CDISC SHARE

CDISC Shared Health and Research Electronic Library

Vision and Scope

Date: 12 September 2009

Version: 0.7

Document History

Issue	Author	Date	Description
0.1	J Evans M Velezis	20 May 2009	First draft prepared for 27th May Technical Meeting
0.2 – 0.4	J Evans	Prior to 3 September 2009	Internal drafts
0.5	J Evans R Kush B Kisler R Facile F Newby	3 September 2009	Second draft
0.6	J Evans R Kush B Kisler R Facile F Newby D Iberson-Hurst	9 September 2009	First draft for September BoD meeting
0.7	R Kush F Newby D Iberson-Hurst	12 September 2009	Final draft for September BoD meeting

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

Table of Contents

1 Introduction

1.1 Overview

The mission of CDISC is "to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare."

Over the past decade, CDISC has fulfilled its mission by publishing and supporting a suite of standards that enable the electronic interchange of data throughout the lifecycle of a clinical research study. Specifically, CDISC has developed standards for use across the various points in the research study lifecycle: Protocol Development (Protocol Representation Model Version 1); data collection (Clinical Data Acquisition Standards Harmonization (CDASH)); exchange of operational data (Operational Data Model); exchange of clinical laboratory data (LAB) and data submission to regulatory agencies (Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM). As adopters have realized the benefits of these standards, it has become apparent that there is a need for a foundational standard to support computable semantic interoperability (CSI) – the predictable exchange of meaning between two or more systems -- across multiple standards including, but not limited to, those developed by CDISC.

In addition to the desire for CSI described above, CDISC's stakeholders have made it clear that there is a pressing need to fill the gaps in the content of the existing standards, to bring those standards closer together while at the same time developing therapeutic area standards; and, they would like CDISC do all of this at an increased pace. In addition, the ability to use EHR data in medical research is becoming increasingly attractive, which emphasizes the importance and value of having common vocabularies/definitions for research and related healthcare data. Therefore, while this project was originally envisioned as a tool primarily for biopharmaceutical product development, it was soon realized that the stakeholders and scope needed to move beyond this to include other areas, including but not limited to academic research and public health reporting to support clinical decisions in healthcare.

These issues are taken as given. The current emphasis is how CDISC best achieves this and what impact any chosen course of action may have.

To address these challenges it is considered by many that an industry-wide shared semantic library would meet these needs and solve the issues CDISC and its stakeholders face. The value proposition for CDISC SHARE (CDISC Shared Health And Research Electronic Library) is the assumption that the creation of a library of shared semantics will enable CDISC stakeholders – global biopharmaceutical companies, academic institutions and clinical research organizations – to achieve multiple benefits including improved operational efficiency around the collection, processing, exchanging and reporting of data, evaluation of drug safety concerns across traditional organizational boundaries, and, in the end, enhanced scientific capabilities and resulting patient benefits and therapeutic efficacies.

At the CDISC Board of Directors meeting in April 2009, a formal commitment was made to execute the inception phase of the SHARE project.

This document is the formal output from the inception phase. It examines the proposition to determine if it is feasible, makes economic sense, and will meet the needs of the stakeholders. Potential business models for CDISC to lead this project are also included.

1.2 Vision

The vision for CDISC SHARE is to build a global, accessible electronic library, which through advanced technology, enables precise and standardised data element definitions that can be used in applications and studies to improve biomedical research and its link with healthcare

1.3 Scope

The scope of CDISC SHARE includes the data elements for all protocol driven medical research and the overlapping areas of Healthcare.

1.4 Limitations / Exclusions

CDISC SHARE will not address clinical data privacy issues since the repository will not contain clinical data.

The proposed solution will not include software applications such as protocol authoring, electronic data capture, or clinical data management.

2 Business Needs and Benefits

2.1 Business Opportunity

The business/market needs and potential benefits of CDISC SHARE center around the value of sharing information among business partners and secondary use of healthcare information. Semantic interoperability (or the ability of computers/systems to exchange data along with the meaning of that data) is at the core of this information sharing. Information sharing supports a number of needs, including safety reporting or Pharmacovigilance, clinical research studies, biosurveillance, patient and disease registries, regulatory reviews of eSubmissions and other such use cases. To compare and/or aggregate data for comparative effectiveness studies or even to compare information on the same therapy or treatment across studies requires data standards.

The challenges that exist today include:

1. Organization-specific data dictionaries, describing variables (both collected and derived) with their related code lists (i.e., value sets)
2. Ongoing maintenance of existing organization-specific data dictionaries is very costly
3. Lack of semantic interoperability due to inconsistent and redundant definition of the variables across organizations
4. Inability to reuse data, particularly outside of their primary purpose
5. Increasing pressure to decrease costs while augmenting the number of new products
6. Lack of ability to aggregate and compare information across studies
7. Diminished public trust of drug safety
8. Regulatory requirements to provide comparative profile of safety and cost-efficacy
9. Increasing globalization of organizations across Asia
10. Can't use healthcare data for research.

The impact of these business problems includes:

1. Missed or late safety signals and efficacy conclusions
2. Larger than necessary trial expenses due to costs for maintaining internal data dictionaries and/or metadata repositories
3. Increasing costs of drug development
4. Larger costs due to repeating trials because data cannot be aggregated
5. Fewer therapies are developed, only potential blockbusters are developed

An electronic library with a set of unambiguous concepts, such as CDISC SHARE, can provide the following potential benefits.

2.2 Provide a Consistent Approach to Standard Definitions

Mapping legacy data that are collected with different terminologies is often impossible, and if possible, meanings may be lost or misinterpreted thus impacting data quality. The use of consistent, standard definitions will obviously improve the quality of information that is exchanged, integrated, aggregated or compared. The standard definitions also facilitate the aggregation, integration and comparisons of information within and across studies since the terminology, code lists and meanings thereof are consistent,

CDISC SHARE should provide a consistent approach to standard definitions which would then allow for the following:

Unambiguous Definitions –The human and computer communication processes are made significantly easier and more effective if identical words have the same meaning, and differing concepts do not use the same word. If healthcare organizations are to exchange information in a meaningful and useful way, a standard and consistent set of definitions is essential.

Definition Quality – Without a rigorous, consistent process of defining concepts, meanings can be ambiguous and/or incomplete, which negatively impacts the quality of the accuracy and quality of information/data.

Target Data Standards – So that data can be aggregated and/or compared, a "target" standard is needed for legacy data mapping and/or to collect information that should be compared in the same way at the beginning of the research study. Hence, new studies should use these "target" standards so that data can be integrated within studies and/or compared or aggregated across studies. The target standards may be based upon both existing and new definitions.

2.3 Improve Access to Standards

CDISC SHARE will be globally accessible and include standard and definitions in a common central library that can be accessed and "consumed" electronically. This will mean that a 'reference standard' will be readily accessible. With respect to clinical research, specifically, CDISC standards are currently available in a number of formats. SDTM, CDASH and ADaM are available in pdf, ODM in html, Protocol Representation is in Enterprise Architect UML, and CDISC Controlled Terminology can be accessed and downloaded electronically via the NCI EVS systems and infrastructure. Many CDISC users have asked for the SDTM domains in a machine-readable format. A repository that contains all of the CDISC standards in a consistent, human and machine-readable format would be a much more usable and effective approach for CDISC users. It will also improve processes and thus reduce costs for CDISC users. The same will be the case for other standards that are incorporated in CDISC SHARE.

2.4 Decrease Costs for Standards Users

Maintenance Costs – Organizations can save maintenance costs if there is a central electronic library of standards with unambiguous definitions that can be leveraged consistently across the entire organization.

Process Improvement – Eliminating the need for mapping legacy data and improving quality always improves processes and decreases costs

Downloadable Metadata - A number of organizations would like to be able to download the standard metadata and use these as the foundation for their own repositories.

2.5 Facilitate Data Re-use (Secondary Use)

Data Aggregation and Mining - Running a clinical research study is quite costly. If data can be aggregated across studies or if legacy data can be mined to answer new questions, the cost savings would be tremendous. Some examples of reviewing aggregated or legacy data are:

6. Evaluate safety issues.
7. Review completed studies to understand effects of placebo or to obtain information on a placebo population.
8. Compare treatments for a similar indication.
9. Assess pre-clinical studies of drugs in different classes to predict probability of failure of studies in human subjects. At a certain probability of failure, the human study would not be run.

10. Assess legacy studies to calculate sample size based on variance of endpoints.
11. Assess the opportunity for new indications.

Sharing Data from EHRs for Research Purposes – Additional efficiencies can be realized by using data directly from EHRs vs. re-entering research information into a separate system. Such data can be used to support clinical research studies, safety reporting, biosurveillance, clinical trial registration, study or patient registries or other research needs.

2.6 Alignment of Clinical Research and Healthcare Standards

The use of EHR data to support research could shorten the time needed for research information to inform healthcare. For this reason it is essential that the data that supports both healthcare and research be aligned. The need to address this alignment is consistent with the CDISC Mission Statement. This is also consistent with collaborations that CDISC has deemed important and increasingly added since 2001 (e.g. with HL7, ISO and the Joint Initiative Council).

2.7 Improve Standards Lifecycle Management

CDISC SHARE should improve the way that standards are developed and maintained more efficient. Such benefits may be manifested in a number of ways or processes as follows:

Initial Development timeframes (in particular new areas, e.g. efficacy) – Development of new therapeutic area standards and aligning with controlled terminology is time-consuming and can take 1-2 years (or longer depending on complexity per domain. The stakeholders need a more timely delivery of new domains and are increasingly requesting new therapeutic standards (e.g., tuberculosis, cardiology).

Approval times – The approval cycle for a new standard is quite lengthy. Depending on how much new content is in a draft release of a standard, reviewers have thirty to sixty days to review and provide comments. Then the standards development team must address each comment, develop consensus about the resolution, and provide a rationale for either changing the standard or not. The approval cycle and subsequent update of the draft standard can take a year or more.

Maintenance of a central repository and also company-specific concepts – In addition to industry standards, each company has their own company-specific content. A large company can have a team of several standards maintenance staff, which are very experienced and expensive.

Governance – The process of maintaining standards often more of an art that draws on the experience and intuition of the staff than a clear repeatable process that can be consistently used for standards-related tasks.

2.8 Enable Computable Semantic Interoperability

There are four pillars of computer semantic interoperability (CSI), which are required but not sufficient to obtain CSI. The four pillars are: 1) a common information model spanning all domains of interest, 2) a computationally robust data type specification, 3) a robust infrastructure for specifying and binding to controlled terminology and, 4) a formal, top-down development process. ***CDISC SHARE is the computationally tractable implementation of the first 3 pillars.***

2.9 Success Criteria

The following are indicators of success of CDISC SHARE (and as such can be seen as the vision for CDISC SHARE):

1. CDISC production standards are all electronically available 24x7 to clinical researchers and standards adopters.
2. CDISC SHARE is established through a collaborative business model with a viable long-term funding mechanism and a cooperative means of organizational governance.
3. CDISC Standards and research concepts that are included in CDISC SHARE are aligned with corresponding healthcare standards.
4. Additional therapeutic area-specific standards domains can be developed more quickly than the current process, after which they are made readily available electronically to CDISC SHARE users (clinical researchers and the healthcare standards users).
5. Data exchanges between clinical research partners are based upon the CDISC and CDISC SHARE standards and definitions.
6. FDA receives eSubmissions with increasingly consistent standards usage.
7. CDISC SHARE uses existing ontologies and terminologies wherever possible, preventing duplication of concepts and redundancies and with a robust content governance methodology.
8. CDISC SHARE supports reviews by clinicians and 'non-IT savvy' users.
9. to increase efficiency across the data chain from collection through submission
10. to improve consistency of current and future CDISC standards
11. to align with other standards such as those developed by HL7 and ISO
12. to increase accessibility to a public, re-usable set of standard definitions valid across the industry
13. to enable the pooling of data across large databases that come from different sources
14. to enable integration of clinical research and clinical data

3 Stakeholders and Stakeholder Analysis

Stakeholders are anyone who could be materially affected by the implementation of a new system, process, or software application. The major stakeholders for SHARE are shown in the matrix below. The questions at the top of the matrix were also addressed for each Stakeholder, to our best estimate/understanding (P= Possibly, N= No, Y=Yes and ? = unknown). The first seven questions came from Dr. Charles Mead in response to advice for structuring the Stakeholder Analysis. Potential competitor and collaborator were added to assess potential partners for the business models.

Affiliation	**Requested the system?**	**Pay for the system?**	**Primary user of system content?**	**Secondary user of the system?**	**Builder of the system?**	**Support the system?**	**Regulate the system?**	**Competitor?**	**Collaborator?**
Academia									
UCSF		**P**	**P**	**P**	**P**	**P**		**Y**	**Y**
Vanderbilt			**P**	**P**				**Y**	**P**
Cleveland Clinic			**P**	**P**					
Duke		**P**	**P**	**P**					**P**
University of Cincinnati			**P**	**P**					**P**
BioPharma									
Wyeth		**P**	**P**						**Y**
Novartis		**P**	**P**						**Y**
GSK	**Y**	**P**	**P**						**Y**
Lilly	**Y**	**P**	**P**						**Y**
Merck		**P**	**P**						**Y**
Genzyme	**Y**	**P**							**Y**
UCB		**P**	**P**						**Y**
Contract Research Organizations (CRO)									
Quintiles	**Y**	**P**	**Y**	**P**					**Y**
Tech Providers									
PHT				**P**					
Pharsight		**P**	**P**	**P**					**P**

Affiliation	Requested the system?	Pay for the system?	Primary user of system content?	Secondary user of the system?	Builder of the system?	Support the system?	Regulate the system?	Competitor?	Collaborator?
Medidata Solutions		P	P	P					P
FDA									
FDA-OC	Y	P	P	Y		P	P		P
FDA-CDER		P	P	P		P	P		P
FDA-Biostats/CSC		P	P	Y		P	P		P
FDA-CBER		P	P	P		P	P		P
NIH									
NLM		P		P			P	P	P
NICHD			P	P					P
NIH		P	P	P					P
NCI		P	P	P	P	P		P	P
Other HHS Groups									
HITSP						P			P
AHRQ		P	P	P			P	P	P
ONC		P		P		P	P	P	P
CDC		P	Y	Y	Y	Y	Y	Y	Y
ANSI; HL7 and other SDOs									
HL7		P	P	Y	P	P	P	P	Y
ANSI				P			P	P	P
NCPDP		P	P	P		P		P	P
SCO		P		P				P	P
IHTSDO				Y				P	P
BioIT Alliance; CAB		P	P	P		P		P	P
ISO, Canada, JWG Glossary			P	P					P
Other									
AMIA				P					P

Affiliation	Requested the system?	Pay for the system?	Primary user of system content?	Secondary user of the system?	Builder of the system?	Support the system?	Regulate the system?	Competitor?	Collaborator?
AHIMA		**P**		**P**					**P**
Gates Foundation									**P**
i2b2		**P**		**P**		**P**		**P**	**P**
OpenEHR								**P**	**P**
Critical Path Institute			**P**						**P**

3.1 Stakeholder Analysis Teleconferences

Approximately 50 teleconferences were held, with over 70 participants. The representatives were contacted by e-mail from R. Kush, which described CDISC SHARE and indicated that CDISC is in the Inception Phase of the CDISC SHARE Project, which includes development of a pilot as well as a Scope and Vision document. The message then stated "We are also gathering input for a stakeholder analysis and would find your expertise and your candid opinions invaluable in helping us to assess the value and potential business models for such a solution to enhance semantic interoperability for clinical research and healthcare information. CDISC does not wish to do this project without the buy-in and support of all of the key stakeholders and without the appropriate collaborations. Your input and that from others involved in clinical research and healthcare arenas will be factored into a Scope and Vision document, which will be presented at the September CDISC Board meeting, where the Board will evaluate the appropriate path for CDISC to take with respect to this project.

A short set of background slides and a document with potential discussion points were sent to the representatives ahead of time (see Appendix B – Stakeholder Analysis Discussion Points and the file *SHARE Stakeholder Analysis Briefing.ppt*). The slides were presented, as appropriate for the audience, by either R. Kush or J. Evans, who were both participated on each teleconference. One led the discussion while the other took notes. The ~ 70 pages of notes provided invaluable information for the sections on Risks, Market Needs and Benefits, Related Current Work, Potential Business Models & Partners and this section on Major Stakeholders.

Key findings from the Stakeholder Analysis that were not included in other sections tended to fall under the following summarized items. *(Further details are provided later in this section under Key Information from Stakeholders.)*

1. All stakeholders interviewed felt that CDISC SHARE is basically a good idea and that it is needed.
2. Many of the stakeholders commented on the fact that this is big, it will not be easy, it will take time and it will be costly. A few cautioned that the scope needs to be managed.

3. CDISC SHARE needs to address clinical research broadly (biopharmaceutical companies as well as academic institutions) and the related areas of healthcare to ensure alignment.
4. CDISC SHARE is not as much a technology issue as a political challenge. CDISC is not a technology company.
5. Adoption will require careful communication about what we are doing, the benefits of CDISC SHARE and why this will bring value to the stakeholders
6. Many current related efforts were mentioned; most of these could either turn into collaborators or they could be competitors, depending on how CDISC proceeds and engages them (or not).
7. Several are anxious to partner in some way; others cautiously offered support at least in terms of collaboration from the perspective of the representatives with whom we spoke; while others are not yet certain if/how to support this effort.

3.2 Online Survey with CDISC Advisory Board

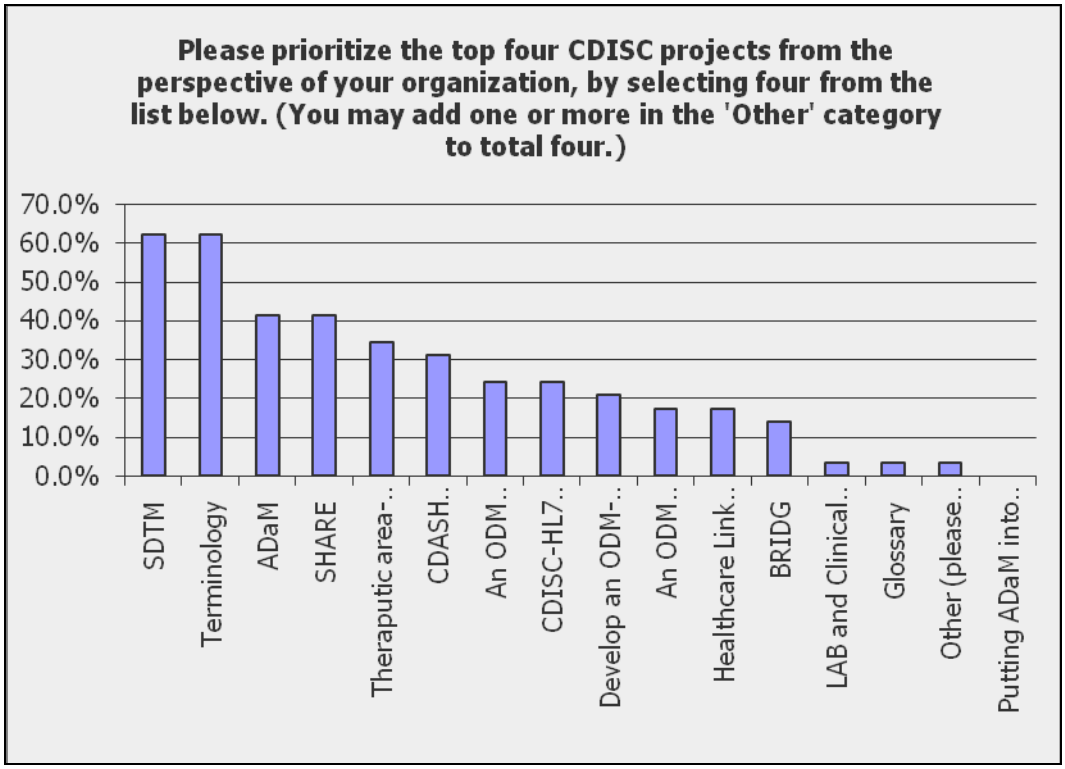
In addition to the teleconferences, because there are 90 CDISC Advisory Board members and time did not permit speaking to all of them, there was a short survey that was sent to these representatives. As is typical of the CAB, we received ~ 30% response rate; however, there were nine (10%) representatives of the CAB and/or the CDISC Board /Membership who were interviewed individually. The 5-question survey provided input into the Quality Requirements prioritization as well as questions asked on the teleconferences.

The results of the online survey are provided below (with the exception of the Quality Requirements prioritization, which is in the Requirements Section of this document).

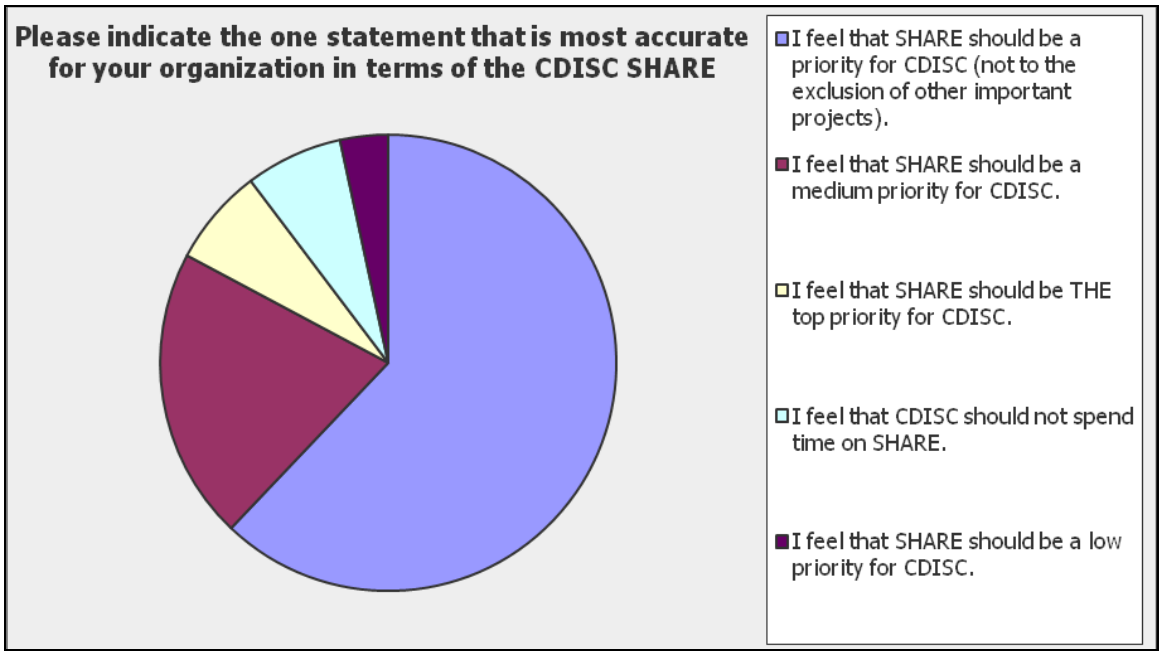
Please prioritize the top four CDISC projects from the perspective of your organization, by selecting four from the list below. (You may add one or more in the 'Other' category to total four.)

Answer Options	Response Percent	Response Count
SDTM	62.1%	18
Terminology	62.1%	18
ADaM	41.4%	12
CDISC SHARE	41.4%	12
Theraputic area-specific standards (with or without CDISC SHARE)	34.5%	10
CDASH Implementation Guide	31.0%	9
An ODM Implementation for CDASH	24.1%	7
CDISC-HL7 Messages (trial design, study data, i.e. SDTM content)	24.1%	7
Develop an ODM-based eSubmission Mechanism (SDTM and ADaM)	20.7%	6
An ODM Implementation for the Protocol Representation Model	17.2%	5
Healthcare Link (EHR-Clinical Research standards and processes)	17.2%	5
BRIDG	13.8%	4
LAB and Clinical Genomics	3.4%	1
Glossary	3.4%	1
Other (please specify)	3.4%	1
Putting ADaM into BRIDG (Statistics Domain Analysis Model)	0.0%	0
answered question		29
skipped question		0

No.	Other (please specify)
1	We would like to see the final version of the metadata implementation guide

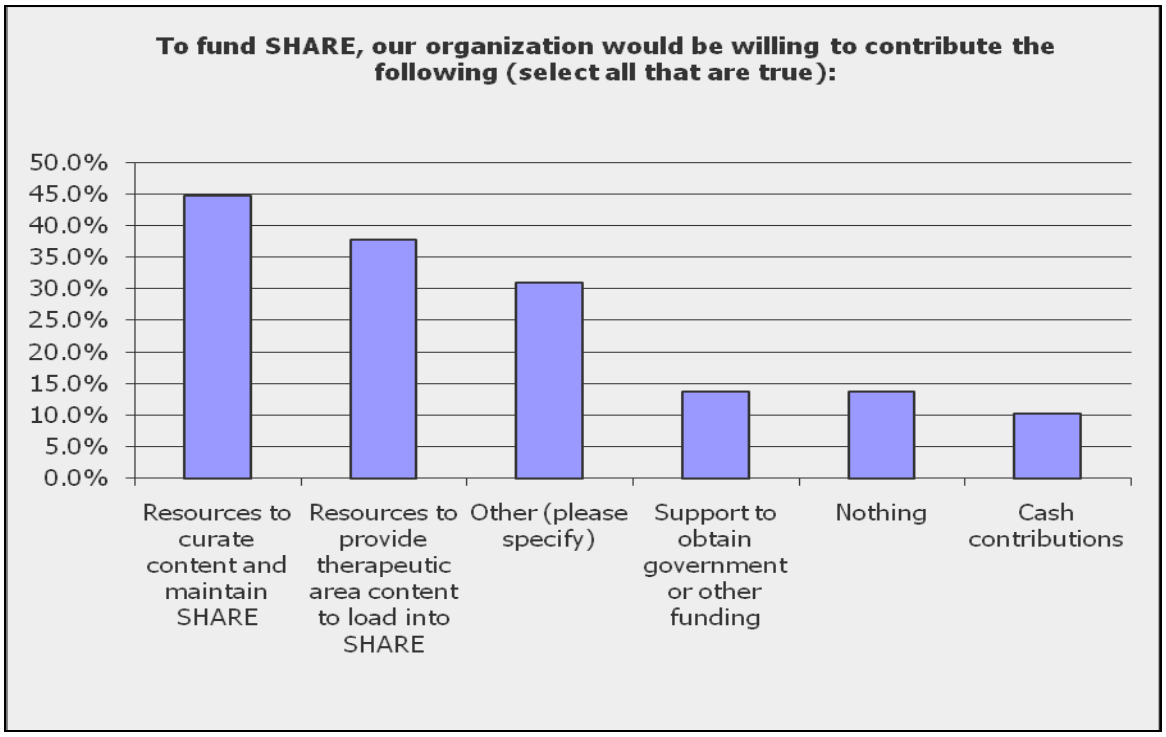

Please prioritize the top four CDISC projects from the perspective of your organization, by selecting four from the list below. (You may add one or more in the 'Other' category to total four.)
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
SDTM
Terminology
ADaM
SHARE
Theraputic area-..
CDASH..
An ODM..
CDISC-HL7..
Develop an ODM-..
An ODM..
Healthcare Link..
BRIDG
LAB and Clinical..
Glossary
Other (please..
Putting ADaM into..

Please indicate the one statement that is most accurate for your organization in terms of the CDISC SHARE (CDISC SHARE Health and Research Electronic Library) project (which has also been called the CDISC SHARE semantic repository or metadata repository). Select one response		
Answer Options	Response Percent	Response Count
I feel that CDISC SHARE should be a priority for CDISC (not to the exclusion of other important projects).	62.1%	18
I feel that CDISC SHARE should be a medium priority for CDISC.	20.7%	6
I feel that CDISC SHARE should be THE top priority for CDISC.	6.9%	2
I feel that CDISC should not spend time on CDISC SHARE.	6.9%	2
I feel that CDISC SHARE should be a low priority for CDISC.	3.4%	1
	answered question	29
	skipped question	0



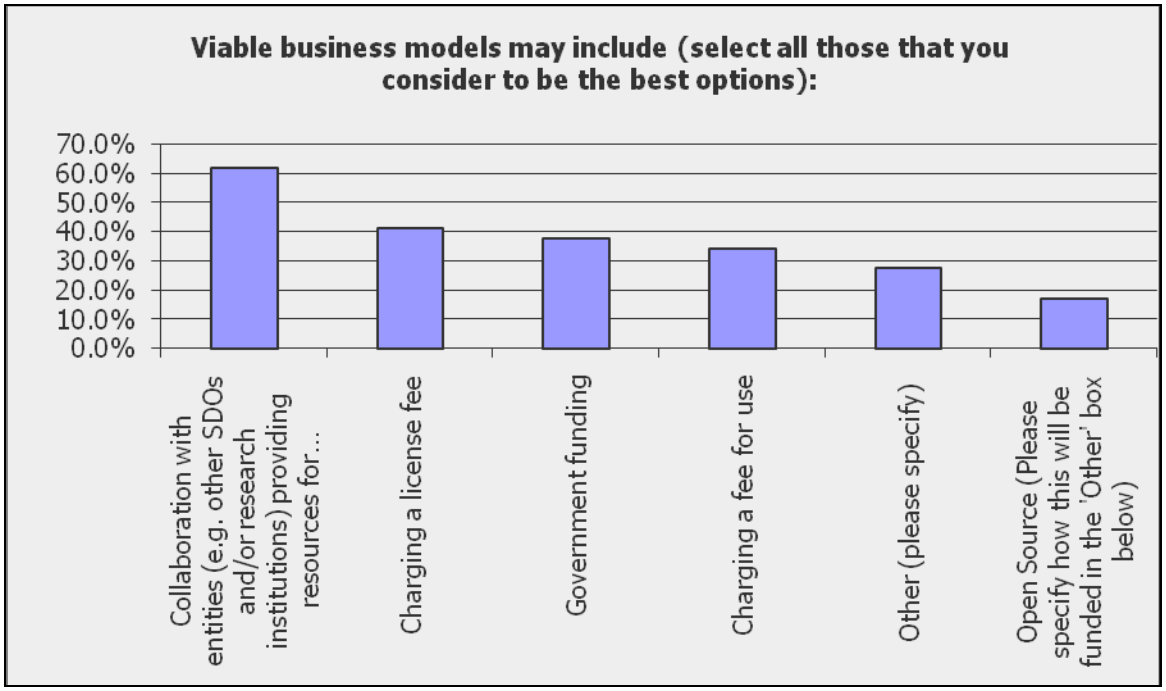
To fund CDISC SHARE, our organization would be willing to contribute the following (select all that are true):		
Answer Options	Response Percent	Response Count
Resources to curate content and maintain CDISC SHARE	44.8%	13
Resources to provide therapeutic area content to load into CDISC SHARE	37.9%	11
Other (please specify)	31.0%	9
Support to obtain government or other funding	13.8%	4
Nothing	13.8%	4
Cash contributions	10.3%	3
answered question		29
skipped question		0

No.	Other (please specify)
1	For our management I would need to define clearly our contribution and the value that we are creating for us and our customers. In general, we would provide resources to build and maintain CDISC SHARE
2	Not in a position to be able to speak for whole organization, but will be happy to see what the company might be willing to do.
3	We are a service provider and as such do not have deep pockets or content to CDISC SHARE. However, we do have very skilled SDTM experts and skilled programmers which we are willing to dedicate as team members, CDISC SHARE participants, technical support & programming, etc... Our interest is being on the inside track of emerging data standards.
4	Provide representation on teams to discuss what would constitute the therapeutic content.
5	Resources to review contents
6	Cash contributions if solid plan that looks achievable. Also would be willing to pay for usage (Software as a service model).
7	CDISC should use membership dues, training/conference profits and its existing operating budget.
8	Not sure at this time.
9	I need to discuss with finance.


To fund SHARE, our organization would be willing to contribute the following (select all that are true):
50.0%
45.0%
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Resources to curate content and maintain SHARE
Resources to provide therapeutic area content to load into SHARE
Other (please specify)
Support to obtain government or other funding
Nothing
Cash contributions

Viable business models may include (select all those that you consider to be the best options):		
Answer Options	Response Percent	Response Count
Collaboration with entities (e.g. other SDOs and/or research institutions) providing resources for maintenance	62.1%	18
Charging a license fee	41.4%	12
Government funding	37.9%	11
Charging a fee for use	34.5%	10
Other (please specify)	27.6%	8
Open Source (Please specify how this will be funded in the 'Other' box below)	17.2%	5
answered question	29	
skipped question	0	

No.	**Other (please specify)**
1	Charge different license fee for CDISC member and non-members
2	Funded by the FDA, i.e. the Regulatory Agencies that are requesting standardized data.
3	Fee for companies included with CDISC membership
4	Some combination of above
5	Is the goal to cover expenses for maintaining CDISC SHARE, or generate profits?
6	Fees charged for use should be to non-contributor organizations only.
7	CDISC should use membership dues, training and conference profits to fund this activity.
	This is what we pay dues / conference fees / etc for.
8	Open Source Comment Collaboration among researchers in large well-funded therapeutic areas may be willing to donate
	time and knowledge in order to save costs in the long run.


Viable business models may include (select all those that you consider to be the best options):
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Collaboration with entities (e.g. other SDOs and/or research institutions) providing resources for...
Charging a license fee
Government funding
Charging a fee for use
Other (please specify)
Open Source (Please specify how this will be funded in the 'Other' box below)

3.3 Key Information from Stakeholders

This section includes key statements from the stakeholders who participated in the teleconferences. In particular, included is information that is NOT necessarily covered elsewhere in the Scope & Vision document (i.e. Market Needs and Benefits, Risks, Related Current Work and Potential Business Models and Partners). The statements below are not attributed in this document in an effort to preserve anonymity. They are organized by the summary points in the beginning of this section. There are many other important statements (over 70 pages of notes) from these calls. However, these seem to be consistent themes.

1. **All stakeholders interviewed felt that CDISC SHARE is basically a good idea and that it is needed.**
    - The SDOs should do this; they are the ones who should develop standards.
    - Makes sense for CDISC to lead this effort (collaborating with others); the more we can talk with others that need to tie in the more it will improve our chances to succeed.
    - CDISC is 'the place to go' for standards in this space, so it makes sense for CDISC to do this.
    - Industry has been asking for something like CDISC SHARE in meetings where FDA was involved; however, in terms of getting data into JANUS, define.xml is the biggest challenge. The review committee won't see HL7 messages for years, but they need define.xml now.
    - CDISC SHARE is important but no more important than other CDISC projects.
    - The timing is right because people are tired of reinventing the wheel.
    - This project strikes me as timely because this is becoming a hot topic within HITSP. HITSP can define requirements but they are not going to endorse a particular product.
    - This is a good idea, but it is also a good idea (as you pointed out) not to have multiple of these efforts, unless we know they are complementary.
    - Anything we can do to get this wild, wild west that we live in corralled would be great. Surprised at how behind this mature industry is in the technology area. This makes so much sense. If we can accomplish this for the industry, it is a win-win.
    - We are in agreement that the industry (including healthcare) needs one place and one tool that we can agree to augment.
2. **Many stakeholders commented on the fact that this is big, it will not be easy, it will take time and it will be costly. A few cautioned that the scope needs to be managed.**
    - A great project that needs to be done; but, ambitious and time-consuming
    - Standards are for the public good and therefore should be funded by the government
    - We haven't been able to get critical mass behind cross-cutting projects like this and they usually become silo projects.
    - How does CDISC shoulder this burden on behalf of humanity? I don't know. Any financial hurdle over a penny will be distasteful to researchers.
    - There is no such thing as a static terminology; we live in a world where there are always changes, and we will need a dynamic system to keep up with the

changes; need to take event and anatomic concepts with evolving technologic and conceptual contexts to integrate these.

- This is the right thing to do. Don't be disheartened. The light is not just going to come on. People are busy.
- Your Board has given you some task! I would love to stay apprised of where this project stands.
- CDISC cannot continue as it has been going; things are falling through the cracks. Also, there is concern that CDISC SHARE will set a precedent of the Board getting into the technical details of CDISC. The Board should provide strategic advice, but not be involved in technical details.
- Don't compromise. Standardising variable names and content is not enough. This work is all about basing content creation on a robust information model. It is all about the data definitions, not on the implementation of these definitions.
- Would not ask my company (big pharma) to implement CDISC SHARE until all of the safety domains are loaded. Companies will need something large enough to implement. 200-500 rating scales could be validated very quickly.
- CDASH would be an excellent place to start.
- This project will be a challenge. How big is the effort? We should not try to 'boil the ocean'.
- Proposed to seed this early and then manage the community development from the seed; don't try to develop a full dictionary that is already vetted and approved. Look at what CDASH did – bit off the elements for safety. If you defined these and made them accessible to users over the web, this could help grow the library almost on a daily basis without a huge amount of standards development in the beginning.
- Assume that the primary content will be fully defined data elements from the CDISC standards models (SDTM, CDASH, ADaM, etc)
- It would help to have clarity from the different aspects of FDA and also to have their 2-page specifications per domain for SDTM.
- Scoping this effort is a huge issue. Are we including all therapeutic areas? Devices?
- There is no value in putting money to develop this without the long-term funding; need to have agreements for the continued funding of this.
- We need to involve more people over time in coming to a point of a reference standard.
- Need to fund the process and the tool.
- The real issue is "What is the cheapest, most likely to endure way of providing the infrastructure for such a database?" It is not a grant. It is not collecting 10 percent of the contribution from 10 players. You could end up with a bad situation. NLM, CDC or NCI are options. But, you really don't want to have to set up an infrastructure-the basic systems infrastructure-for this.
- Philosophically, endorse the general objective. Do it in a way the community can participate in an iterative process. It should allow science to evolve as well as maintain the terminological accuracy. CDISC should not do both the scientific and the terminology standardization. The brokering, networking and facilitation services that CDISC provides is quite valuable, which builds and structures the community.

- This is a monumental task; you must go home and down a bottle of gin every night.

3. **CDISC SHARE needs to address clinical research broadly (biopharmaceutical companies as well as academic institutions) and the related areas of healthcare to ensure alignment.**
    - Agree with goal of aligning with the EHR since physicians will be more likely to use a system that uses words with which they are familiar
    - Actually researchers can manipulate data better than the healthcare side can; keep in mind which one is the 800 pound gorilla.
    - If big pharma owns this, there would be a lack of trust.
    - This should satisfy the needs of government-funded research; then, if it were sufficiently populated, those doing this research should use this or justify why they are not going to use something in particular for your research.
    - We need a collaborative investment for this so that this will work for research data elements.
    - From one HITSP perspective, this individual encourages CDISC to drive the idea of working together. There is a perfect political climate since the lack of this type of thing has tremendous costs. Once we have everyone at the table, we should carve it into bite-sized pieces. The end result may be some kind of system done in such a way where we don't have a massive state machine that has to be maintained every second. "Virtualization" means that we can address these questions: what do I need to know, when do I need to know it and how do I get to it in a reasonable way? Then, the clinical could get to the data the way they need and research could get to the data they need using the same source. Over time, you would hope that differences between research and clinical become smaller and smaller.
    - A slide showing overlapping colored ovals was sent to us, along with slides about secondary data use (collect once and use multiple times) as a goal. The ovals have been called a peacock or a lotus flower. Overlapping with Clinical data are the ovals (peacock feathers), one of which is Research, along with Public & Population Health, Quality & Patient Safety, Clinical Decision Support, and Reimbursement Management. The point is that all of these needs use Clinical Data.
    - From the academic research community, this is an opportunity to bring different parties together to get alignment and support for an important area of biomedical research.
4. **CDISC SHARE is not as much a technology issue as a political challenge. CDISC is not a technology company.**
    - A noble effort and would like it to succeed; however, have been involved in a few too many dictionary efforts to be entirely sanguine; skeptical of the claims to which collaborative technology accelerates work because it still needs to be someone's day job.
    - It cannot be techology alone; getting to a common standard involves entrenched positions and religious wars.
    - The NCPDP standard (Script) allows for RxNorm but it is not happening yet because users are not adopting it.
    - HL7 V3 is essentially dead in the U.S. HL7 should have focused on CDA more.

- Getting the champions and making it happen at a high level are key; the technology can help but it is not the only issue here. Disagreements between people keep this from going forward. Personnel, thinking and relationships are so important.
- The semantic wiki can be a helpful tool but it is not a cure all and it is not highly developed yet for metadata.
- I would like to see CDISC's stamp of approval (weight, cache) behind an effort like this, but CDISC should not be hiring technical people to run this.
- When looking for partners to help accomplish this goal, it either needs to be something like a federal entity that CDISC SHAREs your same principles about open process or they need to be contracts with those who only answer to CDISC and not their own agenda.
- This is not a software development project.
- We should not do this until we have an appropriate funding model. The risks are great for CDISC right now. CDISC needs to be very clear on how this is going to happen, who it will partner with and how it will go forward. CDISC Board should have stood back and looked at the business case/model first; that is where the mistake was made.
- It does not make sense for CDISC to maintain a system, but to maintain the schemas and models that users can download into their own environment.

5. Adoption will require careful communication about what we are doing, the benefits of CDISC SHARE and why this will bring value to the stakeholders

- Is there a certification for those who understand standards? AHIMA may be well-suited to do this. What if people knew about the standards landscape and understood this? This could be like CRA training/certification only for individuals who understand the clinical research standards.
- Suggest you do a panel at AMIA with members of the CTSA community speaking on behalf of standards and mentioning CDISC (i.e. not with CDISC participants).
- Getting involved in CDISC is when the greatest value comes to an organization, not sitting on the sidelines with a 'paper membership'.
- Suggestion to publish the findings from the Stakeholder Analysis quickly after it is finished.
- The Stakeholder Analysis background slides are very good, as is the definition of CDISC SHARE; however, I would welcome a slide on CDISC's achievements in this area and the concept of sharing at the item level and what barriers have been overcome. We should show how this will take advantage of all of the work CDISC has already done.
- Need to define clearly exactly what we are standardizing so that users can see that this will not inhibit creativity or innovation.
- The risks are in communications that are inconsistent, meaning not from CDISC but from Board members and those who misinterpret things. CDISC/Board needs to simplify the communications; the ones that were presented to the Board early on (CMDR) were not understandable.
- We need to be very clear about what we are standardizing: the scientific phenomena vs. the terminological standardization.

6. **Many current related efforts were mentioned; most of these could either turn into collaborators or they could be competitors, depending on how CDISC proceeds and engages them (or not). (See Section on Current Related Work.)**
    - Bigger than USHIK (AHRQ); when USHIK was started, chose a limited scope that was achievable: entering the agreed standards for the US Government, e.g. HIPAA, CHI, HITSP
    - So far, REDcap has sacrificed standards for the sake of delivering a quick solution, but now the users would like to use standards. The research team requirements: 1) make it easy; 2) they want certain functionality, which means they want standards but don't realize it.
    - HL7 CIIC – CDISC SHARE will absolutely fit in, but should not scare the folks with zero exposure to informatics. Also, ACC has its own work, along with AHA.
    - HL7 CIC – very interested in this since they are the ones creating DAMs from the TA-specific standards (e.g. CV and TB).
    - CDC PHIN/VADS is a clearing house where they publish out things as a web publication; it is more of a terminology than a metadata repository.
    - HITSP is looking at PHIN/VADS and USHIK as metadata repositories. (Note that neither of them get to a 'reference standard'; they are more like catalogs.) USHIK (AHRQ) and PHIN/VADS (CDC) have competing products for this role. It would be good to have a place that is up to date that has access to the standards and vocabularies where people could use this as a reference tool. On the consumption side it would be good to have a HITSP IS point to such a repository as the 'source of truth' to get the value sets which are currently in the documents that are around the IS. The discussions were pretty impassioned and people have different feelings. It is not clear whether these two groups would participate in this work that CDISC leads. They are marketing themselves as something similar to what you are doing.
    - Medidata has its own MDR, as does Novartis, and Genzyme is building one.
    - Those working on i2b2 have said they are not even going to try to get people to agree; they have given up. They are building 'ontology mappers' (or ontomappers).
    - The NCI EVS folks are now being approached to include CV standards (NHLBI) and pediatric standards (NICHD); they already house terminology standards for FDA, CDISC and HITSP.
    - FDA is now initiating efforts to develop more CV standards (through the cardio-renal division)
    - MedDRA and Route of Administration – FDA has different requirements and lists than others in healthcare and SNOMED.
    - From the HIT Standards Committee perspective, there seems to be a definite need for this.
    - NCI has had trouble with SNOMED not having capacity to subset or to work quickly enough.
    - This sounds like a cloud computing application. Microsoft and the semantic web may be doing something along these lines.
    - Collaborativedrug.com is doing something along these lines; public access is no charge but collaboration among users has a fee; they are a for-profit group.

- "New standards" occur when companies or individuals do not feel that they have the luxury of time to wait for a standard so they create their own.
- NIH has new leadership (from the human genome project). They will need standards for interoperability. CDISC is well-positioned.
- Item Response Theory: in the outcomes arena, there are questions that exist along with a set of choices. An accumulation of how the responses are related to other responses is maintained in a bank so that one can see the relationships between other questions within a given questionnaire. Such banks are not collaborative; there is a fee for items that are extracted, which is waived if you agree to deposit your results.
- Engaging content experts moving forward is important, e.g. from the Crossing the Chasm meeting (i.e. HL7 CIIC). Also, groups such as the American College of Rheumatology (ACR) and other therapeutic area associations may be very good to engage.
- The Joint Working Group (JIC – ISO, CEN, HL7, CDISC, IHTSDO) has a collaborative tool that is harmonizing glossary terms, the majority of the current content being CDISC glossary terms and definitions.
- MESH Terminology – would it help to include this?
- NLM – the various efforts in this space should be complementary and not conflict with each other.
- HL7 CDA - Templates: Could this be viewed as defining things in terms of templates; you have to look at the template to see multiple observations that relate to a given template. If you approach this looking at the data element level and have not looked at it with the larger constraints of a given template, something may be missing.
- HL7 Detailed Clinical Models – needs to be more clarity around this work and how it relates.
- CDISC should involve AHRQ, with a strong link – the scientific side of the house.
- SNOMED-CT committee is getting more and more organized with a collaborative workbench. It also overlaps with ICD codes although there is not a 1-1 mapping. This could be a win-win since there are not enough resources to get everything updated in SNOMED (with IHTSDO). The general idea is to support interoperability. As you get into data element definitions, you are getting right inside SNOMED-CT. (Run by the IHTSDO)
- UDEF – Universal Data Element Framework. This uses UN definitions of objects and properties.
- OpenEHR is different from CCD/CDA because implementations are formally specified so they can be checked by computers and not worry about interpretation.
- OpenEHR archetypes are defined as "computable specification of a single discrete clinical concept. The example was BP (which could pertain to a person at home taking BP, a BP taken in an ambulance, a BP taken during a stress test and so forth).

7. **Several are anxious to partner in some way (e.g. AHIMA, HL7, OpenEHR, BioIT Alliance, NCI, NICHD, certain pharmaceutical companies and technology providers); others cautiously offered support at least in terms of collaboration from the perspective of the representatives with whom we spoke (e.g. AMIA, IHTSDO, ONC, HITSP, CDC, certain biopharmaceutical companies, FDA); while**

others are not yet certain if/how to support this (e.g. SCO, NLM, certain academic institutions/CTSAs).

- IHTSDO is interested in encouraging the use of SNOMED; happy to help and see what they can do and that appropriate content is in SNOMED.
- AMIA advocates on behalf of standards; would be willing to discuss an advocacy effort with CDISC. Another question was how could we best leverage AMIA in this effot?
- NICHD is leading the National Children's study, which will provide a great opportunity; in addition to NICHD and the CTSA program, can add this to the fleet with respect to our push towards standards and standardization. Now working with NCI, adding pediatrics information (including neonates and infants); willing to work with the CDISC SHARE effort
- Academic researchers thrive on doing things their own way; NLM may be able to encourage this. NCRR is 'anti-leadership' and does not want to tell them what to do (a chess game strategy).
- NLM is interested but does not wish to 'arm wrestle' NCI.
- C-PATH Institute – CAMD project is using CDISC standards for their database
- The SCO (Standards Charter Organization) is trying to bring standards development organizations together.
- The BioIT Alliance needs something very similar to this, if not the same – addressing discovery needs, not just clinical research needs.
- From the Board of ACR, this may be a good way to provide value to the members beyond offering conferences. Rheumatology has some well-developed and consistent data elements.
- CDC reps find CDISC SHARE very compelling. They looked at LexGRID and caDSR in 2007/2008 and spent about 3 months evaluating it. There were challenges in how it was modeled (without an underlying information model) and also duplicity in the content and no governance structure. The BRIDG model is very compelling and would be of interest to them. They would have hesitation putting their resources into something that does not meet their needs. They would like to expand BRIDG into public health and surveillance and may have resources to commit to this. They need further assessment and offered to set up another call. PHIN/VADs is the CDC terminology service; it has more in it than USHIK.
- HITSP is looking for something like CDISC SHARE. They have identified USHIK and PHIN/VADS as possibilities. There would need to be some 'pruning' of USHIK or at least sub-setting to get to a standard.
- AHIMA is willing to put sweat equity into this and offered to help put together a proposal. They also offered to host the tool assuming there is funding for this.
- Microsoft offered to discuss hosting CDISC SHARE as part of their cloud computing capability.
- OpenEHR stated they would love to work with CDISC to involve the research perspective, specifically; they would love to extend their work into the research community. They have a way of showing clinicians 'mind maps' that are easy for them to review. They also have very simple to use tools for reviewers and a process that they are perfecting. They gave up on the wiki approach as it was 'total chaos'.

- OpenEHR/Ocean Informatics are working with the American College of Rheumatologists (ACR) already and offered this to be a potential collaborative project.

3.4 Stakeholder Analysis Conclusion

The stakeholders with whom we spoke were encouraging and generally very positive about CDISC SHARE and about CDISC being an appropriate leader. There were concerns expressed about the scope and a long-term business model along with engaging the right partners. There was broad agreement that CDISC SHARE needs to accommodate clinical research as well as the significant amount of overlapping clinical care/ healthcare terminology and concepts. There were also cautions about how this should be communicated to ensure buy-in and adoption, particularly for the academic researchers. There are numerous efforts that are related to CDISC SHARE. For the most part, there is an opportunity to turn the majority into collaborators (see the section on Current Related Work); however, if not careful, some of these could well end up becoming competitors.

4 Risks

The two original risks identified at the start of the inception phase were:

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.

The following risks have been identified during the Stakeholder interviews:

1. **The project scope is unclear or too big to manage effectively.**
2. **The project proves too costly and time consuming**
    - Start up is a black hole and will cost twice as much as estimated
    - Need to start slow and produce useful results quickly and then build on that.
    - "The overhead in managing industry dictionaries is phenomenal".
    - It takes so long that CDISC loses its reputation of being the CR content experts.
3. **The project is unable to enforce the content rules and therefore the repository content lacks integrity (data becomes duplicative, useless)**
    - Governance process is non-existent or ineffective
    - Don't have sufficient buy-in to the principles of governance.
4. **The project is unable to achieve consensus on target data element definitions**
    - Efforts like this often die because of entrenched positions and religious wars
    - Numerous experiences in not reaching consensus even in small groups
5. **The project is unable to secure involvement of anyone except drug companies**
    - Healthcare standards groups do not see enabling research as a primary goal
    - Office of the National Coordinator (ONC) doesn't appreciate the needs of research and pharmaceutical companies
    - Previous experience in not being able to get critical mass behind a cross-cutting project like this and the project becomes a silo solution.
    - Healthcare and research have traditionally been separate voices.
    - NIH CTSAs (Clinical and Translational Science Awards) are not motivated to endorse any kind of standard.
    - Difficulty in getting one voice from FDA.
6. **Intellectual property (IP) roadblocks**
    - Some biotech companies see the standards as part of their IP, as an integral part of their tools.
    - Some therapeutic areas are highly dependent on proprietary forms.
7. **Collaborative technology doesn't accelerate processes as much as necessary.**
    - Wiki isn't a cure-all, and isn't highly developed for metadata.
    - One organization tried a wiki, but abandoned it after experiencing "pure chaos". They needed more structure than the wiki provided.

- Processes aren't in place to use technology effectively.

8. **Business model / funding isn't viable.**
    - Must include start-up and long-term maintenance
    - Too expensive for small companies
9. **CDISC SHARE is not adopted because it too complicated to learn and/or use.**
    - Isn't easy to implement and use in company computing environments
    - Many people in the stakeholder communities are not familiar with a metadata driven approach.
10. **Political and logistical issues are not addressed sufficiently.**
    - "The biggest challenges will be political and logistical"
    - Must get the broader community behind the project and build coalition/collaboration
    - Legacy issues must be considered
    - People are territorial about current processes and repositories
    - Doesn't align with HITSP recommendations
    - Large stakeholders may not agree to work towards a common goal.
11. **Key stakeholders or enablers don't see the value of this project.**
    - We don't communicate clearly or enough (e.g., Many, if not most users, are not interested in the details of an ISO 11779 based repository. They just want to know what they need to get their job done).
    - Some organizations or constituencies are not sophisticated or mature enough to understand they need CDISC SHARE, even though they want the results that must have standards to achieve.
    - ONC not even requiring standards for the 2011 rollout of doctor's use of EHRs.
12. **The end result doesn't meet the objectives.**
    - Isn't "fit for purpose"
    - Partner has different agenda that doesn't match the project agenda
    - Get locked into using a particular terminology
    - Take too long and miss the window of opportunity, over-engineer, don't project-manage
13. **CDISC loses focus on other important initiatives and these become marginalized.**
    - CDISC doesn't put the right model in place for supporting the project with the righ financial, staffing, and other resources.
    - Flagship products such as CDASH, SDTM and Controlled Terminology become derailed due to lack of proper resourcing for CDISC SHARE.
    - Existing standards are abandoned due to competing priorities with CDISC SHARE.

5 Related Current Work

Based upon the CDISC Collaborations and the Stakeholder Analysis, there are several organizations or initiatives that are doing work that would be considered related to the CDISC SHARE project. These can potentially be either leveraged as collaborators in the CDISC SHARE project or they may become competitors. Please refer to the two columns in the Stakeholder Analysis table with the labels Collaborator? or Competitor? In certain cases, it will depend on how we engage these organizations in CDISC SHARE as to which they will be perceived.

The Business Models section of this document provides certain possible opportunities for the potential collaborators, which would leverage the CDISC proven "Strength Through Collaboration". This section provides an organized listing of a number of these initiatives and/or organizations to facilitate a decision around an optimal business model. Not all potential competitors or collaborators are named in this section since essentially any standards development organization could fall into either of these categories as could any pharmaceutical company or academic institution that develops proprietary standards or metadata repository. We prefer to think of these entities as potential stakeholders and collaborators or users of CDISC SHARE if there is no known information to the contrary.

5.1 Standards Development Organizations (SDOs)

1. Health Level Seven (HL7) – www.hl7.org: has had an Associate Charter Agreement (MOU) with CDISC since 2001. Potential related efforts are the Clinical Interoperability Council (CIC), the Clinical Information Interchange Committee (CIIC), Clinical Document Architecture (CDA) Templates, Regulated Clinical Research Information Management (RCRIM) Workgroup, EHR Workgroup, Terminology.
2. International Health Terminology SDO (IHTSDO) - www.ihtsdo.org: is part of the Joint Initiative Council (along with ISO, CDISC, CEN, HL7); have an MOU with CDISC since mid-2009. Offers SNOMED free for appropriate clinical research, regardless of the country. U.S. has a country wide license.
3. Other SDOs and their collaborations, including ISO, JIC, JWG, CEN, BioIT Alliance and Pistoia Alliance (discovery standards), the SDO Charter Organization (SCO)
4. Universal Data Element Framework (UDEF) – www.udef.com (www.OpenGroup.org)

   The Data Indexing Standard to Reduce the Costs of Applications Integration and to Improve Data Discovery; UDEF provides semantic links, through assigning an intelligent, derived ID as an attribute of the data element, essentially labeling the element as a specific data element concept. When this UDEF ID exists in both source and target formats, it can then be used as an easy analysis point via a match report, and then as the primary pivot point for transformations between source and target.

   The Open Group assumed from AFEI the right to grant public use licensing of the UDEF. Ron Schuldt, Sr. Enterprise Data Architect, Lockheed Martin, originated the UDEF concept based on ISO/IEC 11179 Metadata standards approximately 15 years ago.
5. HITSP - Note that the Health Information Standards Panel (HITSP) is not an SDO; they identify standards to be used to support capabilities and use cases for EHRs. They work with SDOs.

5.2 Non-SDO Organizations or Initiatives

1. OpenEHR - http://www.openehr.org/home.html:

   Technically, *open*EHR is about creating **specifications, open source software and tools** for such a platform. In the clinical space, it is about creating high-quality, re-usable clinical models of content and process - known as **archetypes** - along with formal interfaces to terminology.

2. United States Health Information Knowledgebase (USHIK) - http://ushik.ahrq.gov/registry/index.html?Referer=Index:

   Catalog of US standards, e.g. CHI, HITSP; hosted by AHRQ – Agency for Healthcare Research and Quality; not a reference standard

3. Public Health Information Network (**PHIN**) Vocabulary Access and Distribution System (**VADS**) (PHIN/VADS) - http://phinvads.cdc.gov/vads/WebHelp/Welcome_to_PHIN_VADS.htm :

   " A vocabulary repository and server which allows **CDC**'s public health partners to browse, search, and download vocabulary concepts required for PHIN messaging and applications"; hosted by Center for Disease Control (CDC); not a reference standard.

4. Informatics for Integrating Biology and the Bedside (i2b2) - https://www.i2b2.org/

   An NIH-funded National Center for Biomedical Computing based at Partners HealthCare System. The i2b2 Center is developing a scalable informatics framework that will bridge clinical research data and the vast data banks arising from basic science research in order to better understand the genetic bases of complex diseases. This knowledge will facilitate the design of targeted therapies for individual patients with diseases having genetic origins. The i2b2 is funded as a cooperative agreement with the National Institutes of Health.

5. Research Electronic Data Capture (RedCAP) – http://www.project-redcap.org/

   The **REDCap Consortium** is comprised of 57 active institutional partners from CTSA, GCRC, RCMI and other institutions, and it supports two secure, web-based applications (REDCap and REDCap Survey) designed exclusively to support data capture for research studies.

6. Clinical and Translational Science Awardees (CTSAs) - http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/

   A national consortium of medical research institutions, funded through **Clinical and Translational Science Awards** (CTSA), is working together to improve the way biomedical research is conducted nationwide. Consortium members share a common vision to reduce the time it takes for laboratory discoveries to become treatments for patients, to engage communities in clinical research efforts and to train clinical and translational researchers. CTSA is funded through National Center for Research Resources (NCRR).

7. PhenX – https://www.phenx.org/

   PhenX is a three year project led by RTI International and funded by the National Human Genome Research Institute (NHGRI) to contribute to the integration of genetics and epidemiologic research

PhenX has prioritized 20 research domains related to complex diseases and environmental exposures

Consensus building is being used to develop a recommended minimal set of high priority measures for use in Genome-wide Association Studies (GWAS) and other large-scale genomic research efforts

High priority measures will maximize benefits of future research by enabling cross-study comparisons and analysis

8. National Cancer Institute (NCI) Enterprise Vocabulary Services (EVS) and NCI cancer Data Standards Repository (caDSR)/Mayo–LexGrid – http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/vocabulary

   CBIIT bases its data semantics on controlled terminology supplied by the NCI Enterprise Vocabulary Services (EVS) Project. The NCI EVS represents a set of services and resources that address NCI's needs for controlled vocabulary as well as that of other key stakeholders. The EVS Project is a service of the Center for Biomedical Informatics and Information Technology (CBIIT).

   These services encompass terminology development and coding, terminology licensing, software development and licensing and operations support activities. From its inception, EVS has sought to address the broad spectrum of terminology needs at NCI. EVS provides the base upon which the data semantics of caCORE and caBIG initiatives depends, and houses the CDISC, FDA, HITSP and other controlled terminologies.

6 Major Features and Capabilities

CDISC SHARE is a warehouse of scientific concepts used in biomedical research and healthcare that includes all information about those concepts including: concept meaning, concept definition, variables associated with those concepts, code lists, data types, and relationships between concepts. The high level business capabilities and the business quality requirements are described below. A more detailed look at the business requirements can be found in the related Appendices zip file.

6.1 High Level Business Capabilities

Atomic Definition and Grouping: The solution will provide a definition for each concept and variable. The solution will also provide the ability to group concepts and variables.

Content Curation: The solution will need to provide a mechanism to curate the content and/or upload content from external parties. Prior to the curation process there will be content that will need to be uploaded as individual and/or collections of concepts at one time. There will also need to a mechanism to add, delete or modify concepts already in the library.

Content Accessibility (includes Storage and Retrieval): The solution shall provide 24/7/365 accessibility to the current version of the shared semantics in a view mode. All concepts and its current mappings and annotations should be available real-time. Any content that is in development will be accessible at the point of validation. The solution shall provide a mechanism for searching subsets of the information based on user-defined criteria. In addition to real-time access to the content, the solution will need to provide the capability to export the entire repository for input to an organization-specific repository.

Duplicate Resolution: The solution should prevent duplicate concepts being entered into the repository. In the past, there have been duplicate concepts within organization-specific dictionaries that have led to intra-organization harmonization efforts.

Authoritative Sources: The solution shall as much as possible provide content by reference (not content by value), so that the authoritative source of the concept is controlled. The versioning of content should be maintained in one location and linked in all applicable contexts.

Concept Definitions: The solution will include concept names, concept definitions, and relationships between concepts, and relationships between concepts and variables. The solution will also address how to work with existing standards, variable and concept synonyms, and alternate definitions.

Variables: The solution will include variable names, variable definitions, valid value sets, data types, variable lengths and relationships between concepts and variables, variable grouping, and variable optionality.

Referenced Standards: The solution will contain the content from all CDISC standards including the Protocol Representation Model (PRM), Clinical Data Acquisition Standards Harmonization (CDASH), Study Data Tabulation Model (SDTM), Standard for the Exchange of Non-clinical Data (SEND), Analysis Data Model (ADaM), and Controlled Terminology.

Internationalization/Globalization: The solution will contain information pertinent to biomedical research and healthcare globally. The solution shall provide the appropriate references and/or attributes to enable use of concepts internationally.

Traceability: The solution shall provide a mechanism for tracing the origin of the concept back to its owner, and/or contributing organization. The information shall be provided in order to provide traceability and transparency for the users of the solution.

Governance: The solution will include a plan for governance at three levels:

1. Organizational Governance the governance as it relates to the provision of the entire CDISC SHARE service to a range of organizations (content providers and content consumers).
2. Content Governance the ongoing stewardship, support, and maintenance of the CDISC SHARE content. The governance should provide a mechanism for managing changes (additions, deletions, modifications, or merges) to concepts in the repository as well as concept resolution.
3. Technical Governance the ongoing input into the development of the tools used within the provision of the CDISC SHARE service such that the tools meet the needs of the service.

Change Management: The solution will include a well-documented process that ensures that standardized methods and procedures are used for efficient and prompt handling of CDISC SHARE change requests in order to minimize the number and impact of any incidents on the repository.

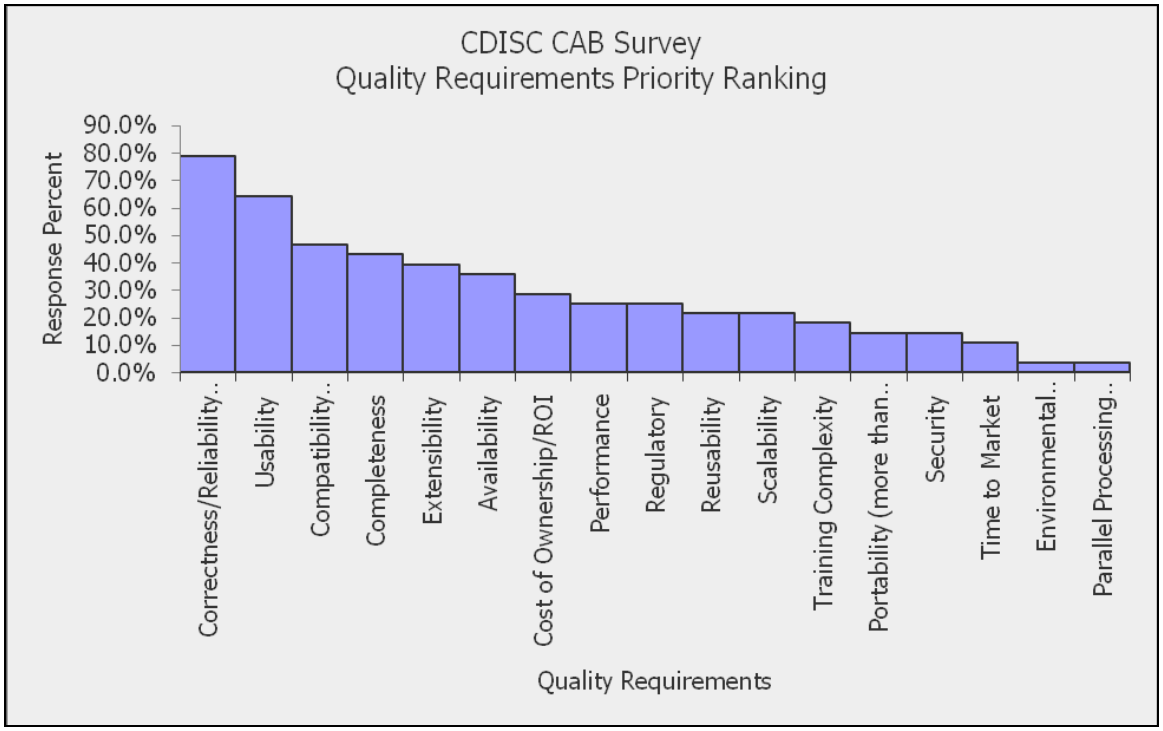
6.2 Quality Requirements

Quality Requirements are those requirements that focus most on the system architecture of a solution. The 90-member CDISC Advisory Board was asked to prioritize the top 5 quality requirements. The results of this survey are shown below. For a definition of each of these requirements see the appendices.

All of the following will be Quality Requirements for the technology to support CDISC SHARE. Please help us prioritize their importance by selecting the top five of the following Quality Requirements for the technology to support CDISC SHARE.

Answer Options	Response Percent	Response Count
Correctness/Reliability for use in mapping data	78.6%	22
Usability	64.3%	18
Compatibility (interaction with other sponsor systems)	46.4%	13
Completeness	42.9%	12
Extensibility	39.3%	11
Availability	35.7%	10
Cost of Ownership/ROI	28.6%	8
Performance	25.0%	7
Regulatory	25.0%	7
Reusability	21.4%	6
Scalability	21.4%	6
Training Complexity	17.9%	5
Portability (more than one operating environment)	14.3%	4
Security	14.3%	4
Time to Market	10.7%	3
Environmental (conditions in which the system must function)	3.6%	1
Parallel Processing (fulfill requirements simultaneously using	3.6%	1

duplicated rather than CDISC SHAREd resources)		
	answered question	28
	skipped question	1



6.3 Detailed Business Requirements

The CDISC SHARE Business Requirements activities included broad input from many stakeholders across the industry, and in particular global biopharmaceutical companies and CROs. Requirements development was distributed across two primary teams: (1) Business Requirements and (2) Content Governance Requirements. Smaller sub-teams were formed as needed to focus on specific CDISC SHARE storyboards and specify the requirements for particular industry roles or "actors." Due to the large amount of material developed the information is presented in a separate document.

7 Potential Business Models and Partners

The Stakeholder Analysis included questions designed to elicit ideas around potential business models for SHARE that will enable a viable, long-term funding mechanism. There were a variety of ideas that were suggested from those who participated in the teleconferences, along with those from the CDISC staff and comparative models. Two questions to the CDISC Advisory Board requested respondents to indicate what their organization would contribute (e.g. resources, funding) and to select what they felt to be viable business models, including offering new suggestions under an 'Other' category. (See the section in this document on the Stakeholder Analysis for these results.) A table has been generated from these sources and ideas, along with pros and cons of the various business model options.

The discussion on business models for the SHARE project is directly related to the scope and vision of the project as well as the timing of development and implementation. There are 3 categories or items that can vary to meet the aims of a project: time, resources and scope. A few examples – with all other things being equal:

1. It's possible to vary the resources and scope of a project to meet a specific project deadline.
2. Accelerating the time would mean reducing the scope and/or increasing resources.
3. Increasing the scope and accelerating the time most likely would mean a significant increase in resource needs.

Business models for the SHARE project would primarily address the resource needs and the business model ideas listed below may be more or less appropriate based ultimately on the time and scope of the production project plan.

An additional consideration for SHARE is how directly SHARE is tied – or not - to the current CDISC organization. Options:

1. SHARE is another CDISC project/standards development team – like SDTM or ADaM – perhaps bigger but run as a CDISC team with all the same governance. Funding/resources could be from current sources or any of the new business models
2. SHARE is set up as a "sub-organization" or division of CDISC. Any of the business models could apply. Under same management/governance as CDISC.
3. SHARE is set up as an independent organization – similar to CDISC but separate – with its own BOD, Management and staffing. Any of the business models could apply.

The decision of this aspect of the project will have a direct bearing on the resources necessary to ensure success – i.e. if SHARE is set up as an independent organization resources would need to dedicated to management, finance, PR, education, HR, events, etc. CDISC already has these resources in place. This certainly does not preclude an independent organization – it just means it would be more time consuming to set up and more costly to run.

The primary decisions we need to make are: Will SHARE be CDISC-run or independent? and What method of funding will be most likely to ensure success?

	Option	***Pros***	***Cons***
0	No change - fund SHARE from current revenue streams	Revenue model is well understood and has been relatively reliable over a 10 year period	Depending on scope/timing of SHARE our current funding may/will be insufficient. May also mean reevaluating our current mix of "products" and redeploying resources off of current projects
1	License/country with a sliding scale based on per capita income or GDP – like SNOMED	Countries that could afford it pay more. Government funding "may" be more dependable	Some smaller or poorer countries may never "sign-up". Government funding may dry up.
2	License/organization with sliding scale based on usage or number of employees – like MedDRA	Organizations that need SHARE would pay to use it. Amount based on size of organization	Could exclude small/poor organizations - especially those in developing countries - would necessitate the right formula for payment
3	**Grant**		
3a	Large grant of sufficient size to live off interest – i.e. $50M invested @5% would yield $2.5M/year	This would allow the SHARE project to have a more or less guaranteed level of revenue to support the project/product. Would also allow for the "product" to be offered free of charge to all. Could be Government(see 4 below) or more likely 1 or more foundation grants = Possibility that many foundations could be enticed to contribute the total thereby sharing the load	Attracting that level of funding would be difficult. An option would be for the SHARE project to just receive the interest off of a ~$50M investment but the "donor" keeps the principal.
3b	For/from individual therapeutic areas from orgs that support that area – we have a bit of that now	Those who want to have their therapeutic elements included basically pay to have that done. Those who are most interested pay. Also allows us to resouce the project with focus on specific areas	Only those who pay get to play
3c	Take a piece of a bigger grant	Allows the project to continue based on many little pieces of bigger grants	Time consuming to track these down and get our piece.

	Option	***Pros***	***Cons***
4	Government funded – like #1 but perhaps limited number of countries and not a license – more like a grant – Stimulus $$	One or more governments contribute to the project with the requirement that the "product" is made freely available to all. This could be something like Stimulus money or FP7 grant.	Getting Govs to sign up may/will be difficult. Unless there is a way to guarantee longterm commitment funding may/would need to be re-upped on an annual basis
5	Government agency funded – get all the various US gov agencies to contribute – AHRQ, ONC, NLM, CDC, NCI, IMI(EU), etc – like #4 but more specific	Like #4 but at a lower level of government. Might be more likely since specific agencies within governments would have more knowledge about what the SHARE project is and what value it brings directly to their agency	More work to track down all the different agencies within all the different countries who could/should have interest
6	Have partners orgs fund % of total – i.e. CDISC, HL7, HITSP, etc	Interested organizations would be "partners" in the effort and share the financial responsibility. Probably mostly non-profits	Dependance on the non-profits to continually fund this on an ongoing basis
7	WHO/ISO donor funding – get them to get the $$	Similar to 6 but get WHO/ISO (others?) to take responsibility for the fund raising and contribution to the project	We'd be at least one step removed from the process of fund generation.
8	Foundations – Rockefeller, Gates, Google, Annenberg, etc. – like #3	Like 3a	Such foundations have indicated that, in the current climate, their portfolio does not include funding standards development-type projects (similar to FP7 issues)
9	Pay to play – contribute so you can influence what gets put in and when	like 3b but open it up to any company/government/organization - only those who pay get their elements included	Would limit the content to only those with the money to get their info included
10	Pay to use – like #2 but on a per use basis – could also pay more up front and get free use forever – Like CRIX	This is the traditional pay per use model. Similar to the CRIX model where we would possibly engage as many as want to to pay a large "foundation" fee in the beginning to allow open access to all their employees/members forever then chage per use or in blocks of use - i.e a discounted price for buying a block of 1000 accesses.	This would require a way to track users and invoice users and would require most/all to pay to use which would limt or eliminate usage by poor/developing countries/organizations

	Option	***Pros***	***Cons***
11	Advertising – Google model	Free, open access but advertisers pay to have ads placed on the access website or within strategically placed areas within the "product". Could be a change to engage a company like Google to show us how this could be done.	Not sure of the logistics of creating an environment where we could sell local/regional/national/international advertising
12	IPO – sell shares in this new “business”	This may look like #9 but we would essentially set up a public company and sell shares to either the public or just to institutions	This may be more complicated. Not sure if any non-profits have ever had an IPO!
13	Have coalition of tool developers pay for it – we develop content – they develop and sell tools to use it	Similar to other options but in this case the technology providers who would seel their tools would fund the development and maintenance	Tool vendors would only fund this if they felt there was sufficient ROI. May not be enough (or perhaps there SHOULD not be many) vendors to support this project revenue needs
14	In Kind "donations"	Any combination of infrastructure (hardware, software, network, etc) as well as development and maintenance resources from any potential contributing organization. Not money - but everything money would buy. Would engage in a meaningful way - all organizations that are serious about seeing this project through to success	Similar to the current CDISC volunteer structure which has it's challenges in both coordinating the resources and not allowing any one person/group to dominate.

8 Pilot

8.1 Purpose

The purpose of the pilot is to address two risks identified at the start of the inception phase:

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 subsequent development work.

8.2 Technology Selection

To address the above risks it was necessary to obtain a piece of technology that supported the aims of the pilot. The selection of the technology commenced with a meeting in San Francisco on the 27th May At that time it was presumed that Tolven would supply the technology for the pilot. A proposal was received from Tolven on the 4th June. On the 5th June Mayo indicated that they also wished to submit a proposal. This was received on the 17th June. At the same time NCI indicated that they would provide financial support for the project. Key technology issues were identified during this period and are summarised below.

Costs from Mayo and Tolven were similar, the Mayo proposal was more considered and supportive of CDISC's aims but feedback from third parties regarding Tolven were not positive. Further conversations with Mayo resulted in their agreement to support BRIDG 2.x semantics.

Thus CDISC had the choice of two comparable technical solutions at equivalent costs. The Mayo option had the advantage of greater cooperation with NCI support.

Question	***Tolven***	***Mayo***
1. Regarding its underlying ontological representations:		
a) -- does the system support RIM semantics?	Yes	No -> Yes
b) -- does the system support BRIDG 2.x semantics?	Yes	No -> Yes
c) -- does the system support ISO 21090 DTs (or HL7R2 ADTs)?	Yes	Yes
d) – does the system support SNOMED?	Yes	Yes
e) – does the system support MedDRA?	Yes	Yes
f) – does the system support CDASH?	Yes	Yes
2. RE system functionality:		
a) – does the system support cross-terminology searching?	Yes	Yes
3. What is your estimation of the time involved in having a system installed and operational?	2 Weeks	2 Weeks
4. What is your estimation of the time/complexity of training -- <<based on previous experience within inexperienced domain experts entering terminology>> -- involved in terms of the goals of the pilot?	8 Hours	8 Hours
5. Please list any other relevant positive or negative issues that have been inadvertently left out of this brief list that you believe would affect the cost,		

efficiency, effectiveness, or overall quality of the use of the system in the context of achieving the two goals of the CDISC pilot project.		

8.3 Method

Using the semantic wiki tool (LexGRID) provided by Mayo, 50 oncology data elements from 5 volunteer organizations (Mayo Clinic, GSK, MD Anderson, Eli Lilly and Genzyme) along with valid value lists have been identified and loaded into the system. The team has been following a process to align equivalent data elements resulting in a single consensus version. As the team undertake the alignment work the process is 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, will be collected and reported. Specifically these metrics will include time needed to prepare, load and add concept references as well as the time needed for the harmonization process.

The wiki has been 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 will permit an assessment of how these support the process of aligning the definitions from the various contributing organizations and permit 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

8.4 Results to Date

The work on the pilot is still ongoing at the time of writing. Significant progress has been made since the start of the project in April. As noted above the various terminologies and source data elements have been loaded. Good progress has also been made on refining the process of harmonization of the elements into a consensus version.

It has quickly become apparent that this process is not as easy as it first appears and that the merge and search functions that are offered to users of the tool are a key component in delivering both ease-of-use and quality in the development process.

There are also issues resulting from the sheer scale of information presented to users and making it easier for users to see what they are working on; to provide tools that subdivide the problem space to make it manageable. There is also a need to provide an ability to collect data elements into meaningful collections, the concept idea.

These issues are now being investigated and improvements to the process implemented within the tool. This will then allow the various metrics to be collected and definitive conclusions to be determined.

At the time of writing it is possible to provide an informal assessment of the pilot outcomes:

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 – Yes, 70% confidence.
2. Can high-quality definitions be created and can ontologies help in ensuring such and avoid duplicate definitions being created. – Yes 50% confidence.

In terms of informing any subsequent development valuable lessons are being learnt. Initial thoughts have been captured in notes included within the appendices. It should be noted that these are preliminary and have yet to be formally reviewed.

9 Appendix A – Quality Requirement Definitions

Term	***Definition***
Availability	The amount or percentage of time that the System is available for use by the users. Availability may be negatively impacted by a variety of events including, but not limited to, user error, hardware failure, external system events, unavailability of support personnel, etc.
Compatibility	The ability of the System under discussion to appropriately interact with others systems in its context.
Completeness	For the domain of the System, the allowable maximum number or percentage of errors of omission.
Correctness	The allowable maximum number or percentage of errors of commission
Cost of ownership/Return on Investment	The total costs (direct and indirect) of owning the System.
Environmental	The environmental conditions in which the System must function
Extensibility	The use of the System in the same context with additional functionality.
Installation Complexity	The combination of direct or indirect costs of the installation of the System
Parallel Processing	The ability of the System to fulfill requirements simultaneously using duplicated rather than shared resources.
Performance	A measure of user expectations of System response times.
Portability	The ability of the System to fulfill its requirements in more than one operating environment.
Regulatory	The specific regulation(s) with which the System must be compliant.
Reusability	The use of the System in a different context with the same functionality.
Scalability	The ability of the System to fulfill its requirements for increasing numbers of users, transactions, etc.
Security	The requirements of the System with respect to access control and/or other context-specific security rules and or regulations.
Time To Market	The statement of the time at which the System must become available to and operable by its intended users.
Training Complexity	The combination of direct or indirect costs for the training of the System's users.
Usability	The measurement of how often, how efficiently, and/or correctly people use the System.

10 Appendix B – Stakeholder Analysis Discussion Points

The following discussion points were used as the basis for the stakeholder analysis calls.

1. Do you feel that an industry-wide metadata repository would fill unmet needs? If so, which ones?
2. How is your organization approaching this issue?
3. Should CDISC take this approach? Should CDISC lead this effort on behalf of other stakeholders? Who else should be included? In what capacity?
4. Are you aware of other projects in this space? If so, what are they and would it be possible to collaborate?
5. Would your organization use this metadata repository if CDISC makes it available? Would you expect it to be free? Or, would you be willing to pay a fee and/or provide a contribution?
6. Do you have suggestions for a funding/business model you feel would be viable?
7. What are the risks from your perspective?
8. What are the benefits from your perspective?
9. What specific requirements would you suggest for SHARE?
10. Do you have any advice/caveats that you would like CDISC to consider?
11. Would you/your organization be willing to delegate resources to help develop such a repository/library? Would your organization provide content?
12. Do you have any comments about the operating environment that such a system should have? Comments on the governance you feel should be in place?
13. Do you have any comments about assumptions or dependencies that should be taken into account for this project?
14. Have we missed anything?

11 Appendix C – Pilot Evaluation, Draft

11.1 Introduction

This document provides a brief outline of the architecture of the CSHARE pilot. It first describes the basic configuration of the environment, including the software and content. It then proceeds to make some *very* preliminary observations of how the components worked, what didn't work and what might be done to correct the problems.

11.2 CSHARE Wiki Configuration

The CSHARE evaluation wiki was based on the Mediawiki software stack. Mediawiki is among the most widely adopted of wiki bases, and is used in a significant portion of the wiki implementations available on the web today, the most prominent of which is Wikipedia. For the purposes of the CSHARE evaluation, the baseline Mediawiki software was enhanced with several extensions, including Semantic Mediawiki – an extension that captures assertions about the classification of and relationships between wiki pages as subject-predicated-object triples, and provides a SPARQL-like language for querying and displaying these triples. Another extension is SMW Halo – an add-on to semantic media wiki that enables Ajax based query of wiki semantics and in-line text annotation. The Semantic Forms extension enables forms based wiki data entry and the LexWiki extension that provides a model and a set of access methods for thesauri, classification schemes and ontologies.

The evaluation wiki was customized where necessary to accommodate specific CSHARE requirements. Examples include enhancement of the SMW Halo search capabilities to provide more sophisticated terminology search, software to load the Excel spreadsheet content into the wiki, tools to map concept codes from and two different coding schemes and enhancements to the Semantic Wiki query tool to allow queries from object to source as well as the built in source to object.

11.2.1 Initial Wiki Load

The CSHARE wiki was loaded with the following terminologies:

- SNOMED CT
- NCI Thesaurus
- HL7 Version 3
- ICD-9-CM
- MeDRA (subset)
- ICD-10
- CDISC CDASH Terminology
- CDISC SDTM Terminology
- CDISC SEND Terminology

After some discussion between Mayo and NCI, it was decided that it would be most important to be able to access the terminology resources listed above as a unified whole. If, for instance, the same concept existed in SNOMED CT, the NCI Thesaurus and MeDRA, we wanted it to be treated as a single unit rather than as 3 separate code system and concept codes. Based on this decision, the NCI extracted a subset of the NCI Metathesaurus, a resource that is derived from the UMLS Metathesaurus, that contained the coding schemes listed above. This extract contained 381250 distinct codes, 88% of which were assigned by the National Library of Medicine as Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs). The remaining 12% of the codes were assigned by the NCI – a portion of which might someday be subsumed by

the UMLS and a portion of which will probably remain unique to the NCI for the foreseeable future.

289741 (76%) of the nodes in this subset had a corresponding code in SNOMED CT. The NCI Thesaurus, which is a terminology resource maintained by the NCI constituted 68,982 codes in this subset, 50,930 of which did not have a corresponding SNOMED CT code. The MedRA subset contributed 43,900 codes, 28,331 of which had corresponding SNOMED CT codes, 5491 of which had corresponding NCI codes and 4165 of which had both.

The remaining terminologies (ICD-9-CM, ICD10, HL7 V3, ICDO (subset), UCUM and NCI-HL7 contributed 41685 codes, 22,783 of which had corresponding SNOMED codes.

The Metathesaurus subset was transformed into LexGrid format and then loaded into the wiki in the LexWiki model format. Tools were made available to map to and from the Metathesaurus and individual code systems. It should be noted that this mapping is by no means 1-1. The Metathesaurus entry for Arthroplathy (C0022408) maps to 5 ICD10 codes, 4 ICD-9-CM codes, 15 MedRA codes, one NCI Thesuaurus code and 8 SNOMED CT codes. The mapping software currently returns only the first matching code.

The wiki was also loaded with the BRIG 2.1 model. This was loaded in two forms – first as an ontology derived from an OWL rendering provided by Cecil Lynch and secondly as a collection of classes and variables that could be mapped to as an organizational model in the harmonization process.

The wiki was loaded with the ISO 21090 Healthcare Data Types. While there was a rendering of the data types which accompanied the BRIDG 2.1 OWL model, it was incomplete and lacked much of the ISO documentation, so we loaded a second image which was derived directly from the ISO 21090 XML Schema specification.

11.2.2 Pilot Content Load

We worked with the CSHARE community to determine the best format for loading the contributed community content. After some discussion and review, we settled on two spreadsheets – one for the data element descriptions and a second for loading the code lists (or value sets). Once the model was determined, we created a formal UML model that was used to map the spreadsheet content into the wiki. The UML model also described how the loaded content was mapped to terminology, data types and the points at which the content would be aligned.

Considerably later in the evaluation process, it was determined that it would be very useful to have both the CDASH and SDTM variable content available in the wiki as well. These were loaded as data elements from the CDISC namespace. We haven't had time yet to evaluate the full usefulness of this load.

11.2.3 Harmonization Process

After a series of iterative discussions, evaluations, prototype steps, etc. a prototype harmonization process was arrived at. This process involved into three steps:

1) Annotation, description and categorization of the individual data elements. This step involved adding names, definitions and semantic categorization to the individual data elements that were supplied by the evaluation community. This step was done by individual community members who were familiar with the use and purpose of the elements.
2) Selecting and sorting the annotated data elements to locate those that were closely related. This step has been referred to as "slicing and dicing".
3) Locating or, if necessary, creating one or more common data elements 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 wiki environment was tailored to attempt to meet the needs of the above process. In particular, it became heavily dependent on a tool, Exhibit, produced by the MIT Simile environment. This tool formed the framework of the "slicer and dicer", and was one of the more successful elements of the prototype, although it would certainly need to be enhanced and streamlined to function usefully in a production environment.

11.3 Preliminary Findings

11.3.1 Terminology Components

The terminology served four roles in the harmonization process:

1) **Classification**: The slicing and dicing process depended "semantic keywords" to determine whether two or more components were related. Formal terminology such as SNOMED-CT, the NCI Thesaurus, etc. provided a sort of "controlled terminology" from which these keywords could be drawn.
2) **Definition**: Terminological resources provided the potential for formally defining the intended meaning of both the community supplied variables and the harmonized data elements. Note that this is not the same as classification, as the purpose of a definition is to provide a formal and precise definition of the particular resource, where a classification is to provide a list of terms that might be used in conjunction with similar related elements.
3) **Value Meanings:** Each of the individual values for enumerated variables need to be linked to a terminological element that provides their intended meaning. As an example, a "1" in a Mayo patient gender variable might mean "male", and needs to be mapped to a corresponding concept code in a standard terminology.
4) **Value Sets**: Value sets represent collections of value meanings. As an example, a value set might represent possible anatomical locations, either in a particular or general context. The ability to determine the nearest value set that contained all of the value meanings for a particular variable turned out to be quite valuable when it came to determining when two or more variables might be related. As an example, we discovered that while the names and descriptions of Genzyme's "Outcome of Adverse Event" and the MD Anderson's "Adverse Event Outcome" were similar, the value set for the former was http://informatics.mayo.edu/cshare/index.php/Category:NCIM_CDISC_SDTM_Adverse_Event_Outcome_Terminology(CL370652) and the latter was http://informatics.mayo.edu/cshare/index.php/Category:NCIM_Adverse_Event_Outcome(1705586), and the set of possible values was quite different (the former listed the final status of the adverse event while the latter listed the *effect*)

While this evaluation is obviously very limited in nature, we observed that:

a) It was difficult to find the set of terminological components that were needed for classification. A search on almost any term name ("lesion size", "disease stage", etc.) yielded tens or even hundreds of possible terminological matches. We believe that there are at least two tasks that must be completed before this sort of terminological annotation becomes viable:

a. Terminology must be pre-vetted for classification. A community subject matter expert needs to create a list of classification "value sets" from which classification elements for a particular domain should be drawn. This needs to be done in such a way that missing elements can be added as needed. It also should be noted that it isn't obvious that it is necessary for these value sets to be drawn from existing terminology, although there will be benefits if it could be
b. Terminology tools need to be considerably more sophisticated than what is available from SMW Halo or even the Mayo extensions. Users need to be able to search by name, definition, code system, parent code, related code, and need to be able to easily display the details of a particular concept – both its textual and its associations with other concepts within selection dialog box.

b) Definitions require a model. A "pile of concepts" are not sufficient to define the intended meaning of a variable or common data element. A model, such as that being developed by the IHTSDO community identifies the various components that are needed to completely define variables while simultaneously limiting the possible selections for the various aspects of the model. The model also normalizes the granularity of various definitions – an issue that was brought out in the Tolven paper discussing the existing caDSR.

c) The ability to map value meanings to common terminology increases the ability to discover overlap. If, for instance, one community maps information to NCI Thesaurus codes and a second to SNOMED CT codes, the mapping work done by the NLM and NCI in the NCI Metathesaurus makes it possible to discover overlap and potential shared content.

d) None of the terminologies carried good value set definitions. While it was often possible to map individual elements. As an example, Eli Lilly provided a rich value set called "Lesion Method of Measure" (http://informatics.mayo.edu/cshare/index.php/EliLilly_Lesion_Method_of_Measurement_Value_Set). While most of the individual values in this set had matching meanings in the terminology space (e.g. "103" maps to SNOMED CT Code 289935006, http://localhost/cshare/index.php/Category:NCIM_Brush_biopsy_action(C0567343)), there didn't seem to be any useful upper level container that represented all of the possible methods. This may be significant, as set members do not necessarily correspond to ontological ordering.

e) Data types played a key role in classification. This said, the ISO 21090 data types appeared to be overkill, as we seemed to be interested in a very limited set (text, date/time, coded, numeric, …), and the nuances such as flavors of null, SET vs. BAG, CD vs CS, PQ vs PQR vs. INT, etc. went beyond what was needed for classification. Note, however, that the *mapping* from variables to common data elements, a step that was discussed but not implemented in this prototype may draw heavily on the details of the 21090 types.

f) The BRIDG model, by and large, was too coarse to add much significant information to what was already known. As with the ISO data types, it appeared that the BRIDG model could play a key role in subsequent model alignment steps, but was of little value from the harmonization perspective.

g) Units, as represented by the HL7 V3.0 UCUM system, played an insignificant role in the harmonization process. It appeared, however, that the notion of dimensionality (e.g. length, area, pressure, concentration, …) might play an useful role in the harmonization

of quantitative variables. Doing this, however, would require the selection of a baseline set of dimensions along with mapping to and from UCUM.

In general, the terminological component added significant value. It is particularly interesting to compare some of the annotations that have been done in the context of the caDSR with those done within this prototype – they are quite similar in coverage and quality. Not unexpectedly, however, the terminology is no "silver bullet". It is both too much and too little, and tools would need to be provided that aided in the selection of the right concept(s) from the terminology when they existed, and in the construction of post-coordinated concepts and sets when they didn't. In addition, tooling which did reasoning across the terminologies would be invaluable – both in discovering similar broader/narrower elements and in comparing pre- and post coordinated terms.

It should be noted, however, that SNOMED CT, the NCI Thesaurus, HL7V3.0 and UCUM each potentially play a different role. SNOMED CT provided broad coverage, for categorization and has the potential to be a primary candidate source for definitions, due to alignment with the IHTSDO model and the strong formal semantics. The NCI Thesaurus was the primary source of value sets, which is not unexpected as the NCI Thesaurus is where the CDISC variables have been recorded to date. The HL7V3.0 terminology provides alignment with HL7 V3 specific messages.

11.3.2 Process Components

The wiki environment served reasonably well as a vehicle for *discussing* the prototype. The availability of all of the terminological components in a single form, the ability to locate specific variables and sets of variables, etc. and the ability to rapidly change the layout and content of forms proved to be very useful.

The wiki environment, however, seemed less than ideally suited for much of the harmonization process. Semantic Mediawiki is a relatively free-form, customizable medium for publication and discussion. It is less than ideal, however, for processing large lists of values, batch mapping, sorting and selecting, etc. It *did*, however, present considerable potential for the purposes of discussion, evaluation and dissemination. We believe that a hybrid model, based in part on enhanced spreadsheets, customized applications *and* Semantic Mediawiki may provide a workable platform for the harmonization process.

It should also be noted that while the Semantic Mediawiki appears to be a useful mechanism for publishing harmonized content, it is probably not the ideal vehicle for communicating formal mappings and/or providing repository services. We would recommend creating an ODM import/export mechanism and a set of enhanced ODM based services for that.

One possible approach, however, might be to back off from the Excel spreadsheets as the primary import format and, instead, consider loading the individual organization forms *directly* into the wiki and doing the extraction and annotation process directly within the wiki.

11.4 Summary

The wiki was loaded with approximately 380,000 terms drawn from 9+ terminologies. While the terminology proved extremely useful in locating potentially similar variables, the process was not nearly as efficient as it could be were proper tooling and more appropriate domains specific subsets available. Many of the variables required more than one code to categorize and/or define, meaning that reasoning capability will be needed to be able to match "pre-coordinated" with "post-coordinated" terms. A formal observables model such as that being developed by IHTSDO would potentially be useful from both a completeness and a appropriate level of granularity aspect. The NCI Thesaurus provided most of the value sets that were found, and would make the best candidate for registering future value sets. The NLM and NCI mappings between code systems appear to provide considerable value.

From a process standpoint, the Wiki provided a useful prototyping tool, but was less than ideally suited for many of the batch sorts of tasks.

12 Appendix D – Business Requirements

For Pharmas and CROs this resulted in four detailed storyboards with requirements analysis for the following roles:

1. Protocol / Scientist
2. Data Manager / Collector / eCRF Developer
3. Analysis Dataset Creation / Biostatistician
4. Data Curator / Integrator / Miner

These storyboards have been included as separate files with the resulting requirements highlighted in Yellow (see files *SHARE BR Storyboard Data Steward-Curator.docx* and *SHARE BR Storyboard Pharma-CRO.docx*)

Additionally, a number of Pharmas have begun leveraging the CDISC SHARE business requirements in understanding and documenting internal business needs for "end-to-end CDISC implementation" and internal metadata repositories, designed to leverage CDISC standards as the foundation. GSK has made a great deal of progress in this area and has made available an internal document, specifying user and functional requirements. Key sections of this document have been included in a separate file (see file *SHARE BR GSK Func Reqs.doc*). The two diagrams at the end of this section depict the scope of the GSK implementation project with:

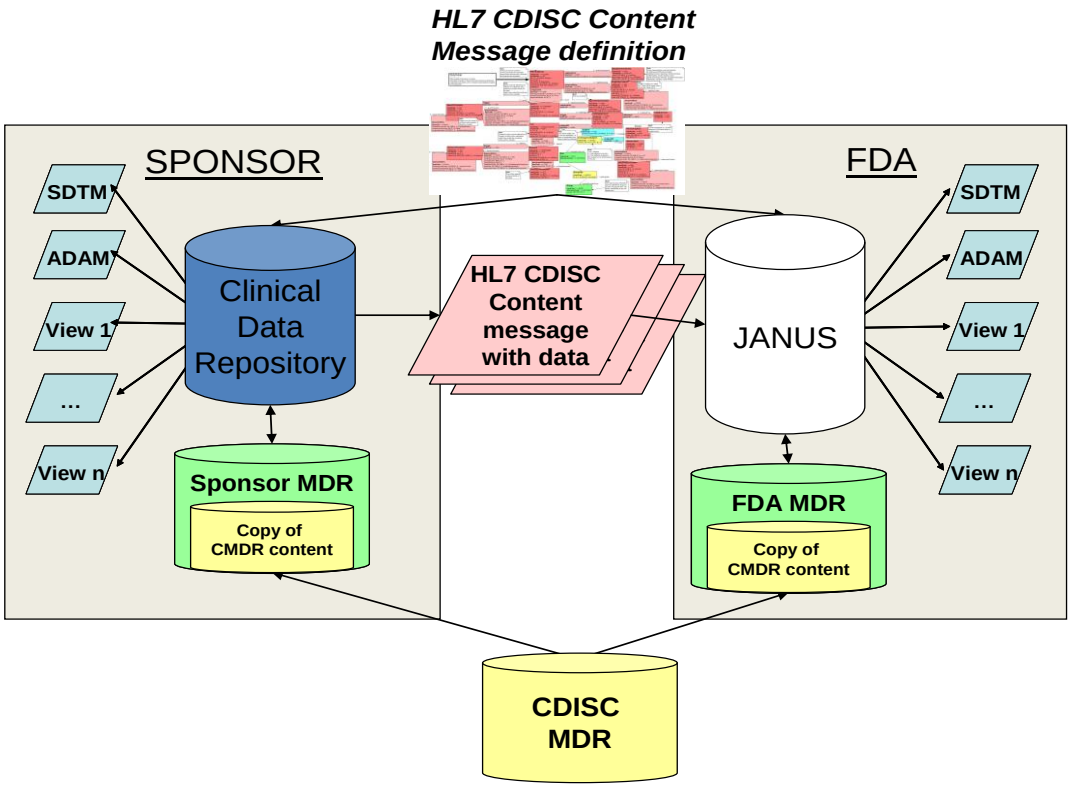
1. The place of Concept at the centre of GSK clinical study standards
2. Moving out from the centre with groupings of concepts to make generic specifications (e.g. the set of data for an AE) and including metadata needed to make operationally usable objects from the specifications
3. At the limit of CDISC implementation scope, providing SCIE study end-to-end tools and processes with the standards-based study-specific specifications needed for them to generate actual study objects (e.g. eCRF pages, operational datasets and SDTM datasets

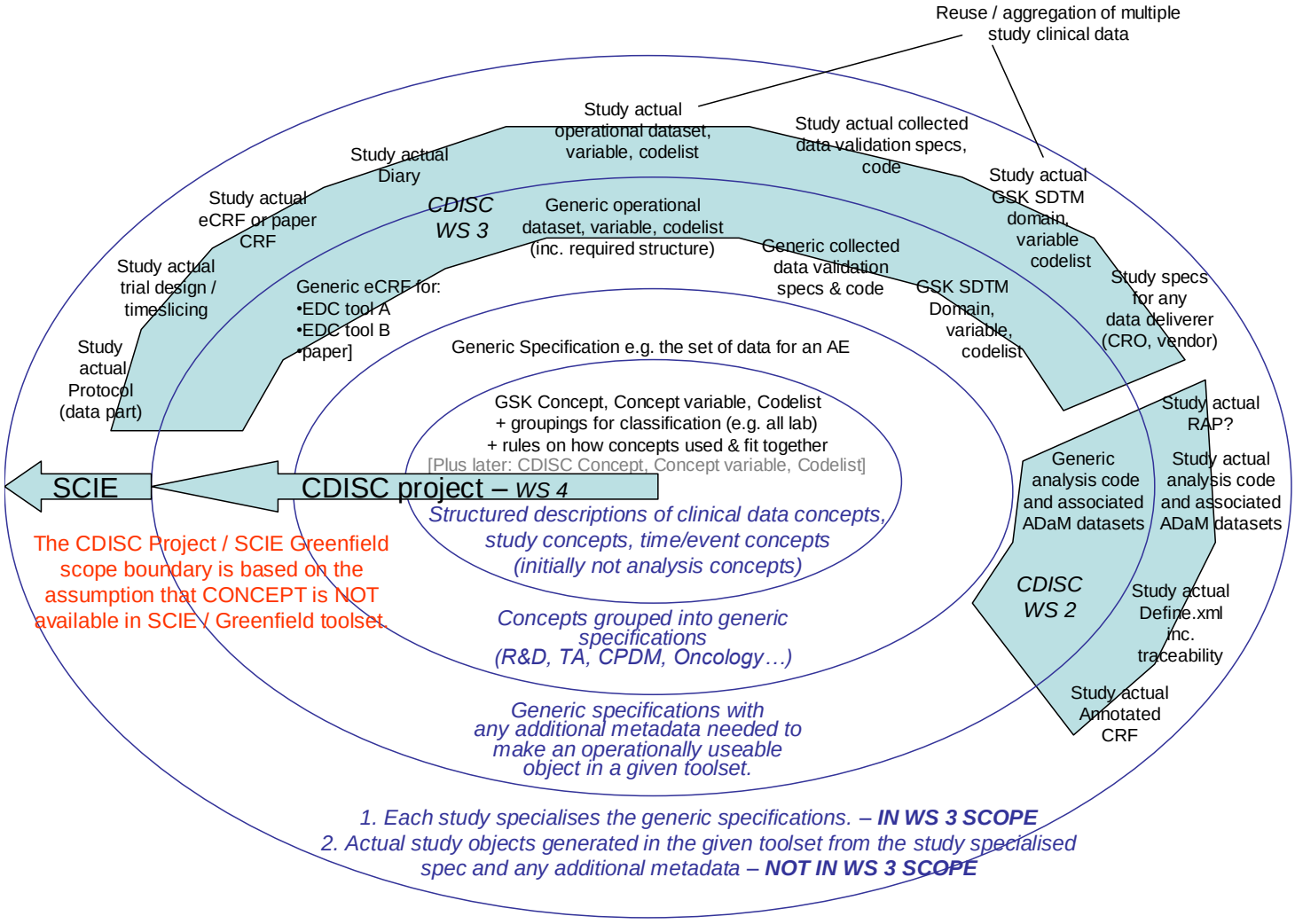
All of the CDISC SHARE business requirements activities have led to detailed discussions around: Variable definition; Attributes needed to describe variables; Concept Definition (both Simple and Complex); Attributes needed to describe concepts; relationships between Concepts and Variables; and Code list, value set binding and Controlled Terminologies. Requirements for Variables and Concepts are found in an attached file (see file *SHARE BR Variables & Concepts.doc*), and a Code list Guidelines document was developed by a team member (see file *SHARE BR Codelist Guideline.doc*).

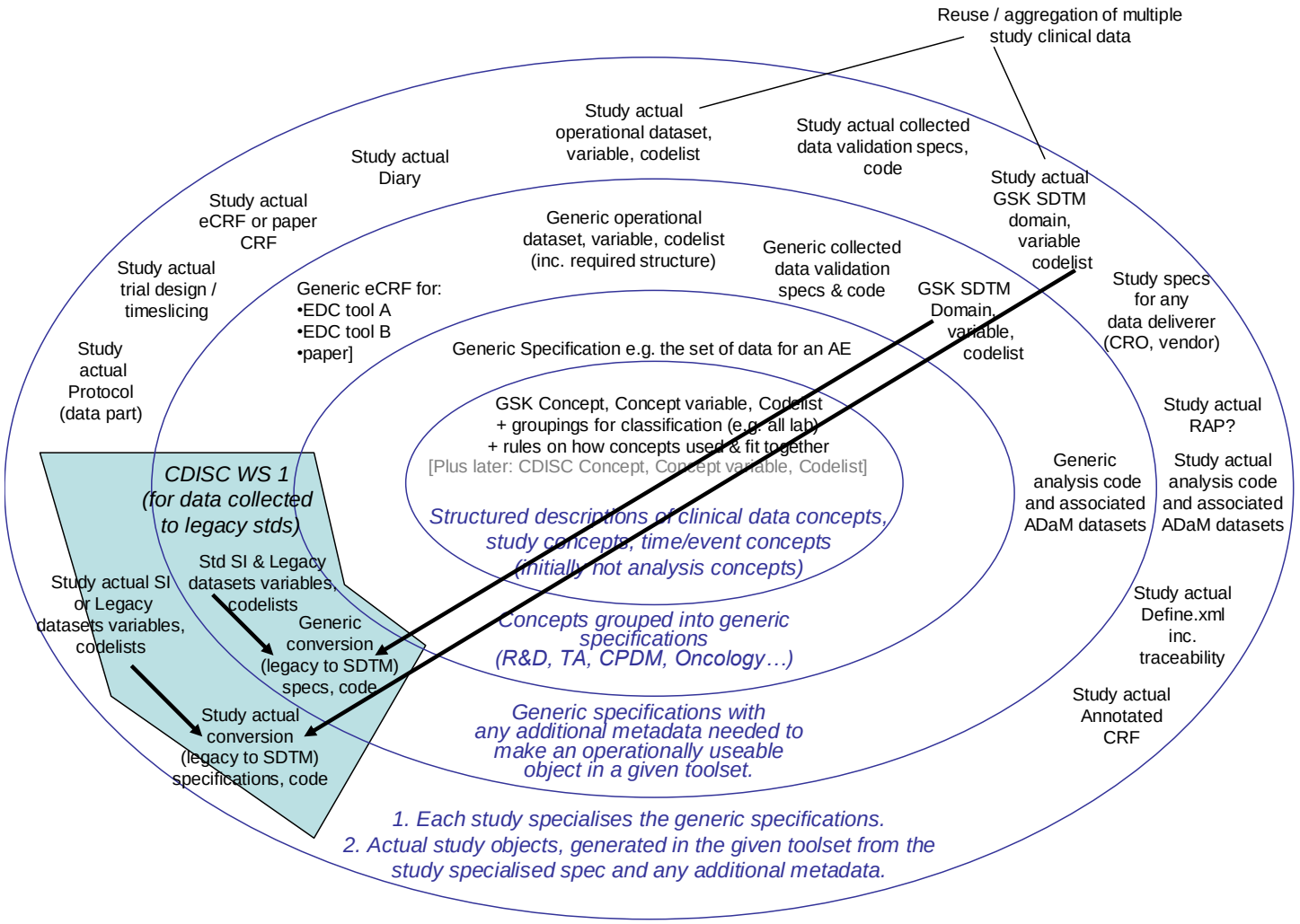
After developing an initial Storyboard for Regulators and Healthcare authorities, it became clear a single depiction would not suffice. There are many varying requirements not only across Federal Agencies / Organizations (e.g. US HHS, FDA, NCI, ONC, CDC) but also within a single Federal Agency such as FDA where different divisions exist (e.g. Office of the Commissioner, CDER, CBER, CDRH) with different needs and approaches. The Stakeholder Analysis in the previous section(s) proved to be a great start in identifying some of these differences, but more detailed requirements are needed. CDISC has begun aligning SHARE requirements with NCI and the planned release for the next generation caDSR. Also, through an ongoing partnership with NCI EVS, CDISC terminology for SDTM, CDASH and SEND is published in the NCI Thesaurus terminology environment. There it is kept aligned with key FDA projects such as Structure Product Label (SPL) and Individual Case Safety Report (ICSR). Additional discussions have been scheduled with other key Healthcare stakeholders (such as HL7 & the Clinical Interoperability Council) in conjunction with the upcoming HL7 Working Group Meeting in September.

The most publicized FDA Use Case is reflected on the right side of the following diagram, where FDA wishes to transport Pharma-CDISC content into the JANUS data warehouse via HL7

Version 3 messages. The current timeframe for such an implementation is 2013. The end goal is for FDA reviewers to extract needed “views” of CDISC standards.




Reuse / aggregation of multiple study clinical data
Study actual operational dataset, variable, codelist
Study actual collected data validation specs, code
Study actual Diary
Study actual eCRF or paper CRF
CDISC WS 3
Generic operational dataset, variable, codelist (inc. required structure)
Study actual GSK SDTM domain, variable codelist
Study actual trial design / timeslicing
Generic eCRF for: •EDC tool A •EDC tool B •paper]
Generic collected data validation specs & code
GSK SDTM Domain, variable, codelist
Study specs for any data deliverer (CRO, vendor)
Study actual Protocol (data part)
Generic Specification e.g. the set of data for an AE
GSK Concept, Concept variable, Codelist + groupings for classification (e.g. all lab) + rules on how concepts used & fit together
[Plus later: CDISC Concept, Concept variable, Codelist]
Study actual RAP?
Generic analysis code and associated ADaM datasets
Study actual analysis code and associated ADaM datasets
SCIE
CDISC project – WS 4
Structured descriptions of clinical data concepts, study concepts, time/event concepts (initially not analysis concepts)
CDISC WS 2
Study actual Define.xml inc. traceability
The CDISC Project / SCIE Greenfield scope boundary is based on the assumption that CONCEPT is NOT available in SCIE / Greenfield toolset.
Concepts grouped into generic specifications (R&D, TA, CPDM, Oncology…)
Study actual Annotated CRF
Generic specifications with any additional metadata needed to make an operationally useable object in a given toolset.
1. Each study specialises the generic specifications. – IN WS 3 SCOPE
2. Actual study objects generated in the given toolset from the study specialised spec and any additional metadata – NOT IN WS 3 SCOPE


Reuse / aggregation of multiple study clinical data
Study actual operational dataset, variable, codelist
Study actual collected data validation specs, code
Study actual Diary
Study actual GSK SDTM domain, variable codelist
Study actual eCRF or paper CRF
Generic operational dataset, variable, codelist (inc. required structure)
Generic collected data validation specs & code
Study actual trial design / timeslicing
Generic eCRF for: •EDC tool A •EDC tool B •paper]
GSK SDTM Domain, variable, codelist
Study specs for any data deliverer (CRO, vendor)
Generic Specification e.g. the set of data for an AE
Study actual Protocol (data part)
GSK Concept, Concept variable, Codelist + groupings for classification (e.g. all lab) + rules on how concepts used & fit together
[Plus later: CDISC Concept, Concept variable, Codelist]
Study actual RAP?
CDISC WS 1 (for data collected to legacy stds)
Generic analysis code and associated ADaM datasets
Study actual analysis code and associated ADaM datasets
Structured descriptions of clinical data concepts, study concepts, time/event concepts (initially not analysis concepts)
Std SI & Legacy datasets variables, codelists
Study actual SI or Legacy datasets variables, codelists
Generic conversion (legacy to SDTM) specs, code
Concepts grouped into generic specifications (R&D, TA, CPDM, Oncology…)
Study actual Define.xml inc. traceability
Study actual conversion (legacy to SDTM) specifications, code
Study actual Annotated CRF
Generic specifications with any additional metadata needed to make an operationally useable object in a given toolset.
1. Each study specialises the generic specifications.
2. Actual study objects, generated in the given toolset from the study specialised spec and any additional metadata.

13 Appendix E – Content Governance

13.1 Introduction

13.1.1 Purpose

The following sections detail the various stages of the CDISC SHARE Governance process and workflow as well as the resulting high-level system requirements. Other business requirements have been included in previous sections of the document.

13.1.2 Background

Due to the complexity of discussions surrounding CDISC SHARE Governance, this work was addressed in a separate team from other CDISC SHARE requirements activities. Primary team members included representatives from NCI EVS and CBIIT organizations, GSK, Eli Lilly, Genzyme, Octagon Research, Quintiles and Novartis. Several CDISC Board and TAC members actively contributed. Throughout the discussions, a number of industry models for metadata governance were considered including but not limited to – NCI caDSR, Lilly, Genzyme and Octagon Research. NCI provided a useful Lessons Learned document *(*see file *SHARE NCI caDSR Lessons Learned.doc)*

Initially, the Governance Team took a similar approach to other requirements teams by developing storyboards for the various roles and activities. Following development of the Data Steward/Curator storyboard (see Business Requirements), it became clear a different approach was needed. Storyboard development was waived in lieu of developing a series of detailed diagrams to depict the CDISC SHARE *Data Element Lifecycle* from data element inception to production and the corresponding governance workflow. This quickly proved to be an effective means to encapsulate the many threads of discussion. Due to the evolution of the Governance process and differences at various stages from Current Standards for Safety Data → Change Management → Future Standards for Efficacy Data, three different diagrams were developed.

The diagrams are included in separate files.

1. *SHARE CG Lifecycle Batch Load CDISC Std.pdf*
2. *SHARE CG Lifecycle Change Management.pdf*
3. *SHARE CG Lifecycle Harmonizing Company Stds.pdf*

13.1.3 Overview

The remainder of this section is organized into 5 main areas: (a) high-level requirements for the initial CDISC SHARE Harmonization Framework; (b) Roles and Responsibilities; and (c) and the three CDISC SHARE Governance workflow stages for receiving, processing, approving and publishing standard metadata aka common data elements. To follow are the 3 primary stages for CDISC SHARE Governance:

1. **Batch Load for Endorsed CDISC Standards** (Stage 1) – this depicts the process for: (a) loading an approved CDISC standard such as CDASH ver1.0 or SDTM ver3.1.1 including terminology content contained in approved standard code lists; (b) aligning it with the initial CDISC SHARE framework and content; (c) addressing discrepancies and processing changes; and (d) finally moving the standard into the CDISC SHARE production environment. In instances where there is a new, substantial release of an existing standard (e.g. SDTM 3.1.1 → SDTM 3.1.2) the delta between the two will be determined and channeled through the CDISC SHARE governance process.

2. **Harmonizing Existing Sets of Company Standards** (Stage 2) – this depicts the process for creating a new CDISC industry standard or "Gold Standard" by analyzing and processing multiple, existing efficacy content and/or company standards for a specific disease category therapeutic area or indication. This stage of the governance is the primary focus of the CDISC SHARE pilot.
3. **Change Management** (Stage 3) – this depicts the process to modify an existing CDISC standard and associated data elements already represented in the CDISC SHARE production environment. This may include an external request to modify a single data element (e.g. data element name, definition and/or codelist values) or multiple data elements as well as adding data elements to an existing standard.

13.2 Harmonization Framework Requirements

Before a metadata or data element governance process can be employed to add new CDISC SHARE content (or change existing content), there needs to be a foundation in place or baseline by which new content can be measured against. In the appendix diagrams this is referred to as the *Harmonization Framework*. It is important that this framework be complete, accurate and robust before current, endorsed CDISC standards are processed. The requirements of this framework include:

- Underlying Meta-Model such as ISO-11179 + extensions
- Underlying Domain Model, beginning with BRIDG ver2.1, with link to healthcare/HL7
- Link to NCI EVS system where CDASH and SDTM codelist terminology and variable names have been coded
- Underlying Ontology such as NCI Thesaurus (NCIt) or SNOMED CT where everything can be semantically represented and defined. This ontology should have an existing hierarchical structure for terminology with a strong base of well-defined terms to work from. A critically important aspect of any supporting ontology (that is often overlooked) is the dedicated expertise and resources to adapt such an ontology to meet CDISC's needs.
- External link(s) to other widely used controlled terminologies (e.g. MedDRA, SNOMED, LOINC) to allow for extension of value sets aka codelists. This may be accomplished through the NCI Metathesaurus or UMLS.
- Alignment with FDA's JANUS warehouse structure

Note: one of the more well-known examples of such a harmonization framework is housed at NCI. The caDSR (Cancer Data Standards Repository) is based on ISO-11179 and leverages NCIt *for the underlying ontology. BRIDG 2.1 is currently represented in caDSR with its underlying semantics represented as part of* NCIt *controlled vocabulary. All CDISC terminology is currently published in NCIt, where it is maintained and kept aligned with US FDA terminology by NCI staff.*

13.3 Roles & Responsibilities

To follow are the primary roles and high-level responsibilities needed for the initial stage of CDISC SHARE Governance. Each will need to be reassessed and evolve over time, particularly as the various Governance workflow stages are tested and implemented. In addition, It will need to be determined whether roles are full-time or part-time and whether they will be supported by a CDISC staff members, volunteers (provided by corporate sponsors such as GSK, Lilly, etc.), part-time paid consultants, partner organization staff members / contractors (e.g. NCI) or a combination thereof.

- **CDISC Working Team & Team Leader** – authors and publishes data standards and is ultimately responsible for a standard's lifecycle, including initial development as well as future releases. Once an endorsed CDISC standard is represented in the CDISC SHARE production environment, future releases should be developed and released directly within the CDISC SHARE framework. At that point, the operational nature of a CDISC working team will be redefined. As an example, current working teams may need to be restructured as specialized teams of Subject Matter Experts.
- **Submitters** (derived from ISO description) – Organization or people who send content to be added to the registry or change requests. Note – Submitters may be from an external organization or from within the CDISC or partner organization supporting CDISC SHARE processes. For organizations submitting externally, there should be a primary company point-of-contact who can become familiar with the CDISC SHARE framework, processes and contacts.
- **Metadata Steward** (derived from ISO and NCI description for "Context Administrator") – individual who is responsible for overall project management, shepherding a large amount of new content through the CDISC SHARE Governance process; responsible for metadata within specific areas and may have responsibilities that cut across multiple areas; oversees the metadata curation performed by Metadata Curators across a content project, functioning as the primary point-of-contact for project level activities and decisions.
- **Metadata Curator** (derived from NCI description) – individual(s) responsible for the bulk of the work within the CDISC SHARE framework and across the governance processes. Since Metadata Curators represent the Key Role, they should be dedicated, full-time CDISC staff members especially with the depth of training required. Curators work across the cradle-to-grave workflow for Data Elements: receives requests or proposals (likely via Steward)…reviews it…processes it…curates it…checks existing content…creates metadata content as well as aligning concepts, definitions and permissible values (new codelist, extending current codelist, sub-setting larger codelists); reviews and/or responds to requests from Submitter or Metadata Steward to create, revise (e.g. adding new values to existing codelist) or locate existing metadata that can meet the submitter's needs; interacts with and advises the submitter/requestor of the metadata to use and to ensure that metadata meets the submitter's needs; uses CDISC SHARE tools to create needed metadata, always reusing existing components as available; furnishes reports to the Context Administrator as requested to support resource determinations; prepares content submission packages when large sets of metadata is being processed to create new standard within the CDISC SHARE framework.
- **Metadata Analyst or Liaison** (as derived from NCI description) – acts as liaison between requesting and submitting groups and Metadata Curator to refine requirements, clarify requests, and identify standards or existing metadata that would be best candidates for reuse; responds with corrected requests when poorly-formed items are requested (such as choice lists that include redundant variables, poorly worded or restrictive definitions, or erroneous content is included); uses APIs or download mechanisms to locate needed metadata for use. This individual does not perform curation activities.
- **Subject Matter Experts** or SME (derived from NCI description) – is an expert in the content being processed. The SME has domain or content expertise that is needed to create and validate metadata in the registry. They respond to requests by the metadata curator (or submitters) to apply expertise to metadata requests - supply definitions, define permissible value sets, verify the use of concepts needed to annotate the variable. This role should include both clinically-focused SMEs as well as Technical / Data Standard SMEs that have an in-depth understanding of data standards, processes, necessary modeling etc.
- **Governance Committee** – is responsible for final adjudication and approval of metadata changes and/or new metadata being moved into the CDISC SHARE production environment

- **Other Roles through CDISC SHARE Partnerships** – there will be additional roles not depicted that may come through CDISC SHARE partnerships. For instance, terminology experts provided by controlled terminology partner(s) similar to those currently made available to CDISC by the NCI EVS organization.

13.4 Batch Load for Endorsed CDISC Standards (Stage 1)

This outlines *Stage 1* of the CDISC SHARE Governance process, focused on currently endorsed CDISC standards for Safety Data. This governance stage is designed to align already approved CDISC standards with the CDISC SHARE Harmonization Framework and accurately represent these standards in the CDISC SHARE production environment. In order to eliminate possible duplication with existing standards, it will be important to populate these standards in CDISC SHARE first before developing new standards within the CDISC SHARE framework. Since the SDTM and CDASH standards are the cornerstone(s) of the CDISC and clinical research data chain (representing data collection on the front-end and regulatory submission on the backend), these foundational standards should be integrated and represented in the CDISC SHARE framework first along with associated controlled terminology. In order to test the CDISC SHARE harmonization framework and accompanying processes, CDASH (considered the simpler of the two standards) should be processed initially in order to assess and refine the CDISC SHARE governance processes. To further simplify content processing in the beginning, the standards can be broken up into more manageable chunks such as by individual domains (e.g. Vital Signs, ECG, etc.). The process and mechanics should be thoroughly evaluated in order to identify and implement workflow improvements as early as possible.

Note: Although there has already been significant harmonization between the SDTM and BRIDG models, it is expected that aspects of SDTM will not align neatly with BRIDG (e.g. TEST and TESTCD fields). As 2 CDISC standards push back on one another (e.g. SDTM and BRIDG), there will need to be an accelerated decision-making process, where changes to either or both standards can be implemented as quickly as possible. This will either result in a change within the harmonization framework to BRIDG and/or a change to a standard that is already being widely used in the global marketplace.

13.4.1 Assumptions

- There is an existing CDISC SHARE harmonization framework in place as outlined in Section 2 and based on the BRIDG model
- CDISC working group has previously authored a data standard
- The standard has been publicly endorsed and is being implemented across the global pharmaceutical research community. For example, SDTM ver3.1.1, ver3.1.2 and CDASH ver1.0
- The standard is comprised of approved Data Elements (or variables), containing standard terminology codelists that may have be harmonized across multiple CDISC standards (e.g. CDASH and SDTM) and/or with external industry standards such as that of the FDA, NCI, HL7, HITSP, ISO etc.
- There is upfront, active collaboration between CDISC working group team members and those responsible for the CDISC SHARE harmonization framework and processes
- Prior to processing a standard through the CDISC SHARE machinery there is a general understanding of the resulting alignment as well as possible discrepancies.
- As a standard is aligned with the CDISC SHARE harmonization framework and moved into the CDISC SHARE environment, that standard will change and be improved. This will impact the standard already being implemented in the marketplace (e.g. SDTM)…the foundation standard in CDISC SHARE (e.g. BRIDG)…or both.

13.4.2 Resulting CDISC SHARE Requirements (Stage 1)

<< CDISC Working Team >>

- Exceptional UI and download capability that allows CDISC working teams to navigate, understand and extract information from the CDISC SHARE harmonization framework (including content, definitions, terminology codelists, structure, etc.)
- Published, downloadable and easy to implement machine-readable format specification (e.g. XML, ODM, SAS or other)
- Intuitive mechanism to utilize and align with required CDISC SHARE format specification in formatting a "batch" request
- Ability to easily submit a properly formatted "batch" request of metadata, representing a full CDISC standard
- Ability to view and track status and processing of the submission as well as change log

Note: CDISC SHARE requirements for CDISC working teams will evolve over time. Initially team members will submit a batch request representing the entire standard. Once the initial standard is represented in CDISC SHARE, the next time around they will only need to submit the delta between the original version of the standard and a new version. For instance, the delta between SDTM ver3.1.1 and ver3.1.2. Once a standard is fully specified and up-to-date in the CDISC SHARE framework and production environment, subsequent releases will be developed as part of the CDISC SHARE process and framework itself, as opposed to being developed externally.

<< Metadata Steward/Curator >>

- Ability to receive incoming batch request and auto-validate its structure
- If format is determined valid and information complete, ability to load the standard and metadata into the CDISC SHARE Staging Area for more detailed processing
- If error is detected, the entire batch is rejected with an error report sent to the submitting team
- Once the standard is in the CDISC SHARE staging area, there needs to be the ability to easily parse the file and decompose / assess the individual data elements of the batch submission
- Ability to manage a detailed Check Point Process to compare individual data elements, structure and codelists of the standard against the CDISC SHARE harmonization framework and production content to verify / check / fix nomenclature, content and structure
- Generation of a detailed comparison report with a comparative list of metadata, concepts, definitions, codelists etc., that allows curators to cross-check content submitted in the batch file against content that already resides in CDISC SHARE. This analysis should include the ability to compare data element names, definitions and codelist values (aka as permissible values or valid values)
- User friendly report noting results of cross analysis: (1) **Exact Match** noting where proposed data elements directly correlate to an existing production data element content; (2) **Near Match** noting proposed data elements that are similar to one that already exists where either data element name, definition or codelist values differ slightly; and (3) **Non-Match** where a proposed data element is not currently represented in the CDISC SHARE production environment. Non-Match data elements will be the simplest to address.
- Ability to quickly verify Exact Match items, so they can be set aside
- Ability to work with and communicate with CDISC subject matter experts to review Near Match items and formulate proposal recommendations to Governance Committee
- After each data element is addressed, there needs to be the ability to create a new "draft" standard for review and sign off by the submitting team
- Ability to format the full standards proposal, including recommendations and actions, for review and sign-off by the Governance Committee

- Communication mechanism between CDISC SHARE staff and Governance Committee representatives in order to address any questions.
- If proposal is approved by Governance Committee, ability to easily load content into CDISC SHARE production environment in Pre-Release status, so it can be thoroughly tested before Final release.

  NOTE: if CDISC chooses to partner with an existing maintenance organization and infrastructure such as NCI, there will need to be a communication mechanism between the maintenance organization and CDISC SHARE staff members to ensure proper hand-off and sign-off the CDISC SHARE production environment standard.

<< Subject Matter Experts (SME) & CDISC Team Members >>

- Ability to monitor progress of the standard, recommended changes etc., as it works it's way through the CDISC SHARE curation and governance process.
- Ability to easily communicate with Data Steward and/or Curator

<< Governance Committee >>

- Ability to efficiently review and sign-off on complete "draft" standard proposal for inclusion into the CDISC SHARE production environment
- If full proposal is approved and/or not approved, ability to communicate back to the responsible Data Steward

<< CDISC User Community >>

- There needs to be an effective public review mechanism for the industry to review updated draft standard and provide comments.
- Mechanism that enables curators and SME's to effectively adjudicate public comments.

13.5 Harmonizing Existing Sets of Company Standards (Stage 2)

This section depicts the process for creating a new CDISC industry standard or "Gold Standard" through the batch-loading, analysis and processing of multiple, existing company standards and industry content. This would likely be around disease-specific categories and indications, taking into consideration Efficacy Data (as opposed to the Safely Data currently addressed by existing CDISC standards).

13.5.1 Assumptions

- There is an existing CDISC SHARE harmonization framework in place as outlined in Section 2 above and based on the BRIDG model.
- Existing, core CDISC standards (e.g. SDTM, CDASH and others) are represented in the CDISC SHARE production environment along with all associated data elements. IMPORTANT: new CDISC standards for Efficacy Data should not be created using the CDISC SHARE mechanism and process until existing endorsed standards are fully loaded, integrated and represented in the CDISC SHARE production environment.
- Some disease-specific efficacy standards have already been developed, endorsed and implemented across the industry. These contain approved standards data elements, including data element name, definitions, code list values as well as domain models. For example, Tuberculosis (TB) and Cardiovascular Disease (CD) Acute Coronary Syndrome. Being driven by US FDA and CDC, the next stage of development is underway for the TB and CD standards. Also, standards development is underway for Kidney Disease as well as Alzheimers and Parkinsons.

- Some existing CDISC production standards, individual data elements and/or associated terminology codelists may be harmonized externally with other industry standards such as FDA, ISO, HITSP or HL7.

13.5.2 Resulting CDISC SHARE Requirements (Stage 2)

<<Submitting Organization >>

- CDISC first puts out a broad industry notification for disease-specific content, and organization receives such notification. IMPORTANT: curators must be involved in this process to ensure the notification does not include duplicate content that already exists in the CDISC SHARE production environment.
- Before this stage of CDISC SHARE Governance begins, external companies need a means to navigate and extract existing content, structure, terminology codelists etc. into their "internal" data dictionaries to harmonize and keep synchronized with different CDISC SHARE version releases.
- Exceptional and intuitive UI that allows Submitting Organizations unfamiliar with the CDISC SHARE framework to easily view, download and understand CDISC SHARE metadata content and harmonization framework (content, structure, relationships between metadata, terminology codelists etc.)
- Published, downloadable and easy to implement machine-readable format specification (e.g. XML, ODM, SAS or other)
- Intuitive mechanism to help align "batch" submission file with CDISC SHARE format specification, including full dataset as well as individual data elements – data element name, definition, codelist values if applicable, etc.
- Ability to easily and securely submit a large "batch" file of company content
- If assistance is required in the above steps, ability for submitter to communicate with Metadata Analyst (or others CDISC SHARE personnel)

<< Metadata Steward / Curator >>

- Ability to receive incoming "batch" submission and auto-validate full dataset, structure as well as individual data elements. It is important to parse out any data elements that do not fit the disease area or indication of interest. This will help alleviate unnecessary processing.
- If an problem or an error is detected, the entire batch file will be held-up (or rejected in its entirety) with an error report sent to the Submitter; there needs to be a communication mechanism between CDISC SHARE staff members and external submitting organizations.
- Once the submitted dataset passes validation and content is determined complete and correct, there needs to be an ability to load the full dataset into the CDISC SHARE Staging Area, keeping each of the various company datasets separated from one another.
- Once in the CDISC SHARE Staging Area, companies should be able to access, view and validate their own datasets, but not that of other submitting organizations.
- The previous 4 steps continue until there are enough representative Company datasets and industry content with which to begin processing and creating a new "draft" standard.

<< CDISC SHARE Pilot Focus - adjudication process to create a new CDISC Standard or "Gold Standard" >>

- After all the necessary batch files have been received and loaded separately into the CDISC SHARE Staging Area, Curators need the ability to effectively cross-reference and analyze the various datasets.

- Curators need the ability to parse out and decompose large numbers of data elements and compare similar data elements from across the company datasets, including data element name, definitions, terminology code list values, etc.
- After combining similar data elements, curators need to the ability to quickly analyze and cross-reference with existing CDISC / CDISC SHARE production content (both Safety and Efficacy content), so duplicate data elements can be parsed out as quickly as possible.
- Data elements should be flagged to reflect the results of processing against existing CDISC SHARE production content: (1) Existing Data Element–Exact Match…should be kicked-out of processing; (2) Existing Data Element–Near Match…should be redirected as a change request to assess need for modification; (3) New Data Element; and (4) Similar Data Element–New Concept Needed…for instance when a new data element concept is need to create a subset of code list values from a larger superset (e.g. Tuberculosis Location vs. Anatomical Location)
- After New Data Elements and Similar Data Elements remain for processing, curator needs the ability to effectively begin harmonization across the various company provided data elements to compare / align data elements names, code list value sets and definitions.
- Links to core terminology and other external terminologies is needed, so existing terms and definitions can be reviewed and processed for consideration; ability to collaborate with terminology partner(s) and experts to help identify applicable terminology or create new terms and definitions if needed.
- Ability to collaborate with clinical subject matter experts (SMEs) to sign-off on clinical data elements, concepts, definitions and code list values.
- Ability to create individual harmonized data elements and concepts to create “draft” standard data element
- Ability to construct a full “draft” standard from the individual data elements, creating the initial version of the new Gold Standard.

<< Metadata Curator / Steward and Data Modeler >>

- Ability to collaborate with CDISC SHARE experts and domain experts to create and/or load disease-specific model (if needed) and align as part of the standard.
- Ability to align disease-specific model with foundational CDISC SHARE foundational model (BRIDG)
- Ability to compile and incorporate feedback from all experts and create “Final Draft” standard, readying it for public review.
- Ability to receive and quickly parse and address public comments, refining new standard and generating proposal for Governance Committee.

<< Governance Committee >>

- Ability to efficiently review and sign-off on complete standards proposal.
- If full proposal is approved and/or not approved, ability to communicate back to the responsible Data Steward.

<< CDISC User Community >>

- There needs to be an effective public review mechanism and user interface for the industry to review full draft standard and provide comments.
- Mechanism that enables curators and SME’s to effectively adjudicate public comments and create “Final” standard

13.5.3 Change Management (Stage 3)

This section depicts the process for modifying an existing CDISC standard and associated data element(s) that is already represented in the CDISC SHARE production environment. This may include an external request to modify a single data element or multiple data elements as well as a request to add data element(s) to an existing standard. There are no additional assumptions for this Stage of Governance. The only additional requirements include a change request and processing mechanism as well as a means to version new releases of CDISC SHARE production content. A similar process is under development and testing by the CDISC Terminology Team to process terminology change requests and versioning. Since Stage 3 of SHARE governance will not need to be implemented until later, detailed processes and lessons learned will be extracted from Terminology Team activities.