PhUSE 2009

Paper CD14

Experiences and lessons learned from a first SDTM submission project

Paul Vervuren, Schering-Plough, Oss, the Netherlands
Bas van Bakel, OCS Consulting, Rosmalen, the Netherlands

ABSTRACT

This paper presents the experiences and the lessons learned from an SDTM conversion project. Within the scope of this project were (1) the late-stage conversion of datasets in internal standard to SDTM, (2) the creation of Trial Design datasets, (3) the development of macros to create define.xml and define.pdf, (4) the actual creation of these data definition files, and (5) the creation of SDTM annotated CRFs. Topics covered are the project strategy and our approach to the development of define.xml and define.pdf. We will present details of the conversion process and give examples of challenging domain implementations. Moreover, we will discuss the results of WebSDM™ checks performed as part a sample submission. Finally, the activity metrics of the project are presented.

INTRODUCTION

After a period of piloting and tool development, a planned NDA submission with relatively few, recent studies offered an excellent opportunity to target for our first SDTM submission. Providing SDTM datasets implicated a full conversion process since our CDMS and reporting/analysis data are entirely based on internal, non-CDISC standards.

For this submission we used reporting datasets as source data for SDTM (fig. 1). Thus, we adopted a 'retrospective development' strategy as described by Kenny & Litzsinger (2005)[1]. Main argument for choosing this strategy was that we wished to remain in sync with the reporting datasets and be able to reuse any derived variables needed for SDTM. Being the basis for all reporting and analyses related to our data the reporting datasets formed the main submission component. Supply of ADaM datasets was considered out-of-scope.

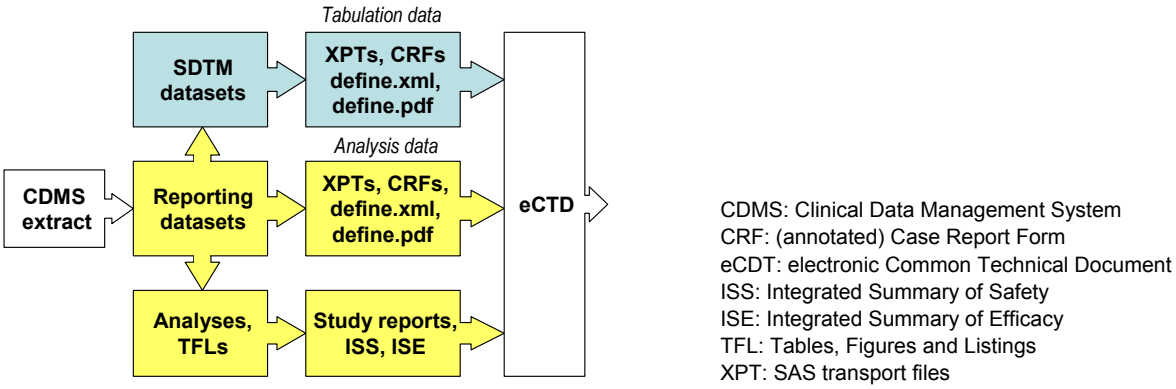

Figure 1: SDTM in the submission data flow from CDMS to eCTD

Piloting prior to this project gave us a good level of experience with SDTM conversion. Moreover, it provided a valuable toolset and work environment, consisting of:

- Structured sample programs (Base/SAS®) for the conversion of 13 standard domains;
- A macro library supporting various routine steps, e.g. creation of ISO8601 date/time variables, addition of MedDRA terms, creation of SUPPQUAL datasets, checking of SDTM metadata and structure compliance;
- An extensible SDTM metadata repository, including CDISC controlled terminology and zero-observation datasets for all standard domains;
- A template mapping sheet.

This gave us a solid basis to be able to provide the required SDTM deliverables within a limited timeframe. Our biggest challenge was the creation of Trial Design datasets and define.xml, with which we had no prior experience.

Within a 10-month period a dedicated team managed to provide all the deliverables and, in addition, submitted an SDTM sample to the FDA. This paper describes the experiences and lessons learned during this project. Consecutively, we will discuss the project approach, the conversion process, the sample submission including WebSDM checks, implementation examples of some challenging SDTM domains and the project activity metrics.

PROJECT STRATEGY

The following considerations were important in determining our project strategy:

- Due to limitation of time, expertise and manpower a pragmatic approach was needed with respect to the development of define.xml.
- Expert input was considered vital in order to speed up and improve the quality of SDTM implementation decisions, give technical advice in the development of define.xml and to support quality review. In addition, we intended to conduct a sample submission in order to get FDA feedback on our SDTM datasets and define.xml.
- The main activities were to be conducted internally. Thus we could leverage existing tools and knowledge and continue building our knowledge and expertise for future projects. Moreover it would ensure that we would be driving any implementation decisions and 'own the data'. This was considered an optimal preparation for FDA or internal questions with respect to the submitted SDTM datasets.
- The team of programmers and statisticians concerned with the reporting process had to focus on the standard reporting and submission components, and were not be burdened with an additional and novel set of deliverables.

APPROACH TO DEFINE.XML

We intended to use the toolset developed in the CDISC SDTM/ADaM pilot project[2], as made available by CDISC, to create define.xml. By using these tools we aimed to save time and avoid the need for in-depth ODM/XML expertise. After having made provisions to read from our metadata environment by modifying the md2odm macro we were quickly able to deliver a working define.xml. However, we were initially not aware that ODM 1.3, used in md2odm, was just a draft version, and that ODM 1.2[3], was the production version of ODM underlying CRT-DDS 1.0[4]. CRT-DDS 1.0 is required by FDA as referenced in the Study Data Specifications document[5] as part of the eCTD specifications[6]. Moreover, ODM 1.2 appears to be expected by WebSDM in order to use define.xml to load SDTM metadata.

Guided by external advice specific modifications were made to the md2odm macro to enable the creation of a define.xml compliant with CRT-DDS 1.0. The main modifications to md2om next to the ODM version were the interfacing to our metadata repository and the creation of the value-level metadata section in define.xml. New functionality added was the automated insertion of page links to SDTM variables on the annotated CRF.

Initially, we considered to supply define.xml as a replacement of define.pdf, but decided at a later stage to deliver both define.xml and define.pdf. Main reasons to change our approach were internal needs for define.pdf as well as the anticipation that FDA would also prefer define.pdf next to define.xml, as expressed in the CDISC SDTM/ADAM pilot report. For this the approach described by Lex Jansen[8] was adopted. Via the SAS XML-mapper a mapping was produced in order to generate SAS datasets representing the information in define.xml. A macro was developed to subsequently read these datasets and generate define.pdf using SAS ODS (an intermediate step via postscript was performed to provide the links (e.g. to CRF pages) in define.pdf the same way as in define.xml).

APPROACH TO CONVERSION ACTIVITIES

Three of five members of the conversion team had no previous experience with SDTM conversion. In order to all begin with the same basic understanding of SDTM we organized the CDISC course 'SDTM Theory & Application' before starting the conversion activities. After this, two possible approaches were envisioned:

- Subdividing domains among the team, and converting domain-by-domain across studies;
- Involve all team members to complete one study first.

Even though the first approach could have provided some short-term efficiencies as domain specialists would have been allowed to focus on 'their' domains, the second approach was chosen. Advantages of completing one study first were: (1) it supported shared learning in the early stage of the project, (2) the completion of one study formed a solid milestone for the team and project stakeholders, and (3) it suited an early external expert review and sample submission.

Activities were kicked-off with a workshop for the team members to annotate the CRF set of the first study. A number of team meetings were subsequently held to jointly discuss the results and make (high-level) modeling decisions. For instance, it was discussed whether the Skin examination data were to be included in the PE domain or in a custom Findings domain (PE was used), and whether the 'Smoking status' question on the SOCIO-ECONOMIC DATA form was best kept with the other questions as part of the custom findings domain created for this page or was to be stored in SU, the standard Substance Use domain (SU was used). After this, mapping sheets were created and conversion programming was conducted for the first study. In all cases mapping and programming was done by the same person. An independent team member QC'd the mapping sheet, conversion program and datasets.

Expert review, conducted by Business & Decision Life Sciences, was performed partly in parallel with data conversion of the first study. First the annotated CRFs and mapping sheets were reviewed, and subsequently the Trial Design datasets and define.xml. At a later stage also the datasets were reviewed.

With this approach a very steep learning curve was achieved in the first months of the project. In 4.5 months we were thus able to submit as a sample to the FDA the complete study data (in SAS transport files), together with the annotated CRF and define.xml for the first study. For a considerable part, conversion of the other studies was a routine activity after this.

CONVERSION PROCESS

In figure 2 a schematic presentation is given of the SDTM conversion process with related sources and deliverables for each study. The distinct process steps involved, numbered 1-7, are described in detail in this section.

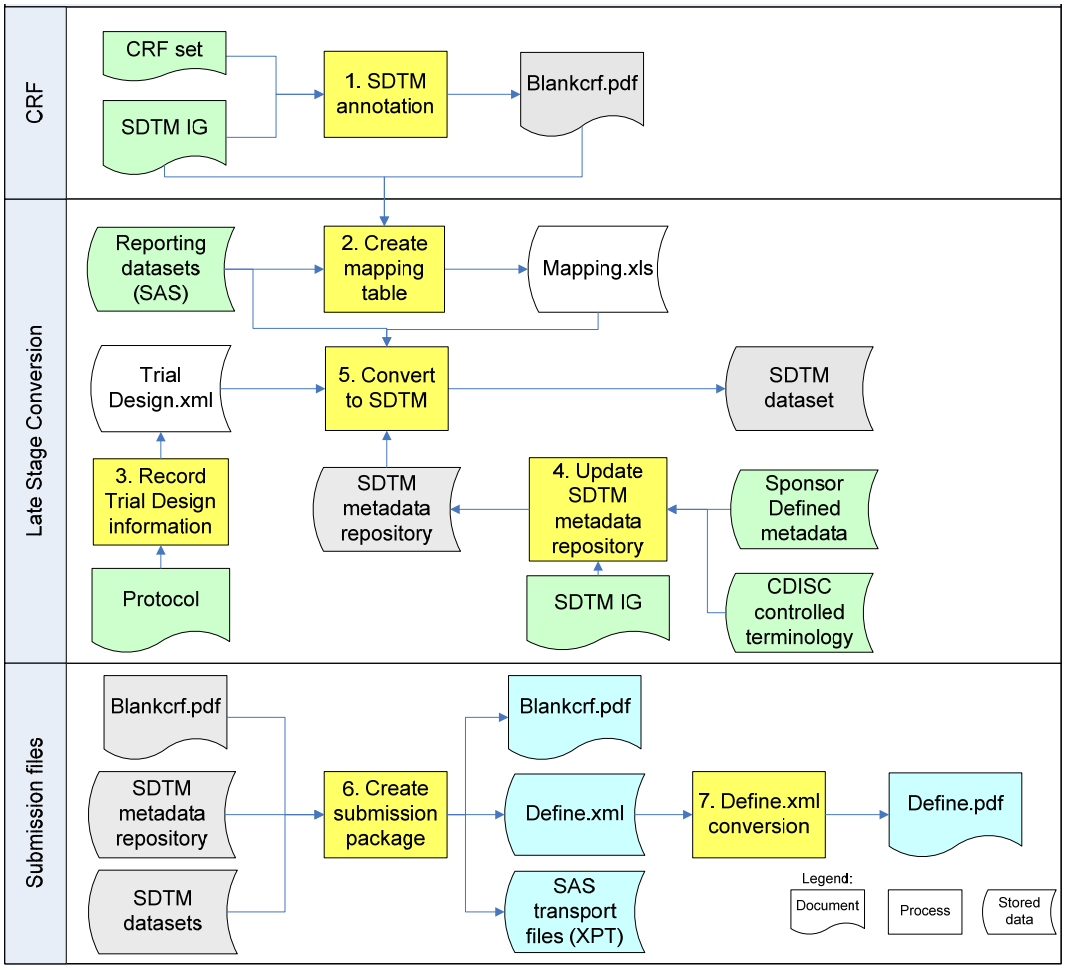

Figure 2: Conversion process: CRF annotation, conversion to SDTM, and creation of submission deliverables.

1. SDTM ANNOTATION

In most studies eCRFs were used, in which case available paper-equivalents were used for annotation. Where needed the CRF set was completed with screenshots of eDiary and ePROs. Many mapping decisions, guided by SDTM-IG 3.1.1, were made during the annotation process.

To be able to automate the insertion of page links to SDTM variables in define.xml, annotation conventions were applied. For example, to distinguish between variables and values, values were put between double quotes (fig. 3). Moreover, domain names were put between brackets (and preceded by the text 'DOMAIN=') and comments like 'Not submitted' were put between curly braces. During the creation of define.xml a check was performed to check whether all annotations were valid and whether expected raw variables were annotated (see "6. and 7. create submission package and convert define.xml to define.pdf").

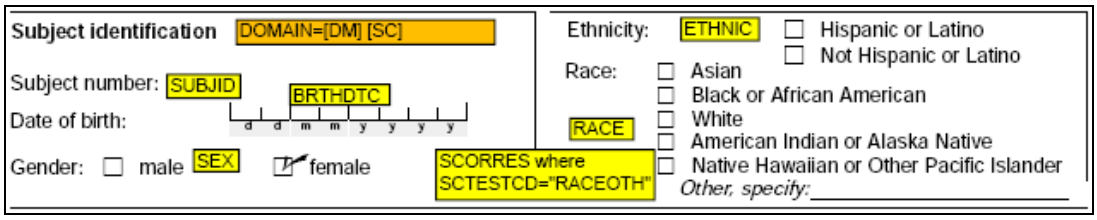

Figure 3: Excerpt from an annotated CRF (blankcrf.pdf)

The similarity of CRFs across studies allowed the reuse of annotations and bookmarks. For this purpose annotations were exported using an Acrobat .fdf file. The .fdf file was edited before importing to remove redundant annotations and to accommodate differences in page numbering. With this .fdf file annotations were imported in the blankcrf.pdf of a next study. Where needed individual annotations were then manually moved to the correct location and annotations for new pages were added. Bookmarks 'by Domain' and 'by Visit' were manually added using Acrobat Professional functionality. Bookmarks for similar CRFs could also be reused by removing all the pages but one of a bookmarked blankcrf.pdf and inserting the pages of a new one. Using the Acrobat function 'Set Destination' bookmarks were reset to the proper pages; redundant bookmarks were removed.

2. CREATE MAPPING TABLE

A mapping table (fig. 4), created in Microsoft Excel, was used to record the SDTM target variable of each source variable, and, if applicable, the derivation rule involved. While high-level mapping decisions were made during CRF annotation, detailed mapping decisions (e.g. application of CDISC controlled terminology, derivation rules, definition of category variables) were made during this step.

Source Libname	Dataset	Variable	Type	Label	SDTM Domain	SDTM Variable	Mapping type	SDTM Mapping Comments
MASTER	BASECHAR	NOP	Char	Number of Protocol	DM, DS, SC, SU, XT	STUDYID		
MASTER	BASECHAR	CENTF	Char	Center Number	DM	SITEID		
MASTER	BASECHAR	COUNTRYF	Char	Country	DM	COUNTRY	_CODE_ country=country	
MASTER	BASECHAR	INVF	Char	Investigator	DM	INVID		
MASTER	BASECHAR	SID	Num	Subject Number	DM	SUBJID		
MASTER	BASECHAR	USUBJID	Char	Unique Subject Identifier	DM, DS, SC, SU, XT	USUBJID		
MASTER	BASECHAR	TCPF	Num	Treatment group as randomized	DM	ARM, ARMCD		Use format TCPF for ARM.
MASTER	BASECHAR	TCAF	Num	Treatment group as treated				NOT MAPPED
MASTER	BASECHAR	AGE	Num	Age (years)	DM	AGE		See derivation DM01 on Derivation sheet
MASTER	BASECHAR	AGECLASF	Num	Age class				NOT MAPPED
MASTER	BASECHAR	SEXF	Char	Gender	DM	SEX		
MASTER	BASECHAR	RACEF	Char	Race	DM	RACE		Use format RACE
MASTER	BASECHAR	RACEOTC	Char	Race other:comment (specification)	SC	SCORRES, SCTEST, SCTESTCD, SCCAT		Record created only if value present. SCTEST = "Race Other", SCTESTCD = "RACEOTH", SCCAT = "DEMOGRAPHIC DATA"
MASTER	BASECHAR	ETHNICF	Char	Ethnicity	DM	ETHNIC		Use format RACE

Figure 4: Example of a mapping table (excerpt)

A column was used to describe the conversion rule or to indicate a variable was not included in SDTM ('NOT MAPPED'). Complex derivations or long derivation texts were described in a separate 'Derivations' sheet.

The mapping table was not driving the conversion process but served mainly as a source of documentation. Reading in the sheet in the conversion program enabled a cross-check of the mapped variables with those extracted from the source dataset. This ensured no CRF variable was overlooked during programming.

3. RECORD TRIAL DESIGN INFORMATION

The details needed for the static SDTM Trial Design datasets TA (Trial Arms), TE (Trial Elements), TS (Trial Summary) and TV (Trial Visits) were looked up in the protocol and statistical analysis plan and entered in XML files using Excel. These XML files were used as source at a later stage in the conversion process in order to generate the SDTM datasets.

4. UPDATE THE SDTM METADATA REPOSITORY

All SDTM metadata was stored in a central metadata repository, which had been designed prior to this project. The repository components were: dataset structure definitions (fig. 5), extended variable metadata (name, label, type, length, use of controlled terminology, origin, role, core; fig. 6) and controlled terminology. In the final step of conversion programming, this repository was used as a reference database to ensure the SDTM datasets created were compliant with SDTM metadata (including use of expected controlled terminology) and structure rules. The repository was also the basis for metadata used in the define.xml, including variable comments. Controlled terminology was defined on a per-study basis. This was to facilitate study-specific definitions like ARMCD. Moreover, it allowed controlled terms in the define.xml to be restricted to values relevant for the study. While sometimes all values need to be presented (e.g. of race, sex), in other cases it is preferable to present only the relevant values (e.g. lab test codes, countries).

	A	B	C	D	E	F	G	
1	Source	Domain	Description	Location	Structure	Class	Purpose	Keys
2	CDISC	AE	Adverse Events	ae.xpt	One record per adverse event per subject	Events	Tabulation	STUDYID, USUBJID, AETERM,
3	SPONSOR	BR	Endometrical Biopsy	br.xpt	One record per biopsy per time point per vis	Findings	Tabulation	STUDYID, USUBJID, VISITNUM,
4	CDISC	CM	Concomitant Medications	cm.xpt	One record per medication intervention episo	Interventions	Tabulation	STUDYID, USUBJID, CMCAT, V
5	CDISC	CO	Comments	co.xpt	One record per comment per subject	Special Purpose	Tabulation	STUDYID, USUBJID, COSEQ
6	CDISC	DA	Drug Accountability	da.xpt	One record per accountability observation p	Findings	Tabulation	STUDYID, USUBJID, DACAT, D
7	CDISC	DM	Demographics	dm.xpt	One record per subject	Special Purpose	Tabulation	STUDYID, USUBJID
8	CDISC	DS	Disposition	ds.xpt	One record per disposition status or protoco	Events	Tabulation	STUDYID, USUBJID, DSCAT, D
9	CDISC	DV	Protocol Deviations	dv.xpt	One record per protocol deviation per subje	Events	Tabulation	STUDYID, USUBJID, DVCAT, D
10	CDISC	EG	ECG Test Results	eg.xpt	One record per ECG observation per time po	Findings	Tabulation	STUDYID, USUBJID, VISITNUM,

Figure 5: Dataset structure definitions in the SDTM metadata repository

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	source	domain	seq	variable_n	variable_label	variat	varia	vari	controlled_term	controlled_term_	origin	role	core	sponso	spo	notes
2	CDISC	DM	1	STUDYID	Study Identifier	Char	20	1			CRF	Identifier	Req		N	
3	CDISC	DM	2	DOMAIN	Domain Abbreviation	Char	2	.	DOMAIN	SPONSOR	Derived	Identifier	Req		N	
4	CDISC	DM	3	USUBJID	Unique Subject Identifier	Char	40	2			Sponsor Defined	Identifier	Req		N	Concatenation of STUDYI
5	CDISC	DM	4	SUBJID	Subject Identifier for the	Char	40	.			CRF	Topic	Req		N	
6	CDISC	DM	5	RFSTDTC	Subject Reference Start	Char	19	.	ISO8601		Sponsor Defined	Timing	Exp		N	Equal to date of first admi randomization for non-tre
7	CDISC	DM	6	RFENDTC	Subject Reference End D	Char	19	.	ISO8601		Sponsor Defined	Timing	Exp		N	Equal to date of last admi of last administration of In will be used. Equal to date
8	CDISC	DM	7	SITEID	Study Site Identifier	Char	20	.			CRF or Derived	Record Qualifier	Req		N	
9	CDISC	DM	8	INVID	Investigator Identifier	Char	20	.			CRF or Derived	Record Qualifier	Perm		N	
10	CDISC	DM	9	INVNAM	Investigator Name	Char	40	.			CRF or Derived	Synonym Qualifie	Perm		N	
11	CDISC	DM	10	BRTHDTC	Date/Time of Birth	Char	19	.	ISO8601		CRF or Derived	Result Qualifier	Perm		N	
12	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
23	CDISC	DM	21	ASRY	All-Subjects-Randomized	Char	1	.	NY	CDISC	Derived	Result Qualifier		Perm	Y	Subject included in the All rule.
24	CDISC	DM	22	ASTY	All-Subjects-Treated Gro	Char	1	.	NY	CDISC	Derived	Result Qualifier		Perm	Y	Subject included in the All
25	CDISC	DM	23	ITTY	Intent-to-Treat Group	Char	1	.	NY	CDISC	Derived	Result Qualifier		Perm	Y	Subject included in the Int

Figure 6: Extended variable metadata in the SDTM metadata repository

5. CONVERT TO SDTM

For each SDTM dataset a separate SAS program was used, in which the conversion rules specified in the mapping table were applied to convert the reporting datasets to SDTM datasets.

In general, conversion programs performed following steps:

A. Extraction of the required data from the source dataset(s) or Trial Design XML files;
B. A check to see whether the selection of source variables was consistent with the mapping table;
C. Domain-specific code to convert source variables to SDTM variables (renaming, recoding to CDISC controlled terminology, conversion to ISO8601 date/time variables, derivation of study day of examination);
D. Addition of visit name from the Trial Visit domain, if applicable;
E. Addition of the SEQ variable;
F. Extraction of comment variables, if present;
G. Metadata and controlled terminology compliance check;
H. Creation of supplemental qualifiers dataset, if applicable;
I. Sorting according to keys and writing to target location.

Steps B, D, E, F, G, H and I were executed with the use of macros. Macros were also used as part of step C: e.g. for (straightforward) recoding, conversion to ISO8601, derivation of study day and splitting up of long text variables into 200 character variables (e.g. COVAL-COVALn).

Conversion programming was concluded with a QC step performed by an independent programmer. Standard checks were:

- Number of observations between source and targets (must be the same or differences must be accountable);
- Conformance with mapping specifications;
- Presence of all raw variables;
- Transfer of text variables (no truncation);
- Correctness of converted variables (numeric to character, recoded variables, date/time variables);
- Correctness of derivations;
- Correct transfer of variables to supplemental qualifiers or comments domain.

6. AND 7. CREATE SUBMISSION PACKAGE AND CONVERT DEFINE.XML TO DEFINE.PDF

With the finalized annotated CRFs and SDTM datasets and the metadata needed for the data definition files readily available, the last step towards a set of deliverables was relatively straightforward. First step was the conversion of the SAS datasets to transport files. Subsequently define.xml was created using the adapted md2odm macro (the appropriate schema files to accompany define.xml were downloaded from the CDISC website[7]). Finally, our define.xml-to-pdf conversion toolset was used to create define.pdf based on define.xml.

An XML editor and a specific check tool for define.xml[9] were used to ensure xml validity and compliance of define.xml with CRT-DDS 1.0. In addition, a semi-automated consistency check was performed between the define.xml and the annotated CRFs, to check whether all CRF items were presented in define.xml and whether all raw variables mentioned in define.xml

were annotated on the CRF. Moreover, a separate check was performed on non-ASCII characters in define.xml, as we learned that such special characters (e.g. `) can prevent WebSDM loading of define.xml. Finally, manual QC checks were performed on define.xml and define.pdf, e.g. completeness and functionality of the menu and links, availability and version of supporting files (schemas, style sheets) of define.xml and correctness of titles and headings.

SAMPLE SUBMISSION TO THE FDA – WEBSDM CHECKS

Before submitting a sample to the FDA, we had received and incorporated expert feedback on all submission components but had not performed WebSDM checks[10]. The define.xml had passed our external expert's validation checks. A sample submission package was prepared, consisting of the define.xml, annotated CRFs, 36 SDTM datasets (as transport files), including 7 Trial Design and 8 Supplemental Qualifier datasets, of the first study we had converted (only a limited number of subjects were included). Eight weeks after the submission, the FDA provided the following feedback:

- Loading of the datasets in WebSDM produced a list of 79 WebSDM validation check error messages.
- There was an undefined problem loading the define.xml into WebSDM. To be able to load the SDTM datasets, FDA had used standard SDTM 3.1.1. metadata. Moreover, their test environment used an older MedDRA version than specified in our define.xml.

WEBSDM ERRORS

The list of 79 WebSDM issues were related to 16 individual WebSDM validation checks (Table 1; check descriptions can be found in [13]). The WebSDM validation checks are built to help FDA reviewers identify potential data quality issues. The error messages therefore do not necessarily point to a data issue. Our first step in dealing with this list was to detect the precise cause of each validation error and determine whether there was a true data problem or a possible SDTM compliance issue.

Table 1: WebSDM check violations related to the sample submission.

Check ID	Severity	Error message	Repair	Cause and resolution
IR4001	High	Null value in column	Yes	VISITNUM missing for Pregnancy Determination data in LB, in line with protocol where these visits had no visit number. Assigned value 'UNSCHEDULED'.
IR4135	High	Missing character result when original result provided	Yes	QSSTRESC empty while QSORRES='NOT APPLICABLE'. Resolved by copying QSORRES to QSSTRESC.
IR4252	High	SDTM Required variable not found	No	Issue of sample submission only, related to absence of IE and SC data for sample subjects.
R4104	High	Rule failed to execute	No	Message not understood. Our sample data seemed to be compliant with SDTM and the validation rule seemed incorrect (as multiple DVTERMs are allowed to be coded to the a single DVDECOD).
IR4000	Medium	No rows in domain table	No	For the sample subjects submitted some datasets contained no data.
IR4009	Medium	Either Original Result or Status should be specified, but not both	Yes	Some lab records with only a sample date present ('Value not available' item was not marked). LBSTAT='NOT DONE' assigned for these records.
IR4108	Medium	Invalid BEFOREAFTER code	No	In CM one subject had a contraceptive method for which start date was not collected. As a result, CMSTRF could not be determined and was set to 'U' (unknown), in line with SDTM-IG 3.1.1 (page 29) but WebSDM only accepts 'BEFORE', 'DURING' or 'AFTER'.
IR4112	Medium	Derived Flag = 'Y' but (Character) Standard Result is null	Yes	Some observations in VS where BMI was missing (VSSTAT was 'NOT DONE'). Resolved by removing derived records with missing result.
IR4253	Medium	SDTM Expected variable not found	No	Issue of sample submission only, related to absence of IE and SC data for sample subjects.
IR4501	Medium	Invalid Subject Visit/Visit Number	Yes	In PE some missing data were included, where we had assigned PESTAT= 'NOT DONE'. but in the absence of the date of examination these records were not represented in SV. Since no actual data were captured these records were removed from PE.
IR4005	Low	No Baseline result	No	One subject did not have a QS baseline visit. This was valid.
IR4117	Low	__ENRF expected when __ENDTC is null	Yes	In several domains (e.g. SU) dates were not collected. There we chose not to include --ENRF (permissible variable). In MH, --ENRF was not applicable since the end reference period was informed consent, not last day of drug intake. Thus, (in SUPPMH) we used MHENRTPT (plus MHENTPT) instead of MHENRF. Only in the custom domain for Pregnancy Data --ENRF was added.
IR4118	Low	__STRF expected when __STDTC is null	No	Violated in several domains in which --STRF was not included because date was not collected (see IR4117).
IR4125	Low	Missing units on value	No	'Number of pregnancies' in custom domain XT, and most of the questions in QS had no unit.
IR4128	Low	Missing units on value	No	See IR4125
IR4130	Low	Start date expected when end date provided	No	For contraceptive history (in CM) the start date was not collected and therefore could not be specified. Based on the end date we did define CMSTRF (='BEFORE').

All except one finding could be readily explained. The message related to WebSDM check R4104 was not understood, and our data appeared to be compliant. This was followed-up with the FDA, who answered that this problem may have been caused by their version of WebSDM (so not a data issue). Three rule violations (IR4001, IR4252, IR4253) were due to missing data, caused by the small selection of submitted subjects. In 6 cases repair actions were performed with respect to the SDTM data, either to correct a true SDTM non-compliance (e.g. missing values of QSSTRESC where QSORRES not missing) or because the resolution would be acceptable and would neutralize a validation error with Severity='High' (e.g. changing a missing VISITNUM by value 'UNSCHEDULED' for pregnancy determination data).

After the sample submission we programmed the WebSDM validation checks (extended with some Janus checks) relevant to our data and performed these checks as part of our QC process. The (irresolvable) check violations that remained, like missing baseline visits and missing unit on lab value, were mentioned in a Reviewers' Guide that was added as part of the submission package.

LOADING FAILURE DEFINE.XML

With respect to the loading failure of define.xml additional information was requested from the FDA, in order to try to resolve the issue. The FDA responded by pointing to a misspecification of the 'Mandatory' attribute related to the ItemRef element identified by OID="TS.TSVAL1" as the possible cause of non-loading in WebSDM. Based on this we implemented the rule described on the CDISC forum ('Mandatory field in Define.xml'):

Perm(issible) variables: Mandatory = "No"
Exp(ected) variables: Mandatory = "No"
Req(uired) variables: Mandatory = "Yes"

We have not been able to conduct a final test to check whether this change had resolved the WebSDM loading issue.

SDTM CONVERSION – EXAMPLE DOMAINS

Conversion to SDTM of common data domains, like adverse events and vital signs, was relatively straightforward. In this section the conversion of some less straightforward domains is discussed to illustrate the team discussions and the thought process underlying our mapping decisions. Firstly, we will discuss the conversion of Skin Examination data to the Physical Examination (PE) domain. Secondly, we will explain why and how vaginal ultrasound (ovary examination) data was converted to the PE domain in one study and a to a custom Findings domain (XV) in another study. In addition, the main considerations with respect to the Trial Arms (TA) and Trial Element (TE) domains are briefly presented.

SKIN EXAMINATION

The occurrence and severity of acne was recorded on the skin examination form (fig. 7). Skin being one of the body systems normally pre-specified on the Physical Examination form, we quickly considered the possibility of using the PE domain for these data. Example 9.4.4 in SDTM-IG, showing records with a PETESTCD of 'SKIN', and PEORRES of 'ACNE' confirmed this thought.

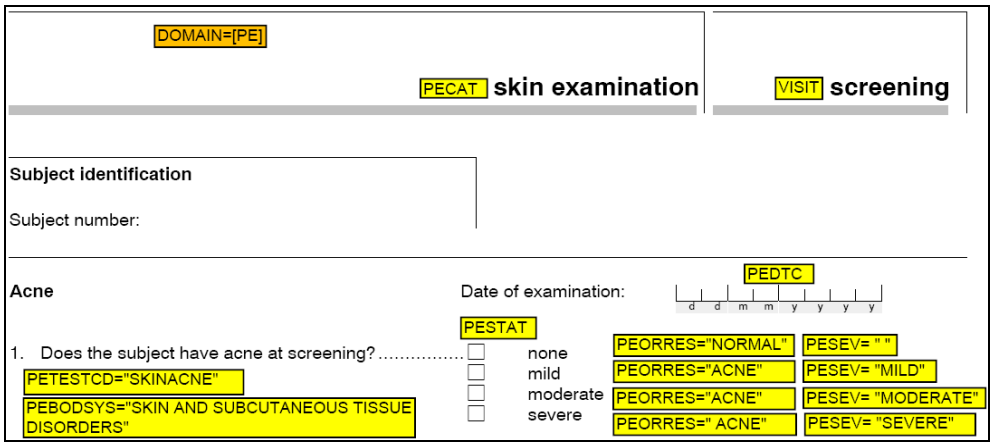

Figure 7: Skin Examination: annotated CRF

However, in our case the test needed to be more narrowly defined since this was not a general skin examination but it focused on the occurrence of acne. This was resolved by including a reference to acne in the test code and test description (PETEST='Skin (Acne)'). For PEORRES, one option considered was to just use the outcomes 'none', 'mild', 'moderate' and 'severe'. However, SDTM-IG prescribes that 'If the examination was completed and there were no abnormal findings, the value (of PEORRES) should be NORMAL'. Thus we mapped 'none' to PEORRES="NORMAL". In case of a mild, moderate or severe acne, PEORRES was assigned 'ACNE' (the abnormal finding). The three severity levels were specified using permissible variable PESEV. When none of the four tick boxes were marked the PESTAT variable was set to 'NOT DONE', in line with the SDTM-IG notes on PEORRES.

TRANSVAGINAL ULTRASOUND

In one study transvaginal ultrasound was performed to check for possible abnormalities in left and right ovaries (fig. 8 shows the CRF). Since this was an examination of a body system with outcomes abnormal or normal, the transvaginal ultrasound data also fitted in the PE domain. Two test codes (OVARYR, OVARYL) were used to distinguish between the left and right ovary.

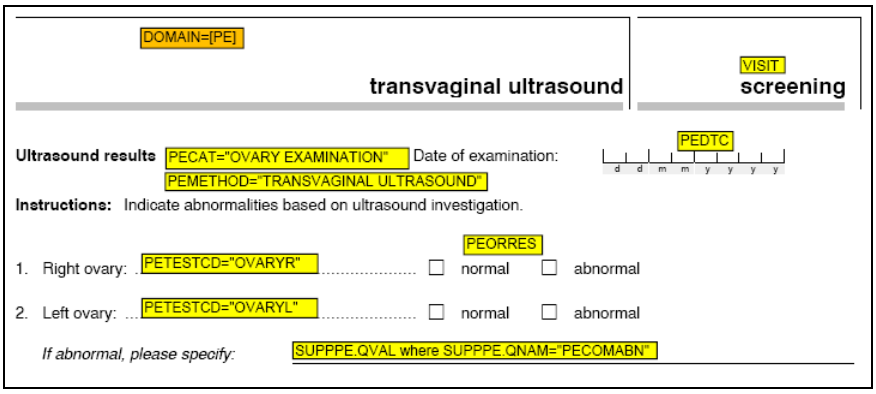

Figure 8: Transvaginal ultrasound (ovary abnormalities): annotated CRF

Abnormalities encountered were specified on the CRF. According to SDTM-IG, the textual description of the abnormality should be stored in PEORRES. However, to specify abnormalities only one field was available for both ovaries. As a result, the specified abnormality could not simply be assigned to the left or right ovary. This was solved by storing this text in a supplemental qualifier dataset and linking it to both the left and right ovary examinations. The values 'NORMAL' or 'ABNORMAL' were stored in PEORRES.

VAGINAL ULTRASOUND

In another study vaginal ultrasound was performed to determine a range of parameters (e.g. follicle size) in order to assess ovarian function (fig. 9). Since these data could not be classified as physical examinations to detect abnormalities of a body system, but contained measurements, derived scores and an evaluation from the investigator, the PE domain was not the correct domain to store this information. Instead, a custom-defined Findings domain 'XV' was used.

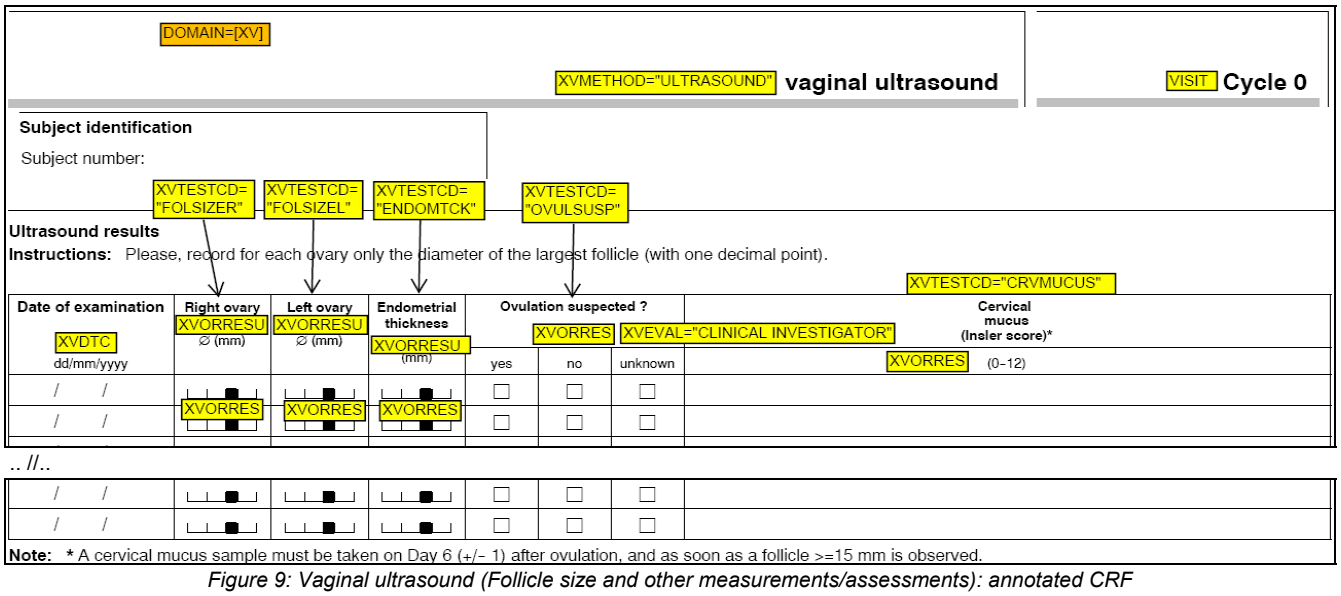

Figure 9: Vaginal ultrasound (Follicle size and other measurements/assessments): annotated CRF

For each measurement or test a distinct test code was used (e.g. XVTESTCD='FOLSIZER' for the largest follicle size in the right ovary). The item 'Ovulation suspected?' was not a measurement but an evaluation from the investigator. To indicate this, we applied the XVEVAL variable. It was discussed to use one test code for the follicle size parameters (e.g. XVTESTCD='FOLSIZE') and use XVLOC to distinguish between left and right. However, since --LOC should be used to specify the location relevant to the measurement (e.g. 'V1' for an ECG lead), we concluded it was not appropriate in this case.

The occurrence of ovulation based on the vaginal ultrasound was independently assessed by an adjudication committee, recorded on an adjudication form (fig.10).

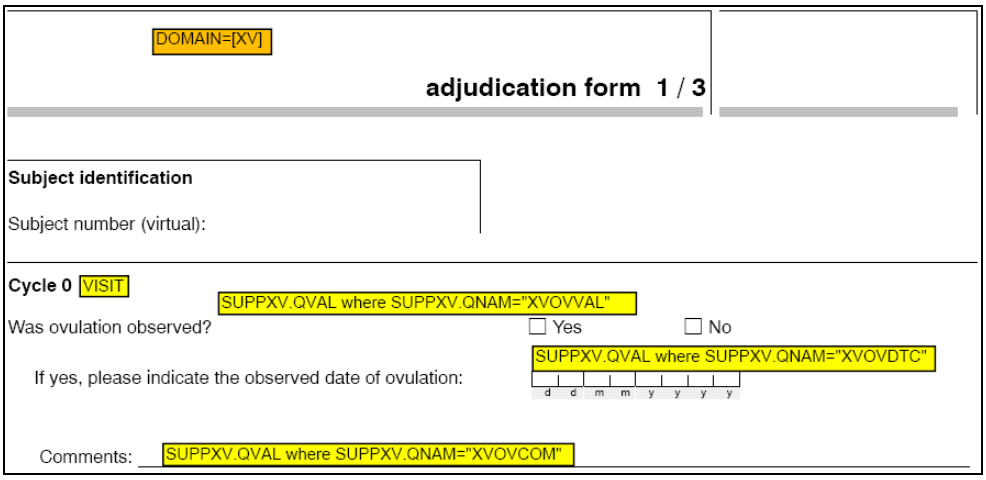

Figure 10: Adjudication form (annotated) used for independent assessment of the occurrence of ovulation

These adjudications were performed per VISIT (i.e. all examinations in a specific cycle were taken into account). For this reason, the results were stored in the supplemental qualifier dataset SUPPXV and linked to all records of the same visit using IDVAR='VISIT'. Because adjudication data represent assigned and subjective data, the Origin (QORIG) and Evaluator (QEVAL) variables in the supplemental qualifier dataset were assigned the values 'ASSIGNED' and 'ADJUDICATION COMMITTEE', respectively.

TRIAL ARMS (TA) AND TRIAL ELEMENTS (TE)

The trial arms, elements and epochs specified in the TA and TE domains can be used in several SDTM domains. We put considerable effort in defining arms, elements and epochs that are both appropriate for the TA and TE domains and can be used in other domains. SDTM-IG leaves considerable freedom how to represent the design of a trial. For example: in one study subjects were planned to be exposed to 13 treatment cycles of 28 days, with 21 days active medication followed by 7 days placebo tablets. With respect to this study we considered the following three options:

A. Generate one element with an expected duration of 364 days (13*28 days) and include this element once in the arm;
B. Generate one element with an expected duration of 28 days and include this element thirteen times in the arm;
C. Generate two elements, one with a duration of 21 days and the treatment description and one with a duration of 7 days and the description 'placebo',

In the data daily intake information was available. However, no distinction could be made between 'active' or 'placebo'. Based on that, option C was discarded as this would lead to problems generating the Subject Element (SE) domain (this domain describes the actual elements related to each subjects).

With the available information it was possible to use either option A or B. Option B was selected as it provided the most detailed (i.e. per cycle) information. One of the CRFs contained questions that were to be answered at the end of each cycle. By choosing option B it would be possible to link each answer to a specific cycle (using the TAETORD variable).

PROJECT ACTIVITY METRICS

Project activity metrics were recorded during the project to support activity and resource planning of future projects. The project was executed between June 2008 and April 2009. The project team consisted of five core members primarily involved in conversion activities. In addition, one ad-hoc member was dedicated to the development of define.xml and define.pdf.

Table 2: Overview of hours spent on various project activities

Activity	Total hours	%
CRF annotation	308	8.0
Define.xml	574	14.9
Mapping	309	8.0
Meetings	267	6.9
Programming	795	20.6
Project Management	626	16.2
Validation	985	25.5
Total	3864	100

An overview of the hours spent by the team on the various activities is shown in table 2. Reading these figures it must be taken into account that:

- CRF annotation includes the team effort on first study, where annotation was used to support the joint modeling/mapping;
- Programming includes general programming activities such as utility macro updates and development (e.g. WebSDM validation checks), generation of the Trial Design domains, and maintenance of metadata tables. General programming activities comprised approximately 15% of all conversion programming.

Expert review is not included in these metrics, but the total external review time amounted to about 50% of what was spent internally on validation.

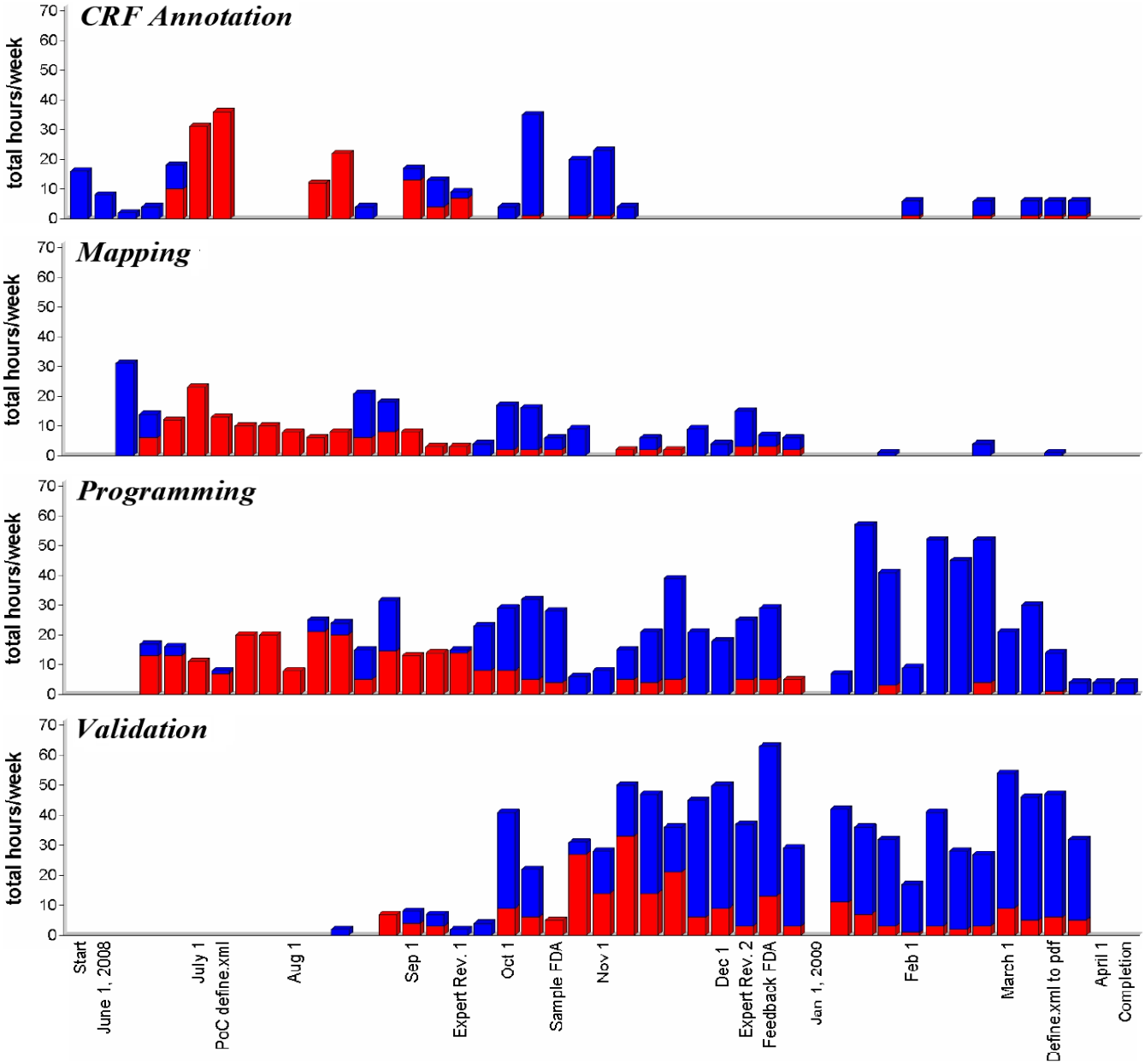

Figure 11: Distribution of hours spent on CRF Annotation, Mapping, Programming and Validation (first study highlighted in red)

Figure 11 presents the distribution of hours spent on four main activities during the course of the project. The first study ('001'), which acted as pilot for the project, is highlighted in red. On the time axis, where each bar represents one week, the main project milestones have been indicated. These are:

1. Proof-of-Concept define.xml;
2. Completion of the first external expert review (mapping and annotation domains first study and technical test define.xml);
3. FDA Sample submission, marking the completion of the SDTM conversion of the first study, including define.xml and annotated CRF;

4. Completion of the second external expert review (mapping and annotation new domains, dataset validation, review Trial Design data, full review define.xml);
5. FDA feedback on the sample submission;
6. Finalization of the tool to convert define.xml to define.pdf;
7. Completion of all conversion activities.

Table 3 presents the metrics of the four main activities per study (including those related to the Integrated Summary of Safety and general activities).

Table 3: Metrics of CRF Annotation, Mapping, Programming and Validation per study

Study information			Metrics per activity							
			CRF Annotation		Mapping		Programming		Validation	
Study number	# CRF pages	# SDTM Domains	Hours	Hours / page	Hours	Hours / domain	Hours	Hours / domain	Hours	Hours / domain
001	45	25	143	3.2	158	6.3	215	8.6	322	12.9
002	46	27	47	1.0	14	0.5	43	1.6	199	7.4
003	25	23	27	1.1	14	0.6	41	1.8	119	5.2
004	20	23	23	1.2	12	0.5	70	3.0	69	3
006	24	25	19	0.8	16	0.6	75	3.0	91	3.6
011	23	23	23	1.0	4	0.2	105	4.6	84	3.7
ISS	n.a.	13	n.a.	n.a.	12	0.9	120	9.2	101	7.8
General	n.a.	n.a.	26	n.a.	79	n.a.	126	n.a.	n.a.	n.a.

The 001 study was used as a pilot. Since the other studies had similar CRFs, many domains in common and followed the same source data standard, much of the annotations, mapping sheets, conversion programs and validation programs developed for 001 could be reused in the other studies. This explains the relatively large number of hours spent on 001 relative to the other studies (table 3).

Programming and mapping was mostly performed by the same person. Omissions or issues in the mapping files were usually solved during programming and hours subsequently assigned to this activity, which explains the relatively high figures for Programming as compared to Mapping. Next to the 6 studies mentioned, the ISS data comprised the data of two more studies (with a different CRF and data definition background as the other studies). This clarifies the relatively high number of hours for mapping and programming for ISS. After study 001 was completed, annotation became rather a routine/technical exercise as the other studies presented very few new CRFs. This is reflected in the similarity of the hours/page figures for Annotation (table 3). New domains and the implementation of Lab controlled terminology seem to explain the higher number of hours on Programming in 004, 006 and 011. New domains, PC, PP, RELREC and EG, with data transferred via electronic data transfer (hence without CRFs that needed annotation), occurred in studies 006 and 011. A great number of new lab tests were introduced in studies 004 (36 new tests compared to 001) and 006/011 (41 new tests compared to 001/004). In total 23 sponsor-defined tests needed to be added to CDISC controlled terminology for LBTEST/LBTESTCD divided over studies 004, 006 en 011. Programmer experience level may also have contributed in part to the observed differences in programming metrics.

Validation covered the review of annotated CRFs, define.xml and define.pdf, mapping sheets, and program code, the technical validation of define.xml and SDTM dataset validation. For the ISS data some statistical summary tables were reproduced using the SDTM datasets. This activity was also assigned to validation. Late changes in the source data, feedback and changes of view caused part of the deliverables to require more than one validation cycle. Altogether this caused validation to consume a large amount of resources. A possible explanation for the decrease in hours needed for the validation of studies that were covered later in time is a gradually increasing efficiency of the validation process.

CONCLUSION

The implementation of SDTM, i.e. choosing the appropriate domains, properly representing all source data and the design of the trials, implementing controlled terminology and following domain assumptions is not a straightforward process and can be time-consuming[11]. We chose to complete the conversion of one study first and spend time with the team to go through a learning curve. This strategy paid off in several ways: it allowed efficient conversion of other studies, motivated the team, and supported the external review process and sample submission. Even though we had no prior experience with the Trial Design datasets, they presented no exceptional challenge during the project. We realize this could be different when designs are more complex and vary strongly between studies.

A sample submission to the FDA drove the implementation of WebSDM checks pertaining to our data, and helped to make some technical improvements to define.xml. In retrospect, one could question whether a sample submission is the most effective way to achieve this, given the indirect communication and long feedback cycle. Solution providers offer services that perform the same checks with direct support. Nevertheless, the sample submission offered clear benefits. Firstly, it presented a milestone that got everyone in the same direction, speeding up activities and decision-making. Secondly, it helped to prepare for the loading of SDTM datasets (next to the analysis datasets) and define.xml in the eCTD by our regulatory department.

Support from an (external) SDTM expert was an important factor for the success of this project. The most significant and distinguishing role for the experts in this project was support with regard to domain selection and modeling. Examples of revisions based on expert input are the implementation of end-of-trial and end-of-treatment data in the DS domain (including the use of the EPOCH variable), the use of a custom Findings domain instead of the SC domain for Socioeconomic data, and the change from an Events to a Findings (custom) domain for reproductive status questions. Since these review activities were conducted largely in parallel with other activities it probably also accelerated the project. Another positive effect of expert input (including confirmation of many things that we did right!) was that it supported the confidence of team members and stakeholders in our approach and deliverables. Finally, it was very helpful to have an expert available for (sample) submission support.

The ideal scenario with respect to the creation of define.xml was that we could treat the toolset from the CDISC SDTM/ADaM pilot as a black box, requiring only a few changes and little knowledge of XML/ODM. This scenario did not become reality, but thanks to the technical skills, dedication and quick learning of our developer and some expert input we were able to produce define.xml within a more than reasonable time (and with extended functionality like the automated insertion of page references). It was no doubt much faster and more cost-effective than producing our own tool from scratch. We recommend anyone considering a similar enterprise to first acquire at least a basic understanding of XML, the ODM model and CRT-DDS requirements (an elaborate overview is given by Molter[12]).

The SDTM conversion tools and metadata repository inherited from earlier pilot activities greatly facilitated the conversion process. After we extended the repository with variable comments for define.xml and made adjustments to support study-specific subsets of CDISC controlled terminology it was possible to repurpose the metadata for define.xml. A metadata-based approach has clear benefits for efficiency, quality and process management[13]. Managing the metadata was performed by one person and with the use of basic tools (i.e. Excel and Base SAS). This fitted the size and purpose of our project. Managing a global metadata repository that undergoes continued change and has multiple users will require a different approach and toolset[14].

REFERENCES

1. Strategies for Implementing SDTM and ADaM Standards. Susan J. Kenny & Michael A. Litzsinger, PharmaSUG 2005
2. CDISC SDTM/ADaM Pilot Project Submission Package, www.cdisc.org/downloads/900171_2008_01_22T1920.zip
3. “ODM 1.2 or ODM 1.3”, CDISC discussion forum: www.cdisc.org/bbs/forums/thread-view.asp?tid=2972
4. Case Report Tabulation Data Definition Specification (define.xml), Version 1.0, 2005. www.cdisc.org/models/def/v1.0/CRT_DDSpecification1_0_0.pdf
5. Study Data Specifications, version 1.4, 2007. www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM163561.pdf
6. www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/ucm153574.htm
7. www.cdisc.org/models/def/v1.0/SchemaFiles.zip
8. Using the SAS® XML Mapper and ODS to create a PDF representation of the define.xml. Lex Jansen, PhUSE 2008.
9. The XML4Pharma CDISC Define.xml Checker, www.xml4pharma.com/CDISC_Define_Checker/
10. Validation Checks Performed by WebSDM™ on SDTM version 3.1.1 Datasets, version 1.5, 2007. www.phaseforward.com/products/safety/documents/ValidationChecksPerformedbyWebSDMtm.Q107.pdf
11. Case Study: Analysis and Metrics of End-to-End Legacy Data Conversions into SDTM, ADaM, and Define.xml. Robert Stemplinger, PharmaSUG, 2008
12. A SAS® Programmer’s Guide to Generating Define.xml. Michael Molter, SAS Global Forum 2009.
13. From CRF Data to DEFINE.XML: Going “End to End” with Metadata. Frank DiIorio & Jeffrey Abolafia, PharmaSUG 2007
14. Managing The Change And Growth Of A Metadata-Based System. Jeffrey Abolafia & Frank DiIorio, PharmaSUG 2008

ACKNOWLEDGEMENTS

The authors first of all wish to thank the members of the project team: Daan de Wildt, Herman Ament, Fred van Dillen, Don Janssen and Gertjan van Maaren. We are also thankful to Renato de Leeuw, Hans van Leeuwen and Peter Stokman, as members of the “CDISC@GCI” steering committee. This project wouldn’t have been possible without the vision and continued support from Kit Roes who sponsored this and earlier CDISC@GCI projects and chaired the steering committee. Finally, we would like to acknowledge Barbara Palladino for her critical contribution in the final phase of the project.

CONTACT INFORMATION

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

Author name :	Paul Vervuren	Author name :	Bas van Bakel
Company :	Schering-Plough	Company :	OCS Consulting
Address :	PO Box 20 5340 BH Oss The Netherlands	Address :	PO Box 490 5240 AL Rosmalen The Netherlands
Work phone :	+31 (0) 412 663921	Work phone :	+31 (0)73 523 6000
E-mail :	Paul.Vervuren[-at-]spcorp.com	E-mail :	Bas.vanBakel[-at-]ocs-consulting.com
Website :	www.schering-plough.com	Website :	www.ocs-consulting.nl / www.ocs-consulting.com