PhUSE US Connect 2019

Paper DH09

ValueListDefs, MethodDefs and WhereClauses: How to Define Data Traceability in a Metadata Repository

Judith Goud, Nurocor, Bennekom, The Netherlands

ABSTRACT

For regulatory agencies it is important to get submissions with clear data traceability to enable verification of the clinical trial results that support the medicinal claim.

CDISC has developed data standards for several clinical data lifecycle stages, and while these standards are aligned they are defined separately for each lifecycle stage. Further they developed the Define-XML standard to support data transportation in a way both machine-readable and interpretable by humans.

Define-XML represents all traceability elements required by regulatory agencies through Item, CodeList, ValueList and Computational Method definitions.

When organizations are using a metadata repository (MDR) to manage their CDISC-based standards it would be beneficial if the MDR can also be used to manage the mappings/transformations between elements from stage to stage and include all elements necessary for submission of a full Define-XML.

INTRODUCTION

Data traceability remains to be a hot topic around the world: a quick search on the internet revealed 106 posts about the topic in the past hour and almost 3000 posts during the last 24 hours. These posts come from all types of industry, from finance to food to pharma to service providers.

The specific interest within the biotechnology and pharmaceutical industry is mainly related to regulatory requirements, however data traceability also allows organizations to be more flexible.

Many publications during the last decade have indicated that to be successful in any market space an organization needs to be able to adapt quickly to change. Applying this to the clinical development lifecycle means that the processes can adapt quickly to changes to the protocol or regulations.

Data traceability can be very useful for this while it can predict very precisely which elements in the data collection, tabulation and analysis are affected by a protocol amendment or need to be modified to satisfy a new or changed regulation.

A good example is when a protocol amendment is approved to add a timepoint to the laboratory measurements after study treatment application. Understanding the data traceability can help a study team determine which tabulation variables and subsequently analysis (datasets) are affected.

As shown in the figure below extra values need to be added to the codelists for LBTPT & LBTPTNUM. This affects LBTPT and LBTPTNUM in the tabulation datasets as well as adds extra lines in the tabulation and analysis datasets and potentially TLFs.

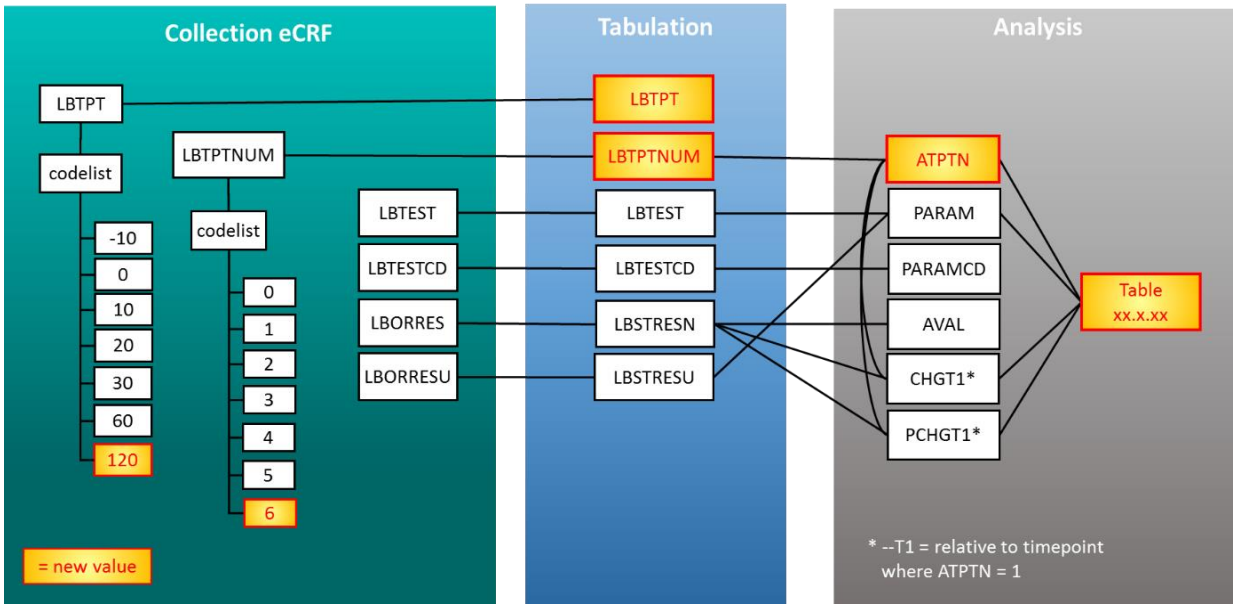


Figure 1 Protocol amendment data traces

Another incentive for biotechnology and pharmaceutical organizations to maintain data traceability is regulatory compliance. For Regulatory agencies it is important to get submission data with clear data traceability so they can verify the clinical trial results that support the medicinal claim. In their guidance documentation they use the concept of traceability, which can be defined as 'providing information how the clinical trial data was collected and what transformations it passed through to get to the trial end results and conclusions'.

The US Food & Drug Administration (FDA) states in their Study Data Technical Conformance Guide[1] in section 4.1.3.2.1:

> "*it is very important that the results presented in the accompanying study report be traceable back to the original data collected*"

Further they dedicate section 8.3 to Study Data Traceability, referencing the

> "*provenance of the data (i.e., traceability of the sponsor's results back to the CRF data)*"

and that for legacy study data (section 8.3.2)

> "*Sponsors should use processes for legacy data conversion that account for traceability*"

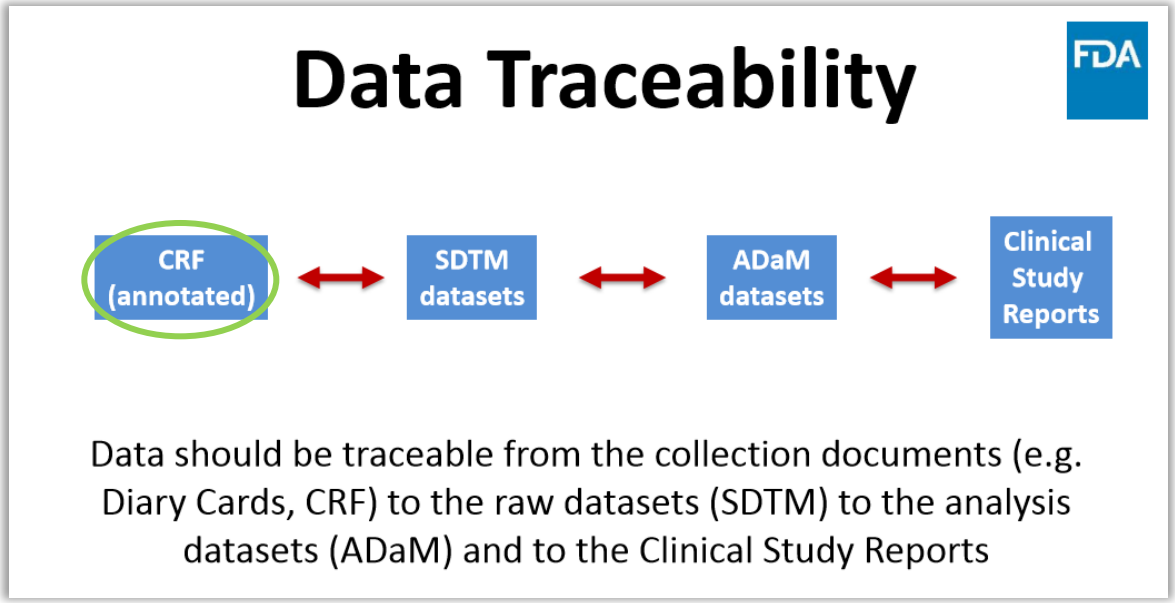
The Japanese Pharmaceutical and Medical Devices Agency (PMDA) Notification on Practical Operations of Electronic Study Data Submission[2] follows the same line of thought and mentions in section 3(1)a(v):

> "*To secure the traceability of data collected in clinical trials, such as from CRFs to the study results for evaluation, it is recommended that the data collected such as from CRFs and other records are summarized into datasets in the SDTM format, and these datasets are used to create analysis datasets in the ADaM format.*
> *When the ADaM datasets are not prepared from the SDTM datasets, …, explain the traceability between the submitted data (such as the procedures in which both datasets were prepared, the relationship between the variables in the database used for the preparation and those in the SDTM and ADaM datasets, and whether there was any information used during the preparation of the ADaM datasets that was not included in the SDTM datasets) in the data guides.*"

Although deemed important, guidance on how to provide this traceability to the regulators is rather limited.
Both the FDA and PMDA only mention the annotated CRF and require this document as part of any submission:

- "*An Annotated Case Report Form (aCRF) is a PDF document that maps the clinical data collection fields used to capture subject data (electronic or paper) to the corresponding variables or discrete variable values contained within the SDTM datasets. Regardless of whether the clinical database is legacy or SDTM compliant, an aCRF should be submitted.*"
  (FDA Study Data Technical Conformance Guide section 4.1.4.6)

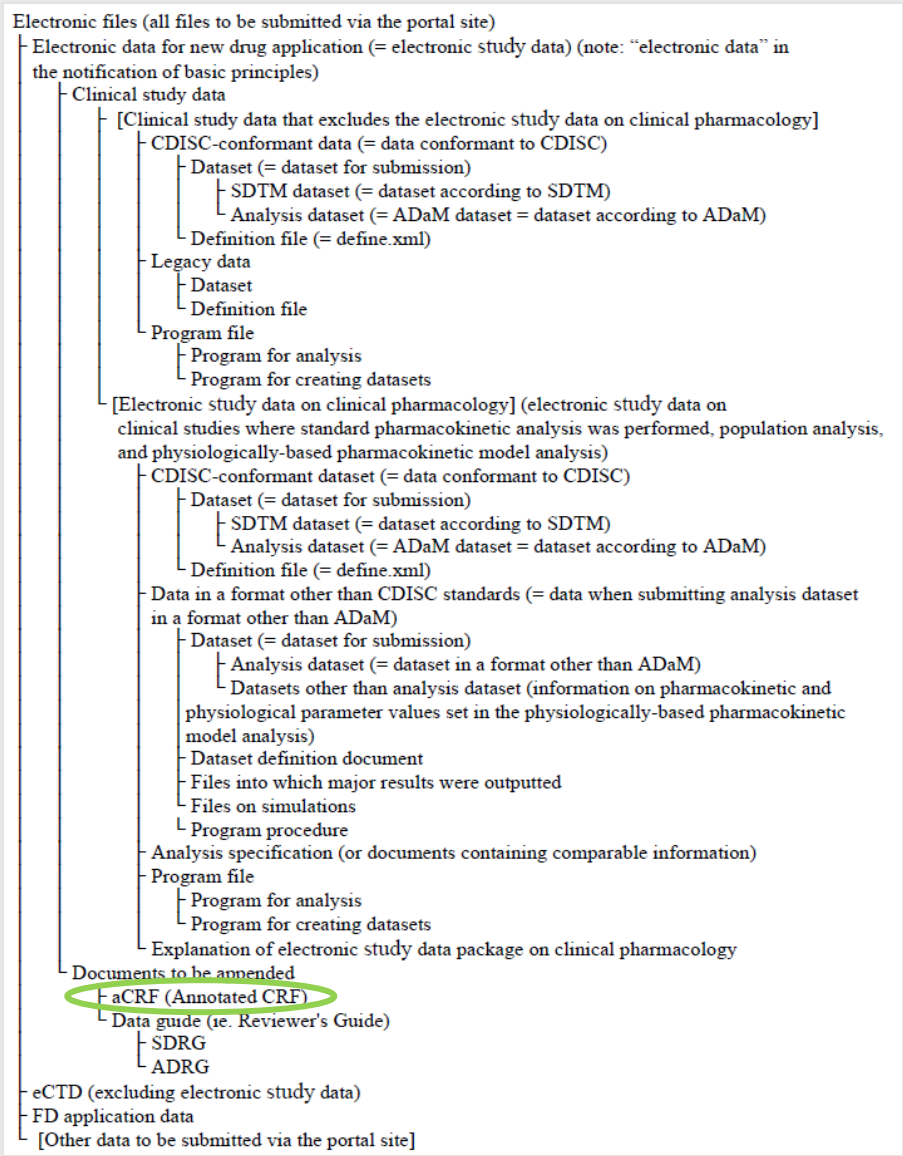
  The following slide from an FDA webinar[3] clearly indicates how the FDA view the annotated CRF as part of data traceability:



Source: CDER SBIA Webinar

- "*In addition to the dataset definition file, submit an annotated CRF demonstrating the relationship between each item of data collected from the CRF and variables included in the dataset.*"
  (PMDA Notification on Practical Operations of Electronic Study Data Submission section 3(1)a(iv))

  The attachment 1 of the PMDA Technical Conformance Guide on Electronic Study Data Submissions lists the aCRF as one of the documents to be appended to the electronic file:



Source: PMDA Technical Conformance Guide Attachment 1

The FDA further mentions that sponsors should provide extensive variable definitions including origin, codelists, derivations and value level metadata.

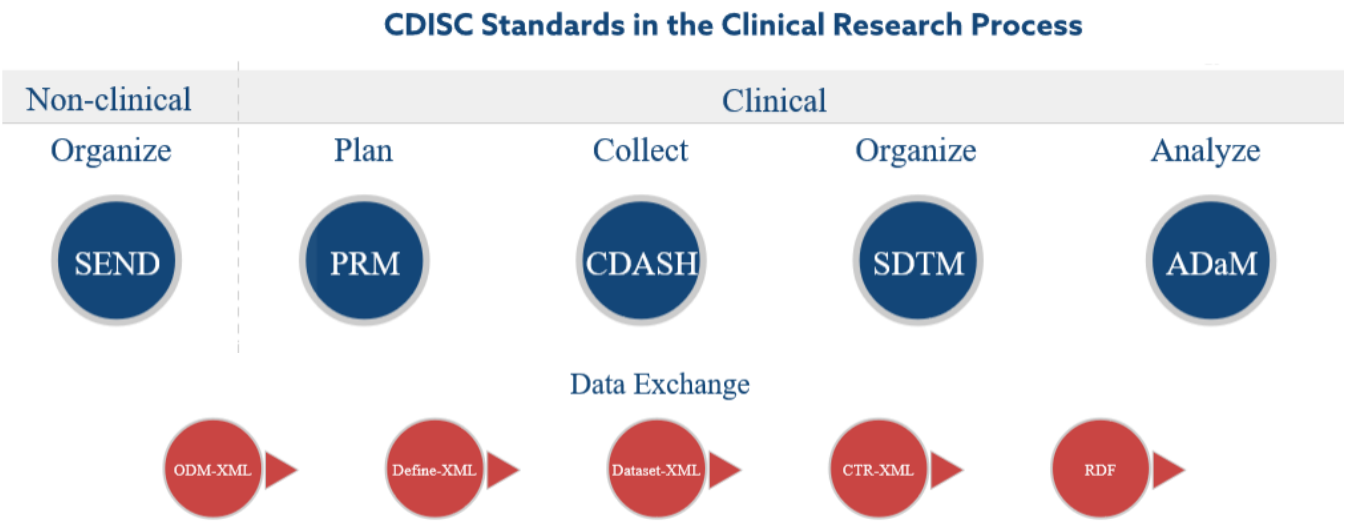
- 4.1.2.2 "*To ensure traceability, all SDTM variables utilized for variable derivations in ADaM should be included in the ADaM datasets when practical.*"
- 4.1.2.6 "*...variables pertaining to the primary and secondary endpoints of a trial, along with their derivations (as applicable), should be provided as well as documented appropriately (i.e., variable-level metadata or parameter value-level metadata) in the data definition file.*"
- 4.1.4.1 "*For the purposes of SDTM and SEND submissions, all Required, Expected, and Permissible variables that were collected, plus any variables that are used to compute derivations, should be submitted.*"
- 4.1.4.5 "*... the sponsor needs to provide complete detail in this file, especially for the specifications pertaining to derived variables. In addition, sponsors should also make certain that the code list and origin for each variable are clearly and easily accessible from the data definition file.*"

STANDARDS

Data traceability is enhanced by using data standards, if you use the same variable name/definition for a certain data collection field across studies chances are the lineage of the data point is also the same.

The Clinical Data Interchange Standards Consortium (CDISC) has developed data standards for several clinical data lifecycle stages, starting at the non-clinical phase and continuing to the analysis results phase.

Further CDISC developed standards for data exchange in a way that is both machine-readable and (through application of a style-sheet) can also be interpreted by humans.



Source: https://www.cdisc.org/standards

Though all CDISC data standards are aligned they are defined separately for each lifecycle stage and most of the Data Exchange standards primarily pertain to a specific clinical lifecycle stage. E.g., the CTR-XML serves as a vehicle to transport trial registry information and Dataset-XML is designed to replace the SAS Transport files for submission.
Therefore, to attain data traceability, additional metadata needs to be defined to bridge the gaps between the lifecycle stages.

The Define-XML standard[4] differs from the other Data Exchange standards in that it is designed to describe submission datasets including data lineage/derivation.
It includes an element for the annotated CRF which provides the traceability between collection and tabulation, while the other elements like Value List Definitions and Computational Methods define any transformations happening from collection to tabulation.
This makes the Define-XML standard the best candidate to satisfy the regulators requirements for data traceability. It provides a framework of what elements need to be captured for presenting the traceability in a standardized way, therefore it can act as a template for the types of metadata to be collected.

DEFINE-XML ELEMENTS FOR DATA TRACEABILITY

The Define-XML standard is an extension from ODM and includes the following elements:

```
<ODM>
  <Study>
  <GlobalVariables>
  <MetaDataVersion>
    <def:AnnotatedCRF>
    <def:SupplementalDoc>
    <def:ValueListdef>
      <ItemRef>
        <def:WhereClauseRef>
    <def:WhereClauseDef>
        <RangeCheck>
          <CheckValue>
    <ItemGroupDef>
      <ItemRef>
    <ItemDef>
      <CodeListRef>
      <def:Origin>
        <def:DocumentRef>
          <def:PDFPageRef>
      <def:ValueListRef>
    <CodeList>
    <MethodDef>
    <def:CommentDef>
    <def:Leaf>
```

* elements indicated in blue are additional compared to ODM

Source: CDISC Define-XML Specification Version 2.0 section 6 Global Element Ordering

The additional elements compared to ODM all have a function regarding data traceability:

- <def:AnnotatedCRF>
  This element holds the annotated CRF which is defined as a “Portable File Format (PDF) document that provides the mapping of data collection fields to the variables or discrete variable values contained within the datasets.” It provides the mapping of data collection fields to the variables in the datasets. It is required when any variable has the origin ‘CRF’
- <def:SupplementalDoc>
  This element can be used to submit supplemental documentation for the mapping of data that was not collected on a CRF, e.g. laboratory or ECG data.
- <def:ValueListdef>
  <ItemRef>
  <def:WhereClauseRef>
  <def:WhereClauseDef>
  <RangeCheck>
  <CheckValue>
  The value list element in combination with the where clause element describes the value level metadata.
  The CDISC Define-XML Specification version 2.0 section 4.4.1 explains value level metadata as that
  “Each Value definition may have a Where Clause attached to it to describe when that Value applies.
  Where Clauses define a condition by using one or more Range Checks.”
- <def:Origin>
  <def:DocumentRef>
  <def:PDFPageRef>
  The origin element provides machine-readable connections between variables, for example, ‘Predecessor’ is used to link ADaM data back to SDTM variables to establish traceability and when variables or values are defined as having the origin ‘Derived’ a MethodDef element must be provided.
- <def:ValueListRef>
  This element ties the ValueList to the field/variable.
- <def:CommentDef>
  The Comment element is made explicit compared to ODM and can be used at any level to provide additional information for the reviewer. At the ItemGroup level is may for example indicate datasets have been split or joined and at the ItemDef level it can provide information about the codelist mapping or a reference to the Reviewers Guide.

EXAMPLE OF A DERIVED VARIABLE

In some countries for privacy reasons clinical studies are not allowed to collect the full birth date of patients.
To accommodate this, the CDISC collection standard defines separate fields for the birth day, month and year. For tabulation the collected fields need to be concatenated/derived into the BRTHDTC field.

Define-XML will hold this information as follows:

- The <def:AnnotatedCRF> will hold a PDF document containing the following page:

Birth Month and Year **BRTHDTC** **BRTHMO** **BRTHYY**	_ _ / _ _ _ _
Sex **SEX**	○ Female ○ Male
Race **RACE**	☐ American Indian or Alaska Native ☐ Asian ☐ Black or African American ☐ Native Hawaiian or Other Pacific Islander ☐ White
Ethnic **ETHNIC**	○ Hispanic or Latino ○ Not Hispanic or Latino ○ Not Reported

```
<ItemGroupDef OID="IG.DM" Domain="DM" Name="DM"
  <Description>
    <TranslatedText xml:lang="en">Demography</TranslatedText>
  </Description>
  <ItemRef ItemOID="IT.DM.BRTHDTC" OrderNumber="1" Mandatory="Yes"
           MethodOID="MT.BRTHDTC"/>
  …
</ItemGroupDef>
```

```
<ItemDef OID="IT.DM.BRTHDTC" Name="BRTHDTC" DataType="date" Length="20"
         SASFieldName="BRTHDTC">
```

```
    <Description>
      <TranslatedText xml:lang="en"> Date/Time of Birth</TranslatedText>
    </Description>
    <def:Origin Type="Derived">
  </ItemDef>
```

- ```
  <MethodDef OID="MT.BRTHDTC" Name="Algorithm to derive BRTHDTC"
             Type="Computation" >
    <Description>
      <TranslatedText xml:lang="en">Derived from BRTHMO and BRTHYY, appending '01'
               for the day and concatenation into ISO 8601 format.</TranslatedText>
    </Description>
  </MethodDef>
  ```

EXAMPLE OF VALUE LEVEL METADATA

The normalized data structure used by datasets based on the SDTM (one record per subject per test code per visit or observation) is efficient for transmitting information. However, in many cases the structure is de-normalized for collection to facilitate data entry and allows for differentiation between tests, e.g. specification of default units or associating a subset unit codelist.

Define-XML will provide information about the traceability from the denormalized to normalized data as follows:

- The `<def:AnnotatedCRF>` will hold a PDF document containing the following page:

Date **VSDTC** **VSDAT**	_ _ / _ _ / _ _ _ _
Time **VSDTC** **VSTIM**	_ _ : _ _
Diastolic Blood Pressure **VSORRES where VSTESTCD = "DIABP"** **DIABP**	___
Diastolic Blood Pressure Unit **VSORRESU where VSTESTCD = "DIABP"** **DIABPU**	mmHg
Systolic Blood Pressure **VSORRES where VSTESTCD = "SYSBP"** **SYSBP**	___
Systolic Blood Pressure **VSORRESU where VSTESTCD = "SYSBP"** **SYSBPU**	mmHg
Height **VSORRES where VSTESTCD = "HEIGHT"** **HEIGHT**	___
Height Unit **VSORRESU where VSTESTCD = "HEIGHT"** **HEIGHTU**	○ cm ○ in
Weight **VSORRES where VSTESTCD = "WEIGHT"** **WEIGHT**	___
Weight Unit **VSORRESU where VSTESTCD = "WEIGHT"** **WEIGHTU**	○ kg ○ LB

- ```
  <def:ValueListDef OID="VL.VS.VSORRES">
    <ItemRef ItemOID="IT.VS.VSORRES.DIABP" OrderNumber="1" Mandatory="Yes">
      <def:WhereClauseRef WhereClauseOID="WC.VS.VSTESTCD.DIABP"/>
    </ItemRef>
    <ItemRef ItemOID="IT.VS.VSORRES.SYSBP" OrderNumber="2" Mandatory="Yes">
      <def:WhereClauseRef WhereClauseOID="WC.VS.VSTESTCD.SYSBP"/>
    </ItemRef>
    <ItemRef ItemOID="IT.VS.VSORRES.HEIGHT" OrderNumber="3" Mandatory="Yes">
      <def:WhereClauseRef WhereClauseOID="WC.VS.VSTESTCD.HEIGHT"/>
    </ItemRef>
    <ItemRef ItemOID="IT.VS.VSORRES.WEIGHT" OrderNumber="4" Mandatory="Yes">
      <def:WhereClauseRef WhereClauseOID="WC.VS.VSTESTCD.WEIGHT"/>
    </ItemRef>
  </def:ValueListDef>
  ```

```
<def:WhereClauseDef OID="WC.VS.VSTESTCD.DIABP">
  <RangeCheck SoftHard="Soft" def:ItemOID="IT.VS.VSTESTCD" Comparator="EQ">
    <CheckValue>DIABP</CheckValue>
  </RangeCheck>
</def:WhereClauseDef>
<def:WhereClauseDef OID="WC.VS.VSTESTCD.SYSBP">
  <RangeCheck SoftHard="Soft" def:ItemOID="IT.VS.VSTESTCD" Comparator="EQ">
    <CheckValue>SYSBP</CheckValue>
  </RangeCheck>
</def:WhereClauseDef>
<def:WhereClauseDef OID="WC.VS.VSTESTCD.HEIGHT">
  <RangeCheck SoftHard="Soft" def:ItemOID="IT.VS.VSTESTCD" Comparator="EQ">
    <CheckValue>HEIGHT</CheckValue>
  </RangeCheck>
</def:WhereClauseDef>
<def:WhereClauseDef OID="WC.VS.VSTESTCD.WEIGHT">
  <RangeCheck SoftHard="Soft" def:ItemOID="IT.VS.VSTESTCD" Comparator="EQ">
    <CheckValue>WEIGHT</CheckValue>
  </RangeCheck>
</def:WhereClauseDef>
```

-
```
<ItemGroupDef OID="IG.VS" Domain="VS" Name="VS"
  <Description>
    <TranslatedText xml:lang="en">Vital Signs</TranslatedText>
  </Description>
  <ItemRef ItemOID="IT.VS.VSORRES" OrderNumber="3" Mandatory="Yes"/>
  …
</ItemGroupDef>
```

-
```
<ItemDef OID="IT.VS.VSORRES" Name="VSORRES" DataType="text" Length="30"
         SASFieldName="VSORRES">
  <Description>
    <TranslatedText xml:lang="en">Result or Finding in Original Units</TranslatedText>
  </Description>
  <def:Origin Type="CRF">
    <def:DocumentRef leafID="LF.blankcrf">
      <def:PDFPageRef PageRefs="11" Type="PhysicalRef"/>
    </def:DocumentRef>
  </def:Origin>
  <def:ValueListRef ValueListOID="VL.VS.VSORRES"/>
</ItemDef>
```

DATA TRACEABILITY IN CLINICAL DEVELOPMENT SYSTEMS

To date most organizations manage their mappings, derivations and value level metadata in spreadsheets.
However, spreadsheets do not support things like change control, versioning, business rules, impact analysis or reuse of elements. All these things need to be accomplished through manual review, creating multiple copies of spreadsheets (and keeping track of them) and team consultation. This introduces human error and creates extra work for study teams to perform manual impact analysis and tracking of where changes need to be implemented.

Increasingly more organizations have recognized an MDR is better suited for these activities and can serve as the single source or truth within an organization. They are managing their standards and study specifications in a central MDR, however these mainly concern standalone lifecycle-stage specific standards. To facilitate clinical trial submissions, it would be beneficial when this MDR can also be used to manage the data traceability from stage to stage.

The advantage of managing the traceability/mapping together with the source and target standards in an MDR is that it can be validated once and reused across studies. When source or target standards are then reused the traceability/mapping comes along, helping study teams to reach (or improve) their target timelines.
The elements from Define-XML mentioned in the previous sections can be used to model traceability into an MDR, allowing the MDR to provide the data traceability information as necessary for the regulatory authorities.

DEFINE-XML ELEMENTS IN A METADATA REPOSITORY

The starting point for regulatory reviewers to look at traceability is probably the submission variable origin. For example, when the origin is Derived they expect a Method in the define-XML and when origin is CRF they would like to know where and how it was represented on the collection CRF.
The define-XML def:Origin element prescribes the values the MDR will need to allow/store for a variable origin. The MDR can use these into a value list to manage entry of correct values for origin.
Further the MDR should allow reuse of the same standard variable across studies while allowing the origin to be defined in the context of a study, e.g. a variable can be collected in one study but may be derived for another study.

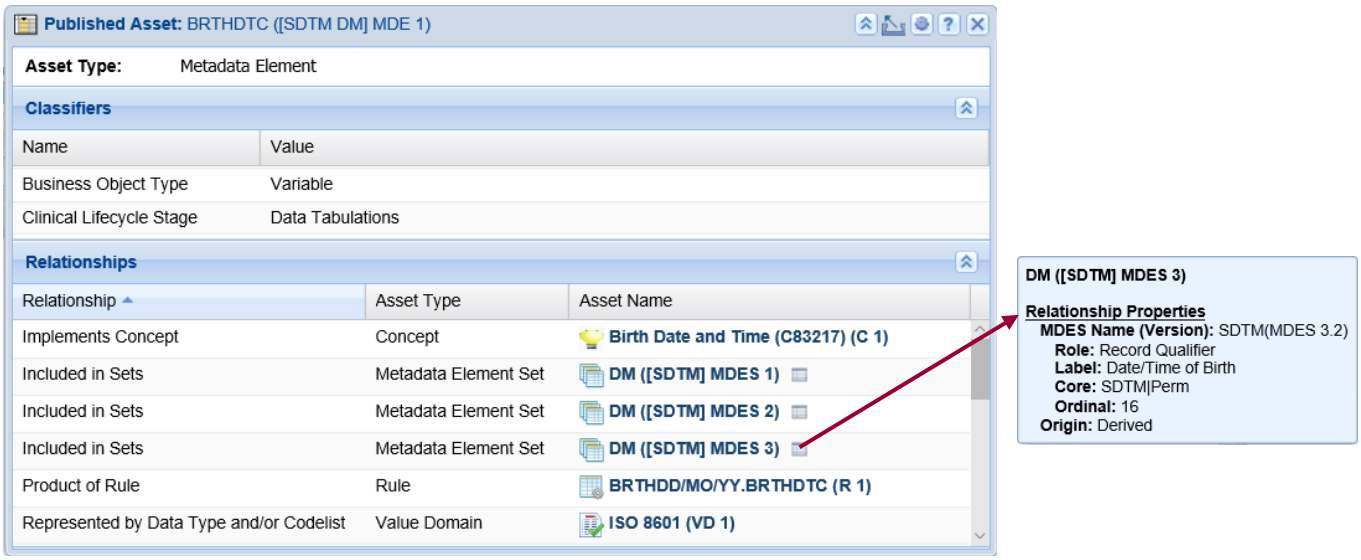


Figure 2 Variable with context dependent Origin

The (reference to an) annotated CRF can to be stored in the MDR within a study specification to populate the def:AnnotatedCRF element. If all metadata around data collection (e.g. CRF setup and layout) and traceability is captured in the MDR, this system can be used to produce an annotated CRF. The same applies for documentation candidate for the def:SupplementalDoc element.

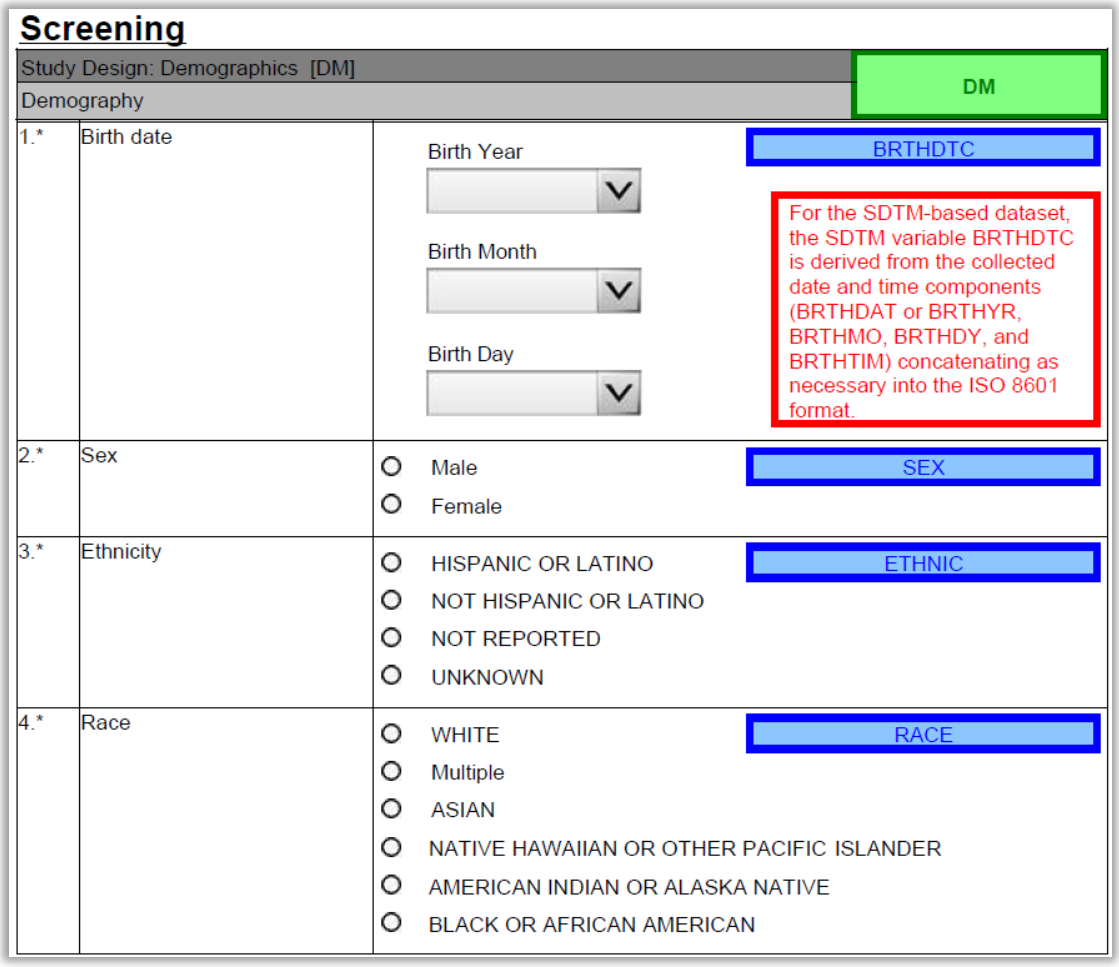
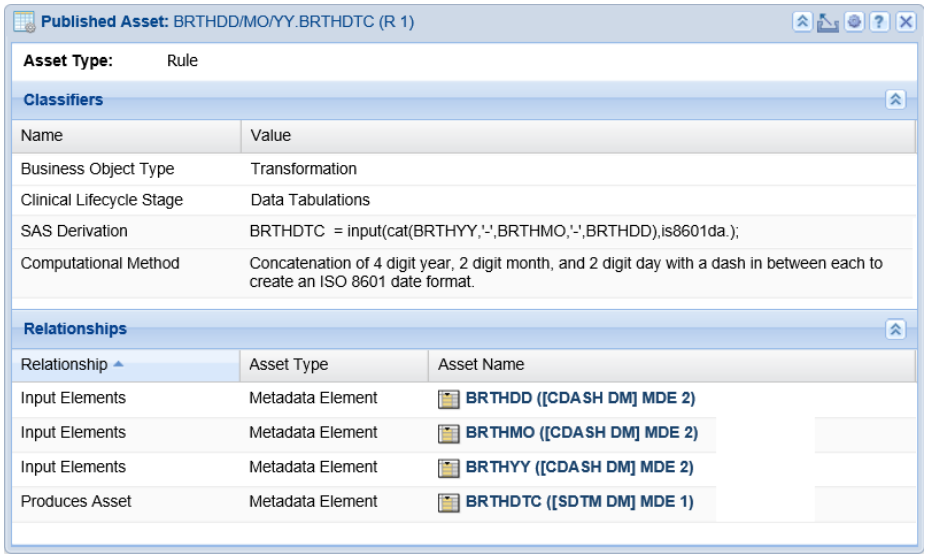


Figure 3 Annotated Demography CRF sample report

Definitions for the derivation are described in the MethodDef element. According to the define-XML standard at least the derivation Type, a Name and a clear Description should be presented about how derived variables are constructed.

The Define-XML allowable values for Type are ‘Computation’ and ‘Imputation’. This indicates the main concern from the regulatory reviewer is to know whether a value was derived based on other values collected at the same time point or conventions that are applied based on values from another timepoint. The MDR may support other values for derivation type to satisfy other use cases (e.g. categorization for standardization or execution), then for production of a compliant define-XML these values will need to be categorized into Computation and Imputation.

Derivation names need to be unique across the submission, which can be ensured when the MDR does not allow duplicate names. The MDR should also enable the name to be associated with a type and a clear method description (or reference to a document storing the description). For internal processing it can be useful to store executable (pseudo-)code with the name and description while then the code can be validated once and reused between studies.



Published Asset: BRTHDD/MO/YY.BRTHDTC (R 1)

Asset Type: Rule

Classifiers

Name	Value
Business Object Type	Transformation
Clinical Lifecycle Stage	Data Tabulations
SAS Derivation	BRTHDTC = input(cat(BRTHYY,'-',BRTHMO,'-',BRTHDD),is8601da.);
Computational Method	Concatenation of 4 digit year, 2 digit month, and 2 digit day with a dash in between each to create an ISO 8601 date format.

Relationships

Relationship	Asset Type	Asset Name
Input Elements	Metadata Element	BRTHDD ([CDASH DM] MDE 2)
Input Elements	Metadata Element	BRTHMO ([CDASH DM] MDE 2)
Input Elements	Metadata Element	BRTHYY ([CDASH DM] MDE 2)
Produces Asset	Metadata Element	BRTHDTC ([SDTM DM] MDE 1)

Figure 4 Derivation example

When transformational metadata is captured in relationship to input and output variables, the content can be visualized as below to facilitate review of the traceability.

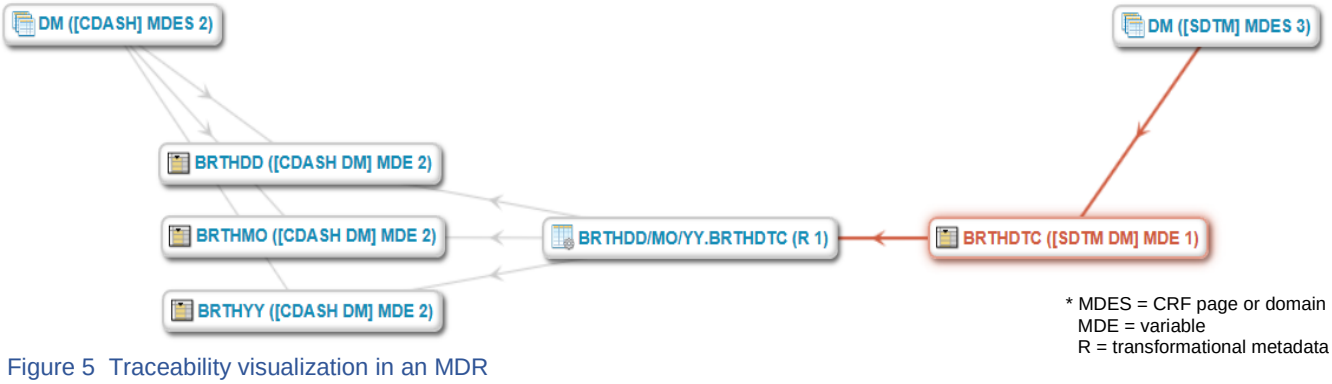


Figure 5 Traceability visualization in an MDR

The next element define-XML provides information about are the requirements to adequately describe Value Level Metadata.

> “Value Level Metadata should be provided when there is a need to describe differing metadata attributes for subsets of cells within a column. It is most often used on SDTM Findings domains to provide definitions for Variables such as --ORRES, --ORRESU, --STRES, --STRESU that are specific to each test code (value of --TESTCD)”.

The following page shows an example where a single collection variable induces values in multiple submission variables, this is commonly known a mapping a horizontal collection structure to a vertical (normalized) submission structure.

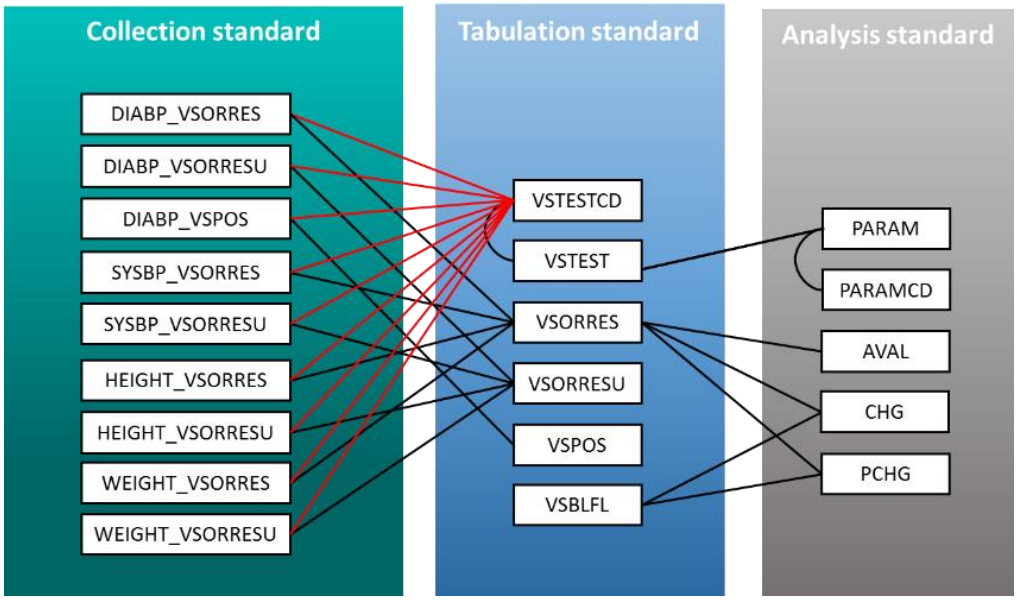
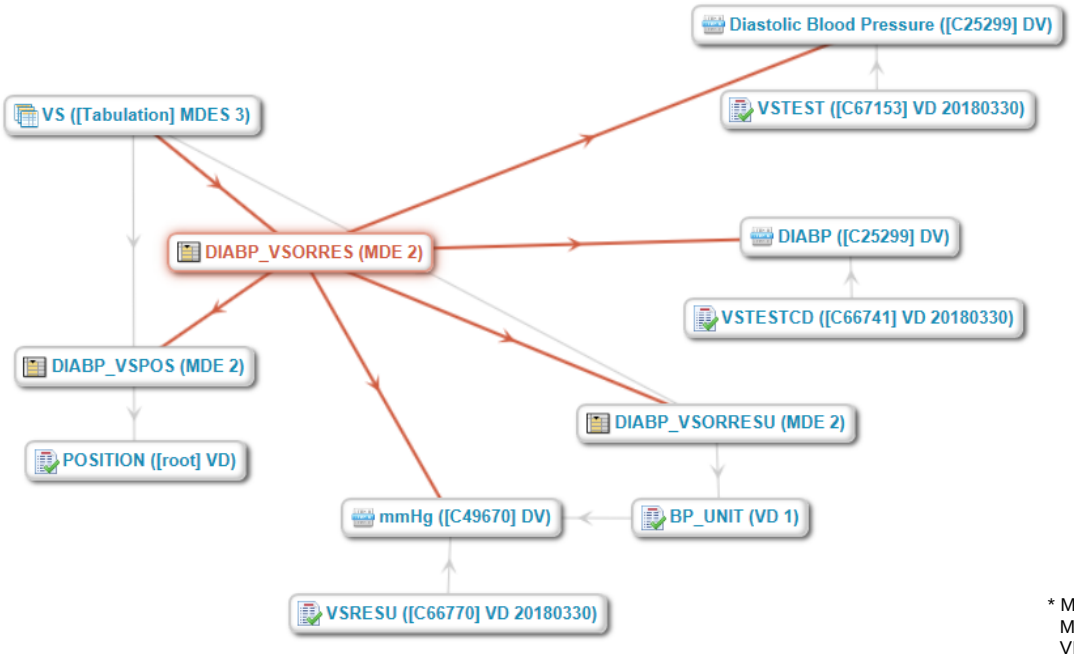


Figure 6 Denormalized (horizontal) to normalized (vertical) structure mapping

According to Define-XML the MDR could/should capture references between the source and target structures including:

- whether the target variable is mandatory according to SDTM (Core = Req (required)) or the organization,
- the order number for value list variables,
- the constraint applicable for each of the values in the list, including a comparator like Equal or Greater Than, one or more check values and whether failing the constraint leads to a rejection (hard) or a warning (soft).

To facilitate reuse of collection, tabulation and transformational metadata in the MDR the attributes above can be captured using references/relationships with context dependent values.
The (SDTM) Core attribute is best captured in the context of a domain, while a variable could be required in one domain and expected or permissible in another domain. The same applies for the order of value list variables that would be captured in relationship to the CRF page.
The constraint information can be stored on the relationship between the collection and tabulation variable or, as many organizations do, be managed using so-called lookup-tables which define the controlled terminology values to be assigned to tabulation variables based on the definition of the collection variable. This lookup information can be captured in the MDR as relationships between the collection variable and the values to be assigned. A generic derivation can then use these links to populate the tabulation variables with the correct values.



* MDES = CRF page or domain
MDE = variable
VD = codelist
DV = codelist entry

Figure 7 Lookup relationships in an MDR

Finally, the MDR should accommodate the provision of Comments on VLM, Domains/Datasets and Variables.
The MDR should provide the possibility to capture comments on each element in the system, to enable clarifications on usage, FAQ, or other information. It is advisable to have a specific attribute to capture comments for define-XML while not all comments are necessary to be submitted.

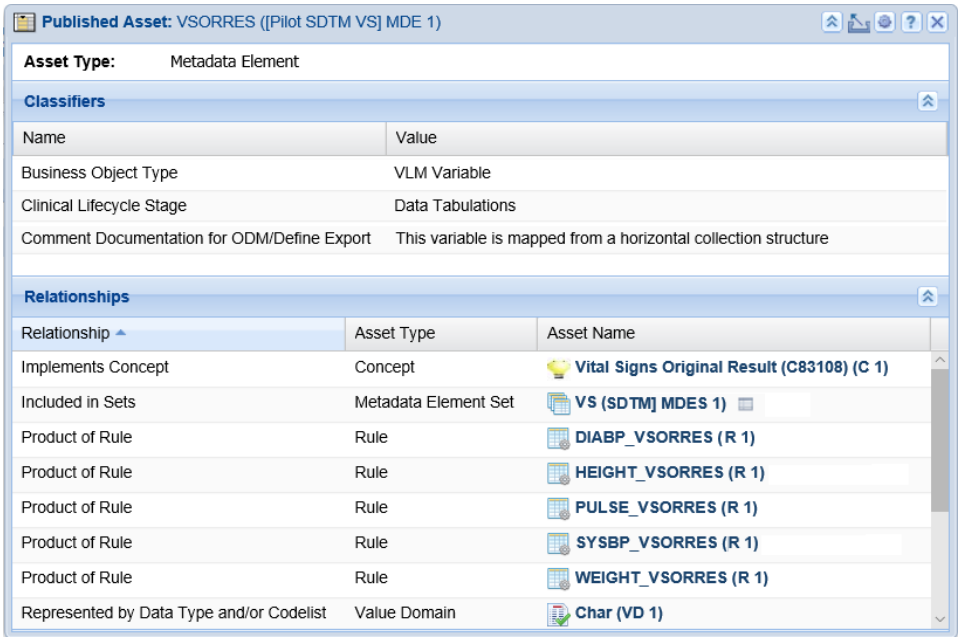


Figure 8 Comment storage example

CONCLUSION

Currently the majority of Biotechnology & Pharma companies manage their metadata using spreadsheets, that do not support control and versioning of data traceability, which must be done manually through review, creating copies (and keeping track of them) and team consultation.
A metadata repository is better suited for these activities and can serve as the single source or truth within an organization.
The repository should hold all information about data element structure and meaning and also enable review of all references and transformations the data element was subject to.
Define-XML provides a good framework of what information should be captured in the MDR to satisfy regulatory reviewers requirements for data traceability by defining the origins, derivation and value level metadata attributes.
These elements should be included in the MDR as context-dependent attributes to facilitate reuse of core variables.

REFERENCES

1. FDA, Study Data Technical Conformance Guide v.3.3, 23 March 2017, https://www.fda.gov/downloads/forindustry/datastandards/studydatastandards/ucm384744.pdf
2. PMDA, Technical Conformance Guide on Electronic Study Data Submissions, 27 April 2015, https://www.pmda.go.jp/files/000206449.pdf
3. Slide from CDER Small Business and Industry Assistance (SBIA) Webinar, 13 July 2017 https://www.fda.gov/Drugs/DevelopmentApprovalProcess/SmallBusinessAssistance/ucm565138.htm
4. CDISC, Define-XML v2.0, 24 April 2014, https://www.cdisc.org/standards/data-exchange/define-xml

CONTACT INFORMATION

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

Judith Goud
Nurocor, Inc.
106 East 6th Street
Suite 900
Austin, Texas 78701
Judith.Goud@nurocor.com
www.nurocor.com