CDISC Implementierung und eSubmission mit SAS

Kurt Häusermann
HMS Analytical Software GmbH
Rohrbacher Str. 26
69115 Heidelberg
kurt.haeusermann@analytical-software.de

Zusammenfassung

Die CDISC-Formate für den Transport von Daten und Metadaten (ODM) und für die elektronische Submission von klinischen Studiendaten (SDTM, define.xml) werden immer häufiger eingesetzt. Pharmafirmen und CROs sind dabei, ihre Prozesse und Datenformate an diese internationalen Standards anzupassen (CDASH). An verschiedenen Konferenzen (PhUSE, SAS Global Forum) wurden Erfahrungsberichte und Einführungsstrategien bereits diskutiert. SAS hat sich seit Etablierung der Standards engagiert, arbeitet aktiv in verschiedenen CDISC Arbeitsgruppen mit. CDISC Standards wurden in SAS Produkten implementiert. Nun wird SAS mit dem *SAS CDISC Toolkit* die Erstellung von elektronischen Submissionen durch die Prüfung weiterer Domains der CDISC Datasets und durch die Erstellung der Metadatendatei *define.xml* bedeutend vereinfachen. Der Toolkit wird Teil von SAS Base sein und dieses Jahr erscheinen.
Die neue *SAS Clinical Data Integration Solution,* unterstützt zusätzlich die *Prozesse* und *Dokumentation* der Erstellung der SDTM-Dateien: Transformationsjobs können mit dem *Clinical Data Integration Studio* mehrheitlich ohne SAS Programmierung zusammengestellt werden, wobei auf die im Metadatenserver gespeicherten CDISC Standards bzw. Firmenstandards zugegriffen werden kann. Dadurch wird die Erstellung der SDTM-Dateien deutlich vereinfacht, die Transformationen, die Quell- und Zieldateien im Metadatenserver dokumentiert und die *define.xml* Datei automatisch erstellt. Schließlich lassen sich neu erstellte Transformationen später wiederverwenden.

Schlüsselwörter: CDISC, SDTM, SAS/Base, SAS CDISC Toolkit, PROC CDISC, SAS Clinical Data Integration

1 Einleitung

1.1 Clinical Data Interchange Standards (CDISC)

Seit ihrer 1997 erfolgten Gründung hat CDISC (Clinical Data Interchange Standards Consortium) erkannt, wie wichtig etablierte Standard Datenmodelle sind, um effizient klinische Daten zu erfassen und auszutauschen. Im Mission Statement wird dies wie folgt formuliert: „*CDISC is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare*“.

Während die anfänglichen Aktivitäten sich hauptsächlich in Europa und den USA entwickelt haben, ist CDISC heute auch in Osteuropa, Südamerika, Indien, China und Japan sehr rege tätig. CDISC ist eine von der Life Science Industrie unterstützte Organisation, die auch mit vielen anderen Standardorganisationen (z.B. ISO, HL7) und vielen Behörden (FDA, EMEA) an der Etablierung neuer Standards arbeitet. Die Standards sind auf der Homepage www.cdisc.org verfügbar.

1.2 Übersicht CDISC Standards

Die CDISC Standards umfassen im Wesentlichen folgende Standards:

- Das Operational Data Model (ODM) ist ein plattform-unabhängiges, XML (eXtensible Markup Language)-basiertes Format für den *Datenaustausch* und die *Archivierung* von klinischen Daten, inklusive *Studiendaten*, *Metadaten* und *administrativen Daten*.
- Das Study Data Tabulation Model (SDTM) enthält die *Studiendaten* für die Arzneimittelzulassung und wird seit 2008 von der FDA bei elektronischen Submissionen akzeptiert. Das SDTM-Datenmodell ist ein tabellarisches Modell, das sich am SAS V5 Transport-File-Format orientiert, das leider in vielerlei Hinsicht eingeschränkt ist. SDTM enthält eine Anzahl fest bestimmter Domänen, Variablen-Klassen, vordefinierte Variablen und Variablenausprägungen (Codelists). Neben dem Standard existiert ein sehr umfangreicher Implementation Guide (SDTMIG), der sehr viele Details des Datenformats beschreibt.
- Das Analysis Dataset Model (ADaM), das für die Übermittlung von Datasets mit prozessierten Daten verwendet wird, orientiert sich ebenfalls am SAS V5 Transport-File-Format.
- Das Lab-Datenmodell (LAB) wurde für den Transfer von Labordaten erstellt. Es kann als ASCII-, SAS-, oder XML-Datei realisiert werden. Das Lab-Modell ist das erste CDISC Datenmodell, das auf dem HL7 (Health Level 7)-Standard beruht.
- Clinical Data Acquisition Standards Harmonisation (CDASH) besteht aus einer Anzahl von Standards (Variablennamen, Variablenbeschreibung, weitere Metadaten der Variablen) für die Datenerfassung. Dadurch wird später die Konversion und Erstellung der SDTM-Dateien vereinfacht, da einzelne Datenelemente bereits die geforderte Struktur aufweisen. Als Datenmodell wird meist ODM verwendet, wobei CDASH die Semantik und ODM die Syntax und die Definition von administrativen Daten liefert.
- Die Case Report Tabulation Data Definition Specification (CRT-DDS, oder kurz define.xml) beinhaltet die Definition der Studiendaten. Diese Definition wurde von der FDA 2005 anerkannt und ist Teil der Electronic Study Data Specifications (eCTD) der FDA. Sie ist Ersatz für die frühere define.PDF Datei, enthält eine detaillierte Beschreibung der Dateien und Variablen und Variablenausprägungen der eingereichten SDTM-Dateien sowie weitere Daten wie Links zu den annotierten Case Report Forms, die als PDF-Dateien Teil einer cCTD sind.

- Controlled Terminology bezeichnet standardisierte Ausprägungen von bestimmten Variablen in SDTM. Für verschiedene Variable verwendet CDISC eine definierte Terminologie, die meist von anderen Organisationen übernommen wurde (z.B. vom NCI National Cancer Institute in den USA).

1.3 Unterstützung der CDISC Datenmodelle durch SAS

SAS kann Daten mit folgenden CDISC Datenmodellen lesen bzw. schreiben:

- ODM V1.2 kann mit der xml-Libname-Engine oder mit PROC CDISC geschrieben und gelesen werden.
- SDTM Dateien können gelesen und geschrieben werden, da der Standard ein SAS Format ist (SAS V5 Transport-Format, kurz xpt) Strukturtests und inhaltliche Tests können für bestimmte Domänen mit PROC CDISC durchgeführt werden. Mit dem neuen Toolset werden die Testmöglichkeiten von PROC CDISC auf weitere Domänen ausgedehnt.
- ADaM Dateien können gelesen und geschrieben werden, da es wie SDTM das xpt-Format verwendet.
- LAB-Dateien können als ASCII-, SAS- oder XML-Datei gelesen und geschrieben werden, XML-Dateien z.B. mit der XML- oder XML92-Libname-Engine.
- define.xml kann mit SAS Bordmitteln erstellt werden. Dies ist jedoch nicht ganz einfach und erfordert entsprechende SAS-Programme. Mit dem neuen SAS CDISC Toolkit können define.xml Dateien direkt erstellt werden.

2 Klinische Studien Life-Cycle und Use Cases für CDISC

Die folgende Abbildung 1 gibt einen guten Überblick, in welchen Prozessen die einzelnen Datenmodelle verwendet werden:

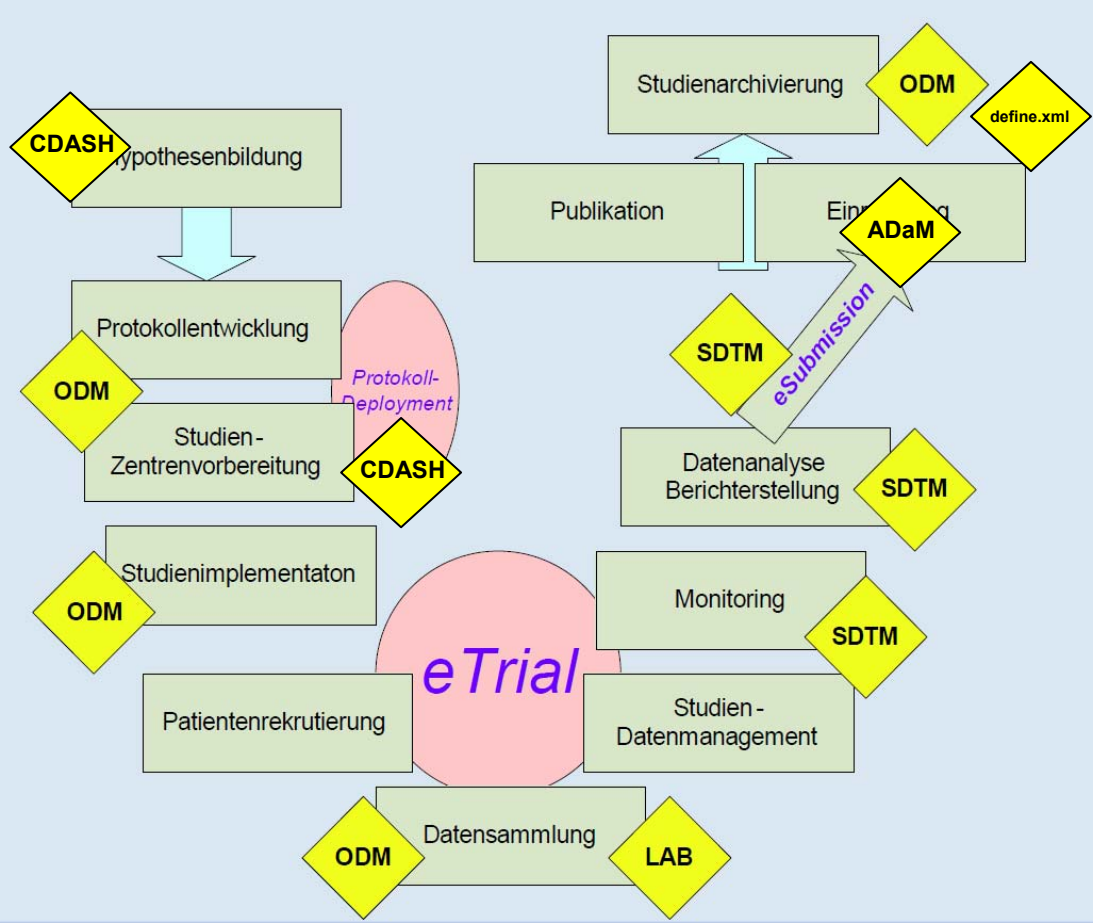

Abbildung 1: Anwendung der CDISC Datenmodelle in einer klinischen Studie aus: Semler et al.: CDISC, Würzburger Archivtage, GMDS, 08.12.2005
ergänzt von K. Häusermann.

- ODM: Protokollentwicklung, Studienentwicklung, Datensammlung und der Studienarchivierung verwendet wird (teilweise mit CDASH Standards)
- LAB: Einspeisung von Labordaten
- SDTM: elektronische Submission zur FDA, Ausgangspunkt für die Erstellung der ADaM-Dateien
- ADaM: elektronische Submission zur FDA
- define.xml: Datendefinition für SDTM bzw. ADaM.

Wir werden im Folgenden die SAS-Unterstützung für XML-Dateien, ODM-Dateien und SDTM/define.xml näher betrachten.

3 XML Dateien mit SAS einlesen oder erstellen

SAS kennt verschiedene Möglichkeiten, SAS Dateien zu lesen und zu schreiben:

- Mit eigenem SAS-Code im Data Step. Dabei muss bedacht werden, dass bestimmte Zeichen in Texten wie „<“, „>“ ,usw. mit einem sog. Escape-Zeichen versehen werden müssen, damit der Text später wieder als XML-Datei eingelesen werden kann (http://www2.sas.com/proceedings/sugi29/036-29.pdf) zuletzt geprüft am 12.06.2009.
- XML bzw. XML92 (SAS 9.2) Libname-Engine. Die Libname Engine bietet eine sehr einfache Möglichkeit, um XML-Dateien zu erstellen bzw. zu lesen.
- SAS XML Mapper: Definition der Struktur der XML-Datei mit XML Mapper. Damit können auch komplexere XML-Dateien mit mehrfach geschachtelten Wiederholungen von Itemgruppen gelesen bzw. erstellt werden (siehe http://www2.sas.com/proceedings/sugi29/119-29.pdf) zuletzt geprüft am 12.06.2009.
- Erstellung einer XML-Datei mit Templates (siehe http://www2.sas.com/proceedings/sugi29/246-29.pdf) zuletzt geprüft am 12.06.2009.

Das folgende Beispiel zeigt, wie eine einfache XML-Datei mit SAS erstellt und wieder gelesen werden kann. In der xml-Libname-Engine können viele Parameter eingestellt werden, die die Generierung des XML-Files bestimmen.

Beispiel: Erstellen einer XML-Datei mit Hilfe der xml-Libname-Engine. Einlesen der erstellten XML-Datei in eine SAS-Datei:

```
libname xmlout xml 'C:\SASTEST\class.xml';
data xmlout.grades;
   set sashelp.class;
run;
data work.grades;
   set xmlout.grades;
run;
```

Erstellte XML-Datei:

```
<?xml version="1.0" encoding="windows-1252" ?>
<TABLE>
   <GRADES>
      <Name> Alfred </Name>
      <Sex> M </Sex>
      <Age> 14 </Age>
      <Height> 69 </Height>
      <Weight> 112.5 </Weight>
   </GRADES>
    …
</TABLE>
```

4 Operational Data Model (ODM)

Das ODM-Datenmodell basiert auf XML und kann folgende Datenbereiche beinhalten:

- *Metadaten*, also Daten, die die eigentlichen Daten beschreiben, das heißt Namen, Label, Datentyp, Daten zur Prüfung der Validität der Daten, Angabe der Ausprägung einer Variable und der Bedeutung der Ausprägungen (Codelisten). Case Report Forms können in Metadaten abgelegt werden, mit Validierungsinformationen für die einzelnen Felder.
- *Daten*, vom Anwender eingegebene Daten
- *Audit-Trail*, Angaben, wann eine Person Daten eingegeben, bzw. modifiziert hat mit Speicherung des alten und neuen Wertes).
- *Administrative Daten,* z.B. StudienOId, MetaddatenOId, usw.

Das ODM-Format kennt einige Elemente, die eine feste Semantik haben, z.B. die Definition von administrativen Größen. Die Bedeutung und Struktur der eigentlichen Daten wird jedoch durch Metadaten bestimmt.

Eine ODM-Datei kann alle oder nur einzelne der oben angegebenen Bereiche enthalten. Sie kann also *nur Metadaten*, etwa die Definitionen eines CRF enthalten. Ein Electronic Data Capture Programm kann diese Datei lesen, dem Anwender die einzelnen Fragen am Bildschirm vorlegen, eine Validitätsprüfung durchführen und anschliessend die erhobenen Daten mit Angaben über die Dateneingabe (Audit-Trail) in einer weiteren ODM-Datei speichern.

Die *define.xml* Datei ist eine erweiterte ODM-Datei, die im Wesentlichen nur aus *Metadaten* besteht.

Das ODM-Datenmodell ist heute insbesondere bei der Datenerfassung verbreitet und es wird von verschiedenen Anbietern unterstützt. Es konnte sich bei regulatorischen Behörden wie der FDA als Ablösung des V5 Transport-File Formats jedoch nicht durchsetzen.

Weiter eignen sich ODM-Dateien für den *Transport von klinischen Daten* zwischen unterschiedlichen Plattformen, z.B. für den Transport zum Sponsor.

Die Syntax der XML-Datei ist in einem sog. *XML-Schema* fest beschrieben. Jede ODM-Datei muss der im XML-Schema beschriebenen Form entsprechen.

4.1 Mit der XML Libname-Engine ODM Dateien lesen bzw. schreiben

Da das ODM XML-Format recht komplex ist, wird man sehr gerne die Implementierung von SAS verwenden wollen, um ODM Dateien zu lesen bzw. zu schreiben.

Beispiel: Einlesen einer ODM-Datei und erstellen einer ODM-Datei mit der XML Engine.

```
filename input 'C:\SASTEST\AE.XML';
libname  input xml xmltype=CDISCODM FormatActive=YES
         FormatNoReplace=NO FormatLibrary="Work";
data work.AE;
   set input.AE;
run;

filename output 'C:\SASTEST\AE2.XML';
libname  output xml xmltype=CDISCODM formatactive=yes;

data output.AE2;
   set work.AE;
run;
```

4.2 ODM Dateien erstellen mit PROC CDISC

Die SAS Prozedur PROC CDISC bietet viele zusätzliche Möglichkeiten für das Erstellen und Einlesen von ODM-Dateien, auf die hier nicht weiter eingegangen werden kann (siehe http://support.sas.com/rnd/base/xmlengine/proccdisc/TW8774.pdf).

5 Study Tabulation Model (SDTM)

Das SDTM-Datenmodell ist, im Gegensatz zu ODM, in starkem Maße durch die Inhalte geprägt. Die in einer Studie anfallenden Daten wurden in *Domänen* aufgeteilt und diese in vier *Beobachtungsklassen* eingeteilt:

- Interventionen (Interventions)
- Ereignisse (Events)
- Ergebnisse (Findings)
- Andere: Demographische Daten, Kommentare, zusätzliche Variablen, Beziehungen zwischen Daten, Aufbau der klinischen Studie

Jede Domäne besteht aus einer Anzahl von sehr genau definierten Variablen, die obligatorisch oder optional sein können. Für bestimmte Variablen existiert eine fest definierte *Terminologie*. Das das physische Datenformat ein SAS V5 Datenformat ist, bestehen verschiedene Restriktionen, die eingehalten werden müssen:

- Die Länge eines Variablennamens darf nicht länger als 8 Zeichen sein.
- Die Länge eines Labels darf nicht länger als 40 Zeichen sein.
- Die Länge einer alphanumerischen Variable darf nicht länger als 200 Zeichen sein.

5.1 Prüfen der SDTM Dateien mit PROC CDISC

Die SDTM-Dateien können für bestimmte Domänen mit PROC CDISC geprüft werden. Diese Prüfung betrifft die formellen Eigenschaften wie Längen der Dateinamen und Labels, als auch die inhaltliche Kontrolle der Daten von bestimmten Domains. Zur Zeit können 13(?) Domains geprüft werden.

Beispiel: Die Prüfung der Domäne AE (Adverse Events) kann mit PROC CDISC wie folgt durchgeführt werden (siehe http://support.sas.com/rnd/base/xmlengine/proccdisc/index.html).

```
proc cdisc          model=SDTM;
   sdtm             SDTMVersion = "3.1";
   domaindata       data = results.AE domain = AE
                    category = EVENTS;
run;
```

Der neue SAS CDISC Toolkit wird eine grössere Anzahl Domänen testen können.

5.2 Erstellung der define.xml Datei mit dem SAS CDISC Toolkit

Zur Zeit ist es sehr aufwendig, define.xml Dateien zu erstellen. Dieser Prozess wird nun durch den neuen SAS CDISC Toolkit bedeutend vereinfacht.

Zunächst müssen die SDTM-Dateien erstellt und deren Korrektheit geprüft werden. Dann werden weitere Informationen bereitgestellt, wie Annoted Case Report Forms als PDF-Dateien, usw.

Liegen alle diese Informationen bereit, kann die define.xml Datei durch den SAS CDISC Toolkit erstellt werden. Anschließend kann die erstellte XML-Datei mit einem XSL-Stylesheet in eine HTML-Datei übersetzt werden, die im Browser angezeigt werden kann.

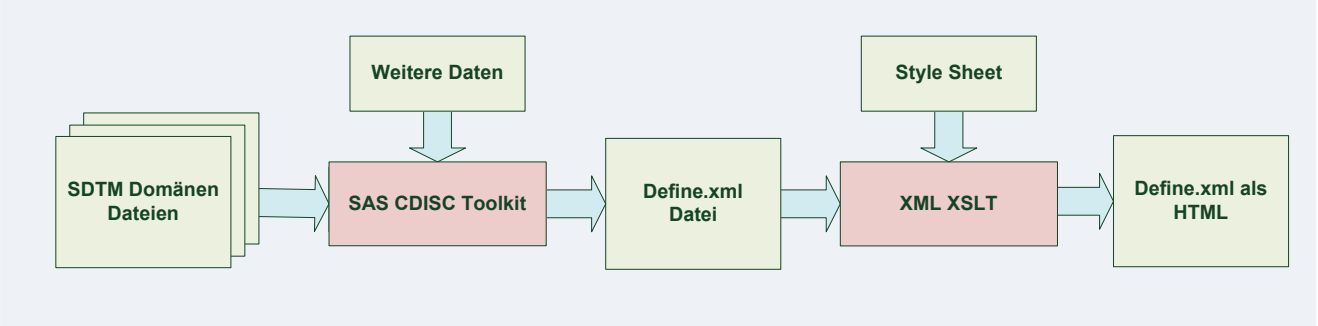

Abbildung 2: Erstellung der define.xml-Datei

6 SAS Clinical Data Integration

Mit der *SAS Clinical Data Integration* 1.2 (CDI), die dieses Jahr erscheinen wird, bietet SAS Pharmaunternehmen und CROs eine neue Möglichkeit, um SDTM-Dateien effektiver zu erstellen. Basis der Lösung ist das im Enterprise Intelligence Umfeld seit langem verfügbare und bewährte *SAS Data Integration Studio*, das um die spezifische Funktionalität für die Integration von klinischen Daten und die Unterstützung der CDISC Standards erweitert wurde. Die gegenwärtige Version unterstützt den SDTM Implementation Guide 3.1.1. (siehe [1] und [2])

Die CDI-Lösung

- ist ein strategischer Ansatz für alle SDTM Erstellungen.
- ermöglicht die Erstellung von automatisierten, wiederholbaren und wartbaren Prozessen
- ist eine metadatengetriebene end-to-end Plattform mit
  - Daten Standardisierung
  - Daten Cleaning
  - Elementen zur Verbesserung der Datenqualität
  - Management der Stammdaten
- beinhaltet eingebaute Unterstützung für CDISC
- ist erweiterbar auf andere Standards

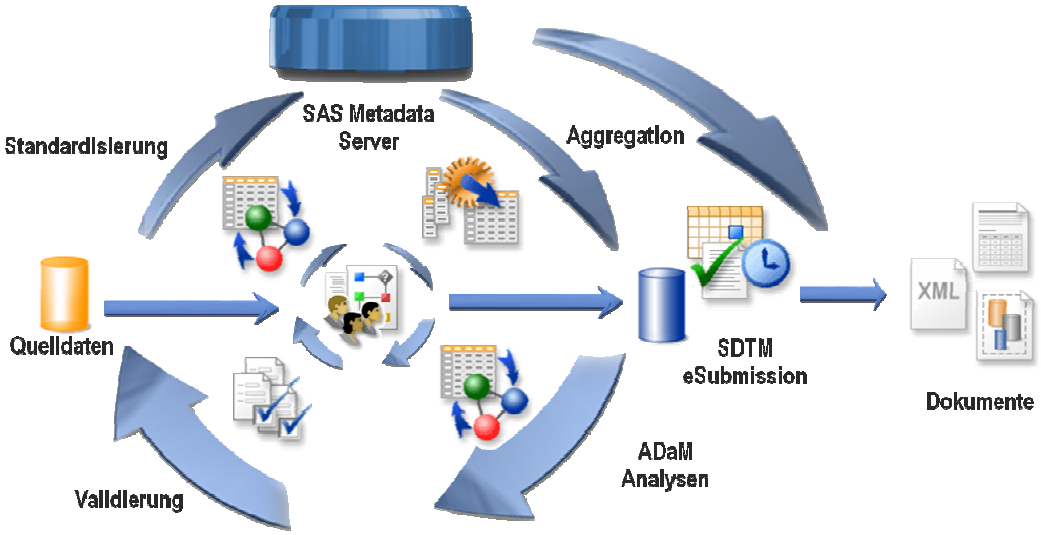

Abbildung 3: SAS Clinical Data Integration Solution

6.1 Metadaten Management

Die Open Metadaten Architektur (OMA) ist ein Kernelement aller SAS Lösungen und bildet auch das Bindeglied zwischen den SAS Komponenten. Die OMA speichert in der SAS Data Integration Lösung die CDISC SDTM Datenmodelle und stellt diese Daten den einzelnen Projekten und Anwendern zur Verfügung. Die OMA enthält die Defini-

tion von SDTM-Domänen, generische Job Templates, CDISC kontrollierte Terminologie und die Variablen, um neue Domänen erstellen zu können. Alle zusammengehörenden Daten werden in der OMA in einem sog. Repository gespeichert. Repositories können in der OMA hierarchisch angelegt werden.

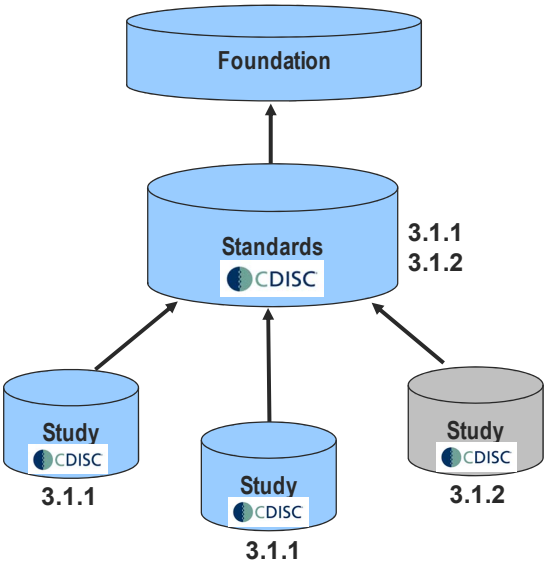

Abbildung 4: Standards-Repository und Studien-Repository

Für die CDISC Standards wird ein Repository angelegt, das alle allgemeingültigen Festlegungen enthält. Für jede Studie wird ein neues Repository angelegt, das die spezifischen Metadaten der Studie enthält. Dazu gehören Anwender, Rechte, Gruppen, zugeordnete Server und Jobs. Für die generellen Festlegungen wird automatisch auf das übergeordnete Standards Repository zurückgegriffen, z.B. für die Beschreibung der Domänen und Variablen. Für CDISC wurden die Standard Beschreibungen für Dateien und Variablen um zusätzliche Attribute erweitert. Das Change Management von einzelnen Objekten wird unterstützt: Änderung bei der Definition von einzelnen Objekten erfolgen über die Checkout/Checkin Funktionen. Die Änderungen werden gespeichert und in einem Audit-Trail eingetragen.

6.2 Clinical Data Integration Studio

Das SAS Clinical Data Integration Studio, kurz DI-Studio, ist ein SAS Client, mit dem man mit einer grafischen Oberfläche Transformationsjobs erstellen, speichern und durchführen kann. Er ermöglicht durch Zugriffe auf die Metadaten die Definition der Dateien. Ein Transformationsjob besteht aus einer Pipeline von Transformationsschritten, die in Dialogen parametrisiert werden können. Die Transformationen gehören zum SAS Standardumfang des DI-Studios oder wurden als DI-Studio Plug-Ins entwickelt, damit CDISC gut unterstützt werden kann. Der Anwender kann bestimmte Funktionen als SAS Programme entwickeln und die Funktion im DI-Studio verfügbar machen. Der Anwender, der die CDISC Jobs zusammenstellt, braucht in der Regel keine SAS Programme zu schreiben.

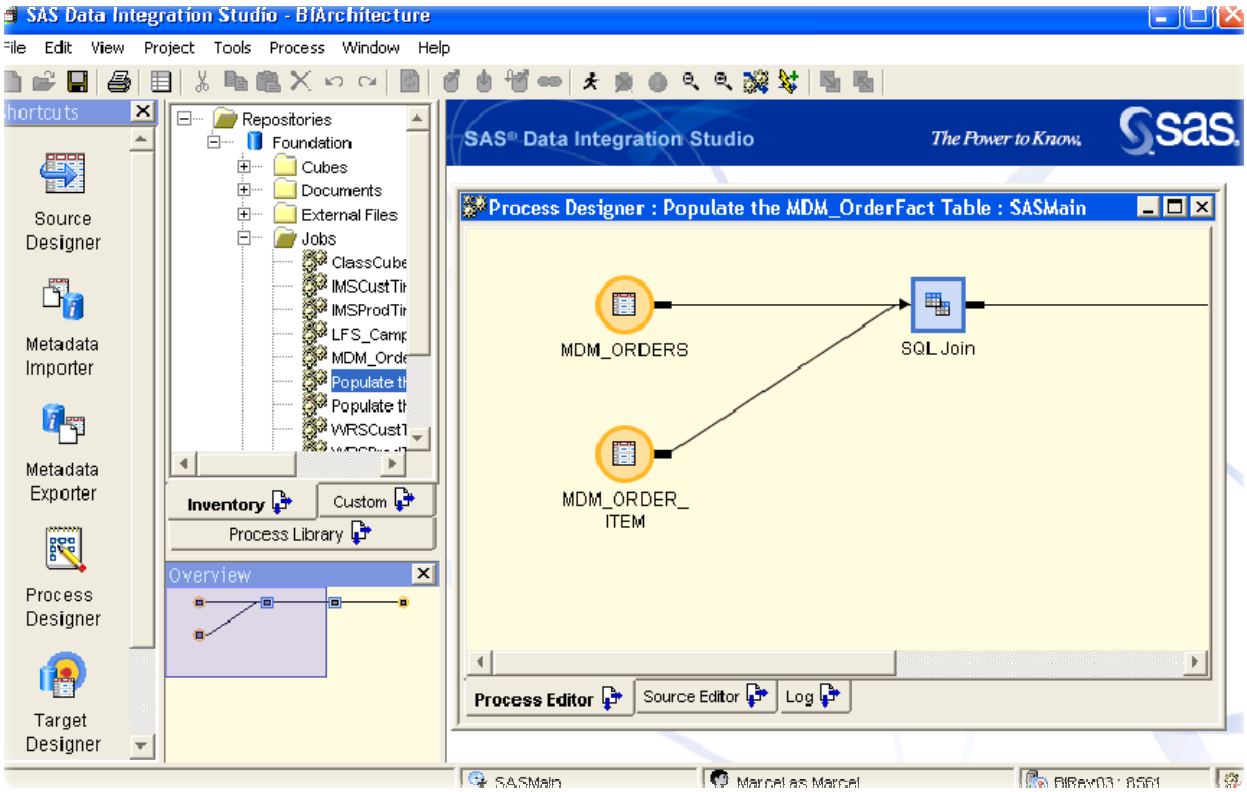

Abbildung 5: Erstellung neuer Transformationsjobs mit dem DI-Studio

Der Anwender hat im DI-Studio einen Überblick über die Transformationsjobs, die Definition der Dateien und der Transformationsschritte eines Jobs.

6.3 Prozess der Erstellung der SDTM Domänen

Der Prozess der Erstellung der SDTM Domänen besteht aus folgenden Schritten:

1. Annotierung der CRF (wie bisher), Namen der Variablen bestimmen.
2. Alle Angaben über Source und Zieldateien und Transformationen in einer Excel-Arbeitsmappe speichern (Mappingdatei).
3. Aufsetzen eines neuen Projektes mit dem Clinical DI-Studio.
4. Erstellung verschiedener Transformationsjobs im DI-Studio.
5. Erstellung der define.xml-Datei.
6. Prüfung der Compliance.

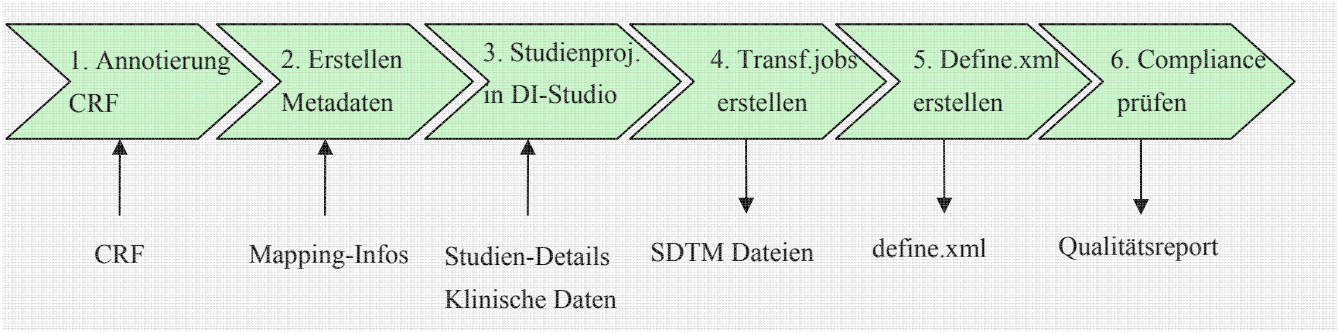

Abbildung 6: Prozess der Erstellung der SDTM Domänen

Schritt 1
Speicherung der annotierten CRF als PDF-Dateien.

Schritt 2
Die Angaben über das Mapping der Source Daten zu den Ziel CDISC Dateien erfolgt zunächst in einer Excel Arbeitsmappe. Alle Angaben über das Mapping erfolgen über dieses Arbeitsblatt. Es enthält weitere notwendige Metadaten.

Schritt 3
Dann wird im DI-Studio ein neuer Studienordner angelegt, so dass das neue Repository auf die Standards zugreifen kann. Es enthält Ordner für die Quelldaten, die Zieldaten und die Jobs. Die Ziel SDTM Datasets werden aus den globalen Standards hineinkopiert. Dann werden projekt- bzw. studienspezifische Metadaten über die Metadaten und Variablen aus dem Excel-Arbeitsmappe entnommen und im DI-Studienordner gespeichert. Diese Metadaten werden u.a. später für die Compliance-Prüfung und für die Generierung der *define.xml*-Datei verwendet. Dadurch wird die Zeit für das Aufsetzen einer neuen Studie auf ein Minimum reduziert. Die Excel-Arbeitsmappe enthält auch weitere Metadaten für die Konversion der Codelisten und die Transponierung.

Schritt 4
Für jede CDISC Domäne wird im DI-Studio mit einem visuellen Designer durch Zusammenstellung verschiedener Schritte mit Drag and Drop ein Transformationsjob zusammengestellt. Aus dieser Pipeline von Transformations-Schritten wird automatisch robuster SAS Code generiert und ausgeführt.

In der Prozess Bibliothek steht eine Vielzahl von Transformationen zur Verfügung. Sie beinhalten z.B. ISO8601 Umwandlungen von Zeitpunkten und Zeitdauern, Umwandlungen mit Look-up-Tabellen und verschiedene Transformationen. Wenn nötig kann diese Bibliothek mit studienspezifischen Transformationsfunktionen erweitert werden und später wiederverwendet werden.

Schritt 5
Anschließend erfolgt die Erstellung der Metadatendatei *define.xml*.

Schritt 6
Am Schluss erfolgt eine Überprüfung der Compliance der generierten Dateien. Diese erfolgt aufgrund der in den Metadaten gespeicherten Globalen CDISC Standards sowie gemäß den Angaben aus der ursprünglichen Excel-Arbeitsmappe.

Ein ähnlicher Prozess kann auch für ADaM erstellt werden.

7 CDISC Toolkit vs Clinical Data Integration Solution

Der SAS CDISC Toolkit, Teil von SAS Base, beinhaltet die grundlegende Unterstützung zur Erstellung der notwendigen Dateien für die eSubmission.

Die SAS Clinical Data Integration Solution ermöglicht dagegen eine *Unterstützung der Prozesse*, vereinfacht die *Erstellung*, *Dokumentation* und *Validierung* der einzelnen Prozessschritte und ermöglicht ebenso die Wiederverwendung von bereits bestehenden Bausteinen. Die Konversion der Daten in das SDTM Modell wird dadurch sicherer, *einfacher, besser organisiert* und ist *mit weniger Aufwand* verbunden.

Literatur

[1] Kilhullen, Michael. „Implementing CDISC Data Models in the SAS® Metadata Server“ Proceedings for the Pharmaceutical Industriy SAS® User Group Conference, 2006, 2007.

[2] van Reusel, Peter, Lambrecht, Mark. „Practical application of SAS® Clinical Data Integration Server for conversion to SDTM data, PhUSE Conference Proceedings, 2008.