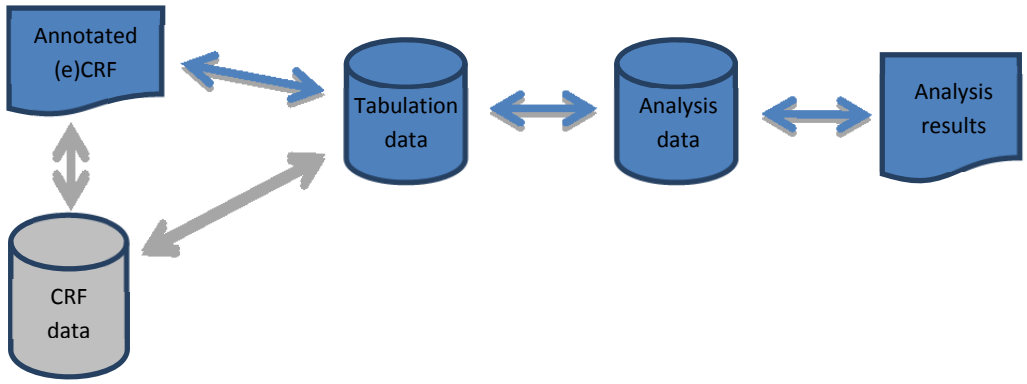
Traceability and Data Flow
Best Practices for Basic Linear Data Flow

01 October, 2014

Scope:

The purpose of this white paper is to describe traceability needs pertinent to regulatory agency review of electronic clinical data. The traceability discussed in this white paper covers the linear data flow of study data from annotated (e)CRF/collected data, through tabulation data and analysis data, and ending at analysis results, as depicted in the diagram below.



For a submission, you will include only the annotated (e)CRF, tabulation data, analysis data, and analysis results, shown above in blue. These components therefore comprise the scope of this document.

We have included in this diagram the CRF (raw) data component to show the entire flow, both internal and external. The internal flow includes the (e)CRF annotated with the CRF data variable names, and the mapping from CRF data to tabulation data, as shown in grey above. In this paper we will not speak any further to the CRF data as part of the data flow, as this not part of the scope. We will instead focus only on the portions shown in blue.

Best practices for traceability will be outlined by defining **what** traceability is, **why** it is important, **how** it can be enacted and **where** more resources for traceability practices can be found.

Definitions and Abbreviations:

The terms and abbreviations defined here are specific to the use cases outlined in this white paper.

- **(e)CRF:** (electronic) Case Report Form.
- **ADaM**: Analysis Data Model. This is the CDISC standard for analysis data.
- **ADRG**: Analysis Data Reviewers Guide. Additional information about this document can be found within the CSS wiki.
- **Analysis data:** Data structured to enable statistical analyses. For example, CDISC compliant analysis data would utilize the ADaM standard.

- **Analysis results:** summary output generated from analysis data. In a submission package, this is the Clinical Study Report (CSR).
- **Annotated (e)CRF:** Version of the (e)CRF annotated with collected and/or tabulation data attributes.
- **CDASH**: Clinical Data Acquisition Standards Harmonization. This is the CDISC standard for collecting clinical data.
- **CDISC**: Clinical Data Interchange Standards Consortium.
- **CRF data:** Data as collected on the (e)CRF. For example CDISC compliant collected data would utilize the CDASH standard.
- **Data transformation:** the process of converting a set of data values from the data format of a source data system into the data format of a destination data system. When we use the term "transformation", we are not referring to a mathematical transformation.
- **define file:** A document containing metadata, submitted in addition to data, usually in the form of an XML or PDF file. Metadata are usually in the form of tables.
- **define.pdf**: A human-readable standard structure for submitting and viewing metadata.
- **define.xml**: A machine-readable standard structure for submitting metadata using the eXtended Markup Language. A schema is used along with the define.xml to display the data in a more human-readable form.
- **Metadata:** Information about data.
- **Regulatory Agency:** A government body responsible for review of submitted study materials.
- **Reviewer:** a user, regulatory or otherwise, who was not involved in the study and is trying to understand and possibly validate the analysis.
- **SDRG**: Study Data Reviewers Guide. Additional information about this document can be found within the CSS wiki.
- **SDTM**: Study Data Tabulation Model. This is the CDISC standard for tabulation data.
- **Tabulation data:** A systematic arrangement of the collected data. For example, CDISC compliant tabulation data would utilize the SDTM standard.

What:

Traceability is the property that enables the understanding of the data's lineage and/or relationship between an element and its predecessor(s). This ultimately results in representing the origin of the data (eCRF) through its final presentation (analysis results), including all the steps in between.

There are two types of traceability commonly used with clinical data: metadata traceability and data-point traceability. The ADaM document[1] defines

- **Metadata traceability** as "describing (via metadata) the algorithm used or steps taken to derive or populate an analysis value from its immediate predecessor"
- **Data-point traceability** as enabling "the user to go directly to the specific predecessor record(s)"

Traceability and Data Flow
Best Practices for Basic Linear Data Flow

01 October, 2014

Why:

Traceability enables the reviewer to easily follow the value of a data point through each step of the data flow and understand the following data and relationships:

- information that exists on the (e)CRF
- information found in the tabulation datasets
- information that is derived within analysis datasets
- the method used to create derived or imputed values
- the data used to create the analysis results

When traceability is successfully implemented, it demonstrates transparency to the reviewers. For example, it allows someone to trace back from the analysis results to the analysis dataset, from the analysis dataset back to the tabulation dataset, and from the tabulation dataset to the annotated CRF.

How:

Traceability is accomplished by clearly establishing the path between an element and its immediate predecessor, ultimately back to the annotated (e)CRF. A combination of metadata traceability and data point traceability enable this.

Metadata traceability is established by describing (via metadata) the algorithm used or steps taken to derive or populate an analysis value from its immediate predecessor. Some recommended best practices for establishing metadata traceability include the following:

- Annotate the blank (e)CRF with the tabulation dataset's variable names and attributes.
- If records are derived in a dataset, create variables to identify them as such. For example, as described in the CDISC SDTM and ADaM documents[1] respectively, derived records in SDTM require the use of the --DRVFL variable, and in ADaM require the use of DTYPE and/or PARAMTYP. This practice of differentiating collected data from derived data will ensure that these records are understood and used correctly.
- Create intermediate datasets to help describe complex derivations in analysis data.
- In the Define file, include details about algorithms used to derive data. Define files can include data-level, variable-level, value-level, and results-level metadata. CDISC[1] provides a lot of material on how to create these files for the standard SDTM and ADaM data.
- Provide additional details in the Study Data Reviewers Guide[2] (SDRG) and Analysis Data Reviewers Guide[3] (ADRG). These documents are particularly useful when describing details that do not translate well to the two-dimensional tables of the define file.

Metadata traceability can also reference other documents. For example, the define file or reviewers guide might mention or link to a specific section of the Protocol or Statistical Analysis Plan. This traceability adds details to help explain what was done and why.

Traceability and Data Flow
Best Practices for Basic Linear Data Flow

01 October, 2014

Data point traceability is the process of keeping and creating variables to provide direct and exact links from one dataset to another dataset or from a dataset to an analysis result. Some recommended best practices for establishing data point traceability include the following:

- Minimize the amount of data transformations across the data flow. For example, in the (e)CRF utilize the same values for subject discontinuation as will be used for analysis. Using controlled terminology from start to finish can help minimize transformations. The CDISC Controlled Terminology[1] team has harmonized terminology across all the CDISC standard models.
- Incorporate the use of data point traceability variables in analysis datasets. A common way to do this is to include the sequence number of the record that was used from the earlier dataset to produce the record in the later dataset. For example, the Basic Data Structure (BDS) described in the ADaM Implementation Guide[1] provides guidance on the use of data point traceability variables to trace back to SDTM datasets.
- Include additional variables that are useful for traceability. The sole purpose of some analysis data variables and rows can be to provide data point traceability. For example, when a vital sign date is imputed to create ADaM variables ADT and ADTF, the original SDTM variable VSDTC, which is not used for analysis, can be included to provide data point traceability.
- Use informative variable labels for variables which do not follow a pre-defined standard convention. Whenever possible, utilize the free text in labels to add additional information about traceability.
- Use analysis flag variables to differentiate data used for analysis from data used for traceability and/or to show connectivity when multiple records are used to derive one record. For example, ADaM uses ANLzzFL as the analysis flags.
- Utilize the numeric analysis data visit variable to match actual time points which might be used in analysis results. Instead of just an ordinal number used for sorting, a numeric analysis visit variable could also be used to support graphical representation. For example, in ADaM, if AVISIT is (Week 1, Week 2, Week 4, Month 2, Month 3), set AVISITN to (1, 2, 4, 8, 12) rather than just simply ordering as (1, 2, 3, 4, 5).

Much of the traceability, as described above, can be created just prior to submission. However, it is more efficient to develop traceability at the same time as developing the data. When traceability is built in as an inherent property at every step, it does not require separate programming or other steps at the time of submission.

Where:

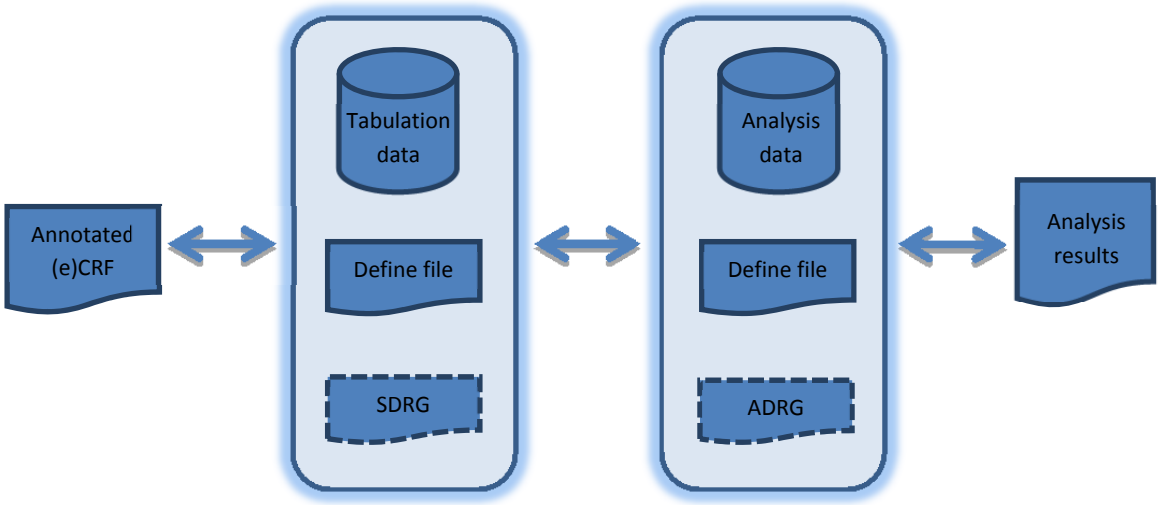
The references mentioned in this white paper can be found at the following links:

1. The CDISC material that can be downloaded for free, including the ADaM document, ADaMIG, SDTM document, SDTMIG, Controlled Terminology for all models, and define.xml, is available on the CDISC website at: http://www.cdisc.org.

2. The Study Data Reviewers Guide (SDRG) template, examples, and instructions can be downloaded for free from the PhUSE CSS Working Groups wiki site at: http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide.
3. The Analysis Data Reviewers Guide (ADRG) template, examples, and instructions can be downloaded for free from the PhUSE CSS Working Groups wiki site at: http://www.phusewiki.org/wiki/index.php?title=Analysis_Data_Reviewer%27s_Guide.
4. Additional references can be found on the PhUSE CSS Working Groups wiki site at: http://www.phusewiki.org/wiki/index.php?title=Summary_of_Traceability_References. This list of traceability references is maintained by the Traceability and Data Flow project within the Optimizing Data Standards working group.

Summary:

In a submission, the following set of material would not only provide the required data and documentation, but also allow understanding of the data flow and traceability:



Note: this diagram applies to any submission that follows a linear data flow, whether using CDISC or not. If using CDISC, replace the "Tabulation data" with SDTM and the "Analysis data" with ADaM.

Best practices for traceability outlined in this document included defining **what** traceability is, **why** it is important, **how** it can be enacted and **where** more resources for traceability practices can be found. Traceability is included as part of the data with data-point traceability, and also in external documents as part of metadata traceability. When traceability is built in as an inherent property at every step, it does not require separate programming or other steps at the time of submission.