Paper DS09

Business-based value in an MDR

Karl Konrad, Merck, Durham, NC, USA
Manori Turmel, Merck, Upper Gwynedd, PA, USA
Brooke Hinkson, Boston, MA, USA

ABSTRACT
Implementation of a Metadata Repository (MDR) is an effective solution to manage the clinical data lifecycle. An MDR must deliver business value and embed efficiencies within an organization that accesses the information. Merck’s MDR capabilities provide robust information management processes to enable informed decision making while defining, implementing, curating and sharing study-level and global-level data standards. A people, process, technology enablement approach facilitates the following:

- Streamlining business customer access to effectively versioned, controlled information
- Supporting data flow visualizations and impact realization
- Harmonization of organizational collaboration and enabling overall clinical lifecycle awareness

We will explore how using a linked graph database further allows for forward-looking capabilities, such as: understanding the interconnected referential aspects of data (e.g., external libraries), linking unstructured data, and streamlining standards up-versioning mechanisms.

INTRODUCTION
Merck has adopted a business value approach to implementing a Metadata Repository. This approach must address the change management aspects of introducing a new, transformative capability as well as balancing the impact to existing business processes, roles and responsibilities and current infrastructure. By using an agile, iterative approach, we are increasingly realizing new capabilities with frequent releases that enable new business processes and value. This paper will explore the planned business value; some of the implementation steps already completed, and describe the near-term and future capability to enable business success with a Metadata Repository.

PLANNED BUSINESS VALUE OF THE MERCK METADATA REPOSITORY

The Merck Clinical Metadata Repository (MDR) was built with the business in mind and to address existing business process and technology pain points. The MDR is seen as a primary source of information for many consumers. The value of the MDR extends beyond providing well-defined metadata, it can drive automation of business processes and provide visibility into and increase awareness of where and how information is used in order to enable informed decision making during information governance processes, such as impact analysis during a change request assessment. A user can log into the MDR, look up an item like AEAGE, and understand how that information is defined and used across the clinical information lifecycle (e.g. from protocol, collection, tabulation, and through submission deliverables) and enabling clinical technologies. By using a semantics-based approach, Merck will be able to dynamically extend the graph model to add new linkages to additional data types, information models, and systems and dependencies as our MDR scope and business drives shift and expand.

Merck recognized that the vision for an MDR as the center of information sharing is a longer term vision that must begin with a practical implementation. The initial focus of the MDR is to concentrate on the collected items and how those are expressed in SDTM. The larger vision to incorporate additional sources of data, as well as additional targets, will be achieved through iterative subsequent releases that expanded the business value the MDR will provide.

SERVING MANY CUSTOMERS

There are multiple customers who directly enter information into the MDR; there are even more customers who will use the definitions within the MDR to achieve and/or support their business processes. The cross-functional standards and application subject matter experts define the end to end data definitions. The Clinical Data Standards group enters and maintains the data collection and SDTM mapping definitions (form and item level definitions). The InForm programmers
utilize the back-end dataset details provided by the InForm data collection model. The Clinical Data Management community manages information stored in reporting datasets to help with site management and query resolution. The Data Management Workbench programmers create the back-end tables used by others. The Safety Reporting group uses the information stored in different reporting datasets to monitor patient safety. The Stats Programming group and SDTM Task Force use the information stored to understand how data flows from collected items to SDTM. Each of these customers has a related but unique set of desired views of the relevant information that supports their specific business processes. The collective customer priorities and expectations about MDR functionality and timing of releases must actively be managed. The business value anticipated from each customer was considered and prioritized into a set of releases.

**TECHNOLOGY**

The technology chosen to implement the MDR should follow the business drivers and align with the operational implementation defined by that business value. Within the industry, there are no obvious success stories to emulate, and the MDR tool landscape continues to evolve. When determining the technology to use to implement the MDR, Merck considered:

- Vision for pursuing an MDR for information management
- Strategy of using an approach with the desired flexibility and growth potential
- Implementation to realize business value quickly and build on success

Merck chose to utilize a technology based on a semantic-data model. The regulatory data target for clinical trials (CDISC) is steadily changing, the science behind some standards is continuing to evolve, and the specific implementation continues to improve. By using a semantics-based approach to the MDR, Merck intends to improve:

- Flexibility, inheritance of properties, relationships to support impact assessment
- Extension of existing information to future information
- Flexibility to attach additional information, such as data definitions from collaborations, specifically when two sponsor companies join forces to jointly submit a drug therapy combination to regulators
- Automation of business processes

When implementing an MDR, it can be tempting to reach too far, and try to solve all of the problems at once. At the onset of the Merck MDR program, we started with far-reaching goals, like having the flexibility to bring in additional standards (HL7 FHIR, etc.) as well as potentially having a metadata model that was infinitely flexible. By keeping the business value as the primary driver, we needed to be more concrete in the implementation and growth. Thus, we started with answering some concrete needs, such as visualizing the relationships inherent in the spreadsheets, while building that solution with a flexible data model that promises to allow those extensible relationships when the business need justifies that effort. Ontology like the one below starts to demonstrate the inherent relationships within the data, without imposing a strict hierarchy:
PLANNING FOR RELEASES

With an MDR program, there is a risk that the scope is so large, and the program tries to meet too many needs within one release, that delivery timelines are very long, often delayed and de-scoped, directly impacting the ability to deliver the anticipated business value. There is also a risk that the longer it takes to deliver a set of features, the more customer expectation will exceed what is actually delivered. The features and functionality of the Merck MDR have been grouped into multiple, smaller releases. Each release has features that are grouped into manageable, pragmatic scope to meet a specific set of customers. This allows the program to be responsive to changing business needs that may require a shift in direction or re-prioritizations. The current roadmap for the Merck MDR is:

- Release 1 included the foundation of the graph-store and ontology definitions that supports the metadata; allows the migration and maintenance of Excel based standard metadata definitions into MDR
PhUSE US Connect 2018

- Release 2 extended the functionality to create new form specifications within the MDR and adds automation to create the Data Management Workbench (DMW) Review Model data definitions
- Release 3 will implement a prioritized subset of study-level specifications and extend the ability to easily update SDTM mapping definitions
- Release 4 will extend the implementation of study-level definitions to include additional data sources, like external data

There is intentional time added between releases to ensure that the appropriate change management practices can be implemented to support customers’ new way of thinking and performing their respective business processes in alignment with the evolving features within the MDR.

LOADING STANDARD MASTER SPECIFICATIONS

Clinical Data Standards at Merck were being maintained in multiple, silo’d, distributed Excel spreadsheets. For example, there are 11 types of spreadsheets that must be harmonized and input into the configuration of our DMW environment to support a study. One of those spreadsheets, the Master Specification, is a domain-based spreadsheet which includes 4 tabs, each tab with numerous columns and rows. Each tab is dedicated to a user community: there is an InForm tab, and DWM tab, a Stats Programming tab, and a consolidated tab that combines information from the other tabs. This Master Specification defines the data collection items for each eCRF form, including code lists, as Global Standards. These Global Standard definitions help design of the EDC system, and drive the implementation of the DMW Review models and generation of SDTM datasets.

Traditionally, one user at a time updates the data definitions within a spreadsheet. The revision history and specification version is tracked in a fifth tab. Currently it is labor intensive and error prone to manually review the revision history text and find the appropriate corresponding change within a particular column and row of a subsequent worksheet. There are five different sets of users who contribute different information at various points in the authoring process.

Transitioning from authoring data definitions in spreadsheets to authoring and editing with the MDR will add business value by streamlining the authoring/editing and review processes and help ensure data quality. The MDR enables the ability to support concurrent authoring by different users contributing their respective content when they are available. The MDR automatically tracks which user authored or edited an item within the Master Specification and when. The MDR provides a summary of changes for reviewers to easily identify and assess what has been modified and by whom and provides the opportunity to comment on the changes in a field adjacent modified item so one can easily associate a comment with the actual item modified before changes are fully approved and promoted to production. An example of the Change Summary level of detail and ability for a reviewer to enter comments in the current Merck MDR is shown below:

The Merck MDR is ingesting the Global Standard definitions of the forms, their items, item level properties, and code lists from the Excel specification files into a semantic graph database. Before the excel file is loaded into the MDR, it follows a 21 step process to clean the information. Some of the cleaning is to facilitate the load process, and many of the steps improve the quality of the information. For instance, there is a simple check to proper casing certain fields, and more involved steps to make some data lineage explicit. Upon ingestion of the standard metadata into MDR, the Global Standards will be maintained within the MDR and the Excel files will be retired from circulation and archived. Loading these definitions into a semantic, ontology-based approach enables:
• Breaking down our current form-based view of standards, and defining item-level granularity enabling flexibility on how we define, manage, and use standards
• Versioning capabilities at a component level. Previously, in excel, versioning occurred at form level only.
• Defining relationships as to where and how information is used. These transitions from undocumented institutional knowledge to documented relationships that can be visualized in a meaningful way to multiple customers.
• Visualization of where and how information is used across the clinical information lifecycle and enabling technologies
• The implementation of the standard information model that is linked to technology implementations that use the information

The MDR provides a way to centralize information which previously was distributed across multiple individual files. A central repository of information allows better searching capabilities, application of consistency checking across definitions and enables streamlined information maintenance and governance processes.

The clinical standards Master Specifications define the eCRF collected Standard Metadata mapped through to Standard SDTM domains in preparation for dataset and analysis activities. The data standards specifications today are defined in alignment to how information is defined and flows across the clinical technologies. Each time there is a change to the way an application needs information defined, the entire Master Specification is updated, versioned, built, tested end to end and then released. Release 3 of the MDR will allow the separation of information definitions from the technology implementation. There are still explicit relationships (links) between the data definitions and the technology implementation, like Item Name, but the standard model will remain consistent even when the application landscape changes. If a new application is introduced, an EDC system for example, the only modification required to the model is to add extensions to support the new system. The foundational items of the information model remain unchanged. Combining metadata from Central Designer, Data Management Workbench, and MDR allows a complete picture of a study while not duplicating all of the properties needed for a specific technology implementation. This separation of information will enable more stability in the global standard by only needing to extend the model where the information is used. Validation efforts can also be streamlined by focusing on the direct areas impacted and not having to conduct full end to end testing, building and validation of a complete build.

The first step in our implementation approach is to recognize that we cannot introduce process changes and technology that negatively impact our ability to support operations. In order to ensure support of the current clinical programs, we have to ingest the data definitions as they are currently defined. Today’s data definitions are driven by tool needs. The first step in transforming that into information is to logically group the Master Specification information in a way that is not possible in spreadsheets. Our next step in the evolution of the MDR is to identify what’s common across all spreadsheets and define a single, common ontology that is then linked to the relevant tools to build and segregate the information definitions. The figure below shows a navigator from the MDR.

This visual insight was not possible when trying to manage information across multiple excel workbooks. Users across all customers now have the ability to visually represent the data definitions and how they are each used across the applications supporting the clinical data flow. The Navigator displays the data collection items on the left, and visually presents how the data flows from EDC, through the Review Model datasets, and where that data item is stored in an SDTM data structure. Each of the different datasets is also associated with a different user community. A navigator like the one in the Merck MDR is key to help communication across these different business groups who each have a different name for the same data item. The new insights brought about by this visualization enable:

• Common language across customers
• Insight into where and how information is used in a global definition
This has been achieved by centralizing the data definitions within MDR and defining the discrete relationships of information to each other and the applications that support the clinical data flow.

**THINGS TO CONSIDER DURING IMPLEMENTATION**

Implementing a new technology platform such as an MDR requires careful consideration of many factors. There will be a human factors component that will require carefully planned change management deployed through trainings, workshops and various communications. There will be a communication strategy that starts at the current cultural and process maturity of the organization. There will be impact to existing process, tools and technologies. All of these must coordinate and work with each other to deliver a successful MDR.

**PEOPLE AND COMMUNICATION**

The concept of the MDR was to enable new business capabilities for information management and governance; to define, use, maintain and share clinical data definitions. Each release is focused on enabling specific, prioritized business capabilities. An example of a capability that is being realized from our initial release (R1) is item level versioning. The powerful technology allows item level versioning, as opposed to versioning of Excel files that were stored in a SharePoint Teamspace with archived historical versions. MDR has changed the way one views the metadata; this requires the users to see the metadata differently.

As we introduce a new tool, it enables new capabilities which require that we think differently about the way we define and manage information. It would be a disservice if we simply recreated our existing business processes in a new tool. Adjusting current thinking and not re-creating past experiences is a challenge and requires close oversight in the development and realization of an MDR. Using the MDR has allowed Merck to start to be more structured in the way it stores and manages that information, but also required that downstream consumers need to change in order to consume the more structured information. As the MDR program developed, a number of downstream macro programs were uncovered that expected data definitions in a specific column of a specific spreadsheet. This discovery process involved multiple communication channels.

The MDR team has planned various ways to handle the change management of moving to the performing business processes within the MDR, from providing training and familiarization opportunities to outreach through various communications. Varying approaches had to be considered due to the number of persons affected, the different skill sets, functional customer groups and time zones. Having the wide range of customers and varied learning styles, we intentionally created and executed a large number of communication options and forums:

- Cross-functional customer engagement to define requirements through interactive, facilitated workshops
- Program Management Office engagement to ensure that Sponsors across the impacted functions are aligned on the roadmap and near-term priorities
- Centralized, open, public awareness meetings to introduce MDR concepts, features, and roadmap; build awareness on background, status and the roadmap; and foster interest and inclusion
- Process walkthrough meetings by functional group to understand process changes where users help to define the new and interim processes
- User forums to demonstrate tips and tricks of MDR to active users, provided by the core MDR development team
- Presentations at governance and department forums to introduce MDR.
- 1-1 meetings as needed to generate senior management engagement, key customer management or work through specific problem solving with SMEs of impacted systems/ processes.
- Mandatory trainings through Merck Learning & Development to providing user access.
- Short ‘How to’ videos of frequently used functions within MDR to help orient new users
- User support material to provide paper documentation for individuals who learn through reading
- SharePoint Community Space to provide a web-based, searchable space for the community
- Formal and Informal User Acceptance Testing to train individuals

**PROCESS**

Managing process changes is a sensitive task when introducing a new system and new way of working. The Global Metadata specifications are one component out of many that make up the Global standards; however, the Master Specifications play a central role in how other components are designed and implemented. Thus, the Master Specifications are central to the standard data collection landscape and can and does impact many other data sources, tools and systems. Understanding how and where the master specs are used is key to a successful transition to use
MDR. In addition to the widely known processes, the team also discovered undocumented and minimally understood processes that have dependencies on the Master Specifications.

A successful adoption of MDR is possible with engagement of customers. The team is continually addressing operational implementations, working with all customers and in this early stage refining, re-defining and clarifying the process steps. This engagement has unveiled several challenges, such as identifying written and unwritten uses of the existing Excel files for downstream functions (conformance checking, UAT, Generating dataset definition files).

People have grown used to the way that an Excel-based world operates, and don’t recognize the struggles of moving between workbooks because it has become part of the way they work. Incorporating MDR iteratively into existing processes involves re-training and delivering new features that may not 100% replace an existing functionality. For instance, it’s easy in excel to hide a column that you don’t want, and just as easy to forget that it’s hidden when you really need it.

**BUSINESS VALUE REALIZED**

Once the business needs were articulated, the first release planned and realized, the actual implementation provided some new insights. Some of the initial, specific, business benefits to address included:

- As an SDTM Programmer, I have to search for information across 4 different tabs of a spreadsheet to understand the traceability
- As a data manager, I am comfortable with the variables I work with, but my InForm and Stats Programmer colleagues don’t seem to understand what variables I’m talking about
- I don’t know if a change to an InForm section affects my SDTM dataset.
- The tab that I work in does not get versioned, and I need to understand when something changed.

Once the information is held in a form that is accessible, other business questions can be answered that weren’t imagined before.

**REMOVING EXCEL**

Excel is an easy tool to use for specifications – it’s loaded on your laptop, it’s flexible and can accommodate row/column data in a flexible way. However, it also allows for non-structured use as well. It’s easy to change the meaning of a column when it’s convenient to do so, but when that spreadsheet changes hands to another person, the inconsistent information can be confusing.

It’s also easy to create additional tabs within a spreadsheet to convey additional properties about an item without confusing the layout of the original tab where an item is defined. However, that flexibility also leads to duplication of effort, where an item name, for instance, is repeated in multiple places. It also leads to additional time and effort when trying to form a consolidated view of the information in a spreadsheet.

Loading those spreadsheets into the MDR provided the opportunity to clean up unstructured information, to remove unnecessary or historic information, and to create relationships across the tabs of a spreadsheet in order to be visually represented in a consistent way. This provided multiple immediate benefits, as described below.

**VISUALIZATION OF DATA FLOW**

One of the challenging aspects of looking at information in spreadsheets is that it’s not obvious how items are related to each other. Loading the metadata and defining relationships within the MDR allowed the opportunity to visualize the flow of data in a new way, as shown below:

The CMENDTCL item starts in InForm, but has 12 different downstream items that it is related to:

- 3 items in RD_CMFM represent the InForm data that is produced as the data is entered
- 3 items in AV_CM are used for safety reporting
- 3 items in RM_CM are used by Data Managers who prefer a more English-sounding item name instead of a 8-character name
• A single item remains in AR_CM, which is an SDTM+ structure
• Finally, the item finishes in either an SDTM 3.1.1 or SDTM 3.1.3 dataset

When a user logs into MDR, they can search for the item name they are familiar with (like CM_STOP_DT_CD), and this map will show them all of the other names and data flow for that item. This helps greatly with communication between InForm developers, Safety, Data Management, and Stats Programming, who all call the same item different things.

Another great use of visualization within the MDR is for SDTM up-versioning. Consider the example below:

The SCANDTC item starts in InForm, is split into two different RM3 datasets (AR_SCAN_CF and AR_SCAN_PR), and then split into SDTM domains. It is very easy for a stats programmer to see the differences between SDTM 3.1.1 mappings and SDTM 3.1.3 mappings. SDTM 3.1.3 changes the output from the CF domain in 3.1.1 to the FA domain. That change becomes clear with a visual representation of the data flow, including where the information was originally collected on the CRF.

IMPLIED SPECIFICATIONS
In any set of specifications, there are functions that everyone “knows” without stating, so it’s easier to just reference the item that is derived, instead of explicitly writing out every contributor to the item. In the case below, ITEMCOMMENTTEXT is a field that combines all of the Not Done flags into a single place:

This derivation was implied in the prior excel specifications. As part of the load into MDR, the derivation was made explicit so that the data flow is clear and obvious. A Data Manager can now search for ITEMCOMMENTTEXT and confirm that there are 4 items in the PTR Form that are collecting Not Done statuses.

CROSS-FORM INFORMATION
Based on a pragmatic approach, the definitions are loaded into MDR based on a form. However, the graph store allows a flexible hierarchy so that items can be considered the current item of focus, and you can understand where and how an item is used across forms. With the MDR in place, it is now possible to confirm and compare how common items are mapped across domains. For instance, the mapping logic for VISITMNEMONIC item appears in 4 domains below:

<table>
<thead>
<tr>
<th>Domain</th>
<th>SU</th>
<th>PROC</th>
<th>DEATH</th>
<th>SM</th>
</tr>
</thead>
</table>
Mapping Logic

The metadata reflects the mapping logic used to derive the VISITMNEMONIC. The information above indicates that the mapping logic is not consistent. The reasons could be historic – at one time the item may have been a direct move from InForm, while now the item has a derivation as part of the item. It could be that there are different needs across the different domains, where VISITMNEMONIC is more complex in PROC and SM, which warrants the additional function.

**OBJECT-LEVEL VERSIONING**

One of the biggest challenges with an excel-based approach is knowing when something has changed, and what is impacted. The MDR provides object-level versioning, with each of the following components:

- Form
- Form Section
- Form Item
- Codelist
- Codelist Item
- Dataset
- Dataset Item
- Dataset Mapping

This independent versioning allows for changes at a low level, and can also lead to confusion. People have become accustomed to referring to all changes within a specification as belonging to a single version of a form. Now, there could be changes to a Form that have no impact to a dataset, and vice-versa. That level of granularity is a new concept that will be propagated across the technology, people, and processes. Consider the following example:

Each object in the navigator has a version number to the right of it, (like CMENDCTL v1). The CMENDTCL item, version 1, is present in the CM form, version 5. Though the form has changed 5 times, whatever prompted that change in the form version was NOT the CMENDTCL form item. There may have been additional items added to the form, or other changes like codelist changes.

These are a few of the examples that are available today to users who are trying to understand the Global Specifications of the data they are collecting and analyzing. These definitions are still at a global level, and the next step in the evolution of the Merck MDR is to connect that information for a particular study, and start to understand re-use and other study-specific implementation definitions.

**NEAR-FUTURE BUSINESS VALUE**

The MDR project continues to deliver business value to a wide range of customers. Consolidating definitions from a number of spreadsheets, defining explicit relationships between information, and adding item-level versioning was a good first step. The number of possibilities for future business value starts to open up after the foundational steps are in place.

**MULTI-COMPANY COLLABORATIONS**

It is important for a company to easily access its own data definitions, understand data flow, and keep track of changes over time. Additional value is obtained when one is able to relate two disparate sets of information. Cross-company compound development collaborations continue to increase; having a mechanism to rapidly and easily assess how data definitions from one company are similar or different from another company’s is critical to support how best to design a collaboration study to facilitate data sharing and reporting and analysis activities by either parties. Even if SDTM is the common dataset standard between two companies, there are plenty of opportunities to collect scientifically sound information that falls outside of the standard SDTM models (e.g., SuppQual items). By using relationships and extensible, flexible metadata storage mechanisms, combining data from different sources will become easier.
Start with clear relationships and traceability from data collection to SDTM, even if that includes multiple transformations. You can then semantically link the data collection items from one company to another, and the challenge of mapping the collected data from another company through to SDTM becomes much easier.

**METADATA ACCESS FOR ALL**

It is important to create a line between functionality that is needed within an MDR, and functionality that is using the information stored in an MDR. Merck will utilize a data connector into the graph store to allow other tools to access the metadata definitions. Ensuring that the MDR information can be accessed by other tools, like Excel and SAS, allows other groups to meet their information needs at a lower cost that building it within the MDR itself. Secondary validation outside of MDR is one instance. For example, it may be important to know that all of the data elements from collection to SDTM have a large enough field length to accommodate the data. It is easier to build a check like this in SAS or Excel than it is to build the check within an MDR. The business value in this case is measured with the ease of implementation and frequency of change of those secondary validations. Most teams will not want to wait 4-6 months for a product release before realizing an additional validation check of the metadata.

Another example of a secondary use would be in UAT testing of a data collection instrument. The specific data collection instrument may change, especially with external data sources, which makes it more expensive to build that checking into the MDR itself. By having a data connector into the MDR, a group can query the expected fields and properties, and then write a simple program to see information held in MDR to confirm if the external data meets the definititional requirements.

**SECONDARY PROCESS SUPPORT**

The Statistical Programming group builds libraries of standard code that work with different versions of datasets. It is important to know which versions of those datasets are in use for a particular study, and at different time points in a study. When standard programs know what versions of datasets are supported, then a programmer can identify which programs may need to be modified as a result of a change in a source dataset. The MDR will play a role in matching study-level implementation with the global definitions of a domain, and subsequently, the standard programs associated with that definition.

**ADDITIONAL DATA TRACEABILITY**

Some implementations of a Stats Computing Environment have effectively utilized header information in a program to create an implied linking between datasets and programs. Using a graph-based data store, and linking in unstructured data like SAS programs will allow a more automated way of associating datasets, dataset items, and programs. Then, when a program is updated to include a new dataset as part of the processing, it is not dependent upon a human being to remember to add that dataset to the header of the program in order to maintain the data flow impact.

**CONCLUSION**

The expectations and possibilities for an MDR can be overwhelming. Using a business value approach to help guide the continuing development, features, and integration points is important for the overall success of the MDR program. This step-wise approach is allowing different business communities to understand the language used for the same data item, is helping to distribute responsibility among the individuals who contribute to a data definition, and is opening new possibilities for improving processes and operations. When the business value lines up with the roadmap of an MDR, that success will feed new and interesting additional business value opportunities which will continue to make a difference in the lives of people through supporting the development of innovative medicines.
CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:

Brooke Hinkson
Merck
126 East Lincoln Avenue
Rahway, NJ 07065
Email: amanda.hinkson@merck.com

Karl Konrad
Merck
5325 Old Oxford Rd
Durham, NC 27712
Email: karl.konrad1@merck.com

Manori Turmel
Merck
351 N Sumneytown Pike
North Wales, PA 19454
Email: manori.turmel@merck.com

Brand and product names are trademarks of their respective companies.