

Into the Fire, Linking CDISC & FHIR

D Iberson-Hurst, Assero Limited & A3 Informatics ApS

ABSTRACT

The FDA have announced that “LOINC codes will be required in NDAs, ANDAs, and BLAs for studies that start after March 15, 2020” [1]. Meanwhile talk continues over integrating clinical research with healthcare and interest expressed in Fast Healthcare Interoperability Resources (FHIR), the new Health Level 7’s (HL7) standard.

This paper (and the associated presentation) will report on research encompassing:

- Mappings from CDISC terminology to LOINC to aid in the population of SDTM datasets.
- Generate SDTM datasets from data held in a database with that data being extracted from an Electronic Health Record (EHR) via FHIR.
- Can the data extracted from EHRs be driven by research artefacts such as forms based on biomedical concepts?

and include:

- The approach taken, and technology used.
- How to formulate the EHR requests and process the responses
- How the stored EHR data can be used to generate SDTM datasets.
- How mappings from CDISC to LOINC can be automated.
- How UCUM might be used to automate data conversions

The presentation will conclude with a summary of results, next steps and potential impact.

INTRODUCTION

This paper reports on a prototyping exercise designed to look at a number of aspects in the integration of CDISC standards, the emerging idea of BCs [2] and the FHIR standards being developed by HL7 [3]. The author wished to see how the interaction of BCs and CDISC terminologies deployed in constructing Case Report Forms (CRFs) could be populated using FHIR resources where those resources used terminologies other than CDISC’s.

To do this a set of test data in an FHIR compliant EHR was created and then accessed using the FHIR Application Programming Interface (API). This data is then used to populate a CRF held within a database containing the CRF metadata. Finally the database can then be queried to produce a simple SDTM dataset containing the RAW data captured.

AIMS

The aims of the work were to create a demonstration web application to prototype and demonstrate the aims outlined below and determine the problems and issues in doing so. These learnings were then to be used in subsequent production application work:

- Can we use HL7 FHIR to obtain patient / subject data based on metadata definitions used in CRF definitions that have been created with Biomedical Concepts?
- Can we map terminologies from EHR data using LOINC [8] & UCUM [9] to the equivalent CDISC terminologies as driven by the CRF metadata?
- Store everything in a database as one coherent set of data, both the definitions (metadata) and the data? For this work, we also wanted to evaluate the use of graph data and evaluate the Neo4j graph database product [7]. The database, acting as a simple data warehouse, would hold the data for multiple subjects.
- Extract a SDTM presentation of the data using domain definitions from an MDR?
- Can all of this be automated from the point where the form definition has been created by a human curator, such that the hard work of capturing the data and performing any necessary mappings is performed by the machine?

PhUSE EU Connect 2018

The demonstration was restricted to the use of a single form from the MDR. This approach was taken simply to reduce the amount of development work given that the exercise was a technology demonstrator. The application was also constructed in a manner that allows for the concepts and key points to be demonstrated.

EHR DATA

The most pressing issue in undertaking this work was access to an FHIR compliant EHR system and the data held therein. A web search led to the HSPC consortium [4]. HSPC provides a simple sandbox facility [5] whereby a FHIR compliant EHR dataset can be quickly created and accessed using the FHIR API. The setup process was simple and painless and a dataset of 68 patients was quickly created that could be accessed via the API.

Being test data there are some limitations and a full range of observations may not be available for every subject and multiple observations for a single subject might not be available.

APPLICATION

GENERAL

As stated above the method employed was to develop a prototype web application that would implement and demonstrate the aims detailed above.

ARCHITECTURE

The general architecture is shown in Figure 1. The demonstration consists of the A3 Informatics Glandon Suite Metadata Repository (MDR) (a web application) [6], the demonstration application (a second web application) that is labelled EHR I/F - referred to as the demonstration application for the remainder of this paper - in the figure and the EHR sandbox application provided by HSPC. The demonstration application interfaces to the MDR using a RESTful API and interfaces to the EHR Sandbox application using a FHIR compliant API.

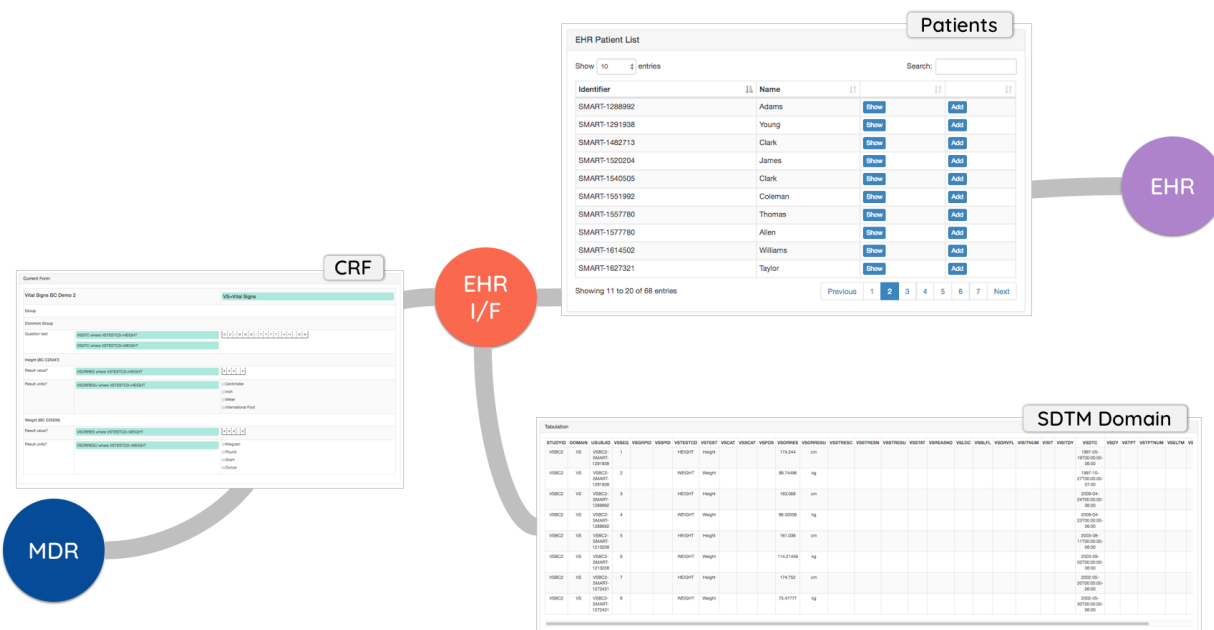


FIGURE 1 - GENERAL ARCHITECTURE

DEMONSTRATION APPLICATION

The demonstration application is a normal web application built using Ruby on Rails. The data for the application is held within Neo4j graph database [7]. The application is divided into four main tabs (screens) and associated views for setting up terminology mappings.

The four main tabs are designed to show the various stages of acquiring and presenting the data and you would not normally expect to see such screens in a production environment. The tabs are:

- selecting a form from the MDR
- acquiring the data
- generating an SDTM domain

PhUSE EU Connect 2018

- display the database graph

On entering the application and the home page, the application initializes itself by reading the set of available forms from the Glandon MDR and the list of patients from the EHR. The user is then able to select a form from the MDR, see Figure 2. The metadata within the form selected will be what the application uses to request data from the EHR.

Figure 2 shows the main application tab. The interface includes a navigation bar with tabs: Study Form, EHR patients, Study Graph, and SDTM Domain. The 'Current Form' section is empty. The 'Available Forms' section displays a search for 'VENTOUX' with two results: 'Ventoux Test Form' version 1.0.0 and 1.1.0. The user is 'Only User [Read, Edit, System Admin]'.

FIGURE 2: MAIN APPLICATION TAB

Once the form is selected, it is displayed in the style of an annotated CRF form within the main application tab. The MDR is responsible for generating the annotations, a task that is achieved by the MDR knowing the relationships between BCs and target SDTM domains. Figure 3 shows a simple form being displayed.

Figure 3 shows the annotated CRF form. The 'Current Form' section displays the 'Ventoux Test Form' with fields for 'Date & time', 'Result', and 'Units'. The 'Available Forms' section shows the search results for 'VENTOUX'.

FIGURE 3: ANNOTATED CRF IN THE MAIN TAB

Having chosen the form, the application expands the database graph by adding in the form metadata and building a study root node to hold the subject data. This metadata defines what needs to be captured from the EHR, based on the definition of the constituent BCs, and is used later by the application for that purpose.

The second tab lists the patients present in the EHR. Each patient has a show button and an add button. The show button just queries the EHR for all of general patient information and displays the response in JSON format. This was added to the demo just to confirm everything was running ok and check FHIR queries. Figure 4 shows the list of patients and a typical FHIR JSON response.

PhUSE EU Connect 2018

Only User [Read, Edit, System Admin]

[Study Form](#)
[EHR patients](#)
[Study Graph](#)
[SDTM Domain](#)

EHR Patient List

Show 10 entries
Search:

Identifier	Name	Show	Add
BILIBABY	Bili	Show	Add
SMART-1032702	Shaw	Show	Add
SMART-1081332	Ross	Show	Add
SMART-1098667	Hill	Show	Add
SMART-1134281	Taylor	Show	Add
SMART-1137192	Williams	Show	Add
SMART-1157764	Hughes	Show	Add
SMART-1186747	Johnson	Show	Add
SMART-1213208	Gracia	Show	Add
SMART-1272431	Graham	Show	Add

Showing 1 to 10 of 68 entries

Previous
1
2
3
4
5
6
7
Next

Patient Data (JSON Format)

```

{
  "resourceType": "Bundle",
  "id": "0b38b6eb-d8e8-41f8-976f-d8a24173117a",
  "meta": {
    "lastUpdated": "2018-09-30T08:34:08.602+00:00"
  },
  "type": "searchset",
  "total": 268,
  "link": [
    {
      "relation": "self",
      "url": "https://api-stu3.hsphconsortium.org/assero/open/Observation?patient=SMART-1272431"
    },
    {
      "relation": "next",
      "url": "https://api-stu3.hsphconsortium.org/assero/open?_getpages=24c014ce-c71d-4344-964f-53145c3001816"
    }
  ],
  "entry": [
    {
      "fullUrl": "https://api-stu3.hsphconsortium.org/assero/open/Observation/SMART-Observation-212-lab",
      "resource": {
        "resourceType": "Observation",
        "id": "SMART-Observation-212-lab",
        "meta": {
          "versionId": "1",
          "lastUpdated": "2017-03-29T11:13:21.000+00:00"
        },
        "text": {
          "status": "generated",
          "div": "
2006-05-09: PSA SerPl-mCnc = 0 ng/mL
"
        },
        "status": "final",
        "category": [
          {
            "coding": [
              {
                "system": "http://hl7.org/fhir/observation-category",
                "code": "Laboratory",
                "display": "Laboratory",
                "userSelected": false
              }
            ],
            "text": "Laboratory"
          }
        ]
      }
    }
  ]
}

```

FIGURE 4: EHR PATIENT TAB

One important point should be noted: As you will see from the screen shots, there are patient names present. Obviously, this is all dummy data. No work has been undertaken regarding patient confidentiality and this aspect was ignored for the purposes of this work.

The add button is a little more specific. It looks at the metadata within the database, extracts the test code from the BCs present and converts it to a LOINC code. It is this conversion that drives the need for the terminology mappings; there is a need to map between CDISC terminology and LOINC because the FHIR observations created within the test data use LOINC coding. Note that FHIR allows other terminologies to be used for the observations. This use of other terminologies is one of the big impediments to integrating the worlds of healthcare and research. The sample extract from one FHIR JSON response shown in Figure 5 indicates the test codes are LOINC while the units for the data values are coded using UCUM.

PhUSE EU Connect 2018

```
"code": {
  "coding": [
    {
      "system": "http://loinc.org",
      "code": "8310-5",
      "display": "temperature",
      "userSelected": false
    }
  ],
  "text": "temperature"
},
"subject": {
  "reference": "Patient/SMART-1272431"
},
"context": {
  "reference": "Encounter/SMART-Encounter-225"
},
"effectiveDateTime": "2002-04-09T00:00:00-06:00",
"valueQuantity": {
  "value": 37.55556,
  "unit": "Cel",
  "system": "http://unitsofmeasure.org",
  "code": "Cel"
}
},
```

FIGURE 5: TYPICAL FHIR JSON RESPONSE

Once the CDISC Test Code is mapped to a LOINC code, the EHR can be queried. The request to the EHR is made using a FHIR query for all observations matching the LOINC code. When multiple results are returned, a very simple method of using the first result from the set returned is used. Obviously, this is not a very sophisticated approach, but this aspect is not the focus of the work. This issue is not an insignificant one. The choice of which observation to use for the purposes of research is very much related to study designs and was considered outside of the scope of this work.

The chosen data is then extracted from the FHIR response and used to populate the database. The approach taken is to copy the BC for the patient (now termed a subject) and attach the data at the correct place within the BC (result, result units, date and time of observation etc). To do this the data from the FHIR resource needs to be extracted so mappings between BC patterns and FHIR resource patterns are maintained. This approach worked well. Ideally, in the future, we want the artefacts, BCs and FHIR resources, to be the same.

The process of adding patients / subjects can then be repeated adding further data to the database.

After adding all the desired patients / subjects we end up with having a database containing the subject data with the data all being linked to BCs and form metadata definitions. We effectively end up with a mini data warehouse.

PhUSE EU Connect 2018

FHIR/SDTM Demonstration (dev): Ventoux (v0.4.0)
Only User [Read, Edit, System Admin]

Study Form
EHR patients
Study Graph
SDTM Domain

Domain List

Show 25 entries

Search:

Identifier	Label		
CE Domain	Clinical Events	Add	Tabulation
CM Domain	Concomitant Meds	Add	Tabulation
DM Domain	Demographics	Add	Tabulation
DS Domain	Disposition	Add	Tabulation
EQ Domain	Electrocardiogram	Add	Tabulation
FA Domain	Findings About	Add	Tabulation
LB Domain	Laboratory	Add	Tabulation
PE Domain	Physical Exam	Add	Tabulation
PE Domain	Physical Exam	Add	Tabulation
QS Domain	Questionnaires	Add	Tabulation
VS Domain	Vital Signs	Add	Tabulation

Showing 1 to 11 of 11 entries

Previous
1
Next

Tabulation

STUDYID	DOMAIN	USUBJID	VSSEQ	VSGRPID	VSSPID	VSTESTCD	VSTEST	VSCAT	VSSCAT	VSPOS	VSORRES	VSORRESU	VSSTRESC	VSSTRESN	VSSTRESU	VSSTAT	VSREASND	VSLOC	VSLAT	VSBLFL
VENTOUX	VS	VENTOUX-SMART-981968	1			TEMP	Temperature				36.05556	C								
VENTOUX	VS	VENTOUX-SMART-99912345	2			TEMP	Temperature				37.0	C								
VENTOUX	VS	VENTOUX-SMART-9995679	3			TEMP	Temperature									NOT DONE	NOT FOUND IN EHR			
VENTOUX	VS	VENTOUX-SMART-1272431	4			TEMP	Temperature				37.55556	C								

The final challenge is to generate an initial SDTM domain and the fourth tab (screen) is designed to do this. The tab contains a list of user domains from the MDR and these domain definitions hold the association to the BCs. This association allows for the variables within a domain to be automatically associated with the data within the BCs. The add button performs this linking operation, reading the definitions from the MDR and then automatically linking the variables with all of the applicable data points. Once this has completed the tabulation can be generated via a single query and a small amount of presentation code.

Note that the association of domain into the database is done manually using the add button for demonstration purposes. This could be automated as the link between Form to BC to Domain is known by the MDR.

The population of the domain is sparse at the moment. No attempt has been made to populate the timing variables but variables such as --CAT, --SCAT, --METHOD, --SPEC, --LOC and --POS etc would come from the BC definitions and captured data if they were present in the BC.

TERMINOLOGY MAPPINGS

As noted above there is a need for the application to manage and process terminology mappings:

- UCUM -> CDISC
- LOINC -> CDISC

The UCUM to CDISC mappings are straightforward in that they are 1-1. The LOINC mappings are less straightforward. Some are 1-1 while others such as the laboratory tests are N-1. Therefore, a graph design was used that could accommodate these needs.

For the complex LOINC mappings the approach as seen in Figure 6 was used

PhUSE EU Connect 2018

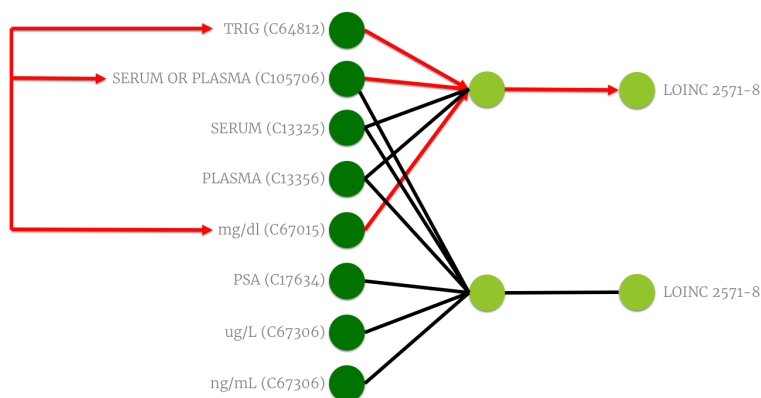


FIGURE 6: MULTIPLE TO ONE MAPPING

Here several CDISC terms are mapped to individual LOINC terms via a mapping node. Note that two mappings overlap in that the same CDISC term can be linked to several LOINC terms. However, it is the combination of CDISC terms that is important. One combination should only match one LOINC term as detailed in Figure 6. Also note that the graph is one way in that the CDISC term combination can be mapped to the LOINC term but a LOINC term cannot be uniquely mapped to a single set of CDISC terms. This is because in the cases noted in the figure the LOINC term does not explicitly state the specimen type, it might be SERUM, PLASMA or SERUM OR PLASMA

The mappings used within the application were loaded from a draft version of a spreadsheet created by the CDISC terminology team that mapped CDISC laboratory terms to LOINC terms. This was an invaluable resource.

GRAPHS

Because the application is a demonstration application the graph visualization screen was added to allow for the internal workings of the application to be viewed to allow for discussion. A straightforward way of achieving this is to provide a visualisation of the actual graph held within the Neo4j database. These visualizations are colour coded using a consistent notation as detailed in Figure 7.

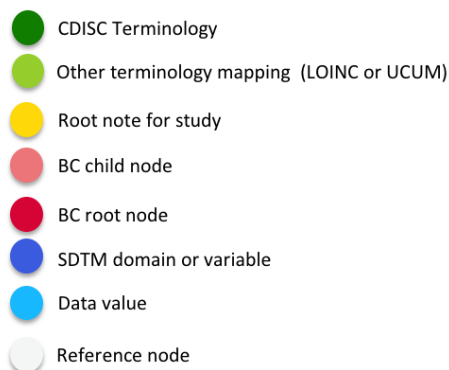


FIGURE 7: KEY FOR GRAPH NODES

Figure 8 shows an example graph after selecting and loading a form. Note we have the central study node linked to the two BCs that are then built up from a set of child nodes (based on BRIDG classes and attributes plus ISO 21090 datatypes) that eventually arrive at the leaf nodes. These are then connected to potential values if they are coded responses and linked to CDISC terminology, in this case the values are the units for Weight and Height. These are then linked to the UCUM equivalent values, the mapped terms.

PhUSE EU Connect 2018

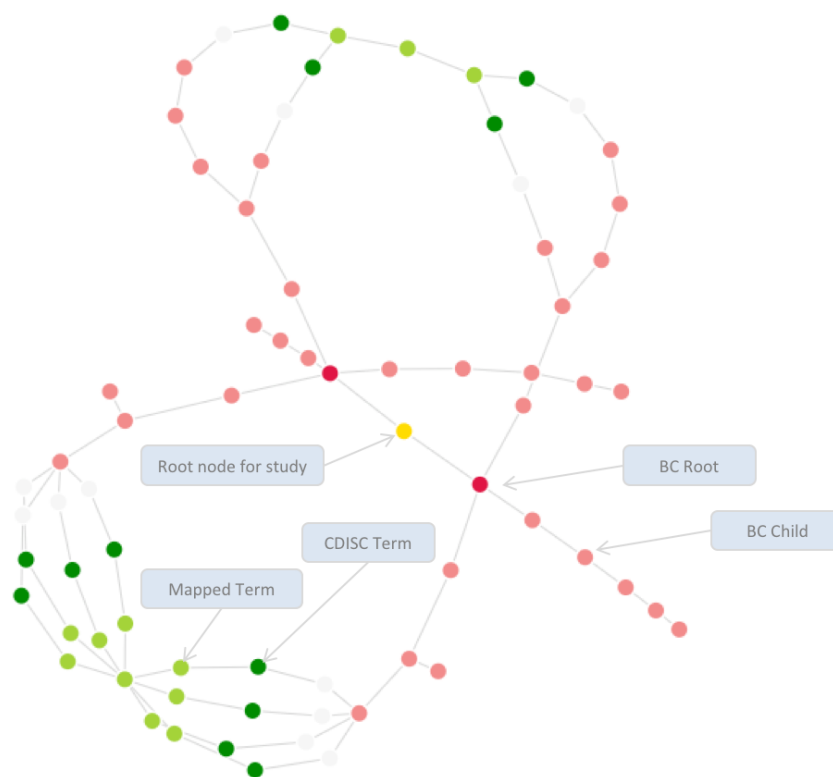


FIGURE 8: FORM AND TERMINOLOGY METADATA

Figure 9 shows the graph after the addition of the data for a single patient or subject. We see the addition of the nodes at the centre for the subject attached to a patient and then a copy of the BCs (cloned from the study node). The BCs are currently replicated but other mechanisms could be employed to simplify the graph. The current approach was taken to speed the implementation and for no other reason, but the approach does create more nodes. At the leaf nodes you see the data points attached to the relevant BC nodes. At the top of the figure is a BC node with four answers (CDISC terms) but one data node indicating the answer extracted from the EHR and linked to the UCUM term.

PhUSE EU Connect 2018

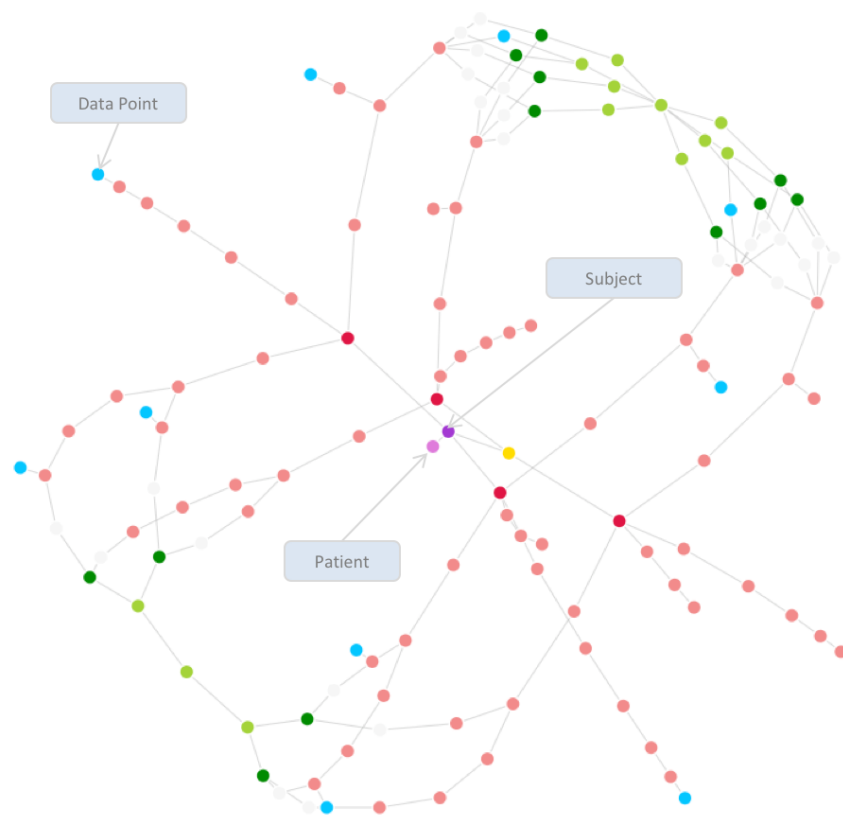


FIGURE 9: METADATA AND DATA

Figure 10 shows a graph after the linking in of a SDTM domain. The metadata in the MDR contained within the domain and BCs allows this to be an automated process. The variables are linked to the BC leaves which in turn have the data values attached. These links then allow the domain to be generated and the values to be placed into the correct place.

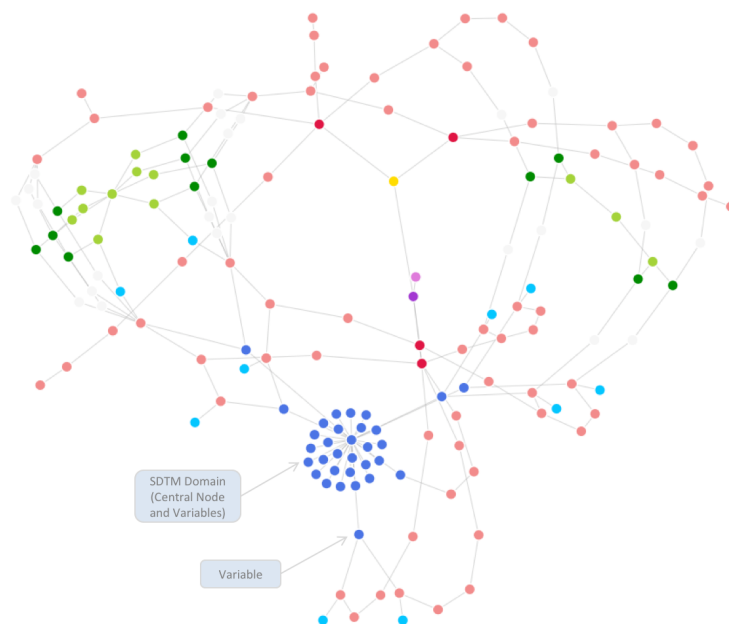


FIGURE 10: DOMAIN METADATA

PhUSE EU Connect 2018

RESULTS AND CONCLUSION

The work met all of the aims, in that data can be extracted from the EHR and used to generate the base SDTM domain. In terms of the effort required to meet the aims it has been very successful with the initial implementation taking three to four days to do this from start to finish but was restricted to handling 1-1 terminology mappings. The N to 1 mapping work took a few more days to implement.

One huge advantage was working from solid definitions for the CDISC Terminology, BCs and Forms. These were all accessible from the MDR via the API. The HL7 FHIR standard was easy to get to grips with and the integration with BCs was readily achieved. Moving them closer to each other is an obvious step for the industry. Generating the domain was also relatively easy.

It is accepted that the work covers some of the easier use cases, but it can already be seen how it can be extended to finding domains in a generic manner. The timing variables will be fun but is actually a different problem, in a way it's the study design problem and how study design relates to observations. The work can be extended to the other domain classes but, as ever, iterate, learn, adjust and repeat.

It can also be seen how the derived variables such as –STRESC, –STRESN –STRESU could be generated automatically, especially given the fact that we have UCUM units involved and leveraged to allow this to be nicely automated. The timing variables will be harder, but I feel many of these should be driven by protocol and study design elements thus driving the need for better study specifications and machine-readable artefacts there.

Pooling data from several studies would be straightforward. Adding more data from another source that employs matching BC definitions simply results in any query returning more results. It also becomes rapidly apparent that placing all the study metadata and data together in a warehouse as linked data makes immense sense. From this store SDTM can be created but much else with it.

NEXT STEPS

The next steps for the work are:

- Extend the range of measurements and observations that can be extracted from the EHR.
- Create the SDTM derived variables automatically.
- Use UCUM to automate data conversions to create SDTM standard result fields etc.

REFERENCES

[1] Federal Register Volume 82, Number 202 (Friday, October 20, 2017). Notices. Pages 48827-48828
[accessed 2018 Sep 28]
<https://www.gpo.gov/fdsys/pkg/FR-2017-10-20/html/2017-22768.htm>

[2] Langendorf. PhUSE US Connect 2018, SI19. Easing Your Pain with Biomedical Concepts
[accessed 2018 Sep 28]
https://phusewiki.org/docs/2018_US%20Connect18/SI%20STREAM%202/si19%20final%20.pdf

[3] Health Level 7 Fast Healthcare Interoperability Resources.
[accessed 2018 Sep 28]
<https://www.hl7.org/fhir/>

[4] HSPC Consortium.
[accessed 2018 Sep 28]
<https://www.hspconsortium.org>

[5] HSPC Sandbox facility.
[accessed 2018 Sep 28]
<https://www.developers.hspconsortium.org/build>

[6] A3 Informatics Glandon Suite MDR.
[accessed 2018 Sep 28]
<https://www.a3informatics.com/tools/mdr/>

PhUSE EU Connect 2018

[7] Neo4j Graph Database
[accessed 2018 Sep 28]
<https://neo4j.com>

[8] LOINC
[accessed 2018 Sep 28]
<https://loinc.org>

[9] The Unified Code for Units of Measure
[accessed 2018 Sep 28]
<http://unitsofmeasure.org/trac>

CONTACT INFORMATION

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

Dave Iberson-Hurst
A3 Informatics ApS
Lille Strandstræde 20C 5.
DK-1254 Copenhagen K,
Denmark

Email: diH@a3informatics.com
Web: www.a3informatics.com

Brand and product names are trademarks of their respective companies.