

KEYRUS
biopharma



insight into value

Stéphane Chollet

stephane.chollet@keyrus.com

Manager Innovation, Tools & Applications

INNOVATIVE CONTRACT
RESEARCH ORGANIZATION

SD01 - KEYDAN, A CLINICAL DATA DE-IDENTIFICATION SOLUTION



Edinburgh 2017
"Digital Innovation in Healthcare"



AGENDA

✓ De-identification

- *Regulatory key dates*
- *Process*

✓ KeyDAN presentation

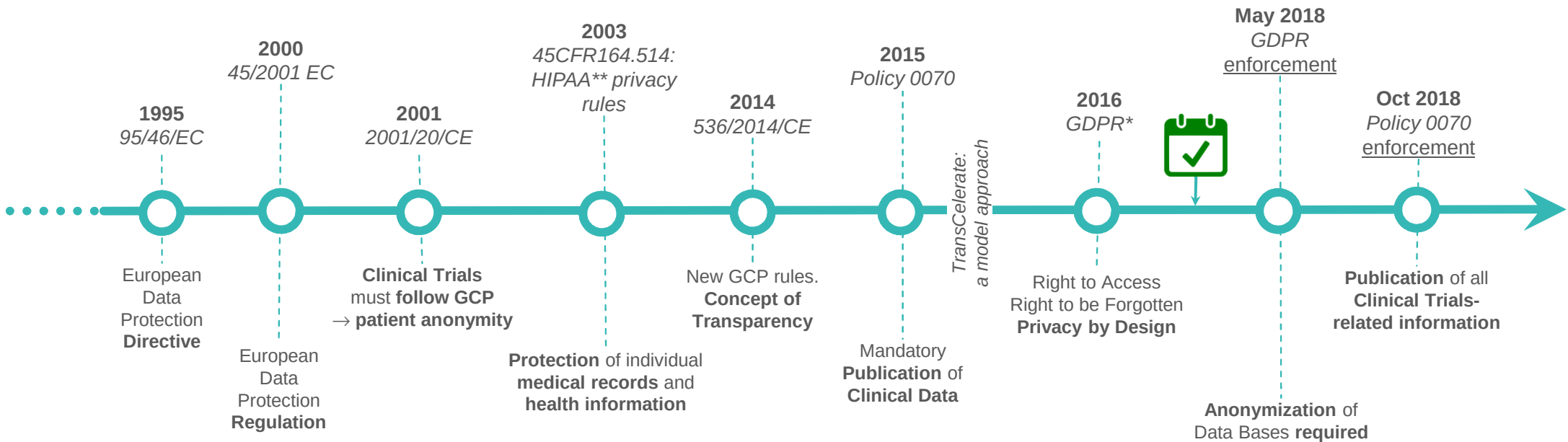
- *Functionalities*
- *Inputs / Outputs*

✓ KeyDAN demo

- *Full process with a legacy database*
- *Specific uses (multiple sources & format, customization, using standard)*



REGULATORY KEY DATES



Increased regulation of

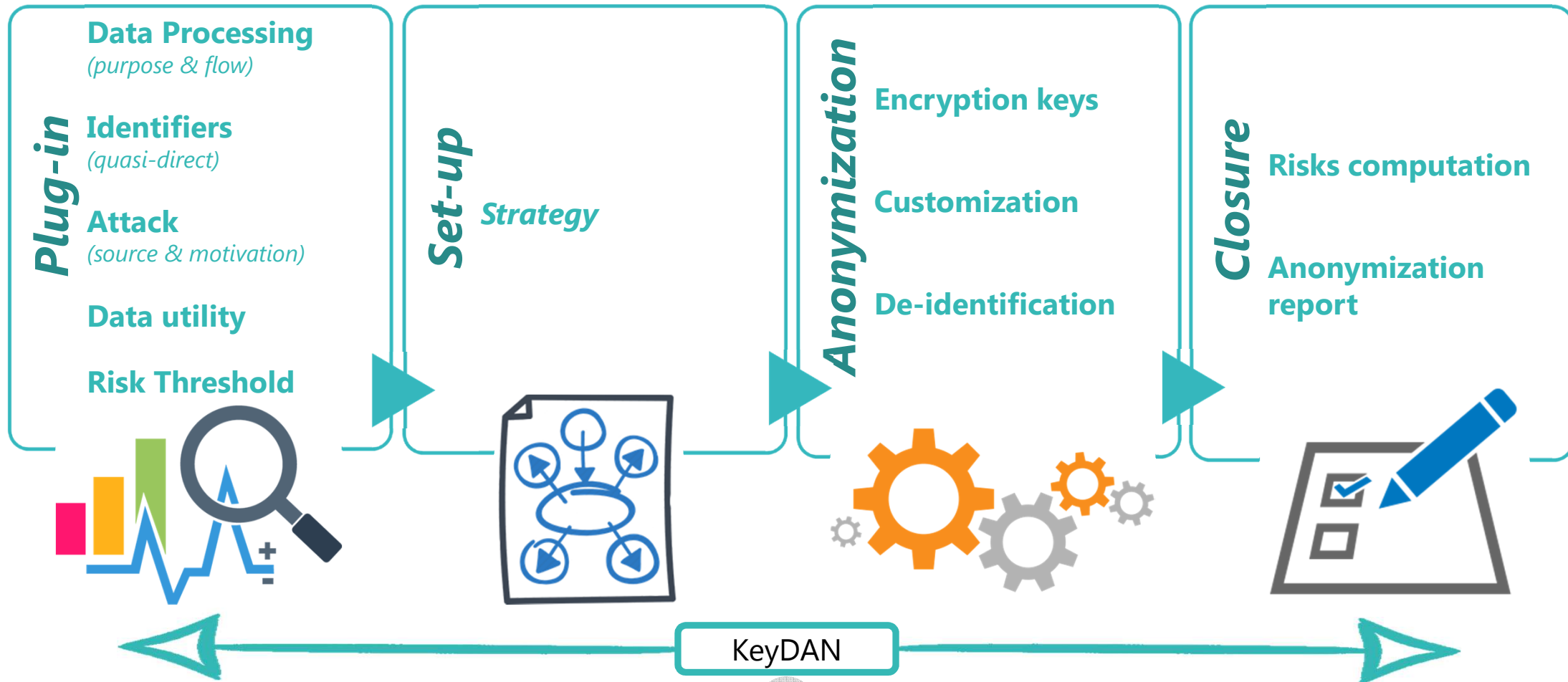
- Subject privacy
- Data security
- Transparency

*General Data Protection Regulation

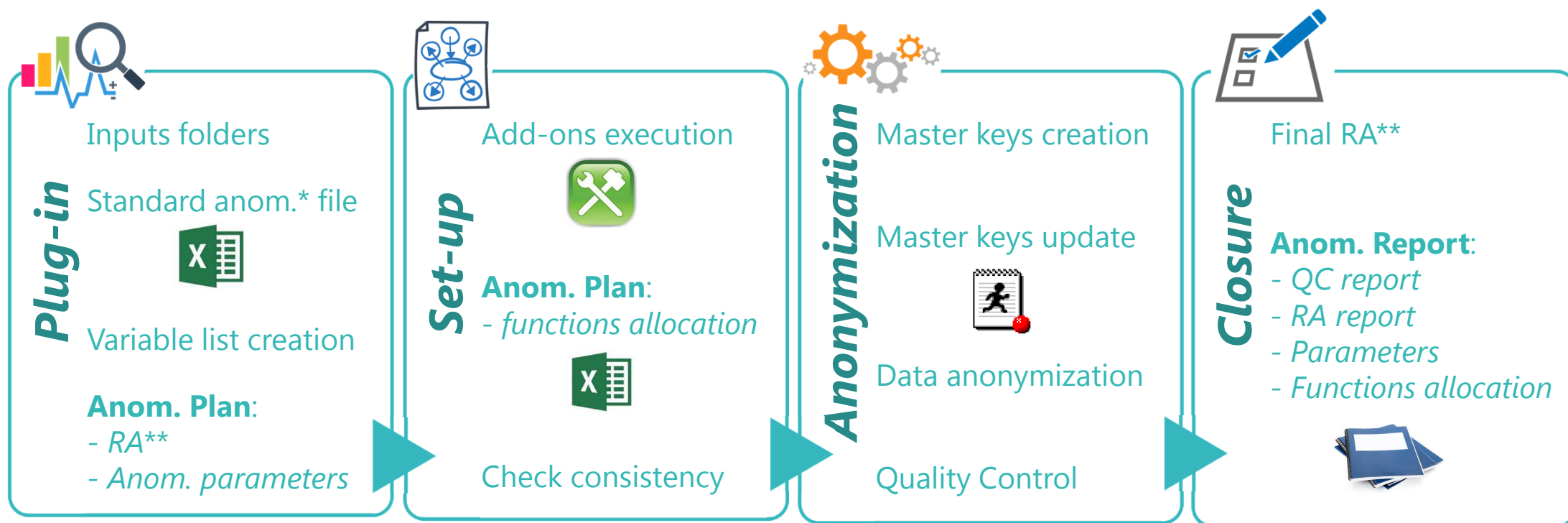
**Health Insurance Portability and Accountability Act

© Copyright 2017 – Keyrus

DE-IDENTIFICATION



KEYDAN PRESENTATION



* ANOnyMization

** Risk Assessment

© Copyright 2017 – Keyrus

STANDARD INSTRUCTIONS



- Use **global standard** (e.g. Phuse one)
- Use **previous** file as standard

	A	B	C	D	E	L	M	N	O
1	Seq_For	Observation	Domain	Variable_Name	Variable_Label	Direct_Quasi_Identifi	DI_Primary_Rule	DI_Alternative_Rules	DI_Comment
28	828	Events-General		--OCCUR	Occurrence				
29	829	Events-General		--STAT	Completion Status				
30	830	Events-General		--REASND	Reason Not Done	Direct	Review and only redact values with personal information	Remove	Importance of the variable depends on the context. This information is not recoded anywhere else and can have an important data utility value for certain analyses.
31	831	Events-General		--BODSYS	Body System or Organ Class	Quasi Level 2	Keep		Only serious and/or visible events should be considered here for further de-identification. Generalization on other variables (e.g. Geography) would be needed.
32	832	Events-General		--BDSYCD	Body System or Organ Class Code	Quasi Level 2	Keep		Only serious and/or visible events should be considered here for further de-identification. Generalization on other variables (e.g. Geography) would be needed.
33	833	Events-General		--SOC	Primary System Organ Class	Quasi Level 2	Keep		Only serious and/or visible events should be considered here for further de-identification. Generalization on other variables (e.g. Geography) would be needed.
34	834	Events-General		--SOCCD	Primary System Organ Class Code	Quasi Level 2	Keep		Only serious and/or visible events should be considered here for further de-identification. Generalization on other variables (e.g. Geography) would be needed.
35	835	Events-General		--LOC	Location of Event		Review and only redact values with personal information		When free text value is entered in "other" value, it should be reviewed and recoded if required during the trial conduct. This is expected to take place during the data cleaning process. This variable is flagged in here as a warning that it may contain PII and if so, they must be redacted.
36				--LAT	Laterality		Review and only redact		When free text value is entered in "other" value, it should be reviewed and recoded if

Phuse SDTMIG.xlsx



	B	C	D	H	K	L	M	N	O
1	MEMNAME	NAME	LABEL	FUNCTION	Direct_Quasi	DI_Primary_Rule	DI_Alternative_Rules	Flag	Comment
71	CE	CESTAT	Completion Status		(vide)	(vide)	(vide)	0	
72	CE	CESTDTC	Start Date/Time of Clinical Event	DATE	Quasi Level 2	Offset	(vide)	1	
73	CE	CESTRF	Start Relative to Reference Period		(vide)	(vide)	(vide)	0	
74	CE	CESTRPT	Start Relative to Reference Time Point		(vide)	(vide)	(vide)	0	
75	CE	CESTPT	Start Reference Time Point	DATE	Quasi Level 2	Offset	(vide)	1	
76	CE	CETERM	Reported Term for the Clinical Event	BLANK	Direct	Remove	Review and only redact values with personal information	2	
77	CE	DOMAIN	Domain Abbreviation		(vide)	(vide)	(vide)	0	
78	CE	STUDYID	Study Identifier		(vide)	(vide)	(vide)	0	
79	CE	USUBJID	Unique Subject Identifier	SUBJ_KEY	Direct	Recode subject ID	(vide)	1	
80	CM	CMCAT	Category for Medication		Quasi Level 2	Keep	(vide)	1	No Action
81	CM	CMCLAS	Medication Class		(vide)	(vide)	(vide)	0	
82	CM	CMCLASCD	Medication Class Code		(vide)	(vide)	(vide)	0	
83	CM	CMDECOD	Standardized Medication Name		(vide)	(vide)	(vide)	0	
84	CM	CMDOSE	Dose per Administration		(vide)	(vide)	(vide)	0	
85	CM	CMDOSFRM	Dose Form		(vide)	Review and only redact values with personal information	(vide)	1	Specific SAS code
86	CM	CMDOSFRQ	Dosing Frequency per Interval		(vide)	Review and only redact values with personal information	(vide)	1	Specific SAS code
87			Intended Dose Regimen		Direct	Review and only redact values with personal information	Remove	2	Specific SAS code

KeyDAN standard



ANONYMIZATION PARAMETERS & DATE FORMATS

	A	B	C	D	E
	Lib_src : libname of the SAS library of source datasets. Source : directory where the source SAS datasets are provided Lib_tgt : libname of the SAS library of anonymized datasets. Target : directory where anonymized datasets will be generated Reference (optional) : path of the reference template to retrieve for a specified source to generate the define_functions sheet.				
1					
2	Lib_src	Source	Lib_tgt	Target	Reference
3	INPUT1	C:\Users\Stephane.Chollet\Desktop\SOURCE\LEGACY	OUTPUT1	C:\Users\Stephane.Chollet\Desktop\OUTPUT\LEGACY	
4	INPUT2	C:\Users\Stephane.Chollet\Desktop\SOURCE\SDTM	OUTPUT2	C:\Users\Stephane.Chollet\Desktop\OUTPUT\SDTM	
5	INPUT3	C:\Users\Stephane.Chollet\Desktop\SOURCE\ADAM	OUTPUT3	C:\Users\Stephane.Chollet\Desktop\OUTPUT\ADAM	

Folders location

	B	C
1	PARAMETER	VALUE
2	Final run:	NO
3	Master keys folder name:	MASTER_K
4	Hypernormal variables methods:	
	Age category method:	COLUMN
5		
	Drop method:	SET TO BLANK
6		
7	Seed:	1
8	Age unit :	YEAR
9	Dates anonymization method :	OFFSET DATE
10	Lower bound:	-35
11	Upper bound:	45
12	Zero:	NO
13	Same rando day:	NO

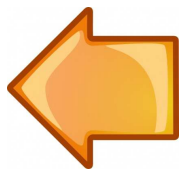
Anonymization Parameters

	A	B	C	D	E	F
1	FORMAT	INFORMAT	TIME	APPLICABLE	EXAMPLE	LABEL
2	DATE7.	DATE7.	NO	YES	10MAR16	DATE7
3	DATE9.	DATE9.	NO	YES	10MAR2016	DATE9_
4	DATE11.	DATE11.	NO	YES	10-MAR-2016	DATE11_
5	DDMMYYN6.	DDMMYY6.	NO	YES	100316	DDMMYYN6_
6	DDMMYYN8.	DDMMYY8.	NO	YES	10032016	DDMMYYN8_
7	DDMMYYB8.	DDMMYY8.	NO	YES	10 03 16	DDMMYYB8_
8	DDMMYYB10.	DDMMYY10.	NO	YES	10 03 2016	DDMMYYB10_
9	DDMMYYC8.	DDMMYY8.	NO	YES	10:03:16	DDMMYYC8_
10	DDMMYYC10.	DDMMYY10.	NO	YES	10:03:2016	DDMMYYC10_
11	DDMMYYD8.	DDMMYY8.	NO	YES	10.03.16	DDMMYYD8_
12	DDMMYYD10.	DDMMYY10.	NO	YES	10.03.2016	DDMMYYD10_
13	DDMMYYP8.	DDMMYY8.	NO	YES	10-03-16	DDMMYYP8_
14	DDMMYYP10.	DDMMYY10.	NO	YES	10-03-2016	DDMMYYP10_
15	DDMMYYS8.	DDMMYY8.	NO	YES	10/03/16	DDMMYYS8_
16	DDMMYYS10.	DDMMYY10.	NO	YES	10/03/2016	DDMMYYS10_

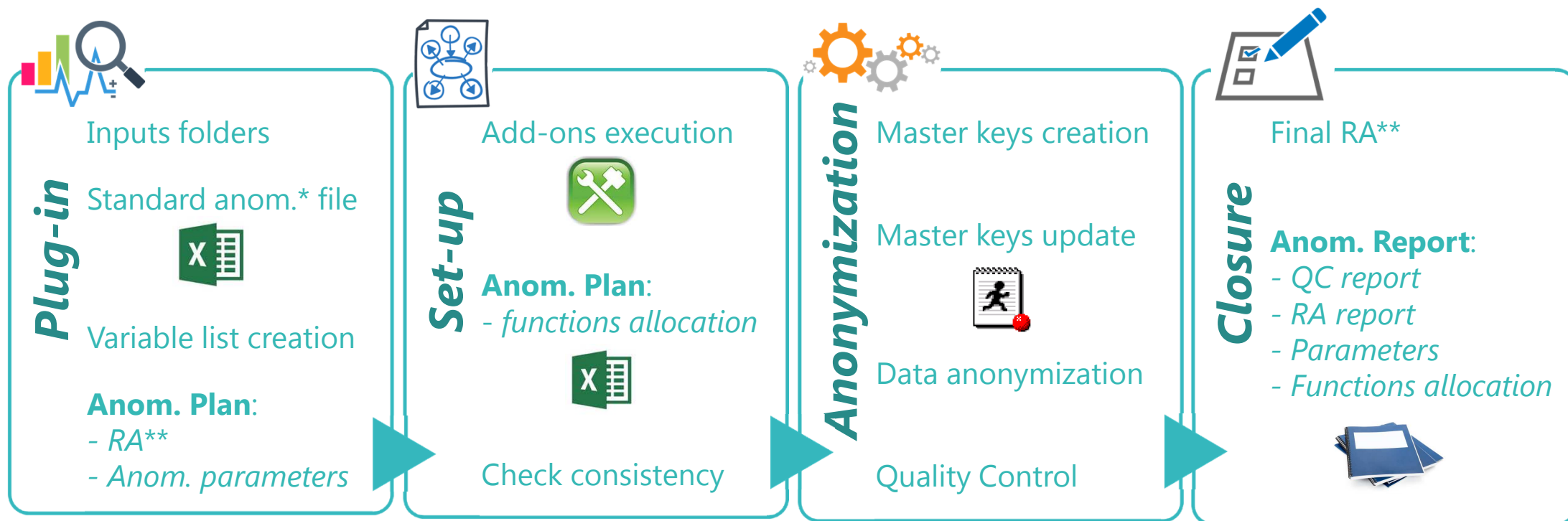
Applicable date formats



- **No need** to **customize** the SAS batch
- An **easy to use** and **evolving** interface



KEYDAN PRESENTATION

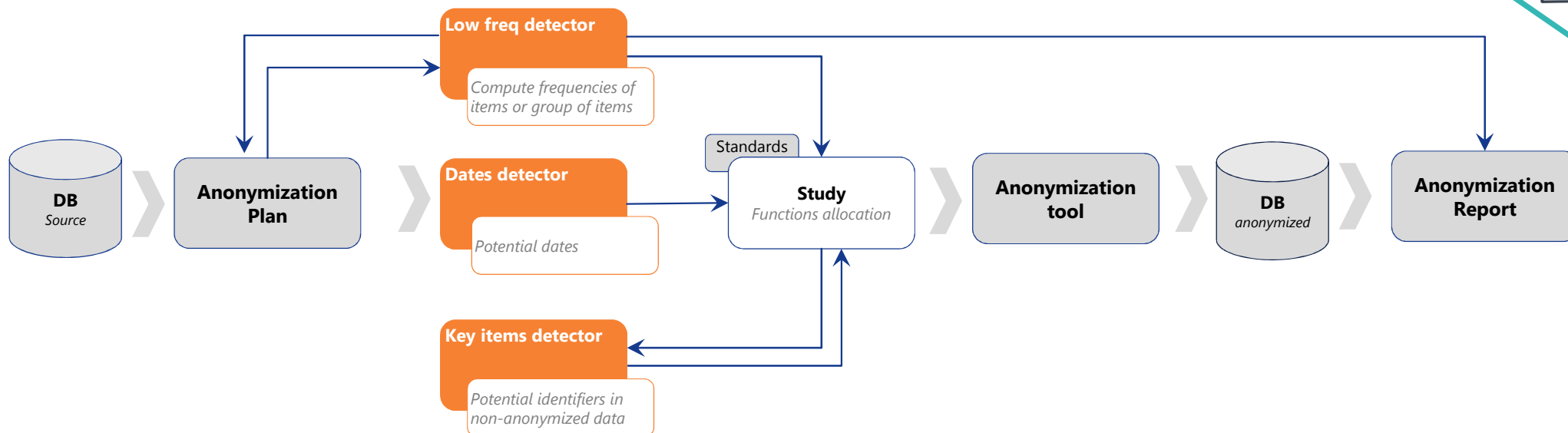


* ANOnyMization

** Risk Assessment

© Copyright 2017 – Keyrus

ADD-ONS

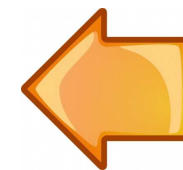


Provide assistance for risk minimization while de-identifying

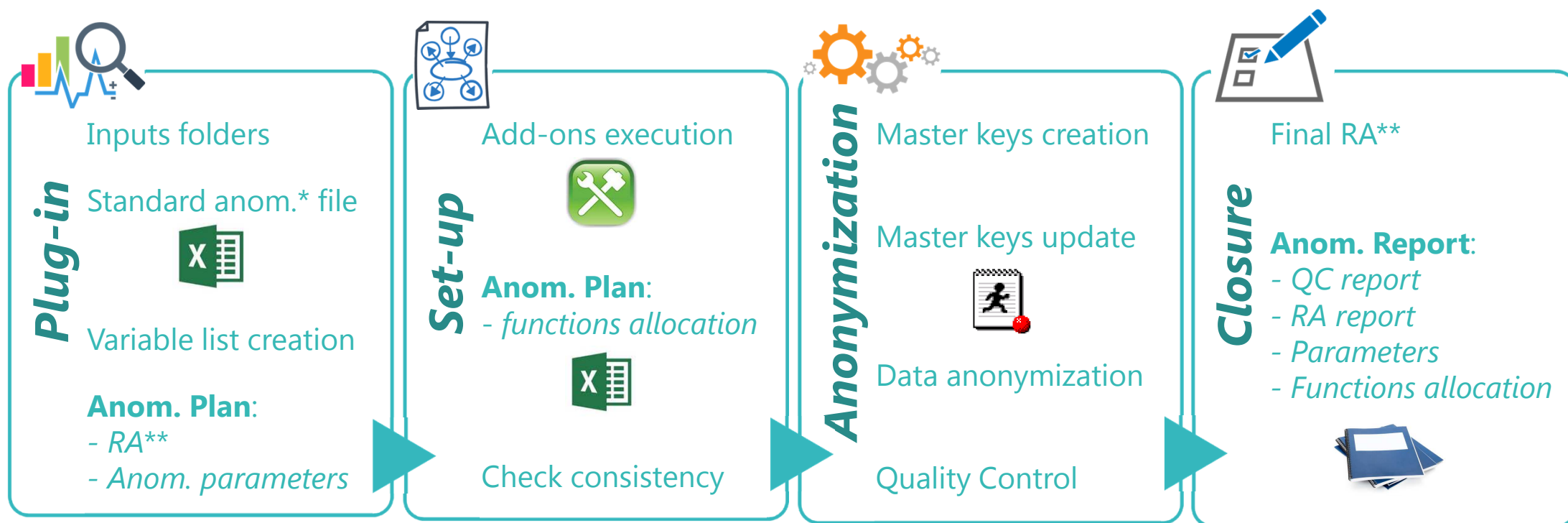
Computes the **re-identification risk** / Looks for **outliers** → **Low-freq**

Does not miss any **strings** that may be **interpreted as dates** → **Dates**

Does not miss any **identifiers present as substrings** → **Key-Items**



KEYDAN PRESENTATION



* ANOnyMization

** Risk Assessment

© Copyright 2017 – Keyrus

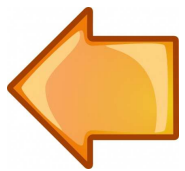
ANONYMIZATION FUNCTIONS ALLOCATION



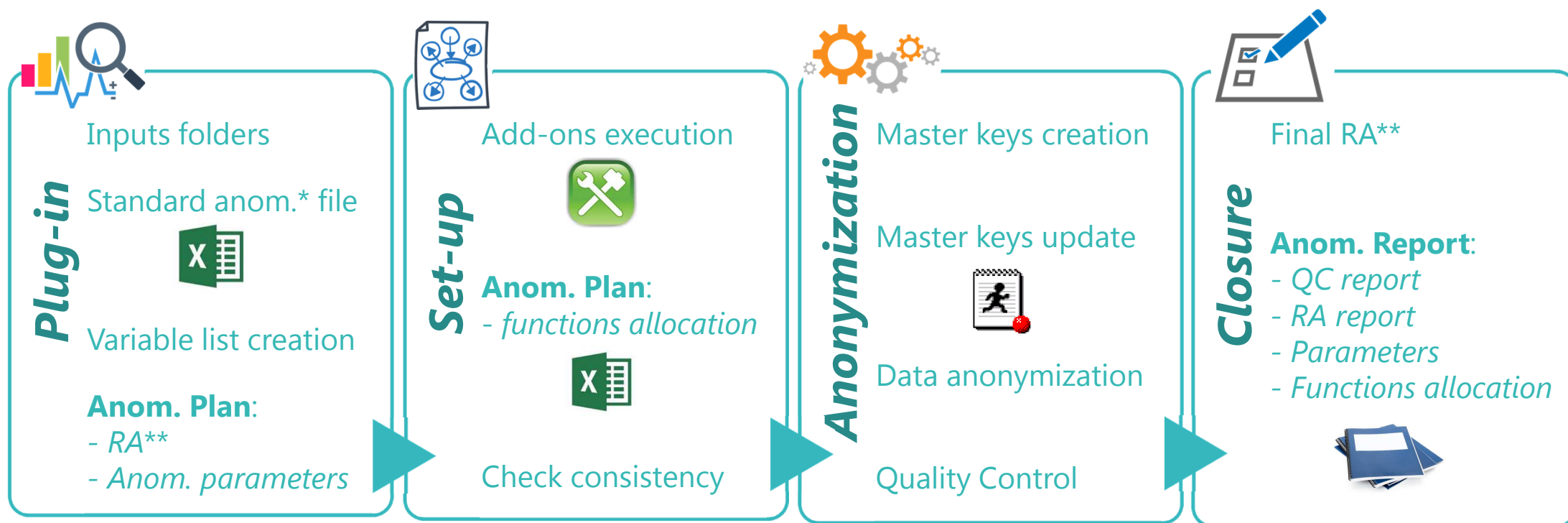
	A	B	C	D	E	F	G	H	I	J	K	L
1	LIBNAME	MEMNAME	NAME	LABEL	TYPE	FORMAT	LENGTH	FUNCTION	LINK	CONDITION	RISK_FLAG	COMMENTS
60	INPUT	DEMO	SITE	Site number	char	\$3.	3				x	
61	INPUT	DEMO	SUBJ	Subject number	char	\$3.	3	SUBJ_KEY			x	
62	INPUT	DEMO	VISID	Visit number	num		8	ANOM_KEY				
63	INPUT	DEMO	BRTHDT	Birth date	char	\$9.	9	DATE				
64	INPUT	DEMO	SEX	Sex	char	\$1.	1					
65	INPUT	DEMO	RACE	Geographical ancestry	char	\$2.	2				x	
66	INPUT	DEMO	RACETX	Geographical ancestry specification	char	\$6.	6				x	
87	INPUT	SIGN	SITE	Site number	char	\$3.	3				x	
88	INPUT	SIGN	SUBJ	Subject number	char	\$3.	3	SUBJ_KEY				
89	INPUT	SIGN	VISID	Visit number	num		8	ANOM_KEY				
90	INPUT	SIGN	INVNAM	Investigator's name	char	\$20.	20	BLANK				
91	INPUT	SIGN	SIGYN	CRF signed?	char	\$1.	1					
92	INPUT	SIGN	SIGDT	Signature date	num	DATE9.	8	DATE				
100	INPUT	VITAL	SITE	Site number	char	\$3.	3				x	
101	INPUT	VITAL	SUBJ	Subject number	char	\$3.	3	SUBJ_KEY				
102	INPUT	VITAL	VISID	Visit number	num		8	ANOM_KEY				
103	INPUT	VITAL	VASS	Assesemnt	char	\$15.	15	DATE				
104	INPUT	VITAL	VND	Test not done	char	\$2.	2					
105	INPUT	VITAL	VSBP	Systolic blood pressure (mmHg)	char	\$3.	3					
106	INPUT	VITAL	VDBP	Diastolic blood pressure (mmHg)	char	\$3.	3					
107	INPUT	VITAL	VHR	Hearth rate (beats/min)	char	\$3.	3				x	
108	INPUT	VITAL	VTEMP	Body temperature (°C)	char	\$4.	4					
109	INPUT	VITAL	VWGHT	Body weight (kg)	char	\$5	5				x	



- **Output** for the Anonymization Report
- Managing **several** data **sources** and **format** (legacy, CDISC ...)
- Use of **standards**



KEYDAN PRESENTATION



* ANOnyMization

** Risk Assessment

© Copyright 2017 – Keyrus

MASTER KEYS CUSTOMIZATION

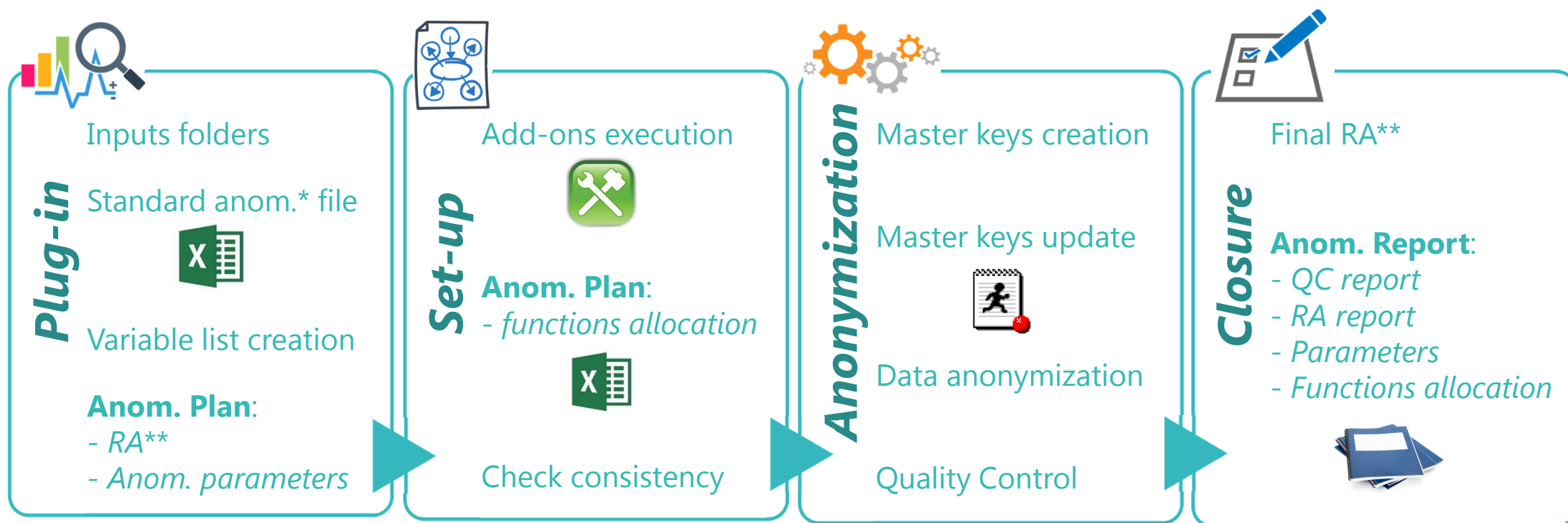
```
1  /*****/
2  /* program name: MASTER_SPEC */
3  /* version      : 1.00 */
4  /* date        : 22SEP2017 */
5  /* programmer   : Stéphane CHOLLET */
6  /*-----*/
7  /* Description: */
8  /*      Study specific anonymization transformation of master keys for MyStudy */
9  /*****/
10
11 □ %MACRO master_spec / DES = "MASTER_SPEC (01.00) - MyStudy specific master keys update";
12
13     /* Mandatory Step */
14     %PUT;
15     %PUT Specific updates to master keys tables execution;
16
17 | PROC SQL UNDO_POLICY=NONE;
18     CREATE TABLE master.usubjid AS
19     SELECT b.usubjid
20           , '0009876-000-' || PUT(INPUT(a.subjid_new, BEST12.), z5.0) AS usubjid_new
21           , LABEL = 'Unique Subject Identifier' FORMAT = $17. INFORMAT = $17.
22           , b.age_max
23           , b.nb_days
24     FROM master.subjid a
25          LEFT JOIN
26          master.usubjid b
27     ON a.subjid = substr(b.usubjid, 13, 5)
28     ;
29
30     /* Mandatory Step */
31     %PUT --> Specific updates to master keys tables execution complete;
32     %PUT;
33 %MEND;
```



- Macro **automatic** detection
- **Maintain consistencies** between variables
- Ability to apply transformations **in addition to the standard functions**



KEYDAN PRESENTATION



* ANOnyMization

** Risk Assessment

© Copyright 2017 – Keyrus

REPORTS (LIST OF) & ANONYMIZATION REPORTS

✓ Anonymization Plan / Report

Disclosure context, identifiers determination, anonymization strategy, risk assessment → Conclusion

✓ Appendix for the Anonymization Report

Datasets creation and content

Folder location of Source(s), Target(s) and Reference anonymization template

Anonymization parameters

Date formats

Function allocation for all datasets/variables (*with the rationale of the change*)

Function allocation summary by function

Function allocation summary by dataset

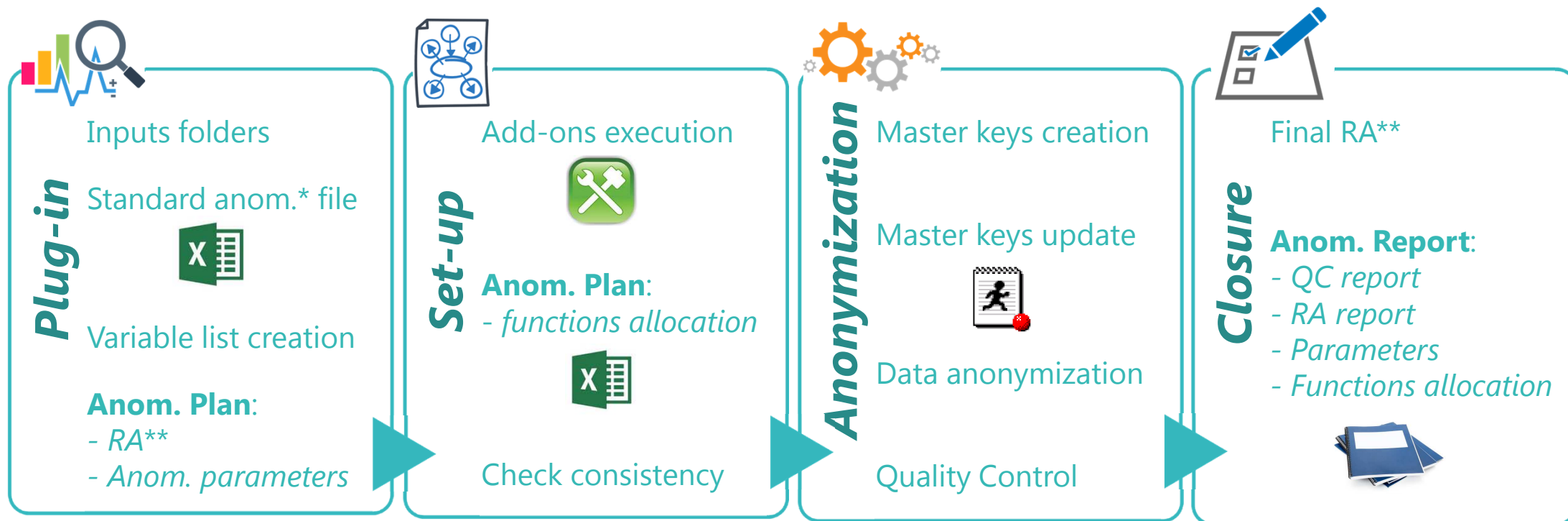
SUBJ_KEY and ANOM_KEY check consistency

DATE and DOB_ANO functions summary

Character dates not processed (→ *specific action needed before disclosure*)



KEYDAN PRESENTATION



* ANOnyMization

** Risk Assessment

© Copyright 2017 – Keyrus

DEMO - SIMPLE EXAMPLE WITH A LEGACY DATABASE



Plug-in

*Input Database
Parameterization* (using low-freq)



Strategy Set-up

Basic (for Subject iD, Center iD and Dates)
Advanced (using dates detector & low-freq)



Anonymization

Data Transformations



Closure

*Risk Assessment
Reports*



ANONYMIZATION - LEGACY

✓ Batch Execution



✓ Data Transformation



VIEWTABLE: Source.Demo							
	SITE	SUBJ	VISID	BRTHDT	SEX	RACE	RACETX
1	GER	326	11	27MAR1938	F	CA	
2	GER	502	11	07NOV1913	M	OT	LATINO
3	LUX	309	11	28MAR1971	F	OT	PYGMY
4	BEL	663	11	27NOV1955	M	CA	
5	BEL	682	11	17DEC2002	M	CA	
6	FRA	812	11	17JUL1965	F	BL	
7	GER	234	11	13SEP1965	F	CA	
8	FRA	839	11	17JUN1966	F	BL	
9	GER	965	11	29DEC1951	M	CA	
10	FRA	517	11	14AUG1993	M	AS	
11	FRA	658	11	28JUL1929	F	CA	

EWTABLE: Master.Subj				
	SUBJ	SUBJ_NEW	AGE_MAX	NB_DAYS
1				3
2	234	9	46	3
3	309	10	98	-2
4	326	1	82	1
5	502	4	74	-1
6	517	3	46	3
7	658	8	18	3
8	663	7	41	3
9	682	11	56	3
10	812	5	10	1
11	839	6	47	2
12	965	2	60	-1

WTABLE: Master.Site			WTABLE: Master.Race	
	SITE	SITE_NEW	RACE	RACE_NEW
1				
2	BEL	B&L		
3	FRA	FRA		
4	GER	GER		
5	LUX	B&L		

VIEWTABLE: Output.Demo							
	SITE	SUBJ	VISID	BRTHDT	SEX	RACE	RACETX
1	GER	1	11	1938	F	CA	
2	GER	4	11	1913	M	OT	
3	B&L	10	11		F	OT	
4	B&L	7	11	1955	M	CA	
5	B&L	11	11	2002	M	CA	
6	FRA	5	11	1965	F	OT	
7	GER	9	11	1965	F	CA	
8	FRA	6	11	1966	F	OT	
9	GER	2	11	1951	M	CA	
10	FRA	3	11	1993	M	OT	
11	FRA	8	11	1929	F	CA	

DEMO - SIMPLE EXAMPLE WITH A LEGACY DATABASE



Plug-in

*Input Database
Parameterization* (using low-freq)



Strategy Set-up

Basic (for Subject iD, Center iD and Dates)
Advanced (using dates detector & low-freq)



Anonymization

Data Transformations



Closure

*Risk Assessment
Reports*

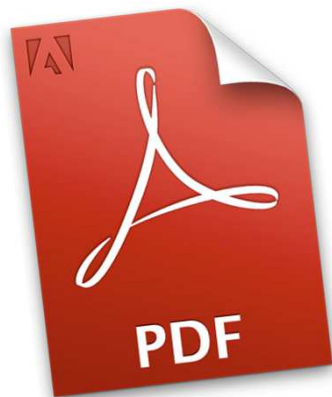


ANONYMIZATION - LEGACY

✓ Risk Assessment



✓ Reports



DEMO - SPECIFICS USES



Multiple sources (*e.g. pulling*) & Format (*Legacy, CDISC, ...*)

Different Sources

Add-ons Use to help with suppQUAL anonymization



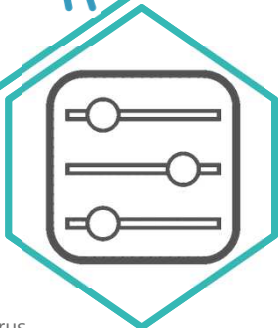
Using standard

*Precompletion of **Functions Allocation Using a Standard***



Customization

Master Keys Customization



Other functionalities

*Generation of the **Same Number of Days***

Repeatability** of the **Process

Destruction** of the **master keys

...



ANONYMIZATION - CDISC

✓ Batch set-up



✓ Data Transformation

✓ where qlabel like '%DATE%'

VIEWTABLE: Srcsdtm.Suppae						
	RDOMAI	STUDYID	USUBJID	QNAM	QLABEL	QVAL
978	AE	DEMO9876-001	0009876-001-01004	HOSPSTD	HOSPITALIZATION START DATE	2006-01-05
1289	AE	DEMO9876-001	0009876-001-01006	HOSPEND	HOSPITALIZATION END DATE	2002-10-29
1290	AE	DEMO9876-001	0009876-001-01006	HOSPSTD	HOSPITALIZATION START DATE	2002-10-19
1505	AE	DEMO9876-001	0009876-001-01007	DEATHDT	DATE OF DEATH	2002-06-13
1506	AE	DEMO9876-001	0009876-001-01007	HOSPEND	HOSPITALIZATION END DATE	2002-06-13
1507	AE	DEMO9876-001	0009876-001-01007	HOSPSTD	HOSPITALIZATION START DATE	2002-06-07
1721	AE	DEMO9876-001	0009876-001-01009	DEATHDT	DATE OF DEATH	2005-10-15
1722	AE	DEMO9876-001	0009876-001-01009	HOSPEND	HOSPITALIZATION END DATE	2005-10-15

✓ where qlabel like '%DATE%'

VIEWTABLE: Tgtsdtm.Suppae						
	RDOMAIN	STUDYID	USUBJID	QNAM	QLABEL	QVAL
978	AE	DEMO9876-001	0009876-000-00098	HOSPSTD	HOSPITALIZATION START DATE	2005-12-28
1289	AE	DEMO9876-001	0009876-000-00151	HOSPEND	HOSPITALIZATION END DATE	2002-11-05
1290	AE	DEMO9876-001	0009876-000-00151	HOSPSTD	HOSPITALIZATION START DATE	2002-10-26
1505	AE	DEMO9876-001	0009876-000-00135	DEATHDTC	DATE OF DEATH	2002-06-16
1506	AE	DEMO9876-001	0009876-000-00135	HOSPEND	HOSPITALIZATION END DATE	2002-06-16
1507	AE	DEMO9876-001	0009876-000-00135	HOSPSTD	HOSPITALIZATION START DATE	2002-06-10
1721	AE	DEMO9876-001	0009876-000-00047	DEATHDTC	DATE OF DEATH	2005-09-27
1722	AE	DEMO9876-001	0009876-000-00047	HOSPEND	HOSPITALIZATION END DATE	2005-09-27

✓ where qnam in ('AEPTCD','AELLT','AELLTCD')

VIEWTABLE: Srcsdtm.Suppae						
	RDOMAI	STUDYID	USUBJID	QNAM	QLABEL	QVAL
5	AE	DEMO9876-001	0009876-001-01001	AELLT	LOW LEVEL TERM	Dyspnea
6	AE	DEMO9876-001	0009876-001-01001	AELLTCD	LOW LEVEL TERM CODE	10013963
8	AE	DEMO9876-001	0009876-001-01001	AEPTCD	PREFERRED TERM CODE	10013972
18	AE	DEMO9876-001	0009876-001-01001	AELLT	LOW LEVEL TERM	Sore throat
19	AE	DEMO9876-001	0009876-001-01001	AELLTCD	LOW LEVEL TERM CODE	10041367
20	AE	DEMO9876-001	0009876-001-01001	AEPTCD	PREFERRED TERM CODE	10034835
30	AE	DEMO9876-001	0009876-001-01001	AELLT	LOW LEVEL TERM	Confusion
31	AE	DEMO9876-001	0009876-001-01001	AELLTCD	LOW LEVEL TERM CODE	10010300

✓ where qnam in ('AEPTCD','AELLT','AELLTCD')

VIEWTABLE: Tgtsdtm.Suppae						
	RDOMAIN	STUDYID	USUBJID	QNAM	QLABEL	QVAL
5	AE	DEMO9876-001	0009876-000-00063	AELLT	LOW LEVEL TERM	
6	AE	DEMO9876-001	0009876-000-00063	AELLTCD	LOW LEVEL TERM CODE	
8	AE	DEMO9876-001	0009876-000-00063	AEPTCD	PREFERRED TERM CODE	
18	AE	DEMO9876-001	0009876-000-00063	AELLT	LOW LEVEL TERM	
19	AE	DEMO9876-001	0009876-000-00063	AELLTCD	LOW LEVEL TERM CODE	
20	AE	DEMO9876-001	0009876-000-00063	AEPTCD	PREFERRED TERM CODE	
30	AE	DEMO9876-001	0009876-000-00063	AELLT	LOW LEVEL TERM	
31	AE	DEMO9876-001	0009876-000-00063	AELLTCD	LOW LEVEL TERM CODE	

A TOOL FOR A PROCESS



Take home messages

Simple, Flexible and Efficient
Consistent with all databases Formats and Structures
Customizable & Evolutive



Risk assessment

Compliance with Legal Requirements

Detailed in SD01 "KeyDAN, a Clinical Data De-identification Solution" paper



Next steps

Add New Transformation Functions
Automate the process with CSR



Contact information

stephane.chollet@keyrus.com

www.keyrusbiopharma.com



You have a question?

You want to know more?

You would like to discuss this with us?

Come and see us at booth #20



KEYRUS
biopharma