PhUSE 2017

Paper DH05

How a Metadata Repository can Facilitate Automated Data Collection Processes

Judith Goud, Nurocor, Bennekom, The Netherlands

ABSTRACT

The processes for eCRF and EDC database creation typically take weeks-to-months to complete. Traditionally these study artifacts have been managed or specified in spreadsheets and other less structured formats by manually reviewing and interpreting input artifacts, such as standards and the study protocol. However, these processes can be significantly improved by programmatically generating these artifacts from a metadata management system starting with the standards, and elaborating them with study-specific details.

As an example, the study eCRF and database metadata can be initially generated from standards, reusing any mappings and transformation logic to the submission datasets. Each of these study artifacts can then be extended to meet study-specific needs. From this metadata, study outputs like the blank and annotated eCRFs and the EDC system specification can be dynamically generated. Not only does this improve overall business efficiency, but it can also increase the quality of submissions and reduce resource needs and cycle times.

INTRODUCTION

A survey to investigate where the industry stands with regards to standards and metadata management and the main drivers for metadata management tool implementation in the Biotechnology & Pharma industry was undertaken with 78 CDISC conference attendees from 55 organizations[1].

Overall results indicated Data Quality was mentioned most often as a driver for implementing a metadata management solution (see figure 1), followed by Regulatory Compliance and Business Process Efficiency as the top 3 drivers.

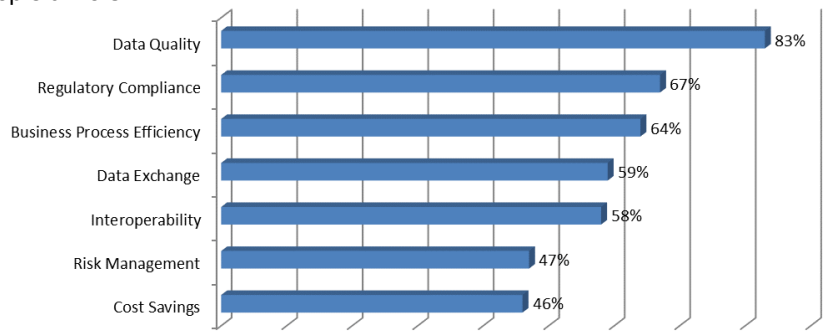


Figure 1. Drivers for Metadata Management (total percentage of respondents indicating each)

Although Data Quality was mentioned most often (84%) as one of the drivers, Regulatory Compliance was the top individual driver reported as the highest priority by 31% of respondents (figure 2).

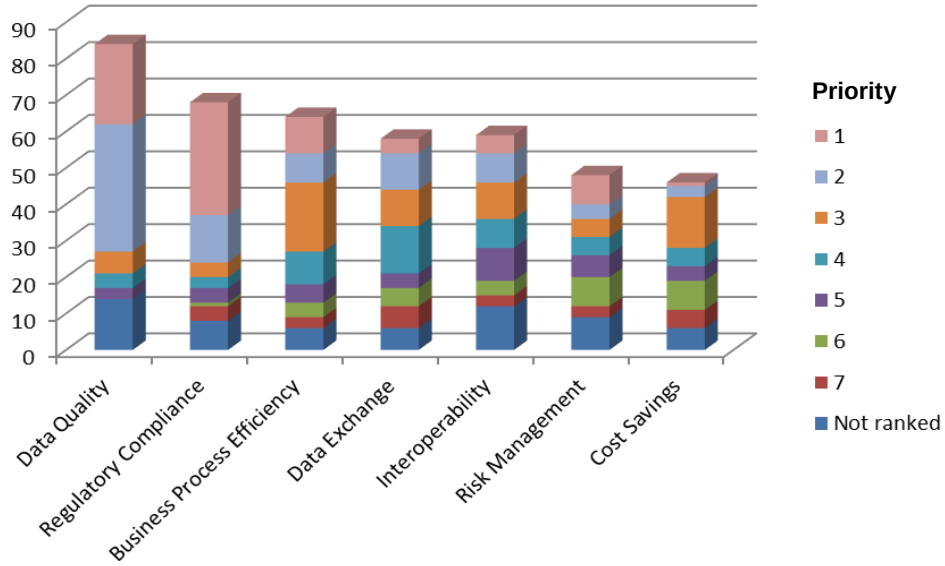


Figure 2. Drivers for Metadata Management Tool Implementation by priority

This picture corresponds with the fact that the Biotechnology & Pharma industry organizations implement technologies to help in the study conduct as well as repositories to manage common data definitions to facilitate Data Quality and data standards adherence to ensure Regulatory Compliance. Business Process Efficiency is the next most important driver identified in the survey

BUSINESS PROCESS CHALLENGES IN THE BIOTECHNOLOGY & PHARMA INDUSTRY

The traditional clinical development lifecycle (CDLC) used by most Biotechnology & Pharma companies follows a fairly linear approach that has worked well in a controlled environment with paper driven processes. It is clearly defined and requires each step to be carried out before moving on to the next. However, it is very slow and rigid, and therefore does not adapt easily to change or continued discovery and innovation.

Although the several new technologies have been introduced with the potential to improve individual processes if used properly, the overall clinical development lifecycle has remained unchanged, while the amount and complexity of data being collected and analyzed has significantly increased.

In response to these data concerns, the Biotechnology& Pharma industry recognized the need for clinical data standards and established the Clinical Data Interchange Standards Consortium (CDISC) for this purpose. CDISC is a standards development organization (SDO) that partners with industry and regulators to develop data standards that increase patient safety, data quality, and more efficient regulatory reviews.

According to a joint analysis between CDISC and Gartner[2], there are many clinical study activities that can be streamlined with data standards and the earlier standards are implemented within the CDLC, the greater the potential benefit. Figure 3 indicates the study activities where CDISC standards can be applied.

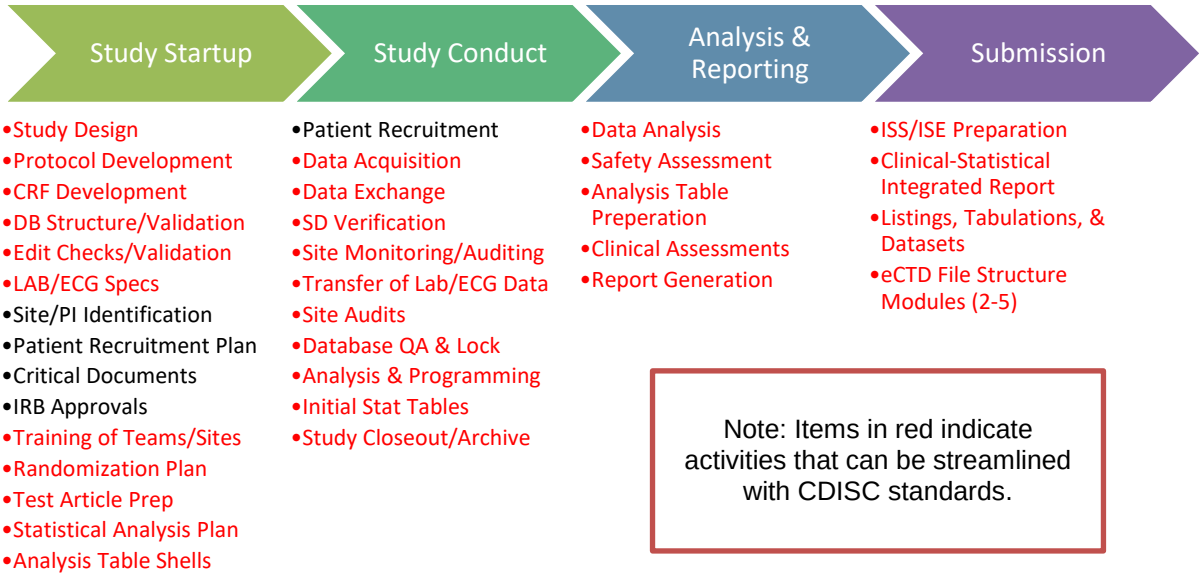


Figure 3. Clinical Study Activities

In the traditional Clinical Data Life Cycle, the creation of clinical artifacts is generally slow, error prone, resource intensive and costly. Organizations develop each artifact from scratch through manual review and interpretation of standards and artifacts created earlier in the lifecycle and/or from other studies. For example (see Figure 4), to create an electronic data capture (EDC) system to collect clinical data from the sites, a data manager and/or case report form (CRF) designer will generally start with the data collection standard and/or eCRF specifications from previous studies that are similar.

They will then read the protocol and manually determine if there are differences in the study definitions for clinical assessments versus those being copied from standard or prior eCRF specifications and make any appropriate changes. They will also use the time and events schedule from the protocol to determine which assessments should be collected at which visits/time points to put the eCRFs in the appropriate order and frequency. All of this is typically documented in an eCRF specification spreadsheet.

Once the spreadsheet is complete with a full database specification, the data manager will create eCRF visualizations either manually or electronically that look similar to what the eCRFs will look like for users through the EDC system browser-based interface. The visualizations are then used for team review and approval. Either during or directly following the eCRF creation process, the data manager will draft edit checks, which are validations for the eCRFs (e.g., enrollment date must not be prior to birth date). These will also go through a review and approval process.

Finally, once the eCRF and edit checks specifications are approved, they are sent to a database developer, who uses the specifications to program the EDC system. Once system development is complete, it is normally tested and approved before it is released along with any other supplemental documentation, such as eCRF completion instructions and training materials.

This entire process can take up to six months from final protocol through database release. This process incorporates the use of computerized systems, rather than paper-based processes, as well as industry data standards. While standards and technologies may have improved the data collection process, they have further complicated the system development process because they were not designed to optimize it.

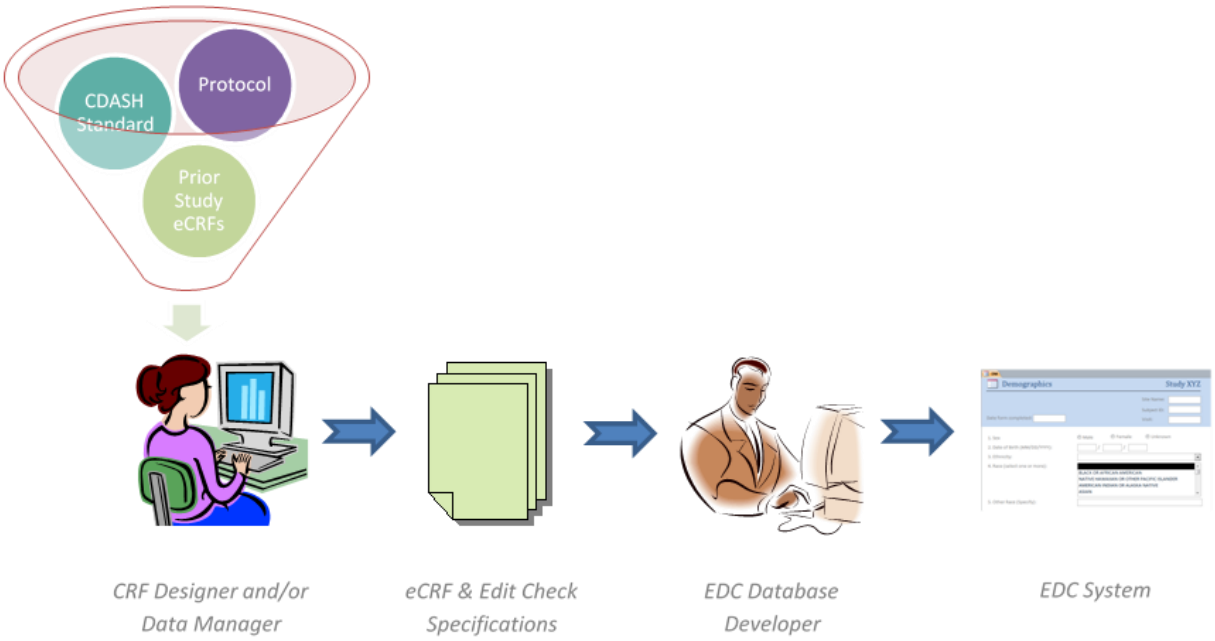


Figure 4. Traditional Approach to ECD System Development

The reason for a lack of automated reuse of information is that today's systems manage only the content of artifacts, and often in an unstructured manner with little or no coupling across many different systems and network locations. Using the previous example, protocols and CDASH standards are normally distributed in PDF and MS Word document formats, while eCRF and edit check specifications are generally MS Excel files, all of which are stored in many different folders across the network. Each artifact is also typically version controlled based on internal functional processes, making it difficult at times to know which version is the latest official version of an artifact.

This becomes even more difficult as the study conduct begins because issues might be discovered that cause amendments to the protocol months or even years later. Because studies are so long, by that point members of the team, processes, and document management systems might have changed, so knowing which file location contains the appropriate artifacts to use in determining how to apply the changes to the eCRFs can be quite difficult.
The data manager must fully understand the current eCRFs, know which version the EDC system is based on, and then review the protocol amendment to decipher what impact it has on the eCRF and edit check specifications. They then must make the appropriate changes to the specifications, obtain reviewer input and approval, and then send the specifications to the EDC developer to make the appropriate changes to the database. The EDC developer then has the same process to go through and issues to contend with as the data manager did in the previous steps. Adding complexity to an already inefficient process, often the existing version of the eCRFs are already being used in the study, so both versions must be implemented and maintained simultaneously.

This makes manual version control more difficult and future changes even more complex because the manual impact analysis of changes must be done to both versions of the eCRFs, edit checks, and EDC system. In the end, this is a resource intensive, manual effort since there is so little automated reuse of information or flexibility in the overall clinical development lifecycle.

IMPROVED BUSINESS PROCESSES

To overcome the challenges with the current clinical development lifecycle, standards and metadata for artifacts created and used throughout this lifecycle should be managed in a centralized repository in order to facilitate their automated reuse. Although the process flow is similar, there is a big difference in how information is managed and used.

As illustrated in Figure 5, rather than the data manager or CRF designer manually interpreting the study protocol, data collection standards, and prior study eCRFs, they have all those definitions already in a centralized repository and can reuse them without having to retype/create them manually. They use the existing content to create a new set of metadata to represent the study specific eCRF and edit check specifications, re-using all the

existing metadata content that applies, and only adding items that are missing. In this case, the complete visit schedule and all assessments per visit will already have been defined in the protocol, so there is no need to recreate them for the eCRFs. Further, the definitions for each assessment will also already exist from either global industry-defined research concepts or by the protocol itself. Each of the industry-defined research concept semantics will also include a data structure. Therefore, all the data manager has to do is put all of these existing components together to build the eCRF, and then add any missing components such as a data structure for non-standard assessments. The same applies for defining the edit checks, as many can be standardized.

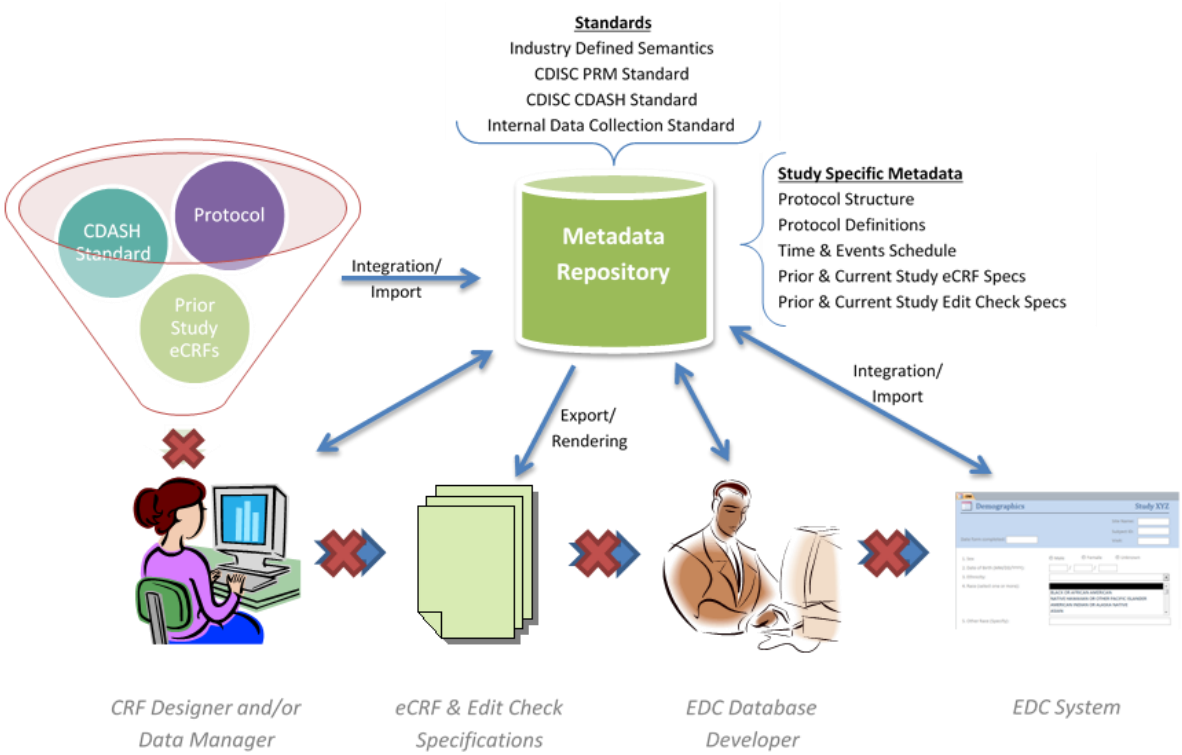


Figure 5. New Approach to EDC System Development

Once the metadata for the study specifications is complete, all of the new content will be reviewed in a metadata governance workflow. Upon approval, the eCRFs and edit checks can be rendered into visualizations for team review and approval simultaneously, rather than sequentially, resulting in a shorter overall cycle time. After approval, the EDC vendor can use the specifications to automatically generate the database instead of creating it manually. This can be done through integration of the repository to the EDC system or export/import of the specifications in a standard format such as the CDISC Operational Data Model (ODM).

INTEGRATION WITH EDC

Further potential of the metadata repository with regards to business process efficiency is by automating setup of Data Collection Processes. For this, the system is to be integrated with clinical applications used for study related activities such as Electronic Data Capture systems.

The below figure pictures a possible integration path between a Metadata Repository and EDC system.

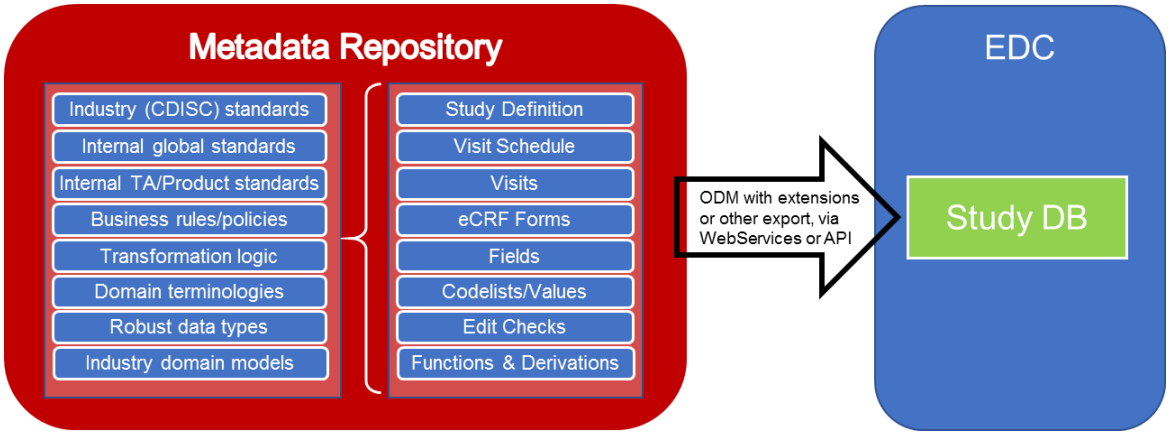


Figure 6. Metadata Repository Integration with Electronic Data Capture system

The standards used to generate a set of study specific data collection metadata need to be enriched with information needed by the EDC system around visit schedules, visits, eCRF forms, fields and codelist terms. This is accomplished by grouping metadata into an appropriate hierarchy and organizing the metadata in such a way that core metadata can be reused while capturing layout and other EDC attributes in relationship to these groups.

Figure 7 shows an example of a hierarchical structure for a Casebook including reusable visits, forms, fields, codelists and codelist values.

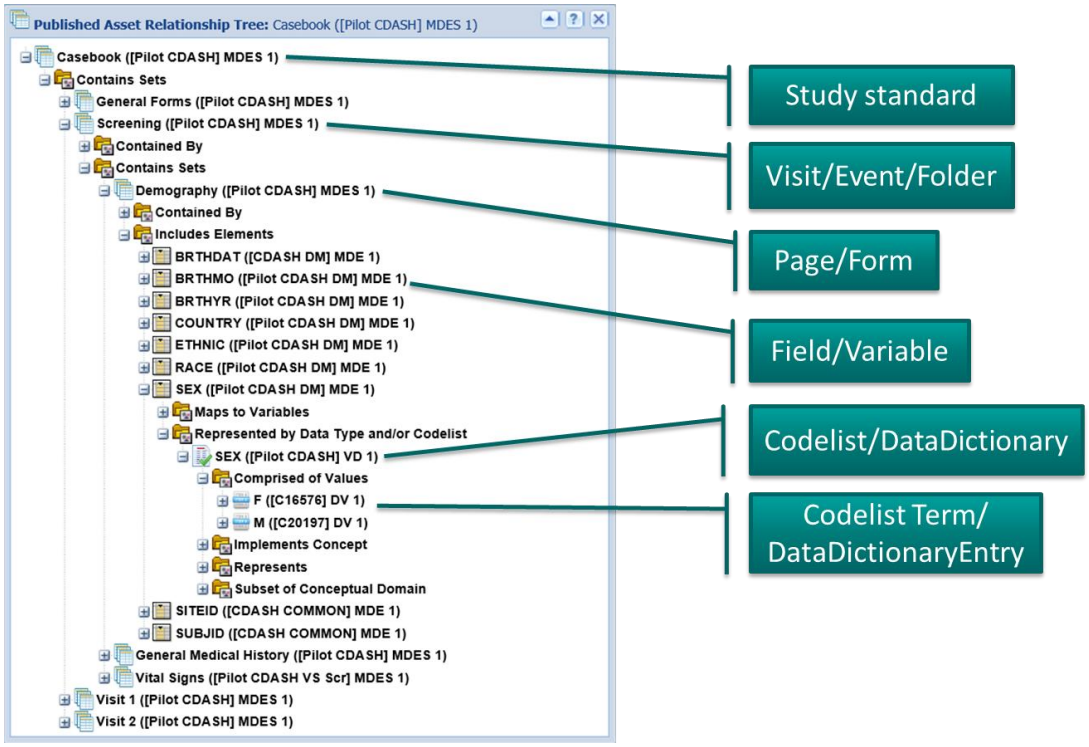


Figure 7. Structural Metadata for a Study Casebook

The metadata fields are grouped under pages/forms keeping the information on the layout like Control Type and ordering of the fields on the form on the relationship (see figure 8). Further layout information like CRF label and instructions, and EDC system information regarding field verification and edit checks, etc. should also be captured to enable complete setup of a study database.

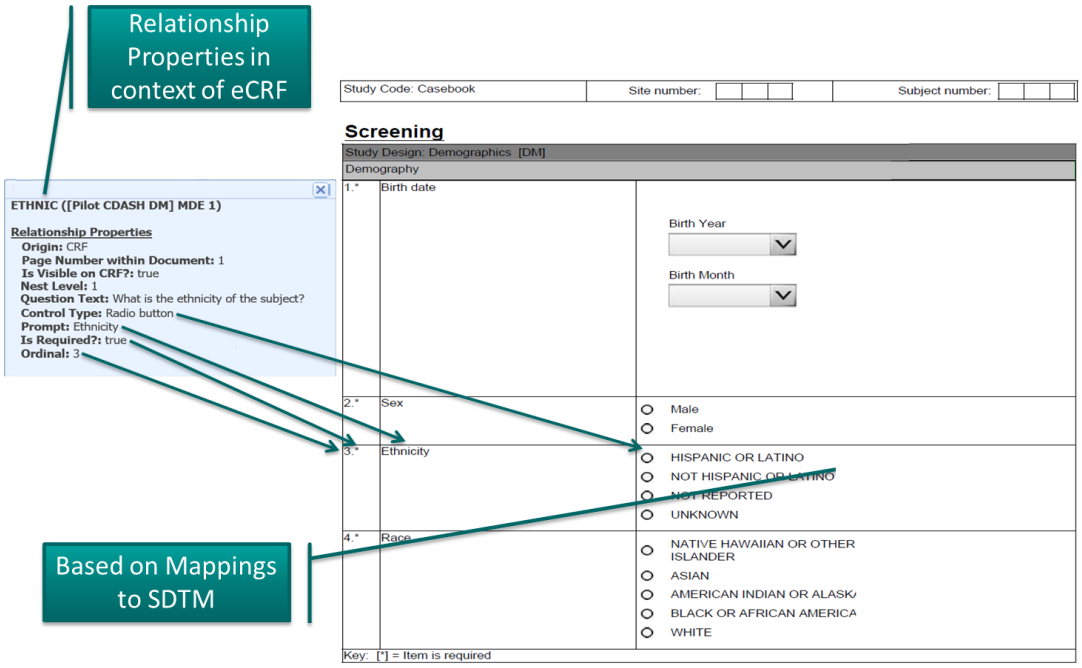


Figure 8. Managing (e)CRF Information in the Metadata Repository

All information managed in the Metadata Repository then need to be transferred to the EDC vendor. Each vendor will have their own input format and method of access that will need to be matched by the Metadata Repository. ODM is one of the CDISC data exchange standards which can be enriched with vendor-specific extensions to convey all necessary information to create a study database.
Figure 9 shows an example ODM file based on the Casebook defined in figures 7 and 8. It indicates the structure of the casebook as well as the ordering of the fields on the eCRF. Other details like label and value lists would reside on the ItemDef and CodeListDef elements.

```
<?xml version="1.0" encoding="UTF-8"?>
- <ODM xsi:schemaLocation="http://www.cdisc.org/ns/odm/v1.3 ODM1-3-2.xsd" SourceSystemVersion="1.0 (7.1.5.108)"
  SourceSystem="SHARE" Originator="CDISC" ODMVersion="1.3.2" FileType="Snapshot"
  FileOID="Semantics.Manager.ODM_1.3.2_Casebook_1.0_1478897301772979722371" CreationDateTime="2017-01-
  10T09:37:59" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xlink="http://www.w3.org/1999/xlink" xmlns="http://www.cdisc.org/ns/odm/v1.3">
    - <Study OID="Study.Casebook.[Pilot_CDASH]_MDES_1">
        + <GlobalVariables>
        - <MetaDataVersion OID="MDV.Casebook.2017-01-10_09:37:59" Name="Casebook MDV" Description="Casebook
          [Pilot CDASH] MDES 1">
            + <ItemGroupDef OID="IG.MDES.Screening.[Pilot_CDASH]_MDES_1" Name="Screening" Repeating="Yes"
              Domain="Screening">
            - <ItemGroupDef OID="IG.MDES.Demography.[Pilot_CDASH]_MDES_1" Name="DM" Repeating="No"
              Domain="Demography">
                - <Description>
                    <TranslatedText xml:lang="en">Demographics</TranslatedText>
                  </Description>
                  <ItemRef OrderNumber="2" Mandatory="No" ItemOID="IT.MDE.SEX.[Pilot_CDASH_DM]_MDE_1"/>
                  <ItemRef OrderNumber="4" Mandatory="No" ItemOID="IT.MDE.RACE.[Pilot_CDASH_DM]_MDE_1"/>
                  <ItemRef OrderNumber="1.2" Mandatory="No" ItemOID="IT.MDE.BRTHMO.[Pilot_CDASH_DM]_MDE_1"/>
                  <ItemRef OrderNumber="1.1" Mandatory="No" ItemOID="IT.MDE.BRTHYR.[Pilot_CDASH_DM]_MDE_1"/>
                  <ItemRef OrderNumber="3" Mandatory="No" ItemOID="IT.MDE.ETHNIC.[Pilot_CDASH_DM]_MDE_1"/>
                  <ItemRef OrderNumber="1" Mandatory="No" ItemOID="IT.MDE.BRTHDAT.[CDASH_DM]_MDE_1"/>
                  <ItemRef OrderNumber="2" Mandatory="No" ItemOID="IT.MDE.SITEID.[CDASH_COMMON]_MDE_1"/>
                  <ItemRef OrderNumber="3" Mandatory="No" ItemOID="IT.MDE.SUBJID.[CDASH_COMMON]_MDE_1"/>
                  <ItemRef OrderNumber="1" Mandatory="No" ItemOID="IT.MDE.COUNTRY.[Pilot_CDASH_DM]_MDE_1"/>
              </ItemGroupDef>
            - <ItemGroupDef OID="IG.MDES.General_Medical_History.[Pilot_CDASH]_MDES_1" Name="MH" Repeating="Yes"
              Domain="General Medical History">
```

Figure 9. Casebook Metadata in ODM Format

Dependent on vendor capabilities the (ODM) file can be pushed from the Metadata Repository via WebServices or APIs to EDC vendor recipients to create the study database ready for use in the clinical study.

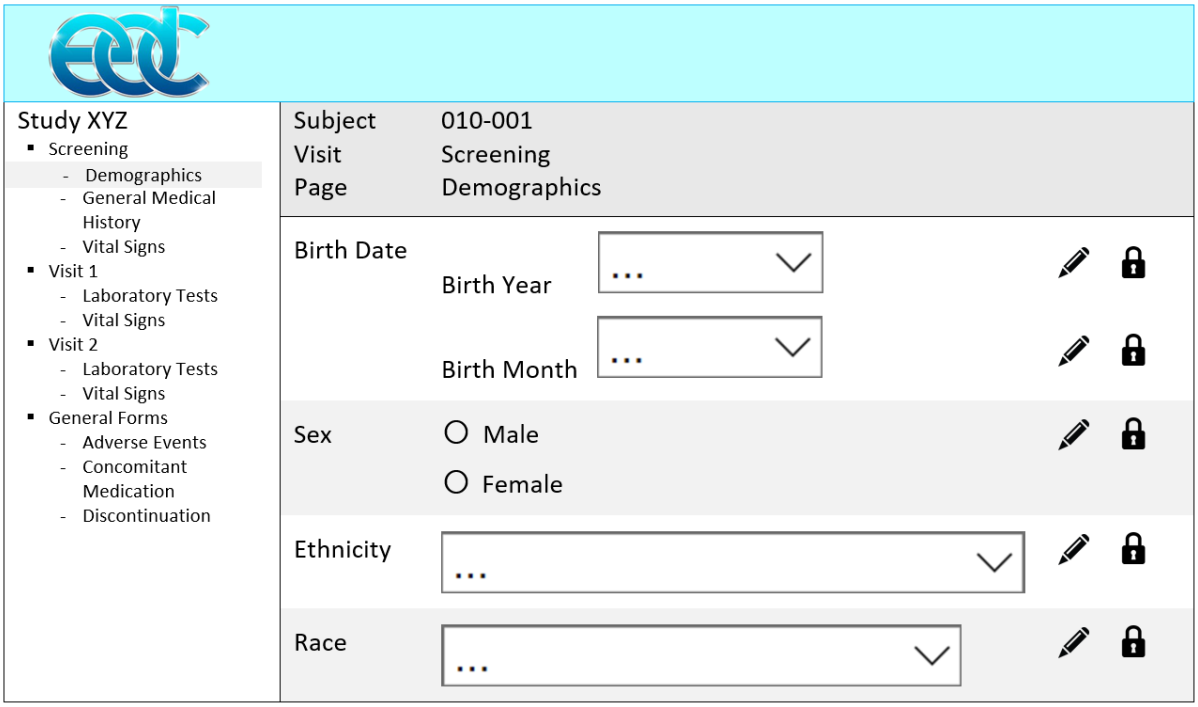
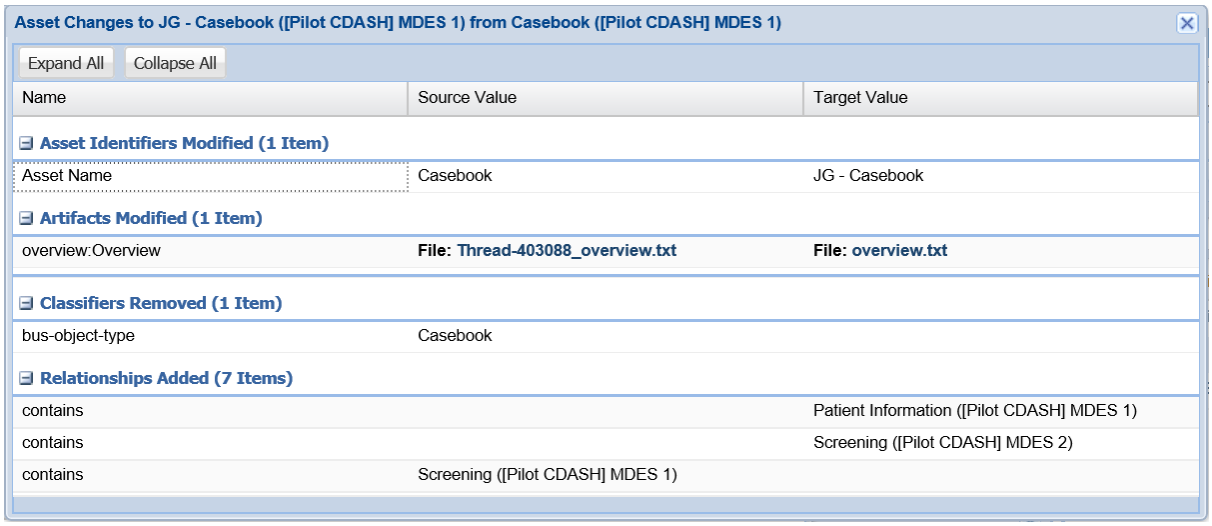


Figure 10. eCRF Metadata in vendor-neutral Format

Any mid-study changes to the EDC database can likewise be implemented in the study casebook within the Metadata Repository and a new Metadata Version can be transmitted to the EDC vendor without interruption of clinical practices.

To ensure standards adherence the metadata from the EDC system can imported back into the Metadata Repository and compared against the original specifications to detect any (inadvertent) changes applied to the EDC database during the study (see figure 11). These changes can then be documented for the submission file.



Asset Changes to JG - Casebook ([Pilot CDASH] MDES 1) from Casebook ([Pilot CDASH] MDES 1)

Expand All | Collapse All

Name	Source Value	Target Value
Asset Identifiers Modified (1 Item)		
Asset Name	Casebook	JG - Casebook
Artifacts Modified (1 Item)		
overview:Overview	**File:** Thread-403088_overview.txt	**File:** overview.txt
Classifiers Removed (1 Item)		
bus-object-type	Casebook	
Relationships Added (7 Items)		
contains		Patient Information ([Pilot CDASH] MDES 1)
contains		Screening ([Pilot CDASH] MDES 2)
contains	Screening ([Pilot CDASH] MDES 1)	

Figure 11. Study Metadata Comparison Report

This type of integration/processing facilitates the following capabilities and benefits:

- **Capabilities**
  - Management of standards across the clinical development lifecycle
  - Mappings to/from upstream and downstream systems and processes
  - Reports for rendering blank and annotated eCRFs prior to database implementation
  - Extraction of eCRF specifications in ODM format
  - APIs that push the specifications to create a draft study in the EDC system
- **Benefits**
  - Greater flexibility with automated impact analysis and inheritance across the clinical development lifecycle, as well as from standards down to study level eCRFs
  - Better utilization of personnel resources with less need for human intervention and interpretation of upstream artifacts (e.g., protocol) and standards
  - Higher quality, more consistent eCRF specifications/systems by using standards and implementing metadata-driven processes
  - More efficient EDC builds that save time and money by automating the selection of CRF elements based on standards and having eCRF visualizations that simplify the review process

METADATA MANAGEMENT SYSTEM CHARACTERISTICS

Currently the majority of Biotechnology & Pharma companies manage their metadata using spreadsheets, however this tool does not support things like change control, versioning, business rules, impact analysis or reuse of elements, which have to be done manually through manual review, creating copies (and keeping track of them) and team consultation. This introduces human error as well as extra work for impact analysis and tracking where data definition changes need to be implemented.
A metadata repository is better suited for these activities and can serve as the single source or truth within an organization. The repository should hold all information about data element structure and meaning, and where a data element is/has been used. It should also enable review of all references and transformations the data element was subject to and link the data element to each specific use case for impact analysis.

A well-designed metadata repository will enable data transparency and business process improvement by:

- Creating, maintaining, governing and using standards consistently
- Maximum reuse of existing artifacts and data assets
- Understanding the impact of changes on other people, processes and systems
- Knowing data exchanged with others is being used or interpreted correctly

CONCLUSION

Regulatory compliance, data quality and business process efficiency are the top 3 drivers for metadata management solution implementation in the Biotechnology & Pharma industry. This paper shows the processes for eCRF and EDC database creation can be significantly improved by programmatically generating these artifacts from a metadata management system starting with the standards, and elaborating them with study-specific details including EDC system required information.
Reusing any mappings and transformation logic to the submission datasets, also study outputs like the annotated eCRFs and the SDTM specification can be dynamically generated[3].

In addition to facilitating business process efficiency, metadata management enables transparency from source to submission as required by Regulatory Authorities[4,5], thus facilitating compliance. Metadata management based on the clinical data standards provides additional value, as several regulators mandate submission of clinical data according to standards developed by CDISC.
Stakeholders can extend existing content rather than having to come up with completely new semantics. A metadata management tool should support the management of semantics and offer a flexible configuration that can manage both the industry-wide clinical data standards as well as study specific metadata.

Data standards-based metadata can also be used to structure data entry and streamline data flows, minimizing Data Quality loss due to entry errors or incorrect data processing. Having a single source of truth for definitions as well as references/transformations needed for analysis and reporting (including an (automated) governance workflow) greatly improves Data Quality by providing a common basis for interpretation and detailed information for data exchange.

REFERENCES

1. Goud J, Metadata Management: The Big Drivers for a Solution, white paper, https://www.roguewave.com/getattachment/488495ac-735e-40a9-b244-51d727595821/Pharma-Metadata-Management-Drivers
2. Rozwell, C., Kush, R.D., Helton, E. Newby, F. and Mason, T. 2009. Business Case for CDISC Standards: Summary, http://www.cdisc.org/stuff/contentmgr/files/0/ff2953ea8dbc8e81080f0e44ba6714c7/misc/businesscasesummarywebmar09.pdf
3. Goud J & Shetty P, How a Metadata Repository enables dynamism and automation in SDTM-like dataset generation, 30 September 2016, http://www.lexjansen.com/phuse/2016/dh/DH05.pdf
4. FDA, Study Data Technical Conformance Guide v.3.3, 23 March 2017, https://www.fda.gov/downloads/forindustry/datastandards/studydatastandards/ucm384744.pdf
5. PMDA, Technical Conformance Guide on Electronic Study Data Submissions, 27 April 2015, https://www.pmda.go.jp/files/000206449.pdf

CONTACT INFORMATION

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

Judith Goud
Nurocor, Inc.
106 East 6th Street
Suite 900
Austin, Texas 78701
Judith.Goud@nurocor.com
www.nurocor.com