

## Implementing CDISC When You Already Have Standards

Sandra Minjoe, Genentech, Inc., South San Francisco, CA

### ABSTRACT

How you choose to implement CDISC will be based on current company standards and their robustness, the amount of control you have at each step of the data flow process, your time vs. resource needs, and your willingness to change.

Working in a company with established standards can both help and hinder CDISC adoption. Because you have standards in place that so many people are used to working with, it can be quite a hurdle to replace them with the CDISC standards. And depending on the structure of the current standards, they may look vastly different than CDISC. However working within a larger organization also means you have people in place whose job it is to manage those standards, and getting those folks to embrace the CDISC standard can really drive it forward.

This paper will describe what we've done and are still doing to implement CDISC, and offer tips that would be helpful to any organization looking to implement CDISC.

### INTRODUCTION

A simplified view of CDISC, from data collection through reporting, can be thought of as:

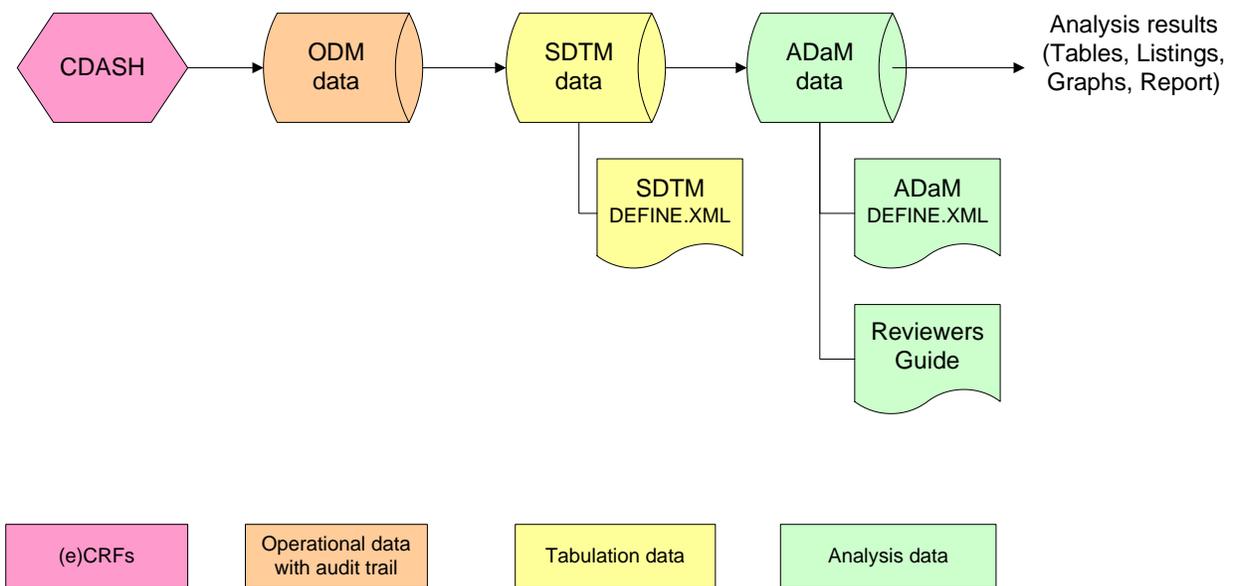


Figure 1. Simplified CDISC flow

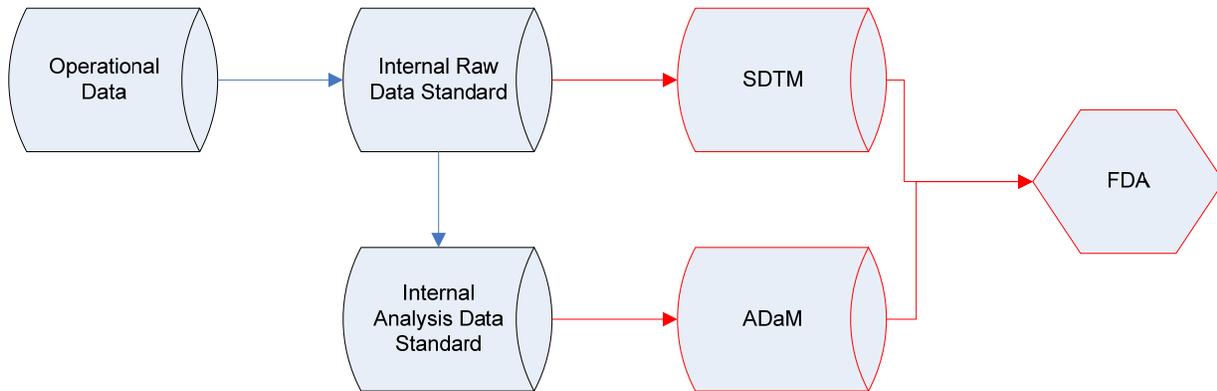
FDA has been threatening to require CDISC data for some time now, and we realized that we needed to start preparing for this eventuality<sup>1</sup>. The CDISC organization has stressed that a greater overall savings can be gained from implementing CDISC as far up-stream as possible<sup>2</sup>.

As you can imagine, implementing CDISC from data collection through reporting requires agreement and support from many levels of management. In a company with standards already in place, it can be difficult to convince management to make such a big change.

## STEP 1: PILOTS

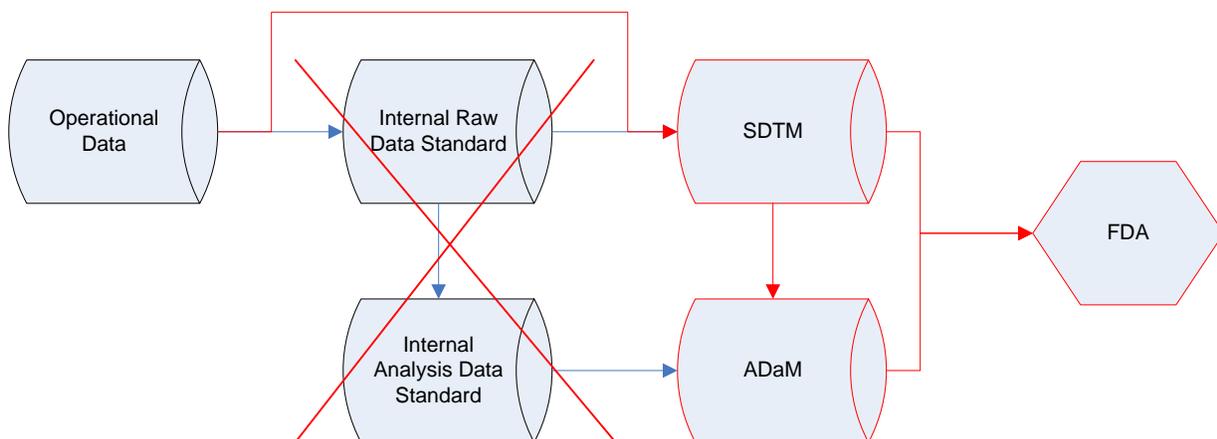
To allow us to best make any recommendation on CDISC implementation to the various levels of management, a few of us who were interested in and/or had some basic understanding of CDISC formed a cross-functional working group. The group included representatives from data management, statistical programming, biostatistics, electronic submissions, regulatory, and information technology. We set about trying to determine the “best” approach for CDISC adoption in our company.

Our first issue was to determine how much work it would be to convert a “typical” study to CDISC, so we’d have an idea what we were dealing with. Our hope was that it’d be a relatively simple and fast task tacked on at the end of the data stream, on the way out the door to FDA. In this way we would not need to change any of our current standards, processes or tools, and we could do this extra task only for the studies that need to be sent to FDA.



**Figure 2. Convert to CDISC on the way to FDA (leaving old systems in place)**

The alternative, to change our internal standards, processes, and tools, would result in a more streamlined process but require changes throughout the organization.



**Figure 3. Convert to CDISC in-stream (replacing current standards)**

To help us make our decision we decided to run pilots on some “typical” studies. We did not want to impact any filing teams, so a team member with experience in that indication was the only study expertise we used. Since we didn’t have a lot of extra resources to give, we contracted out the bulk of the work. We used two different companies, each to perform the data mapping and conversion for one our two pilots. These companies mapped and converted our operational data into SDTM, derived a handful of ADaM datasets, and created the supporting define documents. We used internal employees only to answer contractor questions and QC their work.

We learned a lot from these pilots, including:

1. SDTM leaves a lot open to interpretation. Companies choose to map the same type of information in different ways. Contractors who understand this will lay out the options, make some recommendations based on what they've observed in practice, and have you make the appropriate decision based on the types of data your company collects.
2. Data mapping and conversion appeared to be pretty straightforward to standardize. We estimated at least 80% of any study should be standard, leaving less 20% that really needs more oversight.
3. You must have a systematic way to check that your data conforms to SDTM. Manual checks are not sufficient.
4. It was a lot of work to convert our specs from
  - operational data -> internal raw data standard -> internal analysis data standardinto a define document that read as if we went from
  - operational data -> SDTM -> ADaM
5. You can submit pilot data to FDA. They will provide you feedback on whether or not they would be able to load it, including lists of any error and warning messages.

Our biggest concern was bullet point #4 above. This step is required for FDA to be able to trace analysis results to analysis data to tabulation data to CRF. We knew that by following the data flow as shown in Figure 2 (above) we would have to convert data on the way out the door, but we hadn't considered the additional work in converting specs to a define document. Because of this issue, we realized that we might have to indeed tackle some sort of an in-stream adoption of CDISC.

We found our pilots to be very useful in helping us determine the volume of work to be done. In retrospect it might have been even better to have used more of our internal resources in generating the pilot deliverables, rather than outsourcing so much of the effort.

## STEP 2: CREATING AN ADOPTION STRATEGY

We now had the information from our pilot, but wanted to supplement it with some additional industry research. We spoke to colleagues at various meetings, talked with many vendors, and read through official materials from both the FDA<sup>1</sup> and CDISC<sup>2</sup> websites. We reformed the cross-functional team and set about developing a CDISC adoption strategy recommendation.

### CHOOSING AN ADOPTION SCENARIO

We considered several different scenarios and weighed the pros and cons of each:

- 1. Leave all tools and processes in place and convert to CDISC on the way to FDA**
  - Pros: No change to current standards, tools, or processes. No need to convert studies that won't be filed to FDA.
  - Cons: Time consuming effort to wait until ready to file. Two versions of specs and annotated CRFs would be needed. The data used for in-house analysis would be different than that used by FDA, which would probably make it more difficult to answer their questions during a review.
- 2. Leave data collection and operational data in place and convert to CDISC before analysis**
  - Pros: No changes to data collection tools and processes or to operational data structure. Works for both internal data and data we receive from partners. Data used for in-house analysis is the same as that used by FDA, so answering questions would be straightforward.
  - Cons: Changes to analysis tools and processes would be needed. Two versions of annotated CRFs would be needed. The decision to convert a study would have to happen before determining if we'd use it in a filing.
- 3. Pull data as ODM from our operational system and use CDISC from there through reporting**
  - Pros: No change to data collection tools and processes. The conversion of operational data to SDTM becomes trivial. Only one version of specs is needed. Data used for in-house analysis is the same as that used by FDA, so answering questions would be straightforward.

- Cons: Changes to operational system and analysis tools would be needed. Two versions of annotated CRFs would be needed. Data from external sources would still need to go through some sort of conversion before analysis. The decision to convert a study would have to happen before determining if we'd use it in a filing.

**4. Change CRFs to collect CDASH data, pull data as ODM from our operational system, and use CDISC through reporting**

- Pros: The conversion of operational data to SDTM becomes trivial. Only one version of annotated CRFs and specs are needed. Data used for in-house analysis is the same as that used by FDA, so answering questions would be straightforward.
- Cons: Changes to data collection tools, operational system and analysis tools and processes would be needed. Data from external sources would still need to go through some sort of conversion before analysis. The decision to convert a study would have to happen before determining if we'd use it in a filing.

We realized that making this decision was not trivial: There seemed to be valid pros and cons for each scenario and we needed to somehow prioritize them to choose the right solution for our needs.

We used some parts of a tool from Kepner Tregoe<sup>3</sup> to quantify these pros and cons for us. This allowed us to identify our objectives, determine “must haves” vs. “wants”, and weigh each of the different scenarios against our objectives. And it was a fast process: the cross functional team, led by an outside facilitator, did this in a single 2-hour session. Even though we learned a little more and made a few minor tweaks after that 2-hour session, the ultimate ranking of the scenarios and thus our overall selection did not change.

Basically, the work was captured in a decision-making chart that looked like the following:

Must/ Want Weight	Objectives	Rating		Product of Weight and Rating			
		Scenario A:	Scenario B:	Scenario C:		Scenario D:	
M	Meet FDA requirements and timelines	Y	Y	Y		Y	
M	Increase efficiency of internal processes	N	Y	Y		Y	
7	Facilitate sharing data with business partners		8	56	9	63	9 63
5	Ensure smooth integration with current Dev processes and tools		8	40	3	15	1 5
10	Common data model across study teams (molecules, functions, phases)		3	30	6	60	8 80
...	...						
	TOTAL			XXX		XXX	XXX

**Figure 4. Decision-making chart**

We outlined 4 different scenarios, as described above. Then we described our company objectives, some of which I included above. Once those objectives were all listed, we determined which ones were “must-haves”. Everything that was not a “must have” was then assigned a 1-10 weight. Each scenario was evaluated against each objective to determine how well it would be met. All “must haves” were evaluated as a “Yes/No”, and the rest were given a 0-10 rating. (Note: a “No” for a “must have” was found in scenario A, so that scenario was quickly eliminated.) Then it was just a matter of multiplying the weight by the rating and totaling up each column. The column with the highest total value was the one that best met our objectives.

## CREATING A TIMELINE

Once we decided on the adoption scenario, we needed to figure out the timeline. As the team leader, I reviewed the status of all the CDISC standards we were looking to adopt and made an initial estimate of when each would be ready for our adoption and how long it would take us. I put it into the following layout:

		Resources (estimated Full Time Staff, \$ Purchase)														
		Year 1					Year 2					Year 3				
Function	Task	Q1	Q2	Q3	Q4	\$	Q1	Q2	Q3	Q4	\$	Q1	Q2	Q3	Q4	\$

**Figure 5. Resource estimate chart for each function and task over time**

Once I had this drafted, I then met with the team representatives from each function to review the parts of the table that applied to their function. They provided guidance as to when their function would be able to work on specific deliverables and how long they thought their parts of the adoption would take. There were some specific advantages of putting in this little work up front:

- It allowed me to identify the scope so that each function knew what they were dealing with
- It seemed to help the functional representatives to see the big picture and where their pieces would fit in
- It was quick to do, since multiple people reviewing a single document takes much less time than creating multiple documents and then combining them together

## GAINING APPROVAL

The final document we prepared for senior management contained the following sections:

- **Executive Summary:** a couple paragraph overview outlining the basic adoption proposal, including that the company keep up with the changing standards and modify the plan as appropriate
- **Scope:** describing what the adoption plan covers and what it does not
- **Recommendation:** a short bullet list of the specific CDISC standards to be adopted by which functions
- **Pre-Requisites to Success:** included things like a tight adherence to standards and ability to implement appropriate tools
- **Benefits:** the list of objectives from Figure 4 that would be met with the scenario we chose
- **Investment:** a summary of the resource estimates from Figure 5, followed by the full chart itself (Figure 5)
- **Current Issues:** the list of issues that drove the creating of Figure 4
- **Alternatives:** a 1-paragraph description of each scenario from Figure 4
- **Alternative Discussion:** a summary of the alternative chosen and the reasons why the other alternatives were not chosen
- **Decision Process:** a description of the method we used to make our decision (Kepner Tregoe) followed by the worksheet itself (Figure 4)
- **Team Representatives:** list of all contributors by function
- **Glossary of Acronyms:** helps reviewers who are not familiar with CDISC or potentially even acronyms common in one function but not another
- **Reference Documents:** Copies of documents or partial text from large documents that support the need to adopt CDISC, found mostly on the CDISC<sup>2</sup> and FDA<sup>1</sup> websites

As you can imagine, this ended up to be quite a large document! Ours was 8 pages for the recommendation itself, plus another 11 for the references.

The document was sent around for buy-in from functional heads. The next step was to get on a senior management agenda and present this information in 10-15 minutes.

Our slides were developed to summarize a lot of the information in the recommendation document. Because there was the potential that not everyone in the room would understand CDISC, we spent a couple slides giving the basic overview, including a chart similar to Figure 1. Instead of the Figure 5 resource estimates with more than 20 specific tasks listed, we condensed it down to the following:

Activity	Functions					Timing		
	A	B	C	D	E	2008	2009	2010
	S	P				X	X	
	S				S			X
	P	S		S			X	X

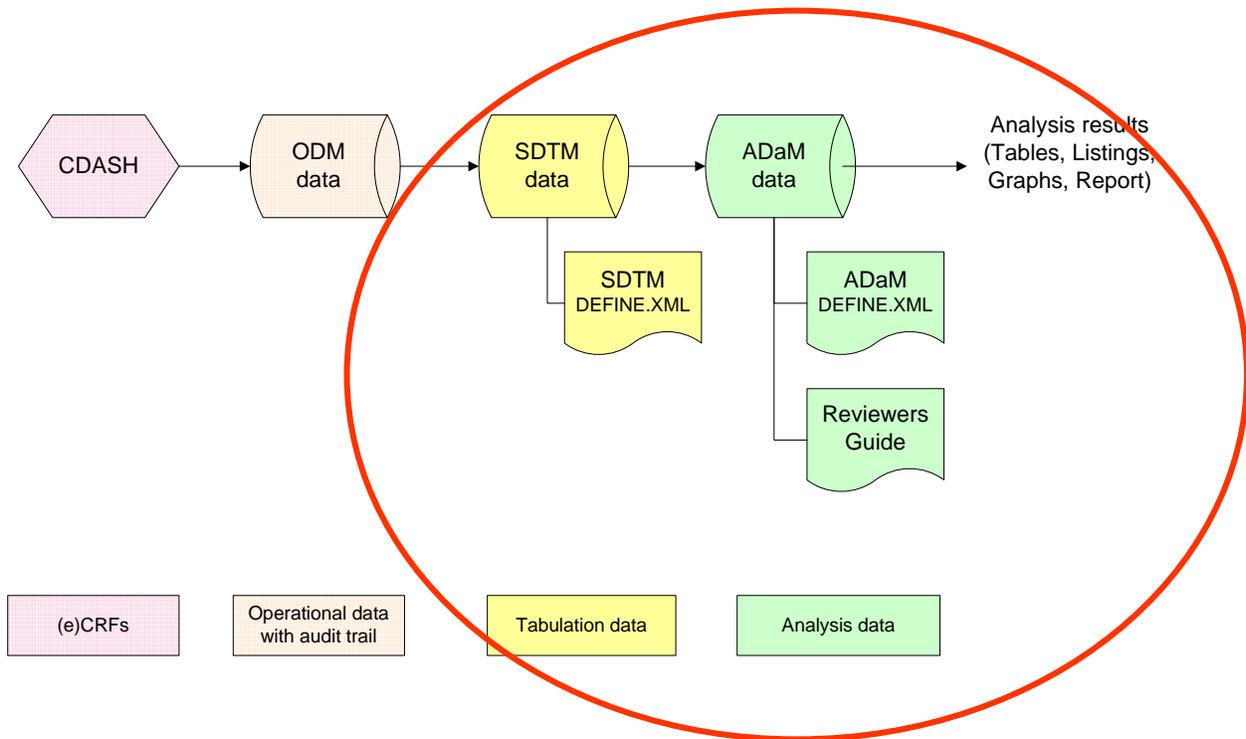
P = Primary, S = Supportive

**Figure 6. Rolled up resource estimate chart for use in slide deck**

We closed the slide deck by also recommending that one person take on the role to drive the CDISC adoption across all functions and ensure that all the functional deliverables would mesh together in the end.

### STEP 3: IMPLEMENTATION

Looking back at Figure 1, implementation for statistical programming and analysis focuses on the right half of that picture:



**Figure 7. Statistical Programming Implementation**

Even while all this cross-functional work was going on, we in the statistical programming and analysis function had

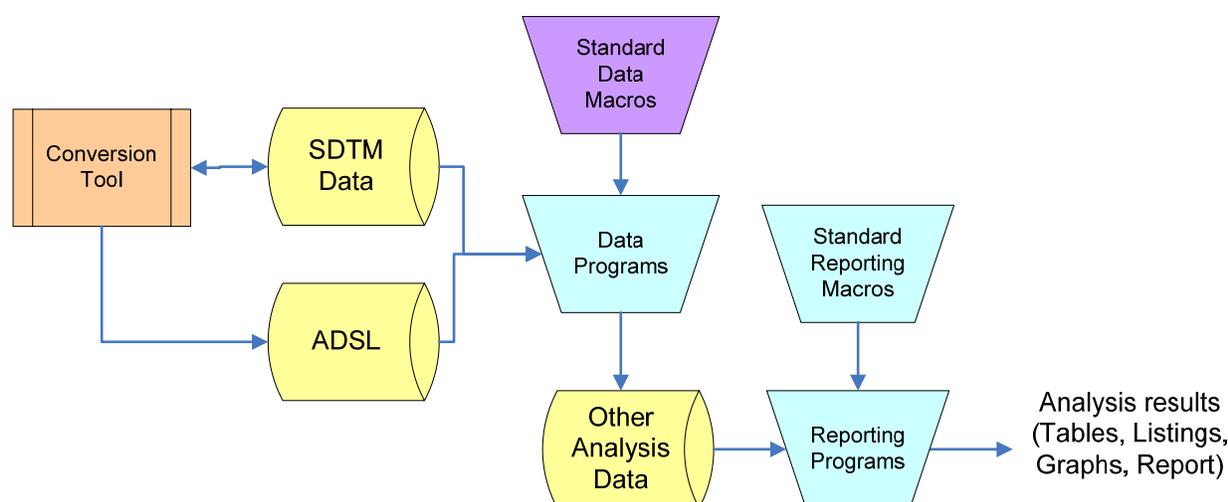
already been investigating CDISC as a new standard. We receive data in many forms, from paper CRFs, an EDC system, partners in their native structure, and outside vendors in previously-agreed-upon standards. Although we are unable to control the structure of the data we receive, once it is brought in-house we could convert it into a single standard structure before analysis. Not surprisingly, the structures we selected were CDISC SDTM for tabulation data and CDISC ADaM for analysis data.

Because the initial focus of our CDISC implementation was after data had already been collected in non-CDISC structures, for our implementation we started by developing:

1. A process and tool for mapping and converting collected data into SDTM
2. New standard analysis dataset structures (including ADaM) and programs, building from SDTM
3. Revised standard reporting macros, based on new data structure and variable names
4. Simple SAS macros to help us work with data in the new structure

Additionally, we also began looking at our filing needs, including define.xml instead of define.pdf that would allow us to submit a true CDISC filing.

Below is a simple diagram showing initial basic steps from the data conversion tool through production of TLGs.



**Figure 8. Summary of Implementation First Steps**

### IMPLEMENTATION DECISION #1: USE TRUE SDTM OR SDTM-PLUS?

In listening to how other companies have implemented SDTM, we've heard again and again about "SDTM-Plus". This term basically means data in the general SDTM structure, with some extra columns to make it more human-workable and analysis-friendly. We needed to decide whether we wanted to use this type of modification to SDTM as our in-stream standard, or if we should go with true SDTM structured data:

#### Use True SDTM

- Pro: We'd be working with exactly what FDA would be using, thus allowing us to easily provide any analysis programs and answer their questions.
- Con: The SDTM structure is much more vertical than our current standard and requires that many variables be put in SUPPQUAL, a potentially difficult transition for many employees.

#### Create an SDTM-Plus

- Pro: It allows you to use some of the features of SDTM without being held to all the constraints. For example, you might include age, sex, race, and treatment group and convert character dates to numeric ones to make SDTM data into SDTM-Plus.
- Con: It would require modifications to specs to create DEFINE.XML and to any programs that need to be sent to FDA, and it would make it more difficult to answer FDA questions.

## **Decision**

We decided that the benefit of working with the same data structure as FDA was the most important need to address. And because we are, after all, SAS® programmers, we developed a few SAS® macros that would allow us to do some standard things to an SDTM dataset to make it easier to use for both us and our friends in Biostatistics.

One of these macros creates numeric dates and times from the character version in SDTM. All these SDTM variable names end in “DTC”, so they were easy to locate. The conversion piece of the program was a matter of string manipulation to parse the character dates and times, decision steps for imputing values as needed, and pushing that into SAS® date and time functions to create resulting numeric dates and times. Kind of fun to write, actually!

Another macro first transposes and then merges all the corresponding SUPPQUAL data onto the relevant domain. SUPPQUAL is set up to make this fairly straightforward, since variable QNAM becomes the transposed variable name, QLABEL becomes the transposed variable label, and QORIG becomes the transposed value.

For us, this combination of true SDTM plus a suite of standard macros allowed us to meet both criteria – using exactly what we send FDA and making SDTM a little easier for us to work with.

## **IMPLEMENTATION DECISION #2: WHAT TOOL TO USE FOR CONVERSION?**

Once we decided what our SDTM model would look like, we needed to figure out how to get all of our data there. Some of the options we considered for data conversion were to purchase an off-the-shelf ETL (Extract-Transform-Load) tool, create a home-grown SAS tool, or farm the work out.

At the time our tool development programmers were busy with other projects, so building our own system wouldn't happen quickly. We determined that SAS® Data Integration (DI) Studio wasn't the best fit for us, especially since there would be a steep learning curve and we'd have trouble finding contractors experienced with SAS® DI Studio who could do the data conversion for us.

In working with our IT department, we discovered that the company was already licensing an ETL tool, and to use it we would not have to pay for any additional users. We learned that a large Pharmaceutical company was already using this same ETL tool for their data conversion into SDTM. Additionally, our in-house pre-clinical group had recently begun using the tool for conversion to their data warehouse (and, as we described it, lab data is lab data, be it mouse or human). Finally, contractors familiar with this more common ETL tool were relatively easy to find, so we could get started right away.

As it turned out, this ETL tool we used did not meet our needs. While we could easily talk with the contract programmers in a common language (SQL), they were not able to develop a system that worked across multiple studies with even slightly different incoming data structures. It took more than 6 months just to convert 2 studies, and we had no expectations of any dramatic time savings for future studies.

In the end, we dropped the ETL tool in favor of developing our own SAS® based tool. We learned a lot about data conversion during this process, and developed re-usable standard data maps, so it was not a total loss.

## **IMPLEMENTATION DECISION #3: HOW MUCH ADAM SHOULD WE USE?**

ADSL (Subject-Level Analysis Dataset) is very well-defined in the CDISC ADaM documentation. That documentation describes the structure (one record per subject) and gives examples of many variable names. FDA seems excited about including the ADSL domain in their Janus database. Conveniently, we were already creating a dataset that contained much of the same information as ADSL and in the same one-record-per-subject format. Implementing ADSL was an easy decision to make, since it basically involved changing a few variable names.

Other ADaM structures have been less well-described in the CDISC ADaM documents. In the past ADaM had defined a set of structures for different types of analyses, such as time to event or categorical, though in the current ADaM release these have been removed. Much of our analyses, such as adverse event and lab summaries, could actually be done with SDTM-plus style data, so we didn't see a huge need for other “true” ADaM data structures.

We've decided to use a combination of ADSL and SDTM-plus for most of our analyses, and short-term we'll continue to use our old standard analysis file structures for everything else. As ADaM delivers more specific direction on data structures, we will add them to our implementation.

## **CDISC IMPLEMENTATION TIPS**

I'm generally happy with how things have been developing in our CDISC adoption, though there are a few things I'd have done differently if I were to start all over again. Combining it all together, I'd recommend the following tips:

- Form a group of CDISC-interested parties across your company, and do a pilot or two to get a feel for the CDISC data structures before making any bold moves
- Form a small cross-functional team of highly placed functional reps and standards representatives and develop a well-thought-out adoption strategy recommendation to senior management, including both functional and cross-functional oversight
- Convince at least one high-level person that this is the right thing to do, and have them act as a CDISC champion
- Become a CDISC member organization and have at least one person from your company get involved in some part of the industry work they do, and you'll get access to all sorts of information not available to the general public
- Hire consultants and contractors sparingly so that, as much as possible, CDISC knowledge is learned by and kept in the organization
- Make decisions based on current resources, including short-term gap solutions when necessary
- Revisit short-term gap decisions frequently and move to a longer-term solution as soon as possible
- Provide CDISC communication regularly to instill a sense of anticipation and a desire at all levels to make the change

## CONCLUSION

Adopting CDISC across a larger company will take awhile, and you're likely to have a few setbacks as you go along. It's not a race against other companies, since each has their own issues to deal with. Our goal should be to ensure our own company will be ready to deliver our data to FDA in the way they will soon be mandating.

Also realize that by adopting CDISC you may also see other benefits, such as streamlined work processes, a better ability to share data with our partners, and less time spent defining data structures between vendors and clients.

Finally, consider the implementation of each standard as a milestone to the overall CDISC adoption plan, and you'll be rewarded with many successes along the way.

## ACKNOWLEDGMENTS

Thanks must go out to Genentech's CDISC Working Group and Data Model Standards team, who did the bulk of the work described in this paper. Thanks also to the many members of CDISC who have so generously answered our questions as we attempted to implement these standards.

## REFERENCES

1. See <http://www.fda.gov/> for official FDA statements. Note that a search of FDA's website on "cdisc" will give you over 1000 hits. The specific FDA document that mentions CDISC is The PDUFA IV Technology Plan is at <http://www.fda.gov/OHRMS/DOCKETS/98fr/FDA-2008-N-0352-bkg.pdf>. CDISC also collects references from FDA on their website (see reference 2 below).
2. See <http://www.cdisc.org/>, specifically the tab for Publications and Presentations, for more CDISC information. Two publications outlining the business case for implementing upstream can be found at <http://www.cdisc.org/news/PR31CDISCBusinessCaseApr2007final.pdf> and [http://www.cdisc.org/pdf/Interchange2005/01-Business\\_case\\_for\\_standardsV6\\_\(2\).pdf](http://www.cdisc.org/pdf/Interchange2005/01-Business_case_for_standardsV6_(2).pdf). The area where you can find FDA-related articles is at <http://www.cdisc.org/publications/regdoc.html>.
3. To learn more about Kepner Tregoe and their decision-making workshops, see <http://www.kepner-tregoe.com/>.

## RECOMMENDED READING

It probably goes without saying that a thorough review of the CDISC standards themselves is recommended. Those standards can be found at <http://www.cdisc.org/standards/index.html>. Note that the documents in the first section of this web page are the most current production versions, though you can find versions in development further down the web page.

The "References" section above also includes many useful resources.

## **CONTACT INFORMATION**

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

Sandra Minjoe  
Genentech, Inc  
1 DNA Way  
South San Francisco, CA 94066  
650-225-4733  
sminjoe@gene.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.