

Enable and Automate: A Technical Roadmap Defining CDISC's Path to End-to-End Automation

Sam Hume, CDISC, State College, PA

ABSTRACT

Enable and Automate is an important pillar of CDISC's new strategy that provides the vision and roadmap to enable standards-driven automation across the clinical research data lifecycle—from study design to analysis results. This paper provides a high-level synopsis of the technical roadmap, describing how the CDISC community will implement and showcase end-to-end automation. Key initiatives include significant standards updates, open-source software tools, and test datasets that will be used to demonstrate the research data pipeline automation. One of the major deliverables is a fully pre-configured example study, with all study components represented as machine-readable metadata, from design through submission. The roadmap outlines a series of epics that we will develop and implement over the first phase of the 360i program.

INTRODUCTION

The roadmap outlined in this paper highlights the key deliverables needed to complete Phase 1 of the 360i program. Phase 2 will start concurrently by focusing on foundational development work, such as the definition of analysis concepts, needed before the phase 2 implementation work can begin. This paper highlights the epics defined for 360i Phase 1.

In addition to providing concrete definitions of the Phase 1 epics, this paper provides a set of principles that will guide 360i development and are important to the correct interpretation of the roadmap. These rules of the road provide insight into how the 360i program will be executed. The next section provides a brief introduction to the 360i program and the Enable and Automate strategic pillar.

360I AND THE ENABLE AND AUTOMATE STRATEGIC PILLAR

The Enable and Automate pillar of CDISC's 360i strategy aims to deliver standards-driven automation across the clinical research data lifecycle - from study design to analysis results. More generally, the CDISC 360i Enable and Automate pillar supports CDISC's strategic focus to improve standards implementation. This focus on standards implementation helps organizations maximize the value generated from their investment in standards. Improving standards implementation involves, in part, demonstrating standards-driven, end-to-end automation across the research data lifecycle. A testable Definition of Done for the Enable and Automate pillar is as follows:

360i has published a complete pre-configured study package with all the components defined in metadata from study design to submission, test data for the study, and open-source software to run the study data pipeline to generate analysis results.

The 360i Definition of Done study package includes:

- Metadata: publish a complete study metadata package that covers the full study data pipeline from study design through TLFs.
- Data: publish the complete set of raw datasets for the test study to execute the automated study data pipeline.
- Software: publish a pre-configured set of open-source software tools that consume the study metadata and test data to execute the study design.

360i extends CDISC's collaborative development model beyond defining standards to include the use of open-source software tools to automate the standards-driven research data pipeline. As part of 360i, CDISC seeks to collaborate with developers so that the open-source software can be used to test standards as they are developed. This will help ensure that new versions of standards work with the software tools in an end-to-end automated research data pipeline.

This does not mean that CDISC will be developing the 360i software. For 360i, CDISC seeks to use existing open-source software and to collaborate with developers to fill gaps and add new features. This is similar to how CDISC develops standards. CDISC volunteer teams do the majority of the standards development work. CDISC would like to work with the developer community, open source and commercial, to ensure that off-the-shelf tools are available to

implement the standards. The availability of common tools will yield numerous benefits, but fundamentally it makes it possible for implementers to maximize the value they get from using the standards.

Two main objectives have been defined to realize the goal of producing standards-driven, end-to-end automation:

1. CDISC develops pre-configured, ready-to-use standards. These standards include all the metadata needed to implement them as they are published in the CDISC Library. Pre-configured, ready-to-use standards mean that less sponsor-specific metadata is needed and this creates many benefits, including:
 - Simplifies standards management and maintenance
 - Reduces variability between studies
 - Simplifies building software tools that automate standards-driven data processes
2. The second objective is to develop standards that support automated processes. Automated processes that use pre-configured standards create several benefits, including:
 - Reduce the variability between studies
 - Increase efficiency and quality
 - Enables standards testing to improve quality

Pre-configured, ready-to-use standards make it possible for industry to work from one definition of the standards. Combining one definition of the standards with standards-driven automation makes it possible to use off-the-shelf software tools to automate the research data pipeline.

The 360i development program has been divided into 2 phases:

1. Phase 1: Study design through SDTM dataset generation and conformance checking
2. Phase 2: ADaM through analysis results including TLF generation

The Technical Roadmap that follows covers Phase 1.

THE PHASE 1 TECHNICAL ROADMAP

This section describes the epics identified for the Phase 1 Technical Roadmap. It highlights concrete epics that will help drive the 360i technical development. Inevitably, the details enumerated in the epics will evolve as we get into the project, and this is a key rationale for embracing an agile methodology. This roadmap does not include Phase 2 work, and that will be covered in another document.

The Phase 1 Technical Roadmap targets a release that automates SDTM dataset generation and is comprised of the following epics:

1. Create study design
2. Generate case report forms (CRFs) and electronic data transfers (eDTs)
3. Generate the SDTM Define-XML
4. Generate annotated CRFs
5. Generate shell datasets
6. Populate the eCRFs and eDTs with raw data
7. Load the operational data store (ODS)
8. Create the oak.sdtm algorithm set
9. Generate the SDTM datasets
10. Create the open rule set for the SDTM datasets
11. Generate a conformance report

These epics will be composed of product backlog items (PBI) in the 360i backlog. The epics work together to form an SDTM generation pipeline, part of the research data pipeline, that will be released at the completion of phase 1. Not all epics must be developed sequentially. Certain epics and PBIs can proceed in parallel which will help the project best utilize the volunteer contributors.

The following sub-sections provide a high-level description of each of the Phase 1 Technical Roadmap epics. The initial scope of the 360i phase 1 work will use the CDISC Pilot Study protocol datasets, also known as the LZZT study. Within the LZZT study, the initial focus will be on a few selected safety domains. After SDTM dataset generation has been demonstrated for these initial domains, additional domains will be added.

1. CREATE STUDY DESIGN

Create Study Design uses existing LZTZ artifacts to generate the study design using the Open Study Builder (OSB) and Study Definitions Workbench (SDW) software applications. These applications will be hosted for team use. The study designs will be exported from both software applications in JSON format. The content of the study design will be consistent with the Digital Data Flow (DDF) Unified Study Definitions Model (USDM).

This epic includes populating the ODM v2.0 study design model with the exported USDM content. Since the ODM study design model is not yet supported by the software applications, initially it will be generated using a Python script and the JSON study design exports. Other artifacts may also be generated if supported by the software applications, such as ODM-based CRFs.

Certain content, such as Biomedical Concepts (BC), can be de-referenced using the CDISC Library API.

The schedule of activities (SOA) will be included in the exported study design.

Inputs	Tools	Outputs
• LZTZ study design XLS • Biomedical Concepts	• Open Study Builder • Study Definitions Workbench • CDISC Library API • Python script	• USDM JSON export • ODM Study Design

2. GENERATE CASE REPORT FORMS AND ELECTRONIC DATA TRANSFER SPECIFICATIONS

The study design generated in the previous epic will include BCs that represent data to be used in the study. The raw data will be represented by CRFs or eDTs, both in ODM format. ODM metadata will be generated by the OSB application assuming that feature is available, otherwise a Python script will be used to generate CRFs and eDTs from the study design.

To generate CRFs, the BCs in the study design must be accessed in the CDISC Library and the associated dataset specializations retrieved. The CDASH dataset specializations will support the generation of ODM-based CRFs and eDTs. A Python script will be used to generate the ODM CRFs and eDTs. ODM based content may be represented in JSON or XML. For ODM in XML, an ODM stylesheet will be used to generate HTML outputs.

Inputs	Tools	Outputs
• Study design exports • Biomedical Concepts • Dataset Specializations	• OSB • Python script • CDISC Library • odmlib • BC Browser • ODM Stylesheet	• ODM-based CRFs • ODM-based eDTs • HTML CRFs • HTML eDTs

3. GENERATE SDTM DEFINE-XML

Similar to CRF/eDT generation, we will generate the SDTM Define-XML using BCs and Dataset Specializations. This time SDTM dataset specializations will be used. The Define-XML will include all the metadata needed to define each ItemDef/Ref as well as the value level metadata (VLM). The generated Define-XML will be rendered as HTML for viewing using the Define-XML stylesheet.

Generating Define-XML using BCs and SDTM dataset specializations has been demonstrated by CDISC at previous Interchange conferences using a SAS application. A new Python application may also be developed to demonstrate Define-XML generation for those without a SAS license.

Inputs	Tools	Outputs
• Biomedical Concepts • Dataset Specializations • Study design in ODM • Study design in JSON	• SAS, Python scripts • odmlib • CDISC Library • BC Browser • XMLMAPs • Define-XML Stylesheet	• Define-XML with VLM • HTML Define-XML

4. GENERATE ANNOTATED CASE REPORT FORMS AND ELECTRONIC DATA TRANSFERS

Using the ODM-based CRFs, eDTs, and the Define-XML generated in previous epics, add annotations to the ODM-based content to reflect how the content is mapped into SDTM. Once the annotations have been added to the ODM, generate an HTML visualization. In future iterations we will consider adding the origins to the Define-XML for traceability.

Note, that if the relationships between the CDASH and SDTM dataset specializations are not yet available in the CDISC Library, then raw data to SDTM mapping metadata may need to be developed.

Inputs	Tools	Outputs
• ODM CRFs & eDTs • Define-XML • Biomedical Concepts • Dataset Specializations	• SAS, Python scripts • odmlib • XMLMAPs • BC Browser • ODM XML Stylesheet • Define-XML Stylesheet • CDISC Library	• ODM with annotations • HTML aCRFs • HTML eDTs • Define-XML with Origins

5. GENERATE SHELL DATASETS

Using one or more of the Dataset-JSON conversion tools, read Define-XML and generate shell Dataset-JSON datasets that include the required metadata. The shell datasets can also be generated using the Dataset-JSON API. Test generating a conformance report that inspects the dataset metadata for standards conformance.

Inputs	Tools	Outputs
• Define-XML • Open rules for SDTM metadata	• Dataset-JSON SAS • datasetjson R Package • Python script • Dataset-JSON Viewers • Dataset-JSON API • CORE	• Dataset-JSON shell datasets in JSON and NDJSON • Conformance report in Excel

6. POPULATE CRFS AND ELECTRONIC DATA TRANSFERS WITH RAW DATA

Using the ODM CRFs and eDTs generated in previous epics, load the raw data into ODM. Given that we have not finalized the source of the raw data, a map that details how the raw datasets map into the ODM may be necessary. We anticipate developing a Python script to load the raw data into the ODM. In a later iteration, we plan to include FHIR data in the raw datasets.

Inputs	Tools	Outputs
• Raw datasets • FHIR data • ODM CRFs & eDTs • Raw data to CRF map	• Python script • odmlib	• ODM with data

7. LOAD OPERATIONAL DATA STORE

For 360i, the Operational Data Store (ODS) is an abstract repository/interface. The exact tools used for the 360i ODS have not yet been determined, and it may be nothing more than an ODM file repository in the early iterations. The main purpose of an abstract ODS is to highlight how sponsors can work directly with the raw data for data monitoring, exploratory analysis, and other analytical needs should they prefer to do so. An instantiated ODS data store may be a graph database, a relational database, a document database, datasets, any other data store, or some combination of data stores. The ODS can accommodate alternative raw data analytics use cases beyond the SDTM and ADaM generation targeted in 360i.

Inputs	Tools	Outputs
• ODM-based raw data • FHIR data • CDISC Library metadata	• Data store Software • CDISC Library	• Raw data query results

8. CREATE OAK.SDTM ALGORITHM SET

Using the oak.sdtm project, select the existing oak.sdtm algorithms needed to convert the LZZT raw data into SDTM. New algorithms may be created to meet the needs of the LZZT study if not all conversions are supported with the existing algorithms. The ODM annotated CRFs and the Define-XML will inform the selection of the appropriate algorithms.

In the future, CDISC plans to load oak.sdtm algorithms into the CDISC Library and include relationships between the CDASH and SDTM dataset specializations. Loading the algorithms into the Library will make SDTM transformations available for the pre-configured standards.

Inputs	Tools	Outputs
• oak.sdtm algorithms • ODM aCRFs • Define-XML	• oak.sdtm	• Algorithm set

9. GENERATE SDTM DATASETS

The oak.sdtm project has developed standard algorithms for transforming raw data into SDTM with an emphasis on using CDASH as the raw data source. To run the algorithms, the oak.sdtm project has developed an open-source R package. For this epic, we will run the R package with the oak.sdtm algorithm set to convert the raw data into SDTM datasets.

Prior to executing the algorithms that generate the SDTM datasets, the raw data must be converted from ODM or the ODS into datasets the oak.sdtm R package can consume. In future iterations, this process may be streamlined so that ODM-based raw data can be converted to SDTM in one step.

If the oak.sdtm R package does not yet support Dataset-JSON v1.1, the datasets generated will be converted into Dataset-JSON datasets.

Inputs	Tools	Outputs
• ODM / ODS data • Shell datasets • oak.sdtm algorithms • oak.sdtm formatted raw data* • Define-XML	• oak.sdtm • *Python/R to transform ODM into oak.sdtm raw datasets • Dataset-JSON Conversion Tools • Dataset-JSON Viewers	• Dataset-JSON datasets

10. CREATE OPEN RULE SET

Prior to generating a conformance report, we will create a set of open rules for checking the converted SDTM domains for the LZZT study. The first step is to select the open rules published in the CDISC Library for the

generated SDTM datasets. Next, any additional conformance or data quality rules will be authored using the Rule Editor.

Inputs	Tools	Outputs
• CDISC Open Rules • Define-XML	• CORE • Rule Editor • CDISC Library	• LZTZ CDISC Open Rule set

11. GENERATE CONFORMANCE REPORT

The final epic in phase 1 is to run the CORE software using the open rule set to generate the conformance report. The initial iterations will generate an Excel-based report.

Inputs	Tools	Outputs
• LZTZ CDISC Open Rule Set • Dataset-JSON datasets • Define-XML	• CORE	• Conformance report in Excel

ROADMAP INFRASTRUCTURE NEEDS

The 360i Enable and Automate work requires some infrastructure to support the development and execution of the automated research data pipeline. For development and collaboration, GitHub provides the needed infrastructure. GitHub provides the means to collaborate on software and standards development activities as well as providing mechanisms for discussions, documentation, and planning activities.

Test data represents an important 360i deliverable that has historically been difficult to provide. Getting the license to use existing anonymized study data has been challenging. The team believes that the availability of a synthetic raw data generator will be an important part of the 360i infrastructure.

There are certain tools that will benefit from a shared, cloud-based configuration, such as the Open Study Builder or the Study Definition Workbench. Ensuring that these systems have the necessary infrastructure for successful use will be important to timely completion of the 360i Technical Roadmap.

We may opt to include APIs for delivery of certain services or content. The CDISC Library is available, and we expect to use the API to access CDISC standards metadata. Other applications that deliver services via an API may need to be hosted, such as the Dataset-JSON API.

Finally, in a volunteer led development effort that incorporates different solutions into one pipeline, the team will need to ensure adequate execution environments are available. For example, the 360i pipeline will definitely need Python and R environments to run. The pipeline will likely also include SAS code that will require a licensed environment to run. Ideally, 360i developers would have the ability to run the research data pipeline locally on their laptop as well as in the cloud.

The infrastructure needs will evolve as the project proceeds, and CDISC will work to keep the infrastructure requirements to a minimum. Ideally, when we deliver an executable study package as an outcome of 360i, individuals will be able to download the package and run it on their laptops.

RULES OF THE ROAD

The Rules of the Road are a set of principles and practices that guide the implementation of the 360i Technical Roadmap. The table below provides a succinct list of the key principles, and the following sections expand on each principle.

Principle	Description
Implementation Focus	Implementation focus is a new way of working that transcends 360i.
Empirical, Iterative Process	We'll implement 360i in small increments and learn as we go.
Make it Work	Make it work, make it right, make it fast.
More Than One Way To Do It	TMTOWTDI, alternative software tools are welcome.
CI/CD	Make it testable. Continuous Integration / Continuous Deployment.
Just Do It	Pull Requests > Suggestions. What gets worked on gets done.
Build Publicly & Visibly	Benefit from network effects and transparency. Collaborate on GitHub.

IMPLEMENTATION FOCUS

With 360i, CDISC has officially shifted its focus to standards implementation. Of course, standards development remains a critical component of what we do, but CDISC will shift the way that standards are developed to include testing them to ensure they support end-to-end automation using existing open-source applications. This will help those organizations implementing the standards realize the full benefits of the standards.

EMPIRICAL, ITERATIVE PROCESS

The 360i technical work will be implemented following an empirical and iterative process. We will work in iterations, gather feedback, and feed the learnings into the backlog. Given the innovative nature of the work, using an agile methodology is critical to our success. We will be learning and refining the work as we go. In past projects with heavy volunteer contributions, a form of kanban provided sufficient flexibility to allow participants with varying levels of committed time to contribute effectively.

MAKE IT WORK

Our initial iterations will focus on making the standards-driven automated research data pipeline work. That means there will be room for improvement and refinement in future iterations. We anticipate that the initial demonstrations will include some workarounds and hacks that were needed to make the pipeline work. Using an iterative approach, we'll take the learnings from our early iterations and feed them back into the backlog to improve the work as we go. Once we get it working and it passes all the tests, then we'll focus on making it right and good. Once we have it right, then we can focus on performance improvements.

MORE THAN ONE WAY TO DO IT

For those former Perl developers, you may be familiar with the tagline, "There's More Than One Way To Do It (TMTOWTDI)." We're adopting that principle for 360i. We fully anticipate that there will be more than one way to implement standards-driven end-to-end automation. Once we get the initial "make it work" approach completed, there will be opportunities to try alternative software tools. This applies to open-source developers and commercial vendors that would like to demonstrate that their solutions can process parts of the automated research data pipeline. Ideally, there will be more than one option for most tools used in 360i.

CONTINUOUS INTEGRATION / CONTINUOUS DEPLOYMENT

Continuous Integration / Continuous Deployment (CI/CD) is a software development practice that automates the process of building, testing, and releasing code. As part of 360i, CDISC will adopt this to develop and release the study packages to include standards metadata, test data, and software. As 360i software tools or standards metadata are modified, a test framework will be triggered to ensure the software continues to function correctly. Testing standards using open-source software tools and test data will improve the standards and make them easier to implement and maintain.

JUST DO IT

As with most volunteer development work, it's better to draft a solution than to make a suggestion or comment. There is no shortage of good ideas, but there's a real shortage of developers to implement them. Taking the Just Do It approach makes it easier for the 360i team to review and accept your work. If you submit a Pull Request for a proposed change to be integrated into the project, it can be reviewed, tested, merged, and deployed all using GitHub.

In a volunteer-driven project like 360i, what gets worked on gets done. This means that the volunteers that put the work in and deliver solutions tend to have the most influence over the direction of the project.

BUILD PUBLICLY AND VISIBLY

Build publicly and visibly means that 360i will follow an open-source approach to development, where updates and improvements are incrementally added to the standards and software in such a way that other developers can review and accept the changes. Furthermore, anyone on the project can review the changes providing process transparency. This transparency and the ability for others to use and contribute to the work creates network effects that increase the value of the work for all. Collaborative standards and software development hold the potential to drive innovation faster and at lower costs.

CDISC also seeks to make the 360i work visible for non-technical participants by scheduling Connectathons, Webinars, and Showcases. There will be many opportunities for non-technical participants to learn about the progress made within 360i and to see the automation in action.

CONCLUSION

This paper summarizes some of the key elements of an initial CDISC 360i Phase 1 Technical Roadmap. The roadmap will evolve once work on the program is underway. As Phase 1 proceeds, we will flesh out a more concrete and detailed roadmap for the Phase 2 deliverables.

For those interested in contributing to 360i, please sign up to volunteer at <https://www.cdisc.org/volunteer/form> and select 360i development team.

RECOMMENDED READING/LISTENING

1. [360i Kickoff Webinar](#)
2. [360i Program Wiki](#)

CONTACT INFORMATION

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

Sam Hume

CDISC

shume@cdisc.org

<https://www.linkedin.com/in/sam-hume-dsc>

<https://swhume.github.io/>

Brand and product names are trademarks of their respective companies.