CDISC Library is the single, trusted, authoritative source of CDISC standards metadata. Through its Data Standards Browser (DSB) and Application Programming Interface (API), users can browse and retrieve normative metadata such as CDASH models, SDTM domain specifications, ADaM data structures, QRS instruments, CDISC Controlled Terminology, etc. Increasingly, CDISC is focusing on enabling end-to-end process automation and simplifying data transformation throughout the data lifecycle.

CDISC makes our data standards available in PDF format. The Data Science team embarked on a proof-of-concept project to digitally transform a few Implementation Guides and make all applicable contents machine-readable, including, but not limited to, informative text, data examples, biomedical concepts, and conformance rules.

The objectives of this initiative are 1) to support CDISC’s Implementation Guides (IG) and User Guides (UG). A Therapeutic Area User Guide (TAUG) is typically comprised of informative content. We carefully review them and make updates to the data model accordingly; 2) to allow access by member organizations and to expand access to all who implement data standards; 3) to build on existing tooling; 4) to ensure CDISC Library serves as the reference for standard implementers to reinforce its role as the central repository for CDISC standards.

With those objectives, a few key results have been identified: 1) the existing metadata pipeline that sources normative content from CDISC Confluence is updated to acquire all remaining content from IG and UG; 2) it supports new standards and it can be retrospectively applied to IGs and UGs already in CDISC Library; 3) content is searchable, discoverable, and bookmarkable; 4) it compliments how normative content is displayed in the DSB; 5) where possible, transform all contents to be machine-readable; 6) ensure content in CDISC Library is accessible through HATEOAS-driven RESTful API.

To present informative content in a digestible manner, the CDISC Library makes use of an ETL pipeline to ingest and transform authored content into a generic structure. This normalized structure stores wiki content in both HTML and a stripped-down Markdown format[2] to support the intended rendering of the content as well as the search functionality inherent in all CDISC Library content. Currently, standards are authored in a Confluence wiki with no differentiation between pages meant for publication and supporting documentation for standard structures. The ETL pipeline processes wiki page labels that are supplied by the standards development team to tag stored documents. These additional metadata tags are used to filter which documents are returned by the API and subsequently displayed in the DSB.

API design was also critical in ensuring the informative content was extensible. Most informative content is well structured into specific sections associated with specific IG structures (domains, datasets, data structures); however, there is some informative content that does not fit into these categories. Additionally, new IG versions often require new sections or documents. Storing documents in an Azure Cosmos DB, allows us to quickly query documents and filter to support any number of document types and structures. The ETL pipeline processes wiki page labels that are supplied by the standards development team to tag stored documents. These additional metadata tags are used to filter which documents are returned by the API and subsequently displayed in the DSB.
The CDISC Library is composed of two components: the DSB and the API. Content displayed on the DSB is all driven by the API. All employees of CDISC Member organizations can create an account to access CDISC Library. Account set-up requires a valid organizational email address to where a verification code can be sent. New users will receive a separate email about obtaining API keys that can be used to programmatically pull metadata from CDISC Library.

Note that IG and UG contents are loaded into CDISC Library as soon as they are reviewed and approved by standards development teams. Release Notes are updated accordingly.

Example IG

The team loaded SENDIG-DART v1.2 to demonstrate the development work in this beta functionality, in conjunction with its public review. Note that this IG is in draft. The content tagging is a work-in-progress by the DART team.

DSB

Here is a short video demonstrating how to locate some of the informative content loaded for SENDIG-DART v1.2. Currently figures & graphics appear as a placeholder.

![Quick video of DSB content](quick-clib-dsb-n...tive-content.mp4)

API

GET /mdr/documents/sendig/dart-1-2/sections will provide a list of supported document sections.

GET /mdr/documents/sendig/dart-1-2/{dataset}/{section} will output the informative content corresponds to a given section for a given dataset. For example, /mdr/documents/sendig/dart-1-2/TP/assumptions output the assumptions for the TP dataset. The output matches the wiki source in TP assumptions of the SENDIG-DART development wiki space.

Discoveries

Database design is essential in determining whether a solution supports a use case and is stable enough to support new use cases. Additionally, while out of scope, using Azure Cosmos DB as a storage environment for documents could potentially support standards authoring in the future, given the right toolset. However, using wiki page labels, CDISC Library can be expanded while standards are authored in wiki documents.

Content validation is becoming essential to verify compatibility. This feature will serve as a detector for non-conformant content so that Metadata Curators can collaborate with Content Owners early in the standards development process.

Conclusions

Informative content in CDISC IGs and UGs provide needed structure and predictability but flexibility is of high importance. For example, the ability to include new content types. These require careful comprehensive review to optimize storage, efficient content retrieval, and, most importantly, usability. Complementary process updates in standards authoring, such as content label metadata, will improve this transformative feature.

Agile has been our development approach, especially for large features like this project. Additional refinements are being developed iteratively. Details will be presented at the CDISC US Interchange in October 2022.

References


CommonMark. “What Is Markdown?” Specifications v0.30, 19 June 2021, spec.commonmark.org/0.30/#what-is-markdown-.