

CDISC SHARE

CDISC Shared Health and Research Electronic Library Pilot Report

Date: 19 January 2010

Version: 1.0



© CDISC, 2010

Document History

Issue	Author	Date	Description
0.1	Pilot Team	9 December 2009	First draft
0.2	Pilot Team	10 December 2009	Updated after TC on 9 th December
0.3	Pilot Team	15 December 2009	Updated – Glossary, abbreviations table added, comments incorporated.
0.4	Rhonda Facile	22 December 2009	Comments incorporated, glossary amended.
0.5	Dave Iberson-Hurst	4 th January 2010	Editorial amendments
0.6	Rhonda Facile	6th January 2010	Editorial amendments
0.7	Rhonda Facile	11 th January 2010	Final comments from team incorporated
1.0	Pilot Team	19 January 2010	Final Report.

CDISC, Inc.
15907 Two Rivers Cove, Austin, Texas 78717
<http://www.cdisc.org>

© Copyright 2010 by CDISC, Inc.

All rights reserved. No part of this publication may be reproduced without the prior written consent of CDISC.

CDISC welcomes user comments and reserves the right to revise this document without notice at any time. CDISC makes no representations or warranties regarding this document. The names of actual companies and products mentioned herein are the trademarks of their respective owners.

CDISC® and the CDISC logo are trademarks or registered trademarks of CDISC, Inc. and may be used publicly only with the permission of CDISC and require proper acknowledgement. Other listed names and brands are trademarks or registered trademarks of their respective owners.

Table of Contents

1	INTRODUCTION	5
2	SCOPE	6
3	GLOSSARY	7
4	ACRONYMS	9
5	METHOD	10
5.1	PROCESS	10
6	RESULTS	12
6.1	TIME NEEDED FOR DATA PREPARATION BY CONTRIBUTING ORGANIZATIONS	12
6.2	LOAD DATA ELEMENTS INTO WIKI	12
6.2.1	Data Element Load Template.....	12
6.2.2	Code List Load Template	13
6.3	DATA ELEMENTS AND CODE LISTS HARMONIZED	13
6.3.1	Table: Summary of DEs Reviewed and Harmonized.....	13
6.3.2	Code Lists.....	14
6.3.3	Code lists Submitted for Adverse Event Toxicity Grade (AETOXGR)	15
7	KEY QUESTIONS	17
7.1	QUESTION 1	17
7.2	QUESTION 2	17
8	LESSONS LEARNED	19
8.1	DATA DEFINITION-CONTENT	19
8.1.1	BRIDG	19
8.1.2	Foundational Terminologies	19
8.1.3	Code Lists.....	19
8.1.4	Data Element Definitions	19
8.1.5	Data Types	19
8.1.6	Preproduction – Templates	20
8.1.7	Strategy for populating the repository	20
8.1.8	Version Control.....	20
8.1.9	Unsolicited Adverse Events.....	20
8.1.10	Short Names and Long Names	20
8.1.11	Data Elements Concepts and Groups.....	21
8.2	WIKI FUNCTIONALITY	21
8.2.1	Landscape (horizontal) View of DEs by Company	21

8.2.2	Code List Harmonization	21
8.2.3	SDTM & CDASH	21
8.2.4	Search Functionality	21
8.2.5	Searching and Viewing Terminologies	22
8.2.6	Horizontal and Vertical Structure.....	22
8.3	GENERAL OBSERVATIONS.....	22
8.3.1	The Need for Active Participation by Contributing Company Representatives	22
8.3.2	Clinical Input	22
9	CONCLUSIONS	23
10	APPENDIX 1: DATA ELEMENTS CREATED	24
10.1	ADVERSE EVENT	24
10.2	LESION MEASUREMENT.....	25
10.3	BLOOD PRODUCTS	26
10.4	ECHO	26
10.5	KARNOFSKY PERFORMANCE SCALE	27
10.6	CHEST X-RAY	27
11	APPENDIX 2: DATA ELEMENT GROUPS CREATED	28

1 Introduction

This document summarizes the observations and findings of the CDISC Share Wiki pilot project and is organized into four sections: Method, Results, Answers to Primary Questions, Other Lessons Learned and Conclusions.

The primary purpose of the pilot was to address the following questions:

1. Can definitions taken from multiple sources be merged into a single version agreed to by all parties and can this be done within a timeframe that makes business sense?
2. Can high-quality definitions be created and can ontologies help in ensuring such and avoid duplicate definitions being created?

A secondary aim of the pilot was to provide any relevant lessons to help in development of the production version of the CDISC Share library.

Another stated goal of the pilot was to collect metrics around the time needed to perform data element harmonization. This proved difficult due to two main reasons, attendance by company representatives on working teleconferences and wiki functionality.

Due to inconsistent attendance on teleconferences, there were many occasions where a subset of the team spent quite a lot of time “guessing” the function/purpose of some of the contributed DEs as the company representative was not present on the call. Even when the company representative was present on the teleconference it was sometimes hard to get an accurate picture of how the CDE was used in practice by the contributing company.

Wiki functionality was also a major issue resulting in the majority of teleconferences being spent learning how to use the wiki, identifying areas that need improvement and suggesting user interface improvements. Mayo made many adjustments to the wiki as a result of this iterative process. However, there remain areas that need enhancement to create software that will be sufficiently robust for this purpose. A discussion of the issues noted is included in section 4.2 Wiki Functionality.

2 Scope

The pilot was limited to a review of a very small set of target common data elements (CDE's) and their properties. For CDEs, we identified short and long names, a (non-robust, non-ontologically based) definition, a data type, a codelist and some "matches" of the CDE against the Metathesaurus and BRIDG. For Groups we did little more than identify a name and associate the appropriate CDEs with the Group.

3 Glossary

<i>Preferred Term</i>	<i>Synonym</i>	<i>Definition</i>
Codelist	Code list, Controlled Terminology (CT), Value List	A set of allowable values for a given common data element. Code lists are for common data elements only.
Code List Value		An allowable value within a code list for a given data element.
Code System		A systematized collection of concepts that defines corresponding codes.
Common Data Element (CDE)	Clinical Data Element, Harmonized Data Element	Data elements that have been created for use between projects, contexts or organizations.
Concept		A unit of knowledge created by a unique combination of characteristics and having single meaning.
Concept Code		A concept unique identifier.
Concept Reference		A concept and its unique identifier that is contained within the NCI Metathesaurus, BRIDG, or other terminology, which maps to either a CDISC SHARE data element or group.
Constituent Data Element		A data element or common data element, which is part of a Group.
CDISC SHARE Library		CDISC SHARE is a global, accessible, electronic library, which through advanced technology enables precise and standardized data element definitions that can be used within applications and across studies to improve biomedical research and its link with healthcare.
Data Element (DE)		A unit of data for which the definition, identification, representation, permissible values are specified by means of a set of attributes. In CDISC SHARE, it comes from contributing sources and is associated with a common data element.
Dataset		A collection of related data records.

Data Type	Data Types	Data types define the structural format of the data carried in the attribute and influence the set of allowable values an attribute may assume (e.g. string, integer, and character). (HL7)
Domain		A collection of logically related observations with a topic-specific commonality about the subjects in a trial, used in the regulatory submission process. The logic of the relationship may relate to the scientific subject matter of the data, or to its role in the trial. Typically, each domain is represented by a dataset, but it is possible to have information relevant to the same topicality spread among multiple datasets. (Source: CDISC Glossary)
Group		A set of data elements or common data elements with a logical, often hierarchical, relationship. For example, "Systolic Blood Pressure unit" logically implies the existence of "Systolic Blood Pressure".
Long Name		A literal identifier of a harmonized data element.
Permissible Value		A value assigned to a coded variable that identifies a particular meaning in the context of that variable. ISO 1179 definition: expression of a value meaning allowed in a specific value domain
Pseudo Common Data Element		A group of contributed data elements, which the CDISC SHARE pilot team decided not to attempt to harmonize. These were grouped using a wiki mechanism that labelled them as CDEs, but these would generally not ever become actual common data elements.
Short Name		An abbreviated literal identifier for a common data element. A short name can also be numeric.
Super Group		A set of data elements or groups that are topically or conveniently related to one another. For example, Systolic and Diastolic Blood Pressure groups might be contained within a Vital Signs super group.
Variable		A single data collection item.

4 Acronyms

BRIDG	Biomedical Research Integrated Domain Group (BRIDG) Model
CDASH	Clinical Data Acquisition Standards Harmonization
CDISC	Clinical Data Interchange Standards Consortium
CDISC SHARE	Shared Health and Clinical Research Electronic Library
SDTM	Study Data Tabulation Model
WHO	World Health Organization
FDA	Food and Drug Administration
LexGRID	Lexical Grid
ICD	The International Classification of Diseases
GSK	GlaxoSmithKline
CTCAE	Common Terminology Criteria for Adverse Events
HL7	Health Level 7
OpenEHR	Open Electronic Health Record

5 Method

Using the semantic wiki tool (LexGRID) provided by Mayo Clinic, 50 oncology data elements (DE) from 5 volunteer organizations (Mayo Clinic, GSK, MD Anderson, Eli Lilly and Genzyme) along with valid code lists were identified and loaded into the system. The team followed a process to align equivalent data elements resulting in a single consensus definition. As the team undertook the alignment work the process was refined and the wiki amended to better support the process. Metrics to evaluate the process and use of the wiki, along with benefits and risks, were collected and reported. Specifically these metrics included time needed to prepare, load and add concept references as well as the time needed for the harmonization process.

The wiki was loaded with various terminologies/dictionaries such as the NCI Thesaurus, CDISC Controlled Terminology (CT), the SNOMED CT, and ICD 9 and 10 along with the BRIDG structure. This permitted an assessment of how these support the process of aligning the definitions from the various contributing organizations and facilitated an assessment of how these support the improvement in the quality of the definitions created and prevent duplicate definitions being created.

The wiki can be found at http://informatics.mayo.edu/cshare/index.php/Main_Page.

5.1 Process

As described in the initial pilot report, after a series of iterative discussions, evaluations, prototype steps, etc. a prototype harmonization process was developed. This process evolved into three steps:

Contribute & Link

This step involved the identification of relevant DEs by the participant organization, loading them into the wiki (contribute) and then adding names, definitions and semantic categorizations to the individual data elements (link). Individual community members who were familiar with the use and purpose of the elements did the linking step.

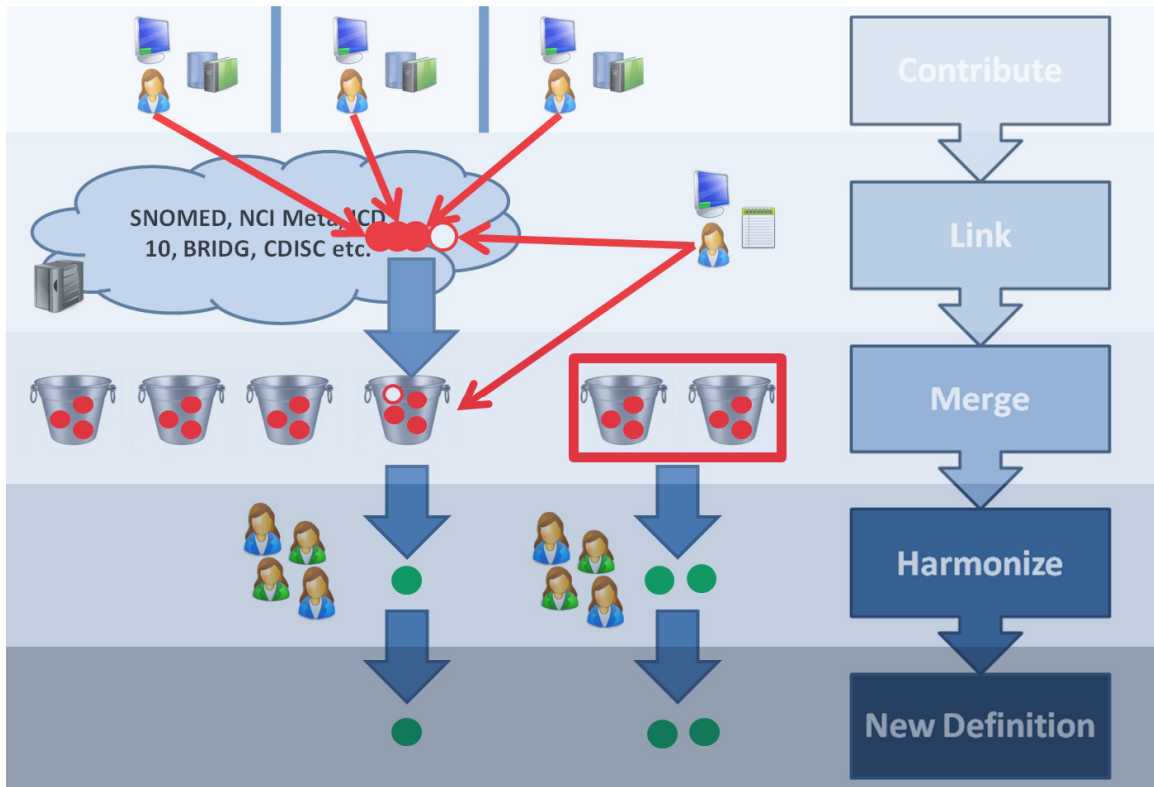
Merge

Selecting and sorting the annotated data elements to locate those that were closely related. This step has been referred to as “selecting and sorting” and it is done using a tool that allows users to narrow their view of the DEs by name, short name, long name, data types, concept tags etc.

Harmonize

Locating or, if necessary, creating one or more common data elements (CDE) that represent the community semantics represented by the selected elements. This step also involved establishing the closeness of the match between the community data elements and common element.

The following diagram outlines the general process followed.



6 Results

6.1 Time Needed for Data Preparation by Contributing Organizations

During the pilot limited information was collected regarding the time needed to identify, extract and prepare data elements for import into the wiki. Time required to perform these tasks varied by organization but generally contributors reported that they spent between 8 and 10 hours pulling the information together.

6.2 Load Data Elements into Wiki

The following table shows the number of data elements, the number of permissible values we loaded for each organization and time spent for loading. As long as the format is standard, the loading time is negligible. There were occasional problems involving special characters and poorly formatted values. These cases it took a little longer to debug.

Organization	Number of Data Elements loaded	Number of permissible values loaded	Time spent loading
Genzyme	111	527	<30 min
GSK	156	1200	<30 min
Lilly	101	1002	<30 min
Mayo Clinic	75	5230	<30 min
MD Anderson	346	959	<30 min

To facilitate the data loading process two templates were used:

6.2.1 Data Element Load Template

Domain	Dataset	DE Short Name	DE Long Name	DE Definition	DE Data Type	Unit	Codelist
Oncology	Lesion Measurement	LSLOC	Lesion Location	Lesion Location			Lesion Location
Oncology	Lesion Measurement	LSMOM	Lesion Method of Measurement	Lesion Method of Measurement			Lesion Method of Measurement

(Shaded and italics = sample data)

6.2.2 Code List Load Template

Codelist	Valid value	Value meaning	Value description	Code system	Concept code
Lesion Location	1	Abdomen			
Lesion Location	10	Aortic arch			

(Shaded and italics = sample data)

6.3 Data Elements and Code Lists Harmonized

The following table summarizes the number of data elements contributed and harmonized. Once the Mayo team made improvements to the wiki and the team members became more proficient in using the software the harmonization process was more efficient. It is important to note that only the last 4-5 calls were spent on the actual harmonization process.

6.3.1 Table: Summary of DEs Reviewed and Harmonized

	Domain	Contributed DEs	Harmonized DEs	Status / Disposition
1.	Adverse Event	172 (+15 SDTM+ 6 CDASH)	12	In progress. 30 min pre-work+3 hours
2.	Lesion Measurement	136	21 + 2 pseudo CDEs	Completed. (25 minutes)
3.	Blood Products	30	8 + 1 pseudo CDE	Completed. (25 minutes)
4.	Echo	23	3	Not completed group decided to leave as is due to horizontal & vertical structure issues. (40 minutes)
5.	Karnofsky Performance Scale	37	3 + 1 pseudo CDE	Completed. (30 minutes pre-work + 25 minutes)
6.	Chest X-ray	21	4 +1 pseudo CDE	Completed. (15 minutes)

For further details see Appendices 1 and 2.

6.3.2 Code Lists

The pilot explored the wiki capability where individual codes can be matched with entries from NCIM. For example, we matched GSK's value of "Cryoprecipitate" with "NCIM Cryoprecipitate (C0443121)" and GSK's value of "Erythropoietin" with "NCIM Erythropoietin (C0014822)", "NCIM Blood erythropoietin (C0920092)" and "NCIM Erythropoietin therapy (C0199970)".

The rationale for matching company code list values to standard terminology is to set things up so that the wiki can then identify matches. For code lists with many values, this would be very time consuming if matched manually. In our attempts to match company CDEs to NCIM CDEs, direct matches were too rare to significantly reduce the volume of manual work.

The opinion was expressed that having a tool take an existing code list, compare it with "external" code lists and report back with a prioritized list of external code lists with a score (big score=close match, small score=poor/no match) would be really useful. Doubt was expressed as to whether external "tailored" code lists can be identified.

It is clear we need tools to facilitate the harmonization of code lists. We did note that where possible, we'd think that CDISC SHARE should use existing code lists when appropriate ones exist.

The discussion also triggered thoughts in the minds of some meeting participants about the different types of code lists CDISC SHARE will need to handle.

1. non-extensible code lists (e.g. "AE severity")
2. extensible code lists which are collections and the addition of additional values does not change the meaning of any existing values (e.g. "UNITS")
3. extensible code lists where addition of new values may affect existing values or their interpretation (e.g. adding the value "Dark Green" to a code lists with existing values "Black", "White" and "Green")
4. extensible code lists where ontologies could have been used but weren't (rather than use MedDRA or SNOMED, values have been made up, resulting in multiple values with essentially the same meaning)
5. hierarchical code lists e.g. race/sub-race.

Due to the limited number of code lists submitted by the participating companies, codelist harmonization was limited. The team did look at submitted values for AETOXGR. This submitted values were clipped out of the wiki and pasted into Excel to accommodate comparison as tools for the codelist harmonization are currently not available.

6.3.3 Code lists Submitted for Adverse Event Toxicity Grade (AETOXGR)

6.3.3.1 Genzyme Toxicity Grade Code List

Code List	Value	Value Meaning
TOXGR Code list/1	1	MILD
TOXGR Code list/2	2	MODERATE
TOXGR Code list/3	3	SEVERE
TOXGR Code list/4	4	LIFE THREATENING
TOXGR Code list/5	5	FATAL

6.3.3.2 Mayo Toxicity Code List

Code List	Value	Value Meaning
CTC Adverse Event Grade Code list	0	Not supplied
CTC Adverse Event Grade Code list/1	1	Not supplied
CTC Adverse Event Grade Code list/2	2	Not supplied
CTC Adverse Event Grade Code list/3	3	Not supplied
CTC Adverse Event Grade Code list/4	4	Not supplied
CTC Adverse Event Grade Code list/5	5	Not supplied

6.3.3.3 GSK Toxicity Code List

Code List	Value	Value Meaning
AETOX Code list/1	1	Grade 1
AETOX Code list/2	2	Grade 2
AETOX Code list/3	3	Grade 3
AETOX Code list/4	4	Grade 4
AETOX Code list/5	5	Grade 5

6.3.3.4 Harmonized Toxicity Grade Code List

Code List	Value	Value Meaning	Value Meaning
	1	Grade 1	MILD

	2	Grade 2	MODERATE
	3	Grade 3	SEVERE
	4	Grade 4	LIFE THREATENING
	5	Grade 5	FATAL

It was noted that the Mayo terms had "0" while Genzyme and GSK did not. Although at time the group could not determine whether either version of the CTCAE (version 3.0 and 4.0) list "0" as an accepted grade*, this difference between submitted lists from Genzyme and GSK prompted a discussion of whether the Wiki should "look forward" and/or be backward compatible. The group felt that the wiki should be forward looking while allowing 2 views (i.e. linking between the old and new).

The group decided to accept the harmonized list above as it is directly from the Common Terminology Criteria for Adverse Events v3.0 (CTCAE), Publish Date: August 9, 2006 - add this reference as an attribute to this code list. The meaning of the CTCAE grade is dependent on the disease area as described in the Common Terminology Criteria for Adverse Events.

** It has been confirmed that Grade "0" is in fact part of the CTCAE v4.0 standard; it is just not in the pdf booklet form of v4. Grade 0 is an official part of CTCAE and NCI reflects this.*

7 Key Questions

As described in the beginning of this report the pilot posed two main questions:

1. Can definitions taken from multiple sources be merged into a single version agreed to by all parties and can this be done within a timeframe that makes business sense?
2. Can high-quality definitions be created and can ontologies help in ensuring such and avoid duplicate definitions being created?

7.1 Question 1

The CDISC Share Pilot project has shown that data from multiple companies can be merged and an agreed to harmonized data element can be created in a relatively short period of time based on a CDE, individual code values to include in a code list, and perhaps CDEs associated in a group. In addition, after an initial investment in time to learn how to use the software and with rules applied that define the quality and quantity of data loaded, opportunity exists to dramatically speed up the time it takes to create standard, harmonized data element definitions.

It has been demonstrated that definitions can be merged in a sensible timeframe and the benefits to an individual organization and to industry as a whole of CDISC SHARE were detailed within the Scope and Vision document.

What we have not shown is if we can agree on a specific and limited collection of objects. For example, we agreed to have a “chest x-ray procedure clinical significance” CDE. But what if, on use, a company asks for something subtly different? Do we say, “No, this is what you need to use” or do we say “gosh, we got it wrong, let’s create another”? The same applies for Groups and code lists. This will have to be addressed and agreed to with rules and governance in the production version of the CDISC SHARE Library. An alternative option is that the company can create their own localized version in their own repository for their own use. Work on the general governance process has been done by the governance subgroup of the CDISC SHARE project. Diagrams of suggested process flow were included in the above-mentioned Scope and Vision Document.

Tied to the governance issue is the question of the speed at which the CDISC SHARE organization responds to a requested change. For example MedDRA has a fast turnaround of 1 week while the WHO has a slow turnaround for additions to WHODrug. CDISC SHARE needs to determine what service level it will provide.

What is clear is that CDISC SHARE will only add benefit if organizations use CDISC SHARE in preference to their own (internal) definitions. Also use by the FDA may well be critical to the success of the project.

7.2 Question 2

The pilot did not directly help in answering this question. However, we know that high quality definitions can be created. Other organizations such as OpenEHR have shown that this can be done. A combination of choosing the right target data elements (atomic rather than composite objects) and the use of ontologies can ensure that hugely beneficial, non-ambiguous, non-duplicative CDEs and Groups can be created. If the CDISC SHARE gets the right contributors – “right” meaning a broad spectrum of organizations actively participating– then CDISC SHARE can be successful.

Supporting ontologies (NCI Metathesaurus, SNOMED, etc.) can play a role in helping to understand the meaning of data elements. However, in the pilot the BRIDG classifications seemed more useful in helping the users to understand the contributed concepts. One aspect that hampered the use of supporting ontologies was the inability to see the entire list in a hierarchical (tree) view and the ability to easily search across the range of ontologies available. This limitation

did not allow for users to see other terms that might be more appropriate for a given data element.

From the start of the pilot we specifically looked for duplicate definitions, as this was a known issue in other repositories, but this proved not to be an issue during the active part of the pilot project.

Finally and unsurprisingly the issue of the quality of the source material is a key factor in developing high-quality harmonized definitions. Requirements for data elements loaded into the production environment must to be clearly defined and rigorously followed to ensure that quality definitions can be developed.

8 Lessons Learned

The following are observations noted by the team during the pilot. They are divided into three broad categories: Data Definition-Content, Wiki Functionality, and General Observations.

8.1 Data Definition-Content

8.1.1 BRIDG

The BRIDG provided a way to classify data elements (e.g., as a result, a method, the name of a product), especially in cases when little more than a name was provided. BRIDG, having a limited number of objects, is relatively easy to map to. Getting companies to map to BRIDG objects provides a relatively quick and easy route to identifying a small subset of contributed CDEs for review.

8.1.2 Foundational Terminologies

These terminologies were not as useful as originally thought in the pilot. Foundational terminologies have too many objects and are not a good route for identifying a small subset of contributed CDEs for review. More needs to be done to assess their utility for naming and defining CDISC SHARE objects.

The wiki “favored” the NCIM so it was not easy to see other applicable terms in other dictionaries. Also, the wiki did not allow the team to see a “tree” of terms that would have facilitated making better choices in the linking stage.

The team thought that these terminologies might be more useful in creating groups, since terms were generally less granular than data elements. For example, date/times for particular observations were generally not in the terminologies.

8.1.3 Code Lists

Code Lists are extremely helpful in understanding data elements, especially when definitions are missing. Very few code lists were submitted by the contributing organizations. Therefore only one code list could be harmonized (AETOXGR).

The wiki did not support the viewing and harmonization of submitted code lists. The harmonization of the AETOXGR code list was done outside the wiki environment in an excel spreadsheet.

In the CDISC SHARE production environment, only one code list could be viewed at any time. The ability to view multiple code lists side by side would have been helpful.

8.1.4 Data Element Definitions

Data element definitions, even incomplete ones, are needed in order to help to understand the meaning of the DEs submitted. Companies often have only short descriptions in the form of metadata while proper definitions are embedded in documents elsewhere. Before submission of CDEs by companies, those companies need to enhance the metadata they extract from their systems.

Data type, and code lists (where applicable), can sometimes act as partial substitutes for definitions.

8.1.5 Data Types

Data types would be helpful. These were missing or incorrect in many cases. Data types should be included where possible, and the guidelines for submitting to CDISC SHARE must have clear guidelines for allowable data types.

Many pilot participants were unfamiliar with complex data types, but learned about them during the course of the pilot. We decided that harmonized data elements would use complex data types. This leads to fewer harmonized common data elements (for example, only one data element with data type TS is need to handle dates, times, and variations), but means that there is likely to be a gap between common harmonized data elements and the data items in sponsor implementations. CDISC will have to decide on an approach to addressing this gap.

8.1.6 Preproduction – Templates

The team agreed that CDISC SHARE must draft guidelines for accepting data to be included in the production software.

Using a standard template and setting minimum requirements for data submission will facilitate both a shortening of load time and ensure data quality and completeness. See page 3 for the templates used in this pilot.

8.1.7 Strategy for populating the repository

The pilot team discussed strategies for quick population of the CDISC Share library. One suggestion is to start with SDTM topic variable values (e.g., names of findings, interventions, and events), then move to other CDEs. It is worth noting that we have not yet tied down the target for CDISC SHARE content. So far we've only been looking at a subset of the metadata for CDEs and for groups. There is much more metadata that could be defined.

8.1.8 Version Control

In working on AE toxicity grading, we realized that there are 2 different versions of CTCAE coding lists, Version 3, and the recently released Version 4. The team observed that rules will be needed for responding to changes in standards referenced by CDISC SHARE. Our initial agreement was that new DEs should be created for a new version of a referenced standard. However, the rules may depend on how the referenced standard handles its versioning.

Note: Interestingly, V3 and V4 codes and decodes are the same: it is the underlying AEs and their coding (moved to MedDRA), which is different between the two standards.

8.1.9 Unsolicited Adverse Events

The team determined that a result of an open ended question with many possible answers (e.g., “What adverse events have occurred?” or “What concomitant medications have you taken?”) would require just one data element. However, each result for a question about a particular item (e.g., “Have you had a rash?” or “Have you taken aspirin?”) would require its own data element, which means that there can be a very large number of these specific questions.

This touches on the rules for DE groups and CDE creation. We have not (yet) specified rules around DE groups thus far.

8.1.10 Short Names and Long Names

Short names were not discussed on the pilot. The maximum length of short names needs to be agreed. If short names are to have any mnemonic value, then naming conventions for constructing short names will need to be agreed. Both short and long names must be unique.

A shortcoming of the pilot wiki tool was that the long name of a harmonized DE was not editable. The fact that the long name had to be decided before any further harmonization work could be done accounts for some of the odd names given to harmonized DEs (e.g., “Blood Product Name (try again)).

8.1.11 Data Elements Concepts and Groups

Discussion is still ongoing as to how best to approach these two levels of data elements.

Too many DEs in a group could be a problem. When 25 people are doing this they need to do this in the same way. Will need content experts to determine how best to include in groups. Are the DEs in a group optional or must all the DEs be used? Rules need be developed to address this.

BRIDG may provide some help in deciding on the “right” size for a group. For example, two attributes of a single BRIDG class probably “belong” in the same group. However, some things that are tightly linked in the real world (e.g., the name of a test and its result) are in different BRIDG classes, so BRIDG does not provide a simple solution.

GSK, has defined the rules for inclusion of a CDE in a group (is it mandatory/optional/conditional to include the CDE in an implementation; is it mandatory/optional/conditional to populate the CDE in an implementation). GSK has also defined similar rules for inclusion of a group in a “super clump”. The first of these is relevant to what we’ve been doing in CDISC SHARE up to now.

8.2 Wiki Functionality

8.2.1 Landscape (horizontal) View of DEs by Company

It would be very helpful to view the DEs from the contributing companies side by side to determine where there is a match or near match. Wiki does not provide that type of view currently. The Wiki tools currently depend on selecting DEs with the classification, such as the same data type; BRIDG item, terminology item, and domain to get potentially related DEs on the same screen. That approach falls down, of course, when these classifications are missing.

Tools must be developed to enable users to organize and order the DEs without having to pull them into excel to match them up. Tools should allow users the capability to “sort” DEs, as well as to filter them.

8.2.2 Code List Harmonization

Code list harmonization functionality needs to be added to the wiki. This currently must be done in a spreadsheet. The spreadsheet process was awkward and inefficient, so a better approach is needed.

8.2.3 SDTM & CDASH

SDTM and CDASH variables could not be used in the harmonization process. Data elements from these standards should be a vocabulary that the wiki can consume and can then be used in the concept tagging process. SDTM and CDASH variables should also be available for inclusion in harmonized CDEs. SDTM and CDASH should also be included in the CDE/Group creation process in exactly the same way as company standards (they will probably be considered to have more weight than the company standards). The issue about how to handle normalization crops up again because of the SDTM structure.

8.2.4 Search Functionality

Search function/algorithms need to be more robust. Team members suggested that the current “Wild card” functionality be enhanced. The tool currently provides the equivalent of a wild card at the end of a text string, but not elsewhere.

Synonym functionality would be extremely useful. This allows tools to recognize that “thrombocyte” and “platelet” are related, ignores differences between American and British spelling and ignore certain word endings (e.g., plural vs. singular).

8.2.5 Searching and Viewing Terminologies

Need to be able to see a search list in its entirety; currently the tool only allows part of list to be viewed in the window.

8.2.6 Horizontal and Vertical Structure

The Wiki doesn't help us when some companies have employed a vertical structure and others have employed a horizontal structure. When a company has employed a vertical structure, then a single variable, such as TESTCD, spawns a multitude of concepts, such as lab tests, questionnaire questions, etc. Thus, when a submitting organization has used a vertical data structure, there must be capability to convert the underlying list of variable values into submitted DEs (e.g. each SDTM LBTESTCD value would be treated as a DE).

The production CDISC Share library should be able to handle both horizontal and vertical data structures. It was noted that this would create a lot of DEs in some cases. Every piece of information deserves its own DE, but there should not be variations of these (e.g. sitting, standing BP).

When a submitting organization has used a vertical data structure, there must be capability to convert the list of names into submitted DEs (e.g. each SDTM LBTESTCD value would be treated as a DE).

8.3 General Observations

8.3.1 The Need for Active Participation by Contributing Company Representatives

As mentioned in the introduction of this report, the team noted that it is impossible to do meaningful harmonization of DEs effectively without the representatives of the contributing companies on the teleconferences to provide background and context.

We have seen that the metadata curator working on the harmonization process relies heavily upon a dialog with others to think through the process and to consider alternate lines of reasoning before coming to a decision. It appears that the process will be easier when we have good definitions for the terms involved, and a curator may then have a better chance of working alone. But even when good definitions exist, there may/will still be a good deal of alternate reasoning to consider in the process, such that a curator will still be less productive if working alone.

8.3.2 Clinical Input

The team has been working under the assumption that if we use multiple assets (i.e. sets) of standards to define the CDISC SHARE standard, we will be "clinically correct". In the longer term, it is recognized that we will need to have clinical input.

Also it should be noted that we have made little use of the myriad of information available other than to acknowledge that these initiatives exist (e.g. OpenEHR, HL7 detailed clinical models, other clinical societies etc).

9 Conclusions

The CDISC Share Pilot project has shown that data from multiple companies can be merged and an agreed harmonized data element can be created in a relatively short period of time. With improvements in software/ tools, templates for accepting data into the environment and better tools for viewing foundational terminologies the process of creating quality harmonized data element definitions can be optimized.

Software needs to be developed that can facilitate a robust search of terms and provide users with views that make comparison of the data elements easy and instinctive. This will make it possible to develop high quality definitions quicker and to detect duplicates. This needs to be coupled with the ability to allow for a “tree” view of the foundational terminologies to facilitate better choices in determining which concept fits best within a given data element.

Rules and templates that govern pre-production of the data before it is loaded into the system must be developed and followed. Templates should be developed to ensure that complete and accurate data is included in order to facilitate a meaningful harmonization of data elements.

It should also be recognized that the pilot represents a subset of what we would need to do when creating CDISC SHARE content. When creating the harmonized content for the production environment there will not be just a scale up in terms of the volume of CDEs and collections of CDEs but there will also be a need to create an increased volume/complexity of metadata for each collection of CDEs.

10 Appendix 1: Data Elements Created

Following are more detailed information on the DEs harmonized by domain.

10.1 Adverse Event

Source DEs	172
DEs Harmonized	<ul style="list-style-type: none"> CDE Adverse Event CTC Grade CDE Maximum CTC Grade CDE Adverse Event End Date & Time CDE Adverse Event Term CDE Adverse Event Start Date & Time CDE Adverse Event Severity CDE Adverse Event Maximum Severity or Intensity CDE Adverse Event Severity or Intensity CDE Action Taken with Study Treatment CDE Relationship to Study Treatment CDE Information about Agent Related to an Adverse Event CDE Nadir Information related to Adverse Event CDE Other Dates Associated with Adverse Events

10.2 Lesion Measurement

Source DEs	136
DEs Harmonized	<p>CDE Criteria used to accession lesion response</p> <p>CDE Date of lesion measurement</p> <p>CDE Lesion identifier</p> <p>CDE lesion Laterality</p> <p>CDE Lesion Measurement Lesion Type (Target, Non-target) –</p> <p>CDE Lesion Measurement Method of Measurement</p> <p>CDE Lesion Measurement Method of Measurement, coded</p> <p>CDE Lesion measurement missing</p> <p>CDE Lesion Measurement Lesion Diameter Try Again</p> <p>CDE Lesion pathology details not individually mapped</p> <p>CDE Lesion Site</p> <p>CDE Lesion site, coded</p> <p>CDE Lesion variables which will not be mapped</p> <p>CDE Lesion Volume</p> <p>CDE Lesion, whether initial cancer source (PRIMARY SOURCE)</p> <p>Presence of palpable lesions</p> <p>CDE Reason lesion measurement is missing</p> <p>CDE Whether lesion has been irradiated</p> <p>CDE Whether lesion is measurable.</p> <p>CDE Whether lesion was irradiated previously</p> <p>CDE Whether there are new lesions</p>

10.3 Blood Products

Source DEs	30
DEs Harmonized	CDE Blood product administration end date CDE Blood product administration start date CDE Blood Product Name (try again) CDE Blood product quantity CDE Blood product quantity unit CDE Blood product reason for transfusion CDE Blood product source CDE Blood products Variables intentionally not mapped CDE Reason for blood product administration

10.4 ECHO

Source DEs	23
DEs Harmonized	CDE ECHO Date of ECHO CDE ECHO Result of Measurement CDE Position of Subject

10.4.1.1

10.5 Karnofsky Performance Scale

Source DEs	37
DEs Harmonized	CDE Karnofsky Performance Scale result CDE Karnofsky Performance Scale date of assessment CDE Karnofsky Performance Scale reason for missing data CDE MD Anderson Karnofsky Variables not mapped
Notes	Discarded 30 MD Anderson data elements or describing scale as these were more related to database building for their specific purpose

10.6 Chest X-ray

Source DEs	21
DEs Harmonized	CDE Chest X-ray abnormality description CDE Chest X-ray abnormality description, coded CDE Chest X-Ray Date CDE Chest X-Ray Normal or Abnormal
Notes	1 pseudo DE

11 Appendix 2: Data Element Groups Created

Name	Transfused Blood Product
Constituent DEs	CDE Blood product name Product quantity Product quality unit
Concept Reference	Category NCIM Blood transfusion (C0085430)
Notes	Good definition for this DE but in general these definition are driven by the use cases received by NCI

Name	Chest Xray
Constituent DEs	CDE Chest Xray Date CDE Chest Xray, normal or abnormal CDE Chest Xray abnormality description CDE Chest Xray abnormality description, coded
Concept Reference	NCIM Chest Radiography (C0039985)
Notes	

Name	Lesion Diameter (generic)
Constituent DEs	CDE Lesion Measurement Lesion Diameter (Try again) CDE Lesion measurement missing
Concept Reference	BRIDG Performed observation
Notes	