Paper DS07

Automated Pipelines for Non-CRF Value Level Metadata, A Meta-programming Approach

Selena Baset, Philipp Ernst and Didier Clément
Hoffman La Roche, Basel, Switzerland

ABSTRACT

Like every step in any clinical trial, data collection is a process that lends itself to harmonization. In Roche, collected data should conform to collection standards that are based on the CDISC-controlled terminologies or their Roche sponsor extension. This means that every variable's permissible value is bound to fall within pre-established enumerated value domains. While dictating the permissible values for a given variable in isolation from others is readily achievable using the existing standard models, dictating the permissible combinations of values for variables collected together is more problematic as it requires maintaining Value Level Metadata (VLM) according to schemas that vary dynamically following the different data collection settings and the applicable standard version. We propose an end-to-end solution to the VLM management problem by designing a generic VLM ontology for capturing machine-readable permutations of permissible values combined with a meta-programming approach to automatically create authoring tools and pipelines to propagate changes following updates to the underlying data collection standards. We demonstrate the end-to-end solution applied to VLM schemas of two different non-CRF models: 1) the ophthalmologic grading parameters and 2) digital measures passive monitoring.

INTRODUCTION

In the clinical trials space, the meticulous collection and management of data are paramount. At Roche, data collection adheres to stringent standards that are aligned with the CDISC-controlled terminologies. These standards ensure that collected variables align with predetermined value domains, maintaining collected data integrity and consistency. However, dictating permissible value combinations for variables collected in tandem presents a daunting maintainability challenge that requires delving into Value Level Metadata (VLM) specifications. This paper introduces an end-to-end solution to the VLM management problem. Our approach combines a generic VLM ontology for machine-readable value permutations with strategies borrowed from meta-programming to automate authoring tools and pipelines, simplifying VLM maintenance following updates to data collection standards. We illustrate this solution through use cases involving non-CRF data models developed within Roche, specifically the models for ophthalmologic grading parameters and digital measures passive monitoring. In the subsequent sections, we will delve into the complexities of the VLM management problem, explore the specific challenges it poses, and detail our proposed solution before concluding with an outlook on potential future application areas.

PROBLEM STATEMENT

Dataset structures and terminology of collected non-CRF data at Roche are very much aligned with CDISC SDTM/SDTMIG. While this approach allows seamless and automated processing of data through their pipeline from data collection to SDTM datasets, selecting correct terminology in the non-CRF data transfer agreements (DTAs) can be complex for the data manager. The extraction of relevant information from data provider documentation (e.g. charters, Scopes of Work) and the translation thereof requires deep data modeling experience, especially for the scientifically increasingly complex types of data such as independent image assessment (e.g. oncology, CNS, ophthalmology), digital health technology/wearable or biomarker data. Guiding data managers by just dictating the permissible values for a given variable in isolation from others has proven not to absorb the complexity of non-CRF data specifications in a satisfactory manner.

Acknowledging the value for users of specifying permissible combinations of values for variables collected together through e.g. CDISC Codetable Mapping Files for data tabulation, Roche similarly maintains VLM for a number of use cases applying two different approaches: 1) maintenance in the metadata repository, or 2) offline maintenance in spreadsheets. Approach 1) is the preferable one as each VLM has a dedicated underlying technical solution for machine readability and connection to the standard source, it is resourceful and not easily scalable to other VLMs as every change in the non-CRF data model it is used for requires retouching of the underlying technical solution. On the other hand, approach 2) is technically more affordable to scale VLM coverage but bears the risk of the VLM not being aligned with metadata repository content and involves a high resource spend in manual maintenance.

We are going to use examples of manually maintained VLM for two non-CRF data collection models to support the problem statement and demonstrate a proposed solution.

Ophthalmic Grading Parameters

Roche is working with more than 10 central reading centers (CRCs) for the independent assessment of ophthalmologic images taken with various modalities. Template charters for each of the CRCs covering the highly specialized read-out designed to answer specific scientific questions on our clinical trials often contain several hundreds of parameters assessed (Table 1), overall leading to a large set of analytes that are very heterogeneously and sometimes scarcely described by the CRC. It requires a considerable amount of resources to clarify definitions of analytes in the CRC charters and to map them to the non-CRF dataset structures and controlled terminology. The risk of inconsistent mapping is mitigated by maintaining spreadsheets explicitly translating the analyte descriptions used by the CRCs into the relevant variables and controlled terminology that are used in the non-CRF DTAs (Table 2). On the other hand, risks associated with manually maintaining these VLM spreadsheets are mostly 1) the control of unauthorized copies circulating across the large Roche-internal organization as well as amongst the CRCs, 2) the challenge to keep the content of the master copy in sync with the metadata repository, and 3) the inability to integrate them into the automated metadata consumption workflow for DTA creation.

Variable	Description	Value
1. Intraretinal Fluid (IRF) per ETDRS region		
1.1	Intraretinal Fluid (IRF) – Within Total OCT volume, given in nanoliter	# Numerical Value NA
1.2	Intraretinal Fluid (IRF) – Within Total Upper Half of ETDRS, given in nanoliter	# Numerical Value NA
1.3	Intraretinal Fluid (IRF) – Within Total Lower Half of ETDRS, given in nanoliter	# Numerical Value NA
1.4	Intraretinal Fluid (IRF) – Outside ETDRS, given in nanoliter	# Numerical Value NA
1.5	Intraretinal Fluid (IRF) – Outside Upper Half of ETDRS, given in nanoliter	# Numerical Value NA
1.6	Intraretinal Fluid (IRF) – Outside Lower Half of ETDRS, given in nanoliter	# Numerical Value NA
1.7	Intraretinal Fluid (IRF) – Within Disc 1mm of ETDRS, given in nanoliter	# Numerical Value NA
1.8	Intraretinal Fluid (IRF) – Within Disc 1mm Upper Half of ETDRS, given in nanoliter	# Numerical Value NA
1.9	Intraretinal Fluid (IRF) – Within Disc 1mm Lower Half of ETDRS, given in nanoliter	# Numerical Value NA
1.10	Intraretinal Fluid (IRF) – Within Disc 3mm of ETDRS, given in nanoliter	# Numerical Value NA
1.11	Intraretinal Fluid (IRF) – Within Disc 3mm Upper Half of ETDRS, given in nanoliter	# Numerical Value NA
1.12	Intraretinal Fluid (IRF) – Within Disc 3mm Lower Half of ETDRS, given in nanoliter	# Numerical Value NA
1.13	Intraretinal Fluid (IRF) – Within Disc 6mm of ETDRS, given in nanoliter	# Numerical Value NA
1.14	Intraretinal Fluid (IRF) – Within Disc 6mm Upper Half of ETDRS, given in nanoliter	# Numerical Value NA
1.15	Intraretinal Fluid (IRF) – Within Disc 6mm Lower Half of ETDRS, given in nanoliter	# Numerical Value NA
1.16	Intraretinal Fluid (IRF) – Within Sector Superior of ETDRS, given in nanoliter	# Numerical Value NA
1.17	Intraretinal Fluid (IRF) – Within Sector Nasal of ETDRS, given in nanoliter	# Numerical Value NA
1.18	Intraretinal Fluid (IRF) – Within Sector Inferior of ETDRS, given in nanoliter	# Numerical Value NA
1.19	Intraretinal Fluid (IRF) – Within Sector Temporal of ETDRS, given in nanoliter	# Numerical Value NA
1.20	Intraretinal Fluid (IRF) – Within Parafoveal Ring of ETDRS, given in nanoliter	# Numerical Value NA

Table 1 - Excerpt (20 out of 198 analytes) of an image charter describing analytes on a grading form for AI-based analysis of OCT (optical coherence tomography) images.

Grading Parameter	Definition	OELOC	OEGRID	OELOC	OELOCDTL	OEMETHOD	OETESTCD	OETEST	Unit	Possible Values
Presence of Hemorrhages and/or Microaneurysms Within ETDRS Grid	Presence and amount of hemorrhages and microaneurysms within ETDRS grid compared to standard photographs	RETINA	ETDRS GRID	RETINA	WHOLE GRID	COLOR FUNDUS PHOTOGRAPHY	MANRYSM	Presence of Retinal Microaneurysm		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Hemorrhage	Hemorrhage will be bright to dark red. Retinal vessels should not be included in any assessment of hemorrhage.	RETINA	ETDRS GRID	RETINA	WHOLE FIELD OF VIEW	COLOR FUNDUS PHOTOGRAPHY	HMRHGE	Presence of Hemorrhage		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Hemorrhage in center 1mm	Presence of hemorrhage within the center 1mm or that extends into the center 1mm	RETINA	ETDRS GRID	RETINA	CENTER 1 MM	COLOR FUNDUS PHOTOGRAPHY	HMRHGE	Presence of Hemorrhage		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Hemorrhage Extrafoveal (outside center 1mm)	Presence of hemorrhage outside the center 1mm or that extends outside the center 1mm.	RETINA	ETDRS GRID	RETINA	OUTSIDE CENTER 1 MM	COLOR FUNDUS PHOTOGRAPHY	HMRHGE	Presence of Hemorrhage		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Hemorrhage at Foveal Center	Presence of hemorrhage at foveal center	RETINA	ETDRS GRID	RETINA	FOVEAL CENTER	COLOR FUNDUS PHOTOGRAPHY	HMRHGE	Presence of Hemorrhage		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Fibrosis in Center 1mm	presence of fibrosis within the center 1mm or that extends into the center 1mm.	RETINA	ETDRS GRID	RETINA	CENTER 1 MM	COLOR FUNDUS PHOTOGRAPHY	RFIB	Presence of Retinal Fibrosis		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Fibrosis Extrafoveal (outside center 1mm)	presence of fibrosis outside the center 1mm or that extends outside the center 1mm.	RETINA	ETDRS GRID	RETINA	OUTSIDE CENTER 1 MM	COLOR FUNDUS PHOTOGRAPHY	RFIB	Presence of Retinal Fibrosis		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Fibrosis at Foveal Center	presence of fibrosis at foveal center.	RETINA	ETDRS GRID	RETINA	FOVEAL CENTER	COLOR FUNDUS PHOTOGRAPHY	RFIB	Presence of Retinal Fibrosis		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of Hard Exudates Within ETDRS Grid	Presence of hard exudates within ETDRS grid	RETINA	ETDRS GRID	RETINA	WHOLE GRID	COLOR FUNDUS PHOTOGRAPHY	HEX	Presence of Hard Exudates		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Presence of dense subfoveal hard exudates within central 1mm	Presence of dense subfoveal hard exudates within central 1mm	RETINA	ETDRS GRID	RETINA	CENTER 1 MM	COLOR FUNDUS PHOTOGRAPHY	HEX	Presence of Hard Exudates		e.g. PRESENT, NOT PRESENT, NOT GRADABLE
Size of Capillary Non Perfusion Within ETDRS Grid	Total area of capillary non perfusion within the macula ETDRS grid, central 42.00 mm2.	RETINA	ETDRS GRID	RETINA	WHOLE GRID	FLUORESCEIN ANGIOGRAPHY	CNPSZ	Total Capillary Nonperfusion Area	mm2	<numeric>, NOT GRADABLE

Table 2 - Excerpt (11 out of 600 rows) of a manually maintained VLM spreadsheet displaying the mapping between analytes generated by applying an ETDRS (Early Treatment of Diabetic Retinopathy Study) grid on a color fundus photography image and their description (blue columns) to the corresponding variables used in the non-CRF DTAs (green columns). Content in the green columns corresponds to terminology maintained in the metadata repository but is not semantically linked.

Digital Measures Passive Monitoring

Another data type Roche is maintaining VLM for is the area of digital measures. Passively monitoring patients' mobility in our clinical trials has become an important piece of study protocols in many indications thanks to the availability of a rapidly growing number of commercially available devices and applications. That leads to large volumes of data being generated and an increasingly heterogeneous landscape of raw data (Table 3), which in turn drives the complexity of the dataset structures and the volume of associated terminology used in the DTAs: the passive monitoring findings dataset contains 15 qualifier or topic variables (Table 4) with an associated total number of 165 controlled terms to describe analytes adequately. The excerpt of the manually maintained VLM for passive monitoring findings shown in Table 5 helps data managers to navigate the complexity mentioned above, but the same risks as outlined for the ophthalmologic grading parameters start to outweigh the advantages VLM can offer.

Data Header	Data Type	Data Description
ActivityDate	date	Date value in mm/dd/yyyy format.
TotalSteps	integer	Total number of steps taken.
TotalDistance	integer	Total kilometers tracked.
TrackerDistance	integer	Total kilometers tracked by Fitbit device.
LoggedActivitiesDistance	integer	Total kilometers from logged activities.
VeryActiveDistance	integer	Kilometers travelled during very active activity.
ModeratelyActiveDistance	integer	Kilometers travelled during moderate activity.
LightActiveDistance	integer	Kilometers travelled during light activity.
SedentaryActiveDistance	integer	Kilometers travelled during sedentary activity.
VeryActiveMinutes	integer	Total minutes spent in very active activity.
FairlyActiveMinutes	integer	Total minutes spent in moderate activity.
LightlyActiveMinutes	integer	Total minutes spent in light activity.
SedentaryMinutes	integer	Total minutes spent in sedentary activity.
Calories	integer	Total estimated energy expenditure (in kilocalories).
Floors	integer	Total number of floors climbed.
CaloriesBMR	integer	Total energy expenditure from basal metabolic rate
MarginalCalories	integer	Total marginal estimated energy expenditure (in kilocalories). Difference between ActivityCalories minus BMR contribution during exercises.
RestingHeartRate	integer	Resting heart rate value.

Table 3 - Excerpt of a data provider's data dictionary for passively monitored physical activity assessed over the duration of one day.

Variable	Label	Description	Codelist Name	Data Type/Length	Sponsor Compliance	Policy
STUDYID	Study Identifier	Roche protocol number	STUDY_NAME_V0	Char 20	EXPECTED	Mandatory
ZPNAM	Vendor Name	Vendor identification short name	VENDOR_NAME_V0	Char 200	EXPECTED	Mandatory
PATNUM	Unique Subject Identifier	Subject identification number for randomized/enrolled subjects		Num 8	KEY	Mandatory
VISIT	Visit Code	Visit identification short name	VISIT_NAME_V0	Char 200	PERMISSIBLE	Optional
PMFTPT	Planned Time Point Name	Name of time point when the passive monitoring measurement is planned to be recorded	SCHEDULED_TIME_V0	Char 40	PERMISSIBLE	Optional
ACCSNM	Vendor Unique Log Identification	Log or session identification number used by vendor at the time of the passive monitoring sensor data creation		Char 40	PERMISSIBLE	Optional
PMFSTD	(Start) Date of Passive Monitoring Measurement Record (Local)	Start date or single date at which the passive monitoring measurement was recorded in the local time zone		Char 11	KEY	Mandatory
PMFSTTM	(Start) Time of Passive Monitoring Measurement Record (Local)	Start time or single time at which the passive monitoring measurement was recorded in the local time zone		Char 10	KEY (potential)	Mandatory
PMFENDD	End Date of Passive Monitoring Measurement Record (Local)	Date until which the passive monitoring measurement was recorded in the local time zone		Char 11	EXPECTED	Mandatory
PMFENDTM	End Time of Passive Monitoring Measurement Record (Local)	Time until which the passive monitoring measurement was recorded in the local time zone		Char 10	EXPECTED	Mandatory
PMFDUR	Duration of Passive Monitoring Findings Recording or Assessment	An ISO 8601 representation "P[yY][mM][dD][T[hH][mM][sS]" of the reported passive monitoring data record episode		Char 25	PERMISSIBLE	Optional
PMFUSTD	Date of Passive Monitoring Measurement Record (UTC)	Start date or single date at which the passive monitoring measurement was recorded according to Coordinated Universal Time (UTC)		Char 11	KEY (potential)	Optional
PMFUSTTM	Time of Passive Monitoring Measurement Record (UTC)	Start time or single time at which the passive monitoring measurement was recorded according to Coordinated Universal Time (UTC)		Char 10	KEY (potential)	Optional
PMFSTZUO	(Start) Time Zone UTC Offset	Time zone UTC offset corresponding to PMFSTD and PMFSTTM.		Char 6	KEY (potential)	Optional
PMFUENDD	End Date of Passive Monitoring Measurement Record (UTC)	End date at which the passive monitoring measurement was recorded according to Coordinated Universal Time (UTC)		Char 11	EXPECTED	Optional
PMFUENTM	End Time of Passive Monitoring Measurement Record (UTC)	End time at which the passive monitoring measurement was recorded according to Coordinated Universal Time (UTC)		Char 10	EXPECTED	Optional
PMFETZUO	End Time Zone UTC Offset	Time zone UTC offset corresponding to PMFENDD and PMFENTM.		Char 6	EXPECTED	Optional
SPDEVID	Sponsor Device Identifier	Device identification and link to the device identification, device in-use and device exposure datasets		Char 100	KEY	Mandatory
ZPMETHOD	Method of Passive Monitoring Findings Test or Examination	Method used for the passive monitoring findings test or examination	METHOD_V17	Char 100	EXPECTED	Mandatory
ZPMTDTL	Detailed Method of Test or Examination	Detailed method used for the passive monitoring findings test or examination	DMMTDTL_V1	Char 200	EXPECTED	Optional
ZPANMETH	Secondary Processing or Analysis Method	Analysis method applied to obtain a processed/summarized result	DMANMETH_V1	Char 200	KEY (potential)	Optional
ZPLOC	Location Used for Measurement	Anatomic location of the subject where the passive monitoring measurement is assessed	LOC_V3	Char 200	KEY (potential)	Optional
ZPLAT	Laterality	Qualifier for anatomical location of the subject where the passive monitoring measurement is assessed	LATERALITY_V1	Char 100	KEY (potential)	Optional
ZPCAT	Category for Passive Monitoring Findings Test or Examination	Classification of the passive monitoring findings test or examination	ZPCAT_V0	Char 100	KEY	Mandatory
ZPSCAT	Subcategory for Passive Monitoring Findings Test	Sub-classification of the passive monitoring findings test or examination	ZPSCAT_V0	Char 100	KEY	Mandatory
ZPTESTCD	Passive Monitoring Test or Examination Short Name	Short name for the passive monitoring test or examination	ZPTESTCD_V0	Char 8	KEY	Mandatory
ZPTEST	Passive Monitoring Test or Examination Long Name	Long name for the passive monitoring test or examination	ZPTEST_V0	Char 40	EXPECTED	Optional
ZPTSTDTL	Passive Monitoring Findings or Examination Test Detailed Name	Detailed name of the passive monitoring findings test or examination	ZPTSTDTL_V0	Char 200	KEY (potential)	Mandatory
ZPACINTS	Passive Monitoring Test or Examination Activity Intensity	Degree or magnitude of passive monitoring test or examination during a physical activity	ZPACINTS_V0	Char 200	KEY	Optional
ZPSLPSTG	Passive Monitoring Test or Examination Sleep Stage	Sleep stage of passive monitoring test or examination during sleep assessment	SLPSTG_V2	Char 100	KEY	Optional
ZPSLPPRD	Passive Monitoring Test or Examination Sleep Period	Sleep period of passive monitoring test or examination during sleep assessment	SLPPRD_V2	Char 100	EXPECTED	Optional
PMFRESN	Numeric Result or Finding in Agreed Units	Quantitative result of the passive monitoring findings test or examination		Num 8	PERMISSIBLE	Mandatory
PMFORESU	Units	Unit of measurement associated with the numeric result	UNIT_V35	Char 40	PERMISSIBLE	Mandatory
PMFRESC	Character Result or Finding	Qualitative or semi-quantitative result of the passive monitoring findings test or examination		Char 200	PERMISSIBLE	Mandatory
ZPSTAT	Completion Status	Text code indicating that the result of the passive monitoring findings test or examination cannot be provided	CMPLSTAT_V0	Char 8	PERMISSIBLE	Optional
ZPREASND	Reason Test Not Done	Reason why the result of the passive monitoring findings test or examination cannot be provided	DMREASND_V1	Char 200	PERMISSIBLE	Optional
ZPCLMD	Collection Mode	Collection mode that describes the technique or procedure that determines how the passive monitoring measurement was conducted	DMCLMD_V1	Char 100	EXPECTED	Mandatory
ZPDVPL	(Digital) Device Placement	The digital device placement status which provides information about the device usage when the passive monitoring test or examination was performed.	DMDVPL_V1	Char 100	KEY (potential)	Mandatory
ZPDVPLAM	(Digital) Device Placement Analysis Method	Analysis method applied to obtain a processed result about the digital device placement status when the passive monitoring test or examination was performed.	DMDVPLAM_V1	Char 200	EXPECTED	Optional
ZPWKSLP	Wake-Sleep Status	The wake-sleep status of the subject at the time of the passive monitoring measurement	ZPWKSLP_V0	Char 20	EXPECTED	Optional

Table 4 - Passive monitoring findings dataset definitions with associated codelists. Relevant codelists associated with variables qualifying an analyte are highlighted with red boxes.

Scientific Denomination	ZPMETHOD METHOD_V17	ZPLOC LOC_V3	ZPLAT LAT_V3	ZPCAT (ZPCAT_V0)	ZPSCAT (ZPSCAT_V0)	ZPTESTCD (ZPTESTCD_V0)	ZPTSTDTL (ZPTSTDTL_V0)	ZPACINTS (ZPACINTS_V0)	PMFORESU (UNIT_V35)	PMFRESC (Codelist to be added to in the next publication)
Steps taken during sedentary physical activity. Steps taken during light physical activity. Steps taken during fair physical activity. Steps taken during vigorous physical activity.	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	PHYSICAL ACTIVITY MEASUREMENT	WALK	COUNT	FOOTSTEP	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	<blank>	<blank>
Duration of sedentary physical activity (in minutes). Duration of light physical activity (in minutes). Duration of fair physical activity (in minutes). Duration of vigorous physical activity (in minutes).	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	PHYSICAL ACTIVITY MEASUREMENT	WALK	DUR	<blank>	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	min	<blank>
Distance traveled during sedentary physical activity (in meters). Distance traveled during light physical activity (in meters). Distance traveled during fair physical activity (in meters). Distance traveled during vigorous physical activity (in meters).	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	PHYSICAL ACTIVITY MEASUREMENT	WALK	DIST	<blank>	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	min	<blank>
Elevation gained during sedentary physical activity (in meters). Elevation gained during light physical activity (in meters). Elevation gained during fair physical activity (in meters). Elevation gained during vigorous physical activity (in meters).	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	PHYSICAL ACTIVITY MEASUREMENT	WALK	CELGAIN	<blank>	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	m	<blank>
Total number of steps during sedentary physical activity. Total number of steps climbed during light physical activity. Total number of steps climbed during fair physical activity. Total number of steps climbed during vigorous physical activity.	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	PHYSICAL ACTIVITY MEASUREMENT	WALK	COUNT	FOOTSTEP	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	<blank>	<blank>
Total number of floors climbed during sedentary physical activity. Total number of floors climbed during light physical activity. Total number of floors climbed during fair physical activity. Total number of floors climbed during vigorous physical activity.	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	PHYSICAL ACTIVITY MEASUREMENT	WALK	COUNT	CLIMBED STEP	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	<blank>	<blank>
Total estimated energy expenditure during sedentary physical activity (in kilocalories). Total estimated energy expenditure during light physical activity (in kilocalories). Total estimated energy expenditure during fair physical activity (in kilocalories). Total estimated energy expenditure during vigorous physical activity (in kilocalories).	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	HEALTH	METABOLIC MEASUREMENT	CALBURN	<blank>	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	kcal	<blank>
Total energy expenditure from basal metabolic rate during sedentary physical activity (in kilocalories). Total energy expenditure from basal metabolic rate during light physical activity (in kilocalories). Total energy expenditure from basal metabolic rate during fair physical activity (in kilocalories). Total energy expenditure from basal metabolic rate during vigorous physical activity (in kilocalories).	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	HEALTH	METABOLIC MEASUREMENT	CALBURN	CALORIES BMR	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	kcal	<blank>
Total marginal estimated energy expenditure during sedentary physical activity (in kilocalories). Total marginal estimated energy expenditure during light physical activity (in kilocalories). Total marginal estimated energy expenditure during fair physical activity (in kilocalories). Total marginal estimated energy expenditure during vigorous physical activity (in kilocalories).	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	HEALTH	METABOLIC MEASUREMENT	CALBURN	CALORIES MARGINAL	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	kcal	<blank>
Resting heart rate value during sedentary physical activity (in beats/min). Resting heart rate value during light physical activity (in beats/min). Resting heart rate value during fair physical activity (in beats/min). Resting heart rate value during vigorous physical activity (in beats/min).	ACTIVITY TRACKING	FOREARM HAND	LEFT RIGHT	HEALTH	VITAL SIGNS MEASUREMENT	HRAV	RESTING	SEDENTARY PHYSICAL ACTIVITY LIGHT PHYSICAL ACTIVITY FAIR PHYSICAL ACTIVITY VIGOROUS PHYSICAL ACTIVITY	beats/min	<blank>

Table 5 - Excerpt (10 out of 450 rows) of a manually maintained VLM spreadsheet displaying the mapping between passively monitored physical activity (typically assessed over the duration of one day) and their description (blue columns) to the corresponding variables used in the non-CRF DTAs (green columns). Content in the green columns corresponds to terminologies maintained in the metadata repository but are not semantically linked.

PRELIMINARIES

In the two following subsections, we briefly cover two key concepts critical to understanding our solution: GDSR, the Global Data Standards Repository, as a semantic model-driven framework and the principles of meta-programming and declarative programming languages. These concepts provide the foundation for the solution approach detailed in the following section.

GDSR AS A SEMANTIC MODEL-DRIVEN FRAMEWORK

Besides being Roche's source of truth for everything clinical data standards, GDSR is a framework for semantic model-driven development. Initially, GDSR was built in-house as a multi-tenant repository for storing metadata standards. The standards, such as industry-wide CDISC or Roche's extensions, are expressed as machine-readable knowledge graphs, stored in a triple store and served to users via a web browser. Over the years, GDSR has grown some more capabilities such as metadata cataloging and versioning, extensive REST APIs for programmable access, automatic integration with version control systems and most recently native authoring and governance tools for content owners with no prior semantic expertise.

With these capabilities acquired, GDSR has become a full-fledged semantic model-driven framework that is increasingly used by different teams within Roche to build semantically-powered FAIR solutions[1] like the one presented in this paper. In essence, the solution development process in GDSR always starts with the ontology creation phase where an ontologist, or an information architect, works closely together with a domain subject matter expert on a set of competency-based questions that will help define the ontology as a formal machine-understandable representation of the domain of interest. With the ontology in place, what follows next is a series of semantic artifacts development that covers all configurations needed to manage and expose metadata as populated knowledge graphs. The artifacts development is usually the responsibility of an information architect who has good expertise with the technical stack of the Semantic Web. Examples of the aforementioned artifacts include the set of SPARQL[2] queries used to expose content downstream and the SHACL[3] shapes used to validate the knowledge graph content and auto-generate user-friendly authoring interfaces.

[1] A FAIR solution to a data problem is one that ensures data generated using that solution meets the FAIR data principles of Findability, Accessibility, Interoperability, and Reusability

[2] SPARQL is the standard query language and protocol for Linked Open Data on the web or for RDF triplestores. SPARQL, short for "SPARQL Protocol and RDF Query Language"

[3] SHACL (Shapes Constraint Language) is a W3C standard for validating the contents of an RDF-style graph database.

The VLM solution proposed in this paper adheres to the development process above but leverages the commonalities between the different manifestations of the VLM problem to standardize and automate most of the artifact configuration steps.

METAPROGRAMMING AND DECLARATIVE PROGRAMMING LANGUAGES

Meta-programming is a programming paradigm in the realm of software engineering where a program has knowledge of itself or can manipulate itself. A metaprogram is therefore a program that writes, analyses or transforms programs including itself. This self-modifying feature of metaprograms allows for significant flexibility in handling runtime code changes efficiently without recompilation. It can also help reduce development time by minimizing the volume of code required to express a particular solution. At its core, meta-programming leverages a higher level of abstraction by defining and referring to software artifacts as data objects. Through the manipulation of these objects, constructing, combining or fragmenting them, the metaprogram can dynamically evolve.

Declarative programming languages are characterized by their ability to describe the desired results of a program execution in a concise and expressive manner that emphasizes "what" needs to be achieved rather than explicitly specifying "how" to achieve it. Whilst not programming languages per se, the W3C standard languages such as RDFS, OWL[4], SPARQL and SHACL have close proximity to the declarative programming paradigm. In the majority of semantic applications where these languages are deployed, they are used to declaratively describe the problem domain, its entities, their relationships and what constraints are applicable to a given solution. The actual imperative logic of how the constraints are applied is then left to generic software components such as OWL reasoners, SPARQL query execution engines, SHACL inference engines, etc.

In the context of this paper, the key note to highlight is that this declarative language approach is the ground on which the GDSR framework described above is built; and is also what enabled the meta-programming-inspired solution to the VLM problem.

PROPOSED SOLUTION

Borrowing from concepts in the metaprogramming paradigm, the solution we propose for managing permissible VLM combinations is one that looks at the problem from a more abstract perspective. Rather than developing software artifacts for a given combination schema, we put in place a meta-schema VLM ontology that captures the essence of what constitutes a VLM schema and would therefore allow for creating different VLM-specific schemas as instances. We then leverage the power of semantic inference to automatically expand the ontology definition and add the classes and properties to describe the new VLM schema of interest. Given that all VLM schema instances have a consistent standard definition as per the meta-schema ontology, automating the development process of the remaining artifacts becomes feasible. Here again, we rely on the semantic inference machinery to automatically generate all necessary configurations including the SPARQL and SHACL necessary for creating the GDSR editor user interfaces and, in the case where legacy VLM combinations were maintained in a tabular sheet, the parser configuration needed to automatically populate the knowledge graph.

Having passed through the automated pipeline above, the legacy tabular form of free text-based value combinations is now fully converted to machine-readable metadata. The resulting knowledge graph is appended to the clinical standards knowledge graph in GDSR where the semantic link to the non-CRF standard variables and controlled terminologies is restored. Besides enforcing data integrity and opening new possibilities for standards compliance assessment, restoring the semantic link to the standard source is a big efficiency booster and a time saver for the data standards manager maintaining the VLM combinations as it practically eliminates the need to manually propagate changes to the tabular VLM sheet following an update to the standard source; which in turn also eliminates the risk of the VLM not being aligned with the metadata repository.

[4] RDFS and OWL are Semantic Web standards for modeling and representing data. RDFS provides basic hierarchy modeling, while OWL enables complex knowledge representation.

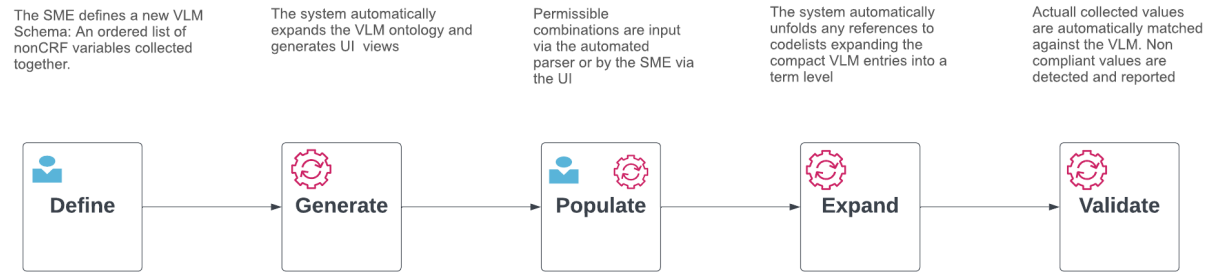

Figure 1. Main steps in the proposed end-to-end workflow for managing permissible value combinations given a VLM schema

The VLM ontology

Just like a metaprogram has knowledge of itself and can manipulate itself, the VLM ontology has a small number of stem classes and properties that describe how the ontology can auto-expand as new VLM combination schemas are defined. This auto-expansion is enabled by a SPARQL inference rule that dedicates how the new classes and properties are created and attached to the ontology.

We can therefore refer to two states of the VLM ontology, 1) an *initial state* that corresponds to domain definitions and stem classes and properties and 2) an *expanded ontology* that corresponds to the totality of the domain definitions, the stem part, as well as all classes and properties automatically created given a new VLM schema instance.

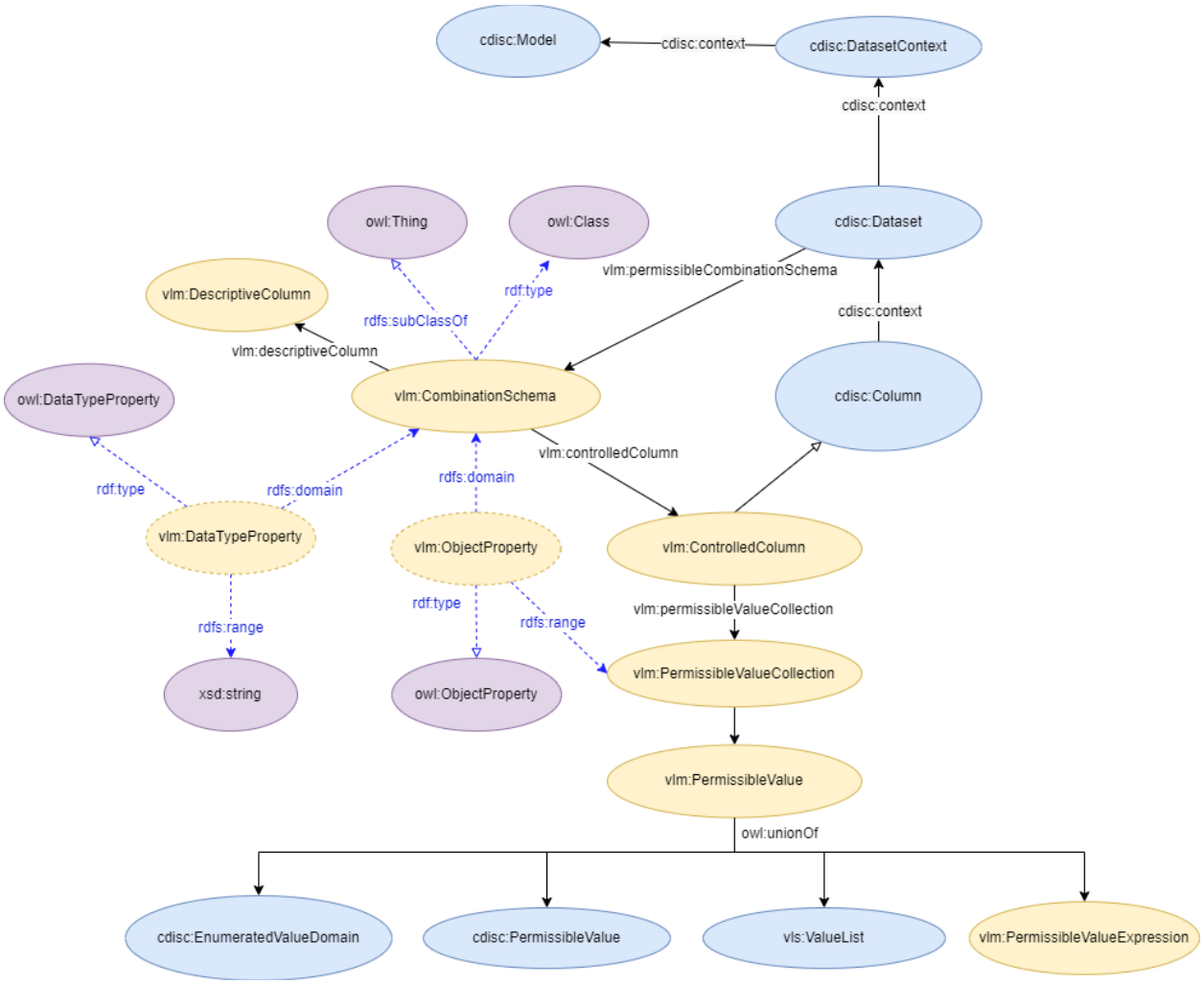

Figure 2. A somewhat simplified representation of the VLM ontology in its initial state before any VLM schema is defined. The dotted classes and properties represent the stem part used for automatically generating the VLM schema-specific parts of the ontology.

A simplified visual representation of the VLM ontology in its initial state is shown in Figure 2. In the initial state, the ontology has no concrete classes or properties to link, for example, the VLM ophthalmic grading parameters to the non-CRF variables in the ophthalmic imaging findings model in GDSR. These concrete classes and properties will be automatically created based on the stem part (the dotted concepts and links) once the data standards manager defines the VLM combination schema for the ophthalmic grading parameters in the user interface. The same automation process applies to the VLM of digital measures passive monitoring or any other future VLM the data standards manager would want to maintain.

User interface creation

As stated earlier, converting the VLM combinations from tabular free text to linked metadata while restoring the semantic link to the GDSR collection standards has a lot of Intrinsic value. From a usability perspective, however, unless we also offer the data standards manager the means to easily maintain the VLM content in its new form, we risk jeopardizing the whole value creation process. Fortunately, the GDSR semantic framework offers a path forward here. As outlined in the preliminary section, GDSR allows for a declarative approach for software artifacts development where starting from the ontology and the configurations provided by an information architect, the GDSR engine would take care of generating the necessary API endpoints and user interfaces to consume the metadata downstream.

In the normal GDSR artifacts development process, the user interface generation is only partially automated. The information architect is still responsible for dictating what constitutes a valid permutation to the knowledge graph and

what the expected UI layout and behavior are. Once provided by the architect, the actual UI view generation is then left to the GDSR engine.

In the VLM proposed solution, however, and thanks to the high level of abstraction applied, the remaining configuration parts where the information architect involvement is needed are fully standardized and are thus subject to a systematic automation process. Recognizing the potential for full automation, we adopted the same metaprogramming approach and created two additional SPARQL inference rules. The first rule is a meta SPARQL generic query that automatically generates other VLM-specific SPARQL queries based on the VLM-specific schema. The second rule generates the corresponding SHACL shapes.

The result is a fully automated end-to-end artifacts creation process starting with the data standards manager expressing a new VLM schema definition and ending with fully functional content authoring components.

Knowledge graph population

With the VLM-specific schema and authoring component automatically created, the next automation piece we looked at was how to ingest the legacy content maintained manually in the tabular sheet into the just-created knowledge graph. For that purpose, we implemented a dedicated parser component, an R solution developed with plumber package and deployed as a small REST application, that upon a call from GDSR would:

1. Dynamically read the newly created configurations for the VLM schema of interest.
2. For every value combination (row in the VLM spreadsheet), perform a series of reverse look-up operations in order to translate from the free-text value into the unique identifier of the GDSR-linked entity (the URI of the corresponding standard controlled terminology).
3. In the marginal case where the reverse lookup operation into the standard terminologies is not successful, perform another lookup into the collection of VLM placeholders. A VLM placeholder is a reusable short text with a unique identifier such as “<Leave blank>” or “Protocol specific”.
4. In the case where no VLM placeholder exists beforehand, register a new one but mark it to be reviewed by the data standard manager.
5. Finally, once the translation is concluded, emit the fully transformed content in the form of a knowledge graph serialization and upload it to GDSR.

CONCLUSION AND OUTLOOK

The issues as exposed in the problem statement and illustrative examples thereof - structural inflexibility, manual maintenance, lack of control of distribution, and misalignment between VLM spreadsheets and underlying source of truth (GDSR) - are addressed and eliminated by the proposed solution described in this paper.

We show how users can create new VLM schema configurations as per upcoming needs. In fact, in addition to the demonstrated two examples of ophthalmologic grading parameters and digital measures passive monitoring findings data, a number of other use cases covering scientifically very demanding data types including independent tumor assessments according to several criteria (RECIST, RANO, Lugano, etc.) or microscopic findings from multiplexed immunohistochemistry assays are lined up for new VLM schema configurations.

Being able to rely on fully semantically linked resources between the VLM and GDSR takes the burden off the users to constantly check the metadata repository for updates to both the dataset structures and associated terminology, resulting in a gain of efficiency in the maintenance of VLM of higher quality.

Surfacing up-to-date VLM at any given point in time through the company’s commonly used access point for reference data in a transparent, easily accessible way will prevent users from storing and distributing outdated personal copies of VLM spreadsheets.

Machine-readability of the proposed solution allows the seamless integration of VLM into the DTA design workflow. Time-consuming and error-prone manual copy/paste steps are eliminated.

Further to the non-CRF data workflow, we foresee a benefit of applying a similar approach adapted for the mapping of eCRF forms to SDTM datasets. Given local data collection at clinical trial sites through eCRFs is often driven by the need for human-readability/-operability of the forms and technical limitations of EDC systems, the raw data cannot always follow state-of-the-art data modeling principles. Therefore, the transformation of that data into SDTM datasets can be very complex to describe in a machine-readable format. The base concepts of the flexible schema configuration shown here could hence open an opportunity to facilitate this step.

In summary, the shown solution supports higher-order FAIRification at a lower resource spend in the end-to-end process of VLM creation, DTA design, data tabulation and poolability for data analysis, and it has got the potential to be expanded into other workflows such as eCRF to SDTM transformation.

REFERENCES

1. FAIR Data Standards in the GDSR (Global Data Standards Repository). PHUSE EU Connect 2019. https://www.lexjansen.com/phuse/2019/pp/PP06.pdf
2. Guarino, Nicola, Daniel Oberle, and Steffen Staab. "What is an ontology?." Handbook on ontologies (2009): 1-17.
3. Hitzler, Pascal. "A review of the semantic web field." *Communications of the ACM* 64.2 (2021): 76-83.
4. Hill, Patricia, and John Gallagher. "Meta-programming in logic programming." Handbook of logic in artificial intelligence and logic programming 5 (1998): 421-497.

ACKNOWLEDGMENTS

The work presented in this paper is the result of the collaborative efforts of the authors together with:

- Magdalena Krochmal, Senior Data Scientist, author of the VLM R Parser component
- Igor Steiner, Principal Data Sciences Product Leader, SME in Oncology non-CRF Data Collection Standards
- Zahra Karimaddini, Senior Data Sciences Product Leader, SME in Digital Measures non-CRF Data Collection Standards
- Shine Shen, Principal Data Sciences Product Leader, SME in Ophthalmology non-CRF Data Collection Standards

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the authors at:

selena.baset@roche.com, philipp_thomas.ernst@roche.com, didier.clement@roche.com