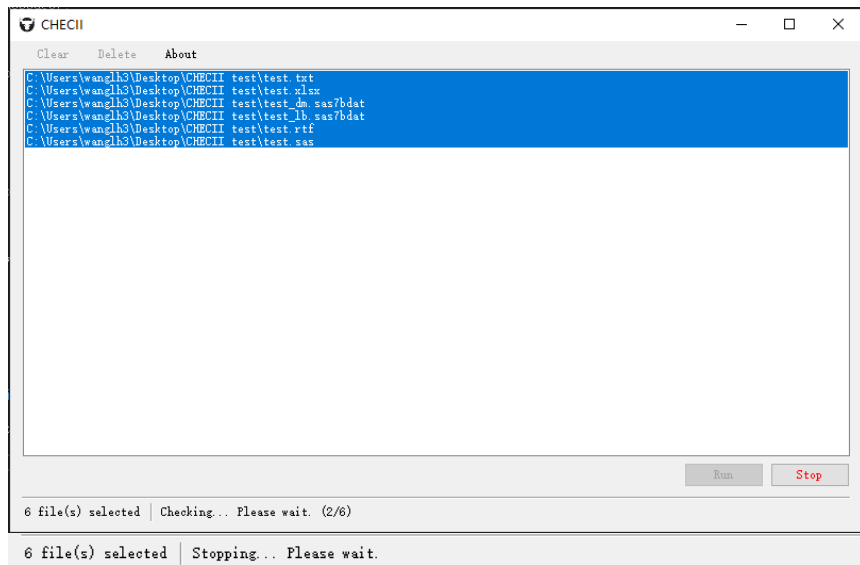
CHECII

- By selecting all the files you want, and dragging them into the box with left click released, you can put those files in CHECII.
- Then you can select multiple items using Ctrl/Shift. Delete the selected items using “Delete” button. Or Clear all the items using “Clear”.



CHECII

- How many items selected is counted at the bottom left corner after your selection.
- After clicking the “Run” button, it will remind you that the check is being run and which item is currently being checked also at the left bottom corner.
- A “Check Completed” will pop up once completed. When you press “Stop” button, it will remind you that the check is stopping and “Operation Cancelled” will pop up once stopped.



CHECII

- Several test files to simulate the possible situations:
- (1) test.txt: A txt file with all ASCII characters. Used to test the situation where there is no non-ASCII character in the file.

No non-Ascii character in this file.

- (2) test.sas: An AE dataset program file from a certain project. Used to test normal program files.

1	Row	Column	nonASCII	Num
2	11	11		1
3	44	4		1
4	47	1		1
5	48	1		1
6	48	2		1
7	48	22	心	1
8	48	23	率	1

CHECII

- (3) test.rtf: A rich text file, used to test the situation for inputting not currently allowed files.

Sorry, this file type is unavailable to check now.

- (4) test.xlsx: An Excel 2007 file more than 5Mb. Used to test the processing time of reading large .xlsx files.

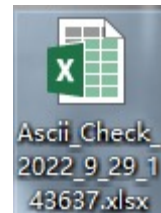
Sheet	Row	Column	nonASCII1	nonASCII2	nonASCII3	nonASCII4	nonASCII5	nonASCII6	nonASCII7	nonASCII8	Num
Sheet1	1	B	...	...							2
Sheet1	1	C	¥								1
Sheet1	1	E	"	"	是	的					4
Sheet1	2	B	对	方	是	个		s	d	f	8
Sheet2	1	B	...	...							2
Sheet2	1	C	¥								1
Sheet2	1	E	"	"	是	的					4
Sheet2	1	M	是								1
Sheet2	2	B	对	方	是	个		s	d	f	8

CHECII

- (5) test_dm.sas7bdat: A DM dataset from a certain project. Used to test normal SDTM dataset with non-ASCII characters.
- (6) test_lb.sas7bdat: A large LB dataset from a certain project, which is more than 33 Mb. Used to test the situation where large SAS datasets are inputted and lots of non-ASCII characters are outputted.

	A	B	C	D	E	F	G
1	Row	Column	nonASCII1	nonASCII2	nonASCII3	nonASCII4	Num
2	1	STUDYID	II				1
3	1	USUBJID	II				1
4	1	INVTNAM					3
5	1	AGEU	岁				1
6	1	SEX	男				1
7	1	RACE	亚	裔			2
8	1	ETHNIC	汉	族			2
9	1	ARMNRS	筛	选	失	败	4
10	1	COUNTRY	中	国			2

- The output is an xlsx file named Ascii_Check followed by the current time stamp, which would appear on your desktop.
- The first sheet would be a summary page reporting the total non-ASCII character number, with hyperlink attached in the next column for ease of jumping to corresponding sheet for more details.

[illegible]



Summary and Future Plan

- This is a small tool designed to simplify the job of filtering non-ASCII characters in various types of files. So that labor and time can be saved, and to make sure there is no simple character errors in the submission.
- Future development of this tool:
 - Enabling more types of files to be checked. For example, .rtf files and .docx files are still under developing. And we may also add .pdf files into it.
 - More types of encoding other than ASCII may be needed. For example, in the latest submission requirement of CDE specification of encoding is needed. We may need to check the encoding we specified ourselves.
 - The report output format may still need to be improved. For example, non-printable characters are hard to tell from those printable. The only way to discern that is by counting according to the “Num” column.

（四）数据审阅说明

为了帮助审评人员更好地理解与使用递交的数据，鼓励申办方递交数据审阅说明。数据审阅说明是对数据说明文件的进一步补充，其内容包括但不限于研究数据使用说明、临床总结报告与数据之间的关系、研究文档（如试验方案、统计分析计划、临床总结报告等）中部分关键信息、所递交程序代码的使用说明、数据集所用编码（如 utf-8、euc-cn 等）及其它特殊情形说明等。数据审阅说明并不旨在取代数据库的数据说明文件，而是通过文档描述的方式来帮助审评人员更准确、高效的理解与使用所递交的数据库、相关术语、程序代码及数据说明文件信息等。数据审阅说明应采用 PDF 文件。

Reference

- The 14th Five-Year Plan for Pharmaceutical Industry
https://www.miit.gov.cn/jgsj/ghs/zlygh/art/2022/art_5d5e4f4a945346c7ab261a9fd2669cb5.html
- FDA “STUDY DATA TECHNICAL CONFORMANCE GUIDE”
<https://www.fda.gov/media/88173/download>
- <https://github.com/dotnetcore/NPOI>
- https://blog.csdn.net/IT_rookie_newbie/article/details/120300980X
- <https://communities.sas.com/t5/Developers/Fastest-way-to-read-SAS-table-in-C/td-p/617724>
- <https://qa.1r1g.com/sf/ask/3675047701/>
- <https://blog.csdn.net/srbhll/article/details/109410353>

Thanks for watching!

Q&A

