PhUSE 2016

Paper DH10

Metadata on the go: Achieving metadata accuracy and consistency continually

Swapna Pothula, SGS Life Sciences, Mechelen, Belgium

ABSTRACT

Metadata enables exchange, review, analysis, automation and reporting of clinical data. Metadata is crucial for clinical research and standardization makes it powerful. Adherence of metadata to CDISC SDTM has become the norm, since FDA has chosen SDTM as the standard specification for submitting tabulation data for clinical trials.

Today, many sponsors expect metadata to be not just compliant to CDISC but also to their own standards. Creating metadata that is consistent and accurate at every point of time from set up until and after the database lock remains a challenge for operational data management.

Metadata repositories help in creating standardized metadata but it is just the beginning and there is a need for more. At SGS Life Sciences, we have implemented an efficient and effective, metadata workflow which not only creates metadata that is CDISC compliant but also enables us achieve consistency and accuracy continually therefore creating high quality metadata.

INTRODUCTION

Metadata is defined to be data about data, but is it that simple? No, there is much more to it and more so in clinical world. Clinical metadata provides conceptual, contextual and process information which not only defines data but also gives insight into the relationship between data. Metadata enables exchange, review, analysis, automation and reporting of clinical data.

Standardization helps exchange and use of metadata across different processes during the life cycle of a clinical trial at the conceptual level but there is a need for flexibility at the contextual level. The context is dynamic. Metadata Repositories (MDRs) address standardization at the conceptual level and leveraging flexibility at the contextual level is what makes metadata more meaningful and usable.

While it is clear that metadata is crucial to create high quality clinical databases, achieving high quality metadata continually remains a challenge for clinical data management.

CHALLENGES ON THE ROAD TO QUALITY METADATA

How do we make sure trial metadata is consistent with CDISC SDTM standards? What if the sponsors have their own standards and are actively involved in the review? How do we balance the diverse sponsor needs? How do we keep up with standards that are changing constantly? How do we make sure that the trial metadata is both accurate and consistent? And how do we do it efficiently and effectively saving both time and costs?

While all of these remain to be the major questions that need to be addressed at the conceptual level, they give rise to many more questions that need to be addressed at the contextual level. These questions trickle down to the role of a programmer who has to find answers and make day to day decisions to provide quality metadata.

Most of the questions have been discussed and addressed quite often at the conceptual level suggesting the metadata driven approach and need for seamless integration of processes and people. But what do they mean for a programmer and how do they translate as day to day tasks for a programmer who actually creates the metadata?

I would like to focus and draw attention to the questions that arise at the contextual level and few scenarios a programmer is confronted with on a day to day basis while creating the clinical metadata.

Switch of environments

Today, sponsors are actively involved in the review of the clinical databases and expect high quality databases and

metadata. Sponsors have their own checks for validating compliance which are run on every snapshot and expect no output.

Since databases are set up in a test environment and with test data, quality metadata would translate to metadata that is consistent with the current data, which is test data. And the moment we go live, we are expected to provide metadata that is consistent with the live data. Most of the times, a snapshot of the database with accurate metadata is expected on the day we go live. How do we make this possible given the time constraints? And it doesn't stop there, live data changes every day, and the snapshots sent to sponsor should always be consistent and compliant.

Standards that change

New versions of standards contribute to overall improvement of quality and broaden the scope of domains. New versions are 'nice to haves' and sponsors will always want them implemented. Upgrading to the latest standards while the trial is ongoing and the database is already set up brings in challenges.

Upgrading to the latest standard doesn't just imply copying the latest version of the metadata standard from the MDR. Since all of the contextual metadata for the trial is set up, a programmer would aim to retain it where applicable and make the upgrades only where needed. How do we do this given the time and cost constraints? How do we achieve compliance both with the standards and the trial in such cases and also be efficient?

Conflict of Standards

Standards are changing and just when we think we have figured out mechanisms to cope with changes, we are confronted with the discrepancies between standards, discrepancies between sponsor and CDISC standards, and discrepancies between 'the' standards.

One Such example would be: Dataset Column length requirement by FDA

We have all seen the 'Variable length is too long for actual data' error on Pinnacle21. The new FDA requirement is to implement variable lengths for each column containing character data. And this is not just applicable per domain but over all domains that are part of the trial database. This implies that the length of a column should be calculated considering the maximum length of the data captured in that column variable occurring in each of the trial datasets.

The VISIT column for example, when the length for this column is calculated over all domains that it occurs in, Pinnacle21 generates errors for certain domains. This is because Pinnacle21 is checking for column length per domain and not over all the domains in which VISIT occurs. Compliance is always questioned when there are discrepancies. Discrepancies as such need to be reported and addressed within very short frames of time and with a rationale. It is not easy to convince sponsor to ignore a Pinnacle21 error.

Non – DM datasets

Datasets that are not generated by the data management but are part of the submission package are the Non-DM datasets. Datasets that are not part of the database when it is set up but are part of the submission package are to be dealt with for most of the trials. Examples of such datasets are PC, PD, PP and so on.

It is the responsibility of the programmer to make sure the metadata for all these datasets is complete and consistent. In case of blinded trials, these datasets are only delivered on the day of lock. Having the Non-DM datasets added to the rest of the datasets and delivering accurate metadata for these datasets on the day of lock is quite a task. What makes it difficult is the fact that you get to see the datasets for the first time on the day of lock when we are always running short of time and finding issues that need to be fixed right away. These datasets only add to the pressure.

How can the consistency be checked for in such cases when the datasets are not part of the database and you cannot run all those checks which you would otherwise run on your database against the standard repositories? Will validating the datasets and Define.xml on Pinnacle21 suffice?

These are some of the few scenarios, every programmer encounters while a clinical trial runs its course. These happen to be more critical for early phase trials where trials last for very short periods of time and need to go through all of the work flows any other trial would but at pace that is 10 times faster. Everything here needs to happen 'on the go' without compromising on quality.

IS QUALITY METADATA CONTINUALLY ACHIEVABLE?

To answer this question and to know how and if we are actually able to create quality metadata continually, we will loop through the metadata workflow set up at SGS. To start with, an overview of the hierarchical metadata repository implementation and processes that create metadata will be provided. Then the usage of in-house developed

applications in conjunction with the open source tools will be explained. Concluding with an elaborate explanation of the work flow itself demonstrating how we do it continually, i.e., being able to create and provide sponsors with quality metadata that is consistent with both standards and data right from the set up until and after the database lock and thus accurate at every point of time. Finally, a scenario will be discussed where we need more than just tools and when it's not all that straight forward.

MDRs

Metadata repositories should be generic, integrated, current, and historical. SGS is a CRO dealing with a variety of sponsors and therapeutic areas. To accommodate the variety of sponsor needs, hierarchical MDRs have been implemented focusing on standardization and reuse. The hierarchical nesting is in the order of CDISC SDTM, SPONSOR/SGS STANDARD, THERAPEUTIC AREA, and TRIAL METADATA as illustrated in Fig.1. We have a SGS standard developed to be implemented for sponsors with no specific standards of their own.

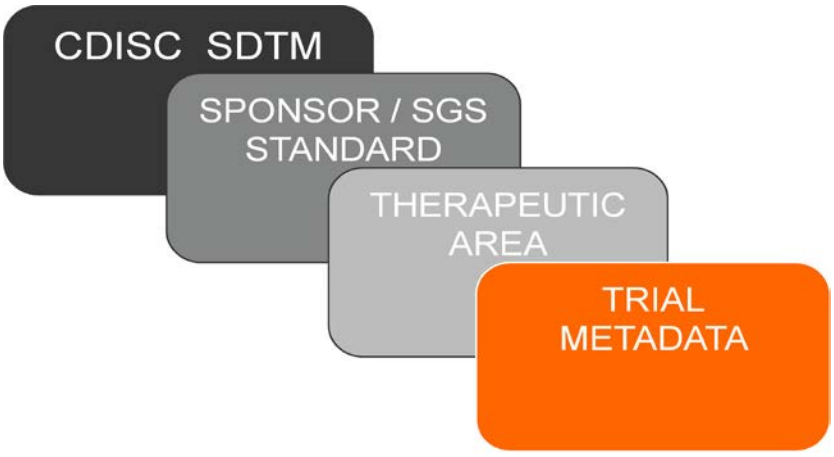


Fig.1 Hierarchical Metadata Repositories

Annotation Tool

Annotation tool is an in-house application used to annotate an (e)CRF. This tool is tailored to custom create annotations taking into account the standard and sponsor requirements. It not only takes into account the physical attributes of the annotations but also guides the programmer to look and find the most applicable codelist values for a given domain making annotations simpler. Fig.2 illustrates the possible values for the EGTESTCD codelist available for the given standard, sponsor and therapeutic area.

The tools' inbuilt search function allows the programmer to look into annotated CRFs from the same (e)Source system, sponsor, standard and therapeutic areas. The filters come in handy when working by example as they can be applied at various levels and in combinations. The tool is designed to make the review process of the annotated CRF user friendly.

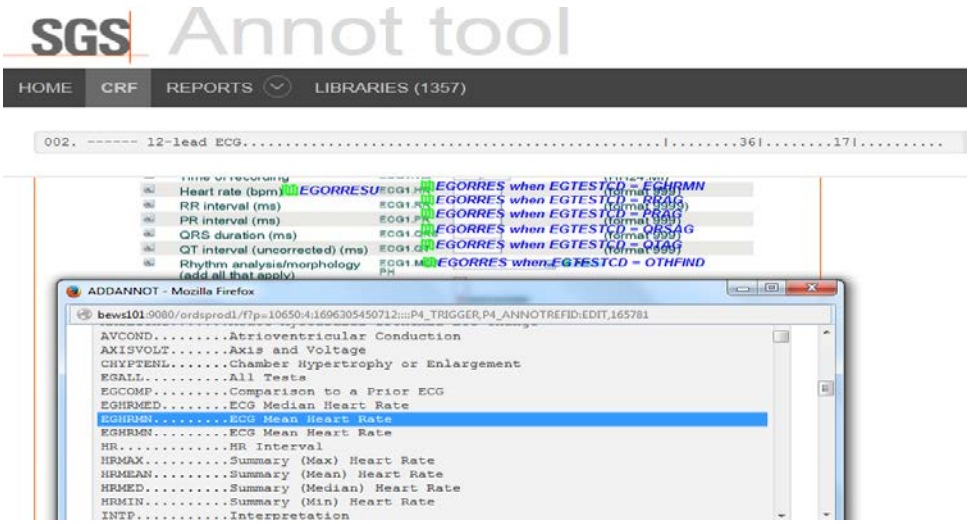
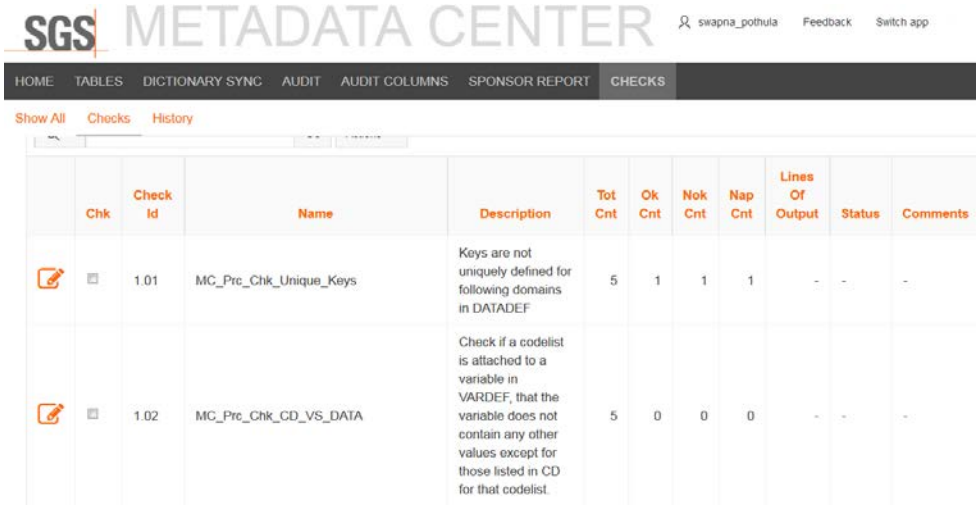


Fig.2 Annotation tool

Metadata Centre

Metadata Centre is another in house tool developed to check for the consistency and accuracy of the trial metadata against all the applicable metadata repositories. This tool is GUI based, integrated into our conversion system available from the moment the trial is being set up in TEST. The checks run on click of a button. All the required parameters such as versions of the applicable standards and therapeutic area for each trial are fetched from the backend.

These checks can be run Ad hoc on their own to provide the output indicating the correctness of the metadata against the current data and applicable standards. Fig.3 illustrates the interactive report generated after the applicable metadata checks have run.



	Chk	Check Id	Name	Description	Tot Cnt	Ok Cnt	Nok Cnt	Nap Cnt	Lines Of Output	Status	Comments
		1.01	MC_Prc_Chk_Unique_Keys	Keys are not uniquely defined for following domains in DATADEF	5	1	1	1	-	-	-
		1.02	MC_Prc_Chk_CD_VS_DATA	Check if a codelist is attached to a variable in VARDEF, that the variable does not contain any other values except for those listed in CD for that codelist	5	0	0	0	-	-	-

Fig.3 Metadata Centre

The output generated by the tool is user friendly and interactive. The programmer is given options to set statuses such as OK, NOTOK and NAP for each of the checks based on the context. The metadata is fine tuned at the trial level using these checks which are simple select statements on the datasets and metadata tables being compared against the applicable standards.

The tool picks the checks applicable for a trial based on all the parameters to generate output at the contextual level. The output not only indicates which check failed, but also provides a brief description of the purpose of the check and a list of records that failed during the check. The tool has quite a lot of functionalities ranging from synchronizing the trial metadata with a new released standard, to generating reports based on the updates made at the trial level in each of the metadata tables making review of metadata efficient both for the programmer and the sponsor.

Pinnacle21

Pinnacle21 is the open source tool used to check for the correctness of the metadata and Define. At SGS, this tool is not used as frequently as we use the 'metadata centre' application. Pinnacle21 is however used ahead of major mile stones to ensure quality.

SDTM Checks

Another in-house tool at SGS is the SDTM checks which run on the data and check for compliance to CDISC SDTM every time there is a new extract from the eSource/(e)CRF and the conversion to SDTM is performed. These checks too are flexible and allow the programmer to add certain exclusions at the trial level. Fig.4 illustrates the SDTM checks output.

SDTM discrepancies - PROD

Drag a column header here to group by that column

Output	SDTM Rule Id	SDTM Rule Label	Domain	Count	Description	Message
output	SDTM_010	SEQ_CONSECUTIVE_SUBJECT	EG_SAS	8	Check that the values of the sequence number variable ([--SEQ]) per subject are unique consecutive numbers starting from 1.	SEQUENCE NUMBER FOR SUBJECT NON-CONSEC
output	SDTM_027	USUBJID_VISITNUM_VISIT_PRES_SV	PC_SAS	5	Check if [VISITNUM] is completed that [USUBJID] - [VISITNUM] - [VISIT] combination exists in [SV].	[USUBJID] - [VISITNUM] - [VISIT] EXPECTED IN S
output	SDTM_044	ORRESU_STRESU_COMPLETED	LB_SAS	2	Check if [--ORRESU] is completed that [--STRESU] is completed.	[--STRESU] VALUE MISSING
output	SDTM_052	ORRES_STAT_NOT_BOTH_EMPTY	IS_SAS	220	Check that [--ORRES] and [--STAT] are not both empty.	[--ORRES] and [--STAT] VARIABLES BOTH EMPT
output	SDTM_052	ORRES_STAT_NOT_BOTH_EMPTY	PC_SAS	999	Check that [--ORRES] and [--STAT] are not both empty.	[--ORRES] and [--STAT] VARIABLES BOTH EMPT
output	SDTM_071	ARMCD_ARM_AVAILABLE_IN_DM	TA_SAS	24	Check that each [ARMCD] - [ARM] combination from [TA] is available in [DM].	[ARMCD] - [ARM] IN [TA] NOT FOUND IN [DM]
output	SDTM_074	ETCD_SCR_PRESENT_TA	TA_SAS	24	Check that at least one record per [ARMCD] with [ETCD] = 'SCR' is present in [TA].	SCR ELEMENT MISSING FOR [ARMCD] IN [TA]
output	SDTM_076	TABRANCH_NOT_COMPLETED	TA_SAS	8	Check if [ETCD] <> 'SCR' that [TABRANCH] is not completed.	[TABRANCH] NOT EXPECTED
output	SDTM_086	ORRES_ISO8601_FORMAT	IE_SAS	2	Check if [--TEST]* contains 'DATE' or 'TIME', or [--TESTCD] ends with 'DTC' or '_D' that the [--ORRES] and [--STRESC] are in ISO8601 --DTC format.	DTC FORMAT IN [--ORRES]/[--STRESC] NOT CO
output	SDTM_087	QVAL_ISO8601_FORMAT	SUPPEC_SAS	26	Check if [QLABEL]* contains 'DATE' or 'TIME', or [QNAM] ends with 'DTC' or '_D' that the [QVAL] is in ISO8601 --DTC format.	DTC FORMAT IN [QVAL] NOT COMPLIANT WITH
output	SDTM_087	QVAL_ISO8601_FORMAT	SUPPEX_SAS	26	Check if [QLABEL]* contains 'DATE' or 'TIME', or [QNAM] ends with 'DTC' or '_D' that the [QVAL] is in ISO8601 --DTC format.	DTC FORMAT IN [QVAL] NOT COMPLIANT WITH
output	SDTM_089	VAL_UPPER_CASE	SUPPEC_SAS	112	Check if no code list is attached to a character variable, that the values of that variable are in upper case.	VALUE NOT IN UPPERCASE
output	SDTM_089	VAL_UPPER_CASE	SUPPEX_SAS	112	Check if no code list is attached to a character variable, that the values of that variable are in upper case.	VALUE NOT IN UPPERCASE

Fig.4 SDTM Checks

The SGS Metadata Workflow

The work flow for metadata creation at SGS is proactive and the deliverables drive the work flow. All of the processes that are part of this work flow are metadata driven. The work flow begins right from the set up of the (e)CRF. The trial programmer is actively involved in the review of the (e)CRF. The compliance of the data that is proposed to be collected, the possible mapping of data and its adherence to standards are paid attention to. Metadata and its usage are the major factors that drive the review process.

Once the (e)CRF is final, the annotations are made in consultation with the MDRs to create SDTM compliant trial specific databases using the annotation tool. The annotation tool is linked to all the applicable MDR's and alerts the programmer of deviations from the standards. Alerts pop up when the programmer for example uses a codelist value that is not part of the CDISC SDTM or sponsor standard. The programmer has to then make a decision at the contextual level after assessing the impact and in consultation with the involved parties to either add the codelist value at the trial level or map it to the closest existing value defined in the applicable standards. On approval and if required these values may be added to the applicable metadata repositories for reuse in other trials.

Ideally metadata should be created before the database is set up. But, this is not always possible and feasible in the practical world. Metadata at SGS is created during the database set up. A set of standard procedures copy all applicable metadata to the trial metadata tables. After this, trial specific procedures are run to set the SGSSTATE variable and also complete the CRF pages where applicable. These procedures need access to information from the annotated CRF therefore all the annotations are made available in the database in the form of a table.

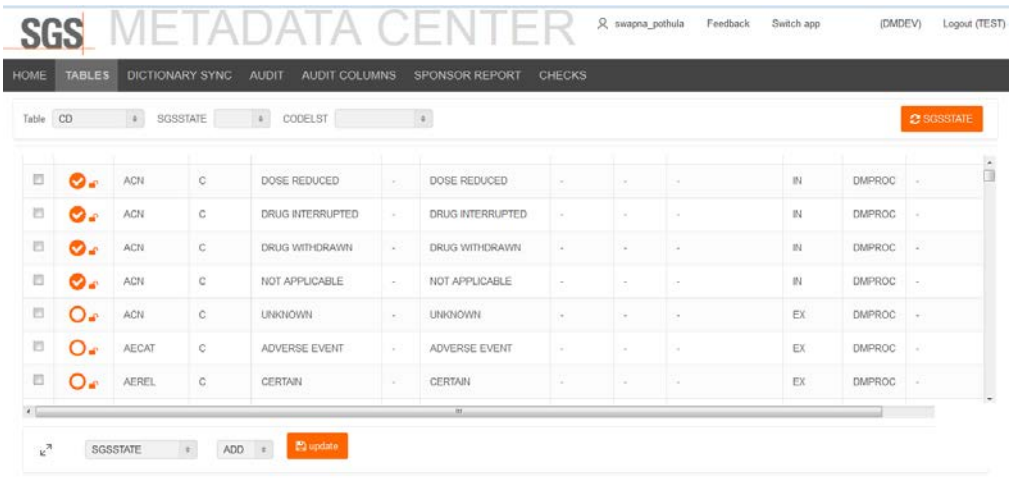


Fig.5 Metadata Centre – IN/EX states

At this point the trial metadata looks like a copy of nested MDRs with SGSSTATE having values IN or EX indicating its inclusion or exclusion in the trial metadata. The metadata views are generated based on these values thereby creating contextual metadata. Fig.5 illustrates the interactive report from the Metadata Centre showing the IN/EX states as assigned by the automated processes with the possibility to change them at the trial level.

It is at this point the metadata checks are run on the trial metadata views (Fig.3) to check for consistency and accuracy. The checks generate output which guides the programmer to further fine tune the metadata at the

contextual level. The output of the checks is interactive and gives the programmer an option to fix them right away if needed, ignore them using NAP when out of context or flag them to be fixed later on. Fig.6 illustrates the status setting functions for a check output on the metadata centre.

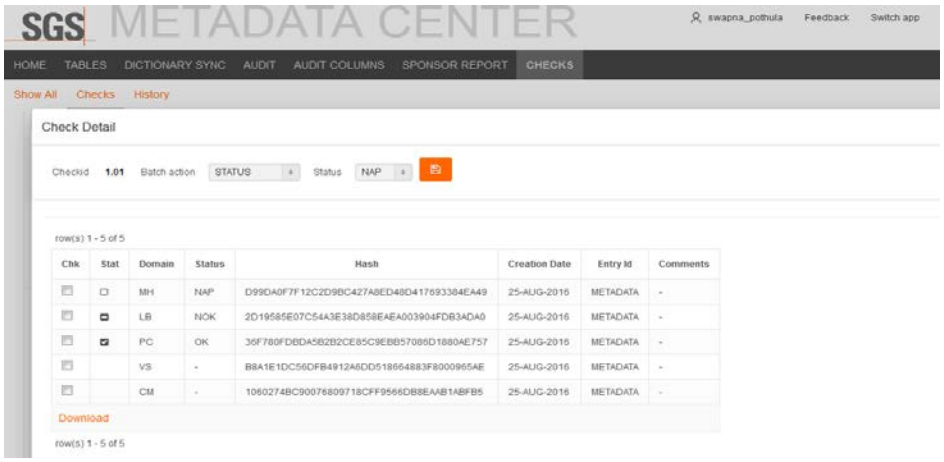


Fig.6 Metadata Centre – OK/NOT/NAP Statuses

The statuses are retained and appear at each run along with the new status should there be a change. This is what gives the programmer a great flexibility and ability to create metadata on the go, on demand and check for compliance continually, even on the day of lock just with a click of a button.

This workflow is a result of concepts and tools implemented at both conceptual and contextual level. The workflow equips and enables programmers at SGS to create and maintain consistent and accurate metadata continually, well almost all of the times.

We do need more than tools

There still are certain situations that arise where we need more than just tools. One such example scenario would be the FDA submission requirement to remove the empty datasets from the submission package. It is not just the exclusion of empty datasets but also exclusion of codelists and valuelists attached to these empty datasets from the trial metadata. While it sounds simple, attention should be paid to codelists that were part of both the empty domains and non-empty domains in the trial database. These codelists should still be retained in the trial metadata and contain only those values that were used in the non-empty domains. How can we check for this and be sure of such an update? The tools can no longer be used to check for these criteria. What we need here would be expertise and experience.

CONCLUSION

There will be more such scenarios and there sure will be need for more than just tools. What we need is a proactive attitude, the so called metadata driven approach, Subject Matter Experts (CDISC/SDTM), Sponsor specific expertise and seamless integration of people and processes that are part of creation, exchange and usage of metadata. More importantly the issues at contextual level need to be communicated and considered while creating or improving processes at the conceptual level**. The key here is to acknowledge that the need for flexibility is as important as is the need for standardization.**

CONTACT INFORMATION

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

Swapna Pothula
SGS – Agriculture, Food and Life
Life Sciences - Clinical Research
Generaal de Wittelaan 19A Bus 5
B-2800 - Mechelen
Email: swapna.pothula@sgs.com
Web: www.sgs.com

Brand and product names are trademarks of their respective companies.