PhUSE 2016

Paper AD04

How to create your Define.xml as early as possible with a click on a button

Roman Radelicki, SGS - Agriculture, Food and Life, Life Sciences - Clinical Research, Mechelen, Belgium

ABSTRACT

Creating the Define.xml (data definition specification) for a trial can be a difficult task, which sometimes requires manual operations that can lead to errors. This paper illustrates the process to automate the completion of the Define.xml based on information available from in-house developed applications used during the setup of the trial and therefore reduces manual operations to a minimum and in the end provides a more accurate Define.xml.

INTRODUCTION

Define.xml transmits metadata for SDTM, SEND and ADaM datasets. It is the metadata file sent with every study in each submission, which tells the regulatory agencies what datasets, variables, controlled terms, and other specified metadata were used. Metadata - from the Greek prefix meta which means after or beyond. Data about data. Data that describes your data.

Which came first, the chicken or the egg? What do you need first? The data so you can define your metadata later on when you've collected all your raw data or do you first define your metadata? The fact is that it's getting more and more important to be able to provide your Define.xml as early as possible, preferably at the start of trial setup. Not just because the sponsor requests it, but because it is used as input for next steps, like defining cleaning rules or creating progress reports. How can you manage this? How can you describe your data if you don't have your data yet?

METADATA REPOSITORY

One solution could be to work with a **metadata repository** (MDR). This is the master source of metadata to be used for study setup. It's a global metadata library based on the SDTM structure, containing all the available domains, variables, value lists and code lists that can be used for a specific implementation of SDTM, SEND or ADaM.

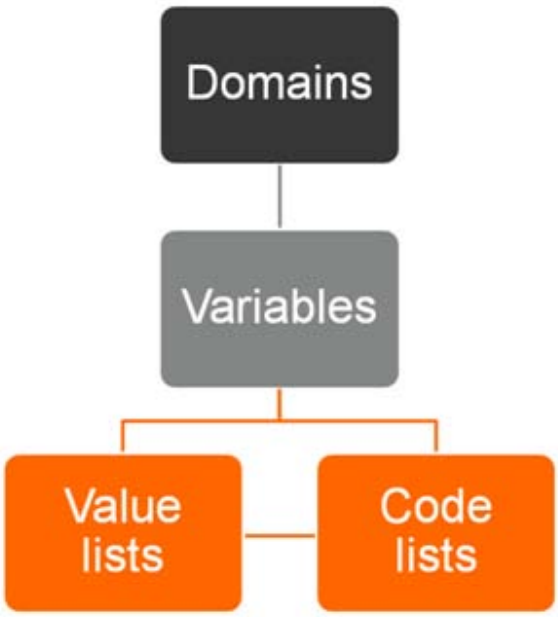

Fig. 1 Metadata repository

Fig. 1 illustrates the structure of the MDR. At the top level you will find all available domains for a specific implementation of SDTM. If we move further down, you can see all the available variables per domain. At the bottom you will find the value level metadata and the code lists.

HOW DO WE USE THE MDR TO EVENTUALLY PROVIDE A COMPLETE DEFINE.XML FILE

As a starting point we use the complete master metadata library and make it more and more trial specific when information becomes available during the setup of a specific trial. Therefore the complete MDR is copied to trial level and modified as needed.

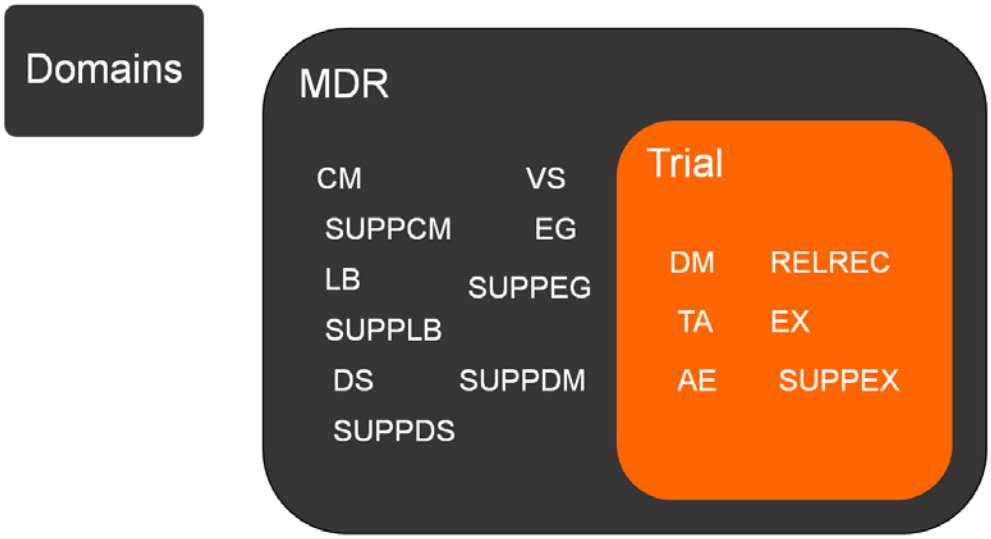

Fig. 2 Selection of domains applicable for a specific trial

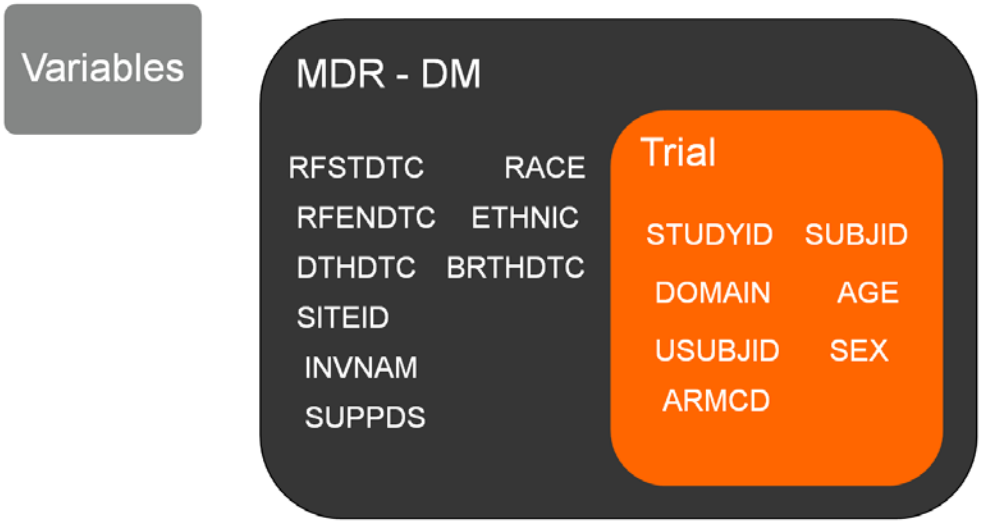

Fig. 3 Selection of variables for domain DM for a specific trial

Fig. 2 and **Fig. 3** illustrate how we select, from the complete MDR, only the domains and variables that are applicable for a specific trial.

DATASET	DESCRIPTION	CLASS	STRUCTURE	PURPOSE	KEYS	DSORDER	SGSSTATE
TA	Trial Arms	Trial Design	One record per planned Element per Arm	Tabulation	STUDYID, ARMCD, TAETORD	001	IN
DM	Demographics	Special-Purpose	One record per subject	Tabulation	STUDYID, USUBJID	009	IN
CM	Concomitant/Prior Medications	Interventions	One record per recorded intervention occurence or const...	Tabulation	STUDYID, USUBJID, CMTRT, CMSTDTC, CMSCAT	013	EX
EX	Exposure	Interventions	One record per protocol-specified study treatment, const...	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC, EXTPTNUM, EXREFID	015	IN
AE	Adverse Events	Events	One record per adverse event per subject	Tabulation	STUDYID, USUBJID, AEDECOD, AESTDTC	020	IN
DS	Disposition	Events	One record per disposition status or protocol milestone p...	Tabulation	STUDYID, USUBJID, DSDECOD, DSSTDTC, DSSCAT	023	EX
EG	ECG Test Results	Findings	One record per ECG observation per time point per visit p...	Tabulation	STUDYID, USUBJID, EGTESTCD, VISITNUM, EGTPTREF, EGTPTNUM	031	EX
LB	Laboratory Test Results	Findings	One record per lab test per time point per visit per subject	Tabulation	STUDYID, USUBJID, LBTESTCD, LBSPEC, VISITNUM, LBTPTREF, LBTPTNUM	036	EX
VS	Vital Signs	Findings	One record per vital sign measurement per time point pe...	Tabulation	STUDYID, USUBJID, VSTESTCD, VISITNUM, VSTPTREF, VSTPTNUM	053	EX
RELREC	Related Records	Relationship	One record per related record, group of records or dataset	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, RELID	060	IN
SUPPEX	Supplemental Qualifiers for EX	Relationship	One record per IDVAR, IDVARVAL and QNAM per subject	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, QNAM	074	IN
SUPPLB	Supplemental Qualifiers for LB	Relationship	One record per IDVAR, IDVARVAL and QNAM per subject	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, QNAM	080	EX
SUPPVS	Supplemental Qualifiers for VS	Relationship	One record per IDVAR, IDVARVAL and QNAM per subject	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, QNAM	103	EX

Fig. 4 database view of selected domains

Fig. 4 shows the modified MDR for domains on trial level in our database. Indicating a domain is applicable for your trial is simply achieved by the use of the SGSSTATE flag (last column in **Fig. 4**). SGSSTATE 'IN' will include the domain in the Define.xml and SGSSTATE 'EX' will exclude the domain. This method of in-or excluding domains is used for variables, value lists, code lists and computational methods as well.

SDTM-IG 3.2

Annotated Case Report Form
- Tabulation Datasets
- Value Level Metadata
- Controlled Terminology
- Computational Algorithms
- Comments

Tabulation Datasets for Study SGSTRIAL (SDTM 3.2) (SDTM-IG 3.2)

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
TA	Trial Arms	Trial Design	One record per planned Element per Arm	Tabulation	STUDYID, ARMCD, TAETORD	TA.xpt	
DM	Demographics	Special-Purpose	One record per subject	Tabulation	STUDYID, USUBJID	DM.xpt	
EX	Exposure	Interventions	One record per protocol-specified study treatment, constant-dosing interval, per subject	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC	EX.xpt	
AE	Adverse Events	Events	One record per adverse event per subject	Tabulation	STUDYID, USUBJID, AETERM, AESTDTC	AE.xpt	
RELREC	Related Records	Relationship	One record per related record, group of records or dataset	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, RELID	RELREC.xpt	
SUPPEX	Supplemental Qualifiers for EX	Relationship	One record per IDVAR, IDVARVAL and QNAM per subject	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, QNAM	SUPPEX.xpt	

Go to the top of the define.xml

Fig. 5 Define.xml

From the trial level metadata database tables SGS has an application that will create a Define.xml. **Fig. 5** illustrates the domain part of the Define.xml based on the tables from **Fig. 4**.

We use information collected via in-house developed applications to be able to automatically make the metadata trial specific.

IN-HOUSE DEVELOPED APPLICATIONS

ANNOTATION TOOL

This is a tool that will guide the user in annotating a CRF based on the applicable standard. The user will be able to copy annotations from the annotation library or will be guided to add a new annotation. Worrying about the layout of your annotations like font, font size, color is a thing of the past. The tool will apply the correct layout of the annotations based on the client standard or SGS standard. **Fig. 6** shows all the available annotation types and their layout settings for a specific standard.

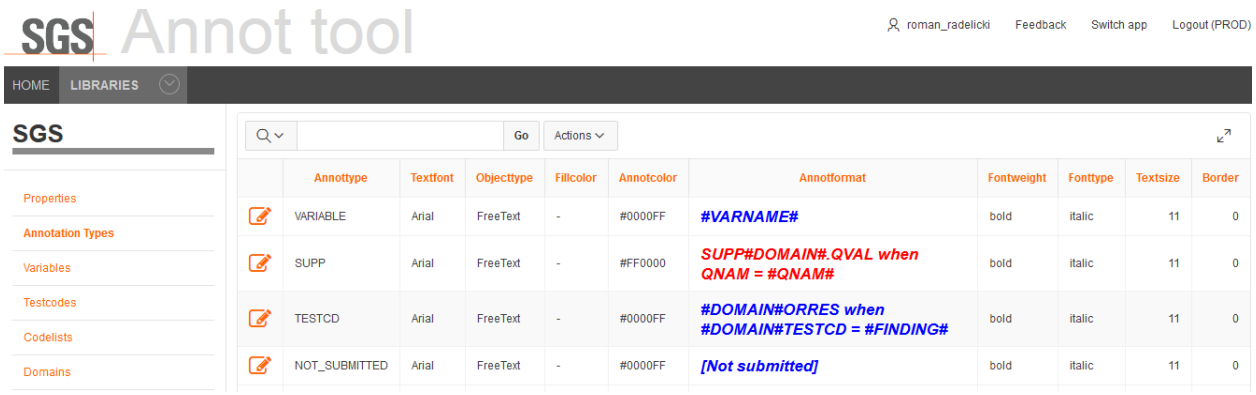

Fig. 6 Annotation types and layout settings

After loading a PDF version of the eCRF into the tool, the user can start annotating the document and will be guided by the tool. E.g. if a form was already annotated for another trial, the user will be able to copy the annotation and modify if needed. Should a new annotation be added on a page, a wizard will guide the user in creating the correct annotation.

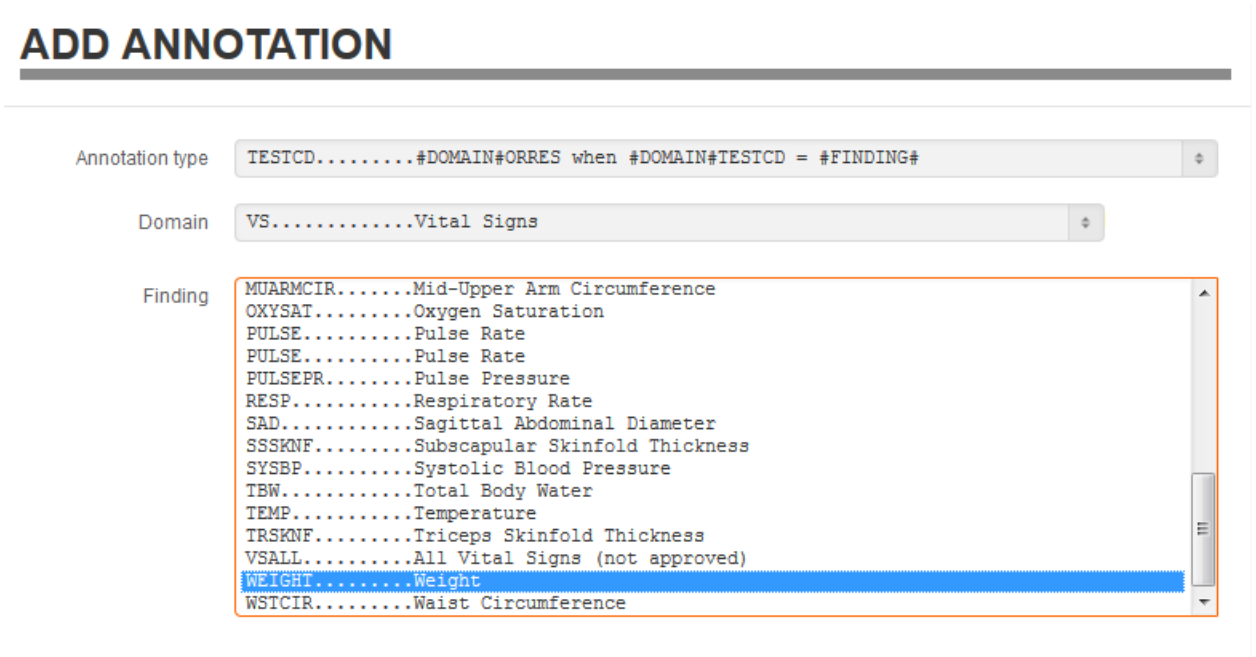

Fig. 7 add an annotation

Fig. 7 shows the wizard that allows the user to add a finding on the Vital Signs form. First the user will select the annotation type, in this case a finding. Then a domain is selected from all finding domains available for this applicable standard. The last step is to select the correct test code needed for the annotation from a list containing all available test codes for domain VS for that applicable standard.

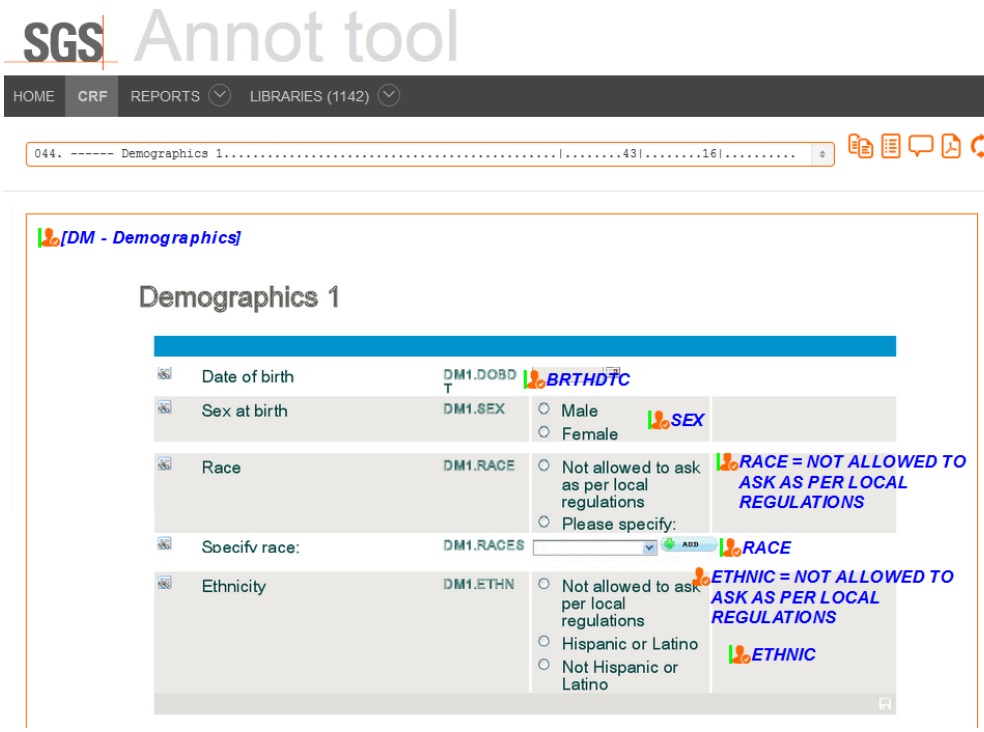

Fig. 8 Annotated Demographics form in the annotation tool

Fig. 8 illustrates how an annotated Demographics page will look like in the annotation tool. Another advantage of the tool is the review process. Per annotation on the eCRF a status can be applied that allow the reviewers to indicate the applicable annotation has been reviewed or approved. **Fig. 9** shows all available status settings in the annotation tool. When all annotations are reviewed and approved, the user will be able to export the annotations to the PDF and share it with involved parties.

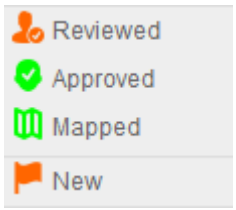

Fig. 9 Status in the annotation tool

All this will result is a more accurate and qualitative annotated eCRF. In the background all annotations are captured electronically in the database and can therefore be used in other processes like the creation of the Define.xml or in the mapping of the CRF data to SDTM data. We now know the domains, variables, value lists, code lists used in the annotation and this enables us to make the metadata more specific by stetting the SGSSTATE flag as explained in **Fig. 4**

SDTM MAPPING TOOL

Based on the annotation tool and on the MDR, mapping of the CRF data to SDTM is performed by the programmer using the SDTM mapping tool.

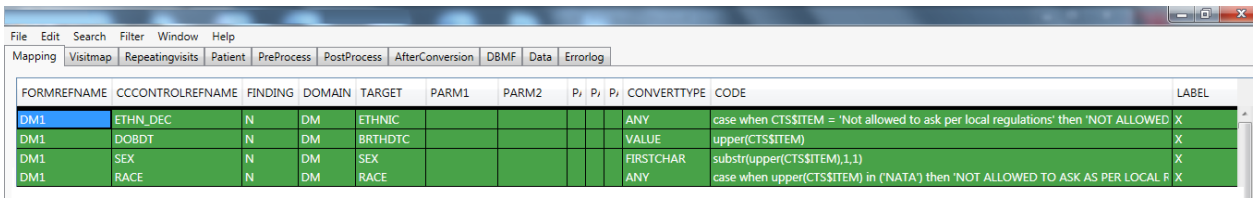

Fig. 10 SDTM Mapping tool

Fig. 10 illustrates the use of the mapping tool. eCRF data, usually SAS™ datasets, are loaded into our database. All input fields, radio buttons, checkboxes from the eCRF are captured and presented to the programmer like a pivot table. Per form the programmer is able to map each column to a specific SDTM variable. The first line in **Fig. 10** DM1.ETHN_DEC corresponds with the question 'Ethnicity' on the eCRF in **Fig. 8**. The programmer is guided to either copy the complete mapping from another trial if this form was mapped before or to create a new mapping by selecting the applicable domain and variable from a standard specific dropdown list.

The key to be able to provide a Define.xml as complete as possible even before the first patient enters the trial is to make sure test data was entered with all possible questions and answers populated on the CRF. That enables us to select all applicable domains, variables, values lists and code lists specific to that trial and create already a Define.xml which is of high quality and will look like the Define.xml which will be used at database lock.

DATASET	VARNAME	VARLABEL	DATATYPE	LNGTH	ORIGIN	VARROLE	CORE	VARORDER	CODELST	VALUEOID	CRFPAGE	SGSSTATE
DM	STUDYID	Study Identifier	text	40	Protocol	Identifier	Req	001				IN
DM	DOMAIN	Domain Abbreviation	text	2	Assigned	Identifier	Req	002	DOMAIN			IN
DM	USUBJID	Unique Subject Identifier	text	40	Derived	Identifier	Req	003				IN
DM	BRTHDTC	Date/Time of Birth	datetime	19	CRF	Record Qualifier	Perm	016			44	IN
DM	SEX	Sex	text	1	CRF	Record Qualifier	Req	019	SEX		44	IN
DM	RACE	Race	text	200	CRF	Record Qualifier	Exp	020	RACE		44	IN
DM	ETHNIC	Ethnicity	text	200	CRF	Record Qualifier	Perm	021	ETHNIC		44	IN

Fig. 11 database view of selected variables of a domain

All the data captured in those tools will enable us to not only decide whether a domain or variables should be included, additional automatic updates are possible. **Fig. 11** shows you all the selected variables in the DM domain. Based on information from the annotation tool we are able to determine the origin for a certain variable. BRTHDTC, SEX, RACE and ETHINC are annotated hence the ORIGIN column is set to CRF. Not only do we know these variables come from the CRF, we also know from which page they are coming from and therefore we can populate the CRFPAGE column as well.

WHAT ARE WE STILL MISSING?

Off course the Define.xml is a living document and is subjected to changes during the course of a trial. After the DTA's (Data Transfer Agreement) are in place, external data should be added to the datasets and will therefore cause updates to the Define.xml. Randomization and blinded data which is only available at lock will trigger the needed updates. Changes in a specific SDTM implementation guide will cause modifications in the metadata. The length of the variables in the metadata, when implemented early in the trial, might change with every snapshot of data you need to provide to the sponsor. All these changes are programmed in the SDTM Mapping tool mentioned above and can therefore automatically update your metadata with the new available information.

CONCLUSION

The use of the in-house developed applications not only facilitate, speed up and improve the quality of certain tasks that are performed like e.g. the annotation and the mapping of the eCRF, It also provide us with valuable information needed to automatically create accurate metadata and in the end the Define.xml. The use of different sponsor or version specific MDR's allows us to easily integrate new standards into our different applications without the need of reprogramming and provides us with the flexibility we need today.

REFERENCES

http://www.cdisc.org/define-xml

CONTACT INFORMATION

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

Roman Radelicki
SGS – Agriculture, Food and Life
Life Sciences - Clinical Research
Generaal de Wittelaan 19A Bus 5
B-2800 - Mechelen
Email: roman.radelicki@sgs.com

Web: www.sgs.com/cro