PDXIntegrator Documentation

Release 1

Peter Robinson

Mar 26, 2019

Contents:

1 Setting up PDXIntegrator **1**
- 1.1 Setting up drugbank . . . 2
- 1.2 Running the simulation command . . . 2

2 Drugbank **3**
- 2.1 Our usage of drugbank . . . 3

3 NCIT **5**
- 3.1 Parsing the NCIT ontology file with OWL-API . . . 5

4 Clinical/Patient Module **7**

5 Clinical/Tumor module **11**

6 Model creation module **13**

7 Model quality assurance **15**

8 Model study module **17**

9 Demonstration of simulating cases and performing SPARQL queries **19**
- 9.1 SPARQL Queries . . . 20
- 9.2 Development plans . . . 21
- 9.3 Visualization . . . 22

10 PDXIntegrator **23**

CHAPTER 1

Setting up PDXIntegrator

PDXIntegrator is a Java program that requires at least Java version 8.

Note that for now, we are using a private fork of the ontolib library for parsing the NCIT obo file. Go to https://github.com/monarch-initiative/OLPG and enter the following commands (notify Peter if you need access).

```
$ git clone https://github.com/monarch-initiative/OLPG
$ cd OLPG
$ mvn install
```

This will put the OLPG library files into your local maven repository (.m2 directory) and let you build PDXIntegrator.

PDXIntegrator is provided as a maven project. The quickest way to set it up is to clone the project from the GitHub site at https://github.com/TheJacksonLaboratory/PDXintegrator and then to use maven to build the project. In the following, we show a command that will display a help message–if you see this, then you have successfully built the program.

```
$ git clone https://github.com/TheJacksonLaboratory/PDXintegrator
$ cd PDXintegrator
$ mvn package
$ java -jar target/PdxIntegrator.jar

[ERROR] no command

Program: PdxIntegrator (Common Knowledge Graph PdxModel for PDXNet)
Version: 0.0.2

usage: java -jar PdxIntegrator.jar command [-c <arg>] [-d <arg>] [-o <arg>]
Available commands:

download
    java -jar PdxIntegrator.jar download [-d directory]: Download NCI files to␣
↪directory at -d (default="data").

simulate
    java -jar PdxIntegrator.jar simulate [-d directory]: Requires NCI files in␣
↪directory at -d (default="data").
```

(continues on next page)

(continued from previous page)

```
map
    java -jar PdxIntegrator.jar map [-d directory]: todo.
```

1.1 Setting up drugbank

Before running the `simulation` command, the Drugbank XML file needs to be downloaded and processed. See drugbank.

1.2 Running the simulation command

Currently, we are building out a complete RDF model for the PDX Minimal Information standard (https://www.ncbi.nlm.nih.gov/pubmed/29092942). Our strategy for development is to generate random PDX cases, write the corresponding RDF code to file (with the `simulate` command). Then, to test the query-ability of the model, we ingest the RDF file and query it using SPARQL queries (with the `query` command). This is inteded to allow collaborators to view the model and make suggestions for improvement. The end game would be to develop this code to allow ETL of PDX data into this model. The following sections show how to set up and run the simulation

CHAPTER 2

Drugbank

We propose to use the Drugbank resource to track medications. According to the Drugbank website (2018-02-26), The latest release of DrugBank (version 5.0.11, released 2017-12-20) contains 10,999 drug entries including 2,504 approved small molecule drugs, 942 approved biotech (protein/peptide) drugs, 109 nutraceuticals and over 5,108 experimental drugs. A recent Nucleic Acid Research database article provides further details. Drugbank is licensed under a a Creative Common's Attribution-NonCommercial 4.0 International License, and therefore can be used freely for non-commercial applications.

2.1 Our usage of drugbank

We used the Java JAXB tool `xjc` to create Java classes that mirror the XSD XML schema definition for Drugbase (the schema is available here: https://www.drugbank.ca/releases/latest). To use PDXIntegrator, you will need to download the Drugbase XML file (which requires free registration). Then follow these steps to unpack the XML file and to remove the space character from the file name.

```
$ unzip drugbank_all_full_database.xml.zip
    Archive:  drugbank_all_full_database.xml.zip
    inflating: full database.xml
$ mv full\ database.xml fulldatabase.xml
```

Now we can extract the contents of the XML file for use by PDXIntegrator. The following command does that and creates a new file in the `data` directory that is also used by PDXIntegrator to store downloaded files.

```
$ java -jar target/PdxIntegrator.jar drugbank --drugbank fulldatabase.xml
```

Adjust the path to `fulldatabase.xml` as necessary.

This will generate a new file in the data directory called drugbank.tab. The contents of this file look like this.

```
Denileukin diftitox     173146-27-5     DB00004 ANTINEOPLASTIC AND IMMUNOMODULATING␣
↪AGENTS
Etanercept      185243-69-0     DB00005 ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS
Bivalirudin     128270-60-0     DB00006 BLOOD AND BLOOD FORMING ORGANS
```

(continues on next page)

(continued from previous page)

```
Leuprolide      53714-56-0      DB00007 ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS
Peginterferon alfa-2a   198153-51-4     DB00008 ANTINEOPLASTIC AND IMMUNOMODULATING␣
↪AGENTS
Alteplase       105857-23-6     DB00009 BLOOD AND BLOOD FORMING ORGANS
Sermorelin      86168-78-7      DB00010 SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX␣
↪HORMONES AND INSULINS
Interferon alfa-n1      74899-72-2      DB00011 ANTINEOPLASTIC AND IMMUNOMODULATING␣
↪AGENTS
```

For the purposes of this demonstration program, we will use the antineoplastic agents. ToDO–if we choose to stick with DrugBank, we can exploit the hierarchy and will need to emit some RDF to represent the hierarchy that is recorded in the XML file (for now, we will treat these medications as literals for demo purposes).

CHAPTER 3

NCIT

We will use the NCIT thesaurus to provide IDs for the concepts in PDXnet. /

3.1 Parsing the NCIT ontology file with OWL-API

The National Cancer Institute (NCI) has developed a thesaurus (NCIT) of many kinds of cancer-relevant data. The Monarch Initiative has developed a version of this ontology as an OBO file. This Wiki provides details of this effort.

PDXIntegrator will download the OBO file with the following command.

```
$ java -jar target/PdxIntegrator.jar download
```

The ncit.obo file will be stored in a newly created directory called `data`. If the file is already present, PDXIntegrator will emit a warning message (delete the file if you want PDXIntegrator to download a fresh copy of the file).

PDXIntegrator uses the code in the class `NcitOwlApiParser` to parse the Neoplasm terms of NCIT (only). Currently, it only saves the IDs and the labels (although it may be useful in the future to also ingest the synonyms in case this code will be used to drive an autocomplete. This is sufficient to cause the simulated patients to have real NCIT diagnosis codes for the cancers.

For the SPARQL queries, it will be necessary to include the subclass definitions of NCIT to enable queries that use the NCIT hierarchy.

CHAPTER 4

Clinical/Patient Module

The PDX-MI provides a clinical module that is divided into two submodules: "clinical/patient" and "clinical/tumor." "Clinical/patient" requires information about the patient from which the engrafted tumor originates, including age, sex, ethnicity, and disease diagnosis.

Field	Rec	Description	Example	PDXNet
Submitter Patient ID	D	Unique ID. Tumor id if not supplied	PAT-123	CenterID:PatientID
Sex	D	Patient Sex	female	NCIT:C16576
Age at diagnosis	D	Age at diagnosis	30	binned in 5 year age groups
Age at collection	D	Age at specimen collection	30	binned in 5 year age groups
Submitted Diagnosis	E	Initial Clinical Diagnosis	invasive breast cancer	invasive breast cancer
Diagnosis	E	Initial Clinical Diagnosis Term	invasive breast cancer	NCIT:C6257
Consent to share	E	Patient Consent to share data		PDXNET:consent_ACADEMIC_ONLY
Ethnicity	D		caucasian	currently string literal
Race	D			currently string literal
Current treatment drug	D		everolimus	CHEMBL83
Current trtmnt protocol	D		afinitor;10 mg/day	
Prior treatment protocol	D		afinitor;10 mg/day	
Prior treatment response	D			RECIST
Virology Status	D		HIV-/HBV-/HCV+/HTLV-/EBV+	NCIT

Table 1. Rec: Recommendation; E: essential; D:desirable. Desireable fields will be shown as 'Not Reported' if no

data is provided.

1. **Submitter Patient ID**. Used when possible for identifying multiple models from the same patient. Patient 123 from JAX would be shown on the PDXNet website as JAX:PAT-123. Will use supplied PDXNet abbreviations to prefix any ID. If no ID is supplied, the ID will be based on Tumor ID.

2. **Sex**. Patient sex. We will use this field to record biological sex. We will use the NCIT terms:

- Female (Code NCIT:C16576): A person who belongs to the sex that normally produces ova. The term is used to indicate biological sex distinctions, or cultural gender role distinctions, or both.
- Male (Code NCIT:C20197): A person who belongs to the sex that normally produces sperm. The term is used to indicate biological sex distinctions, cultural gender role distinctions, or both.

3. **Age at Diagnosis**. Patient age at time of initial clinical diagnosis in years. For display : to reduce the possibility of patient identification, PDX-MI recommends grouping ages into 5-year groups, although more granular groupings may be used in cases such as pediatric tumors if approved by a contributor's Institutional Review Board. Here, we have implemented binned age groups as follows.

```
PDXNET:PAT-1511  PDXNET:ageBinLowerRange 55 ;
    PDXNET:ageBinUpperRange      59 ;
    (...) .
```

In the simulation code, we simulate using 5 year bins, but any ranges could be used in real code. For now, the age is understood to be in years, and if there is a need to be more precise we would need to change the model. **TODO** it may be better to include the word "year" in the predicate name?

Note that there is a mistake in the PDX-MI, which uses a six year range instead of a five year range: 30–35 (binned in 5-year age groups)

4. **Age at Collection**. Patient age when specimen was collected in years. Will be binned for dispaly as above.

5. **Submitted Diagnosis**. The initial clinical diagnosis provided as free text.

6. **Diagnosis**. Initial clinical diagnosis. Submitted diagnosis mapped to NCIT term Note this represents the initial diagnosis and may be less precise than the histological diagnosis used in the second module. We will take the diagnosis codes from NCIT. The following shows an example triple for an individual with a diagnosis of Central Nervous System Histiocytic Sarcoma (NCIT:C129807).

```
PDXNET:PAT-1511 PDXNET:hasDiagnosis   NCIT:C129807 .
```

7. **Consent to share data**. Patient consent. Reporting on consent is essential. We are using the following codes.

- PDXNET:consent_NO
- PDXNET:consent_YES
- PDXNET:consent_ACADEMIC_ONLY

TODO 1 define what we mean by these categories and add more categories as necessary. **TODO** 2. Define relation to NCIT term for Consent (NCIT:C25460).

8. **Ethnicity** Patient ethnicity. **TODO** we need to decide which reference terminology to use. One option is to adopt the NCIT terminology, which includes

- Ethnic Group (NCIT:C16564) (which includes concepts such as Hispanic etc.)

9. **Race** Patient race. **TODO** we need to decide which reference terminology to use. One option is to adopt the NCIT terminology, which includes

- Race (NCIT:C17049) (with multiple subclasses specifying various populations)

10. **Current treatment drug** Patient treatment at time of specimen sample. We would like to have a resource to that represents classes, ingredients, brand names, dosage forms, etc., in a computable manner. There are several contenders, including ChEMBL and RxNorm (a standardized drug nomenclature maintained by the National Library of Medicine), but I think that DrugBank is the best option because it is comprehensive, it combines detailed drug data with comprehensive drug target information, it has an open source (Creative Common's Attribution-NonCommercial 4.0 International License) license, and it is easy to use. A newer resource DrugCentral can also be used to map between many resources. **TODO** this is an issue that will require thought and consensus building. Please communicate ideas/comments to Peter. Currently, the PDXNet simulation is showing data from DrugBank as a literal (String).

```
PDXNET:PAT-248 PDXNET:currentTreatmentDrug  "Leuprolide[DB00007;53714-56-0]" .
```

The String currently shows the name (Leuprolide), the DrugBank ID (DB00007), and the CAS id (53714-56-0). If we decide to go with DrugBank, then probable the triple should be formed like this.

```
PDXNET:PAT-248 PDXNET:currentTreatmentDrug  drugbank:DB00007 .
```

By adding other information from DrugBank to the RDF data available in our query engine, it would be possible to formulate expressive queries about PDX models that have been treated by drugs that correspond to some overall treatment category (e.g., Leuprolide corresponds to L02AE - Gonadotropin releasing hormone analogues), have certain indications (e.g., Leuprolide is indicated for Advanced Prostate Cancer), interact with certain drugs (e.g., Allicin; The therapeutic efficacy of Allicin can be decreased when used in combination with Leuprolide), etc.

11. **Current treatment protocol (dose; details)** There is currently no ontology that I know of for representing dosages. There are many ways of representing dosages, e.g., 10 mg/day or 5 mg b.i.d. **TODO** discuss what methodology would work best for PDX centers.

12. **Prior treatment protocol** The medication data should be represented as above. The surgery data could be represented using MedDRA codes (a rich and highly specific standardised medical terminology to facilitate sharing of regulatory information internationally for medical products used by humans), but MedDRA does not have an open license and it may be difficult to reuse/redistribute, and so if we want to use MedDRA we would need to come to an agreement with them. MeSH would be an option, although MeSH is not always ontologically well structured, but there are a large number of terms. The NCI thesaurus has a hierarchy of terms for Intervention or Procedure, including Cancer Diagnostic or Therapeutic procedure, including terms for operations such as Mastectomy (NICT:C15277).

    This is probably sufficient for our needs, and I would suggest we use this.

TODO – decide if the NCIT codes are sufficient for our needs. I suggest that we examine the subhierarchy underneath the term Cancer Diagnostic or Therapeutic Procedure (Code C79426).

13. **Response to prior treatment** progressive disease (RECIST1.1) These items can be represented in the NCIT, which has a subhierarchy for Clinical Course of Disease (Code C35461), which includes items such as "Complete remission", "Progressive disease" and many more. Currently, the PDXIntegrator uses the following five terms

- notAssessed
- completeResponse
- partialResponse
- stableDisease
- progressiveDisease

TODO Decide on whether we want to limit this category to a small number of terms (like the above), to allow any term from the NCIT Clinical Course of Disease subhierarchy, or choose some other scheme. Currently, I am using the PDXNET namespace for these terms in the RDF code, but we should use the NCIT namespace once we have decided where to take this.

14. **Virology status** Probably the NCIT subhierarchy of Viral infection (Code C3439), (which includes these viruses and many more) would be best. We can represent this in RDX using a scheme such as this.

```
PDXNET:PAT-248 PDXNET:virologyStatus  NCIT:C141405 .
```

where NCIT:C141405 is the code for Hepatitis B Virus Positive (Code C141405). Note that we may either want to use the terms for virus infection (which is a clinical diagnosis) or for serology (as in this example, with the term coming from the Laboratory Finding subhierarchy of NCIT). It depends on how we want to model this. **TODO** Determine the terminology and the depth of detail we want to capture.

CHAPTER 5

Clinical/Tumor module

The following table shows the recommendations from the PDX-MI manuscript.

Field	Rec	Description	Example	PDXNet
Submitter Tumor ID	E	Tissue ID	TUM-123	Unique ID
Primary Tumor Tissue	E	Primary Tumor Tissue	breast	UBERON code
Primary, Met, Recurrence	E	Disease Progression	Recurrence	enumeration
Specimen Tumor Tissue	E	Sampled Tissue	breast	UBERON code
Tissue Histology	D	Histologic Diagnosis	invasive ductal carcinoma	NCIT code
Tumor Grade	D	Tumor Grade	grade 3	AJCC Grade
Stage; T N M	D	Tumor Stage		AJCC TNM Stages
Diagnostic Markers	D	Clinical BioMarkers	ER+, PR+, HER2+;	
Treatment Naive Patient	D		yes/no	enumeration
Tumor Sample Type	D	Collection Procedure	biopsy	enumeration
Subline of		Subline of model	PDX-123	Model Identifier
Subline reason		Why a subline	Lost Cisplatin Resist.	String

1. **Submitter Tumor ID** Display as CenterID:TumorID to act as a primary key. If not provided, patient ID can be used. Tumor 123 from JAX would be shown on the PDXNet website as JAX:TUM-123

2. **Primary tumor tissue** The tissue of the primary tumor. For now, the simulation shows a random UBERON code for an anatomical entity. We will use the uberon cross-species anatomy ontology [8] that is developed by Monarch Initiative (M. Haendel, C. Mungall).

3. **Primary, metastasis,recurrence** Is the specimen tissue from the primary tumor, a metastasis or a recurrance For now, we are using PDXNet entities, but we should use the NCIT terms for these items. This would allow users to enter a more specific NCIT term such as Distant metastasis (C18206), which is a child of Metastasis (C19151)

4. **Specimen tumor tissue** Tissue from which the specimen was collected. Same as Primary tissue if the tumor is not metastatic. Uberon as above. **TODO** For melanoma do we want to capture specimen location?

5. **Tissue histology** This is the pathologist's diagnosis and may often represent a refinement of the clinical diagnosis given in the Patient/Clinical module. Should use the same terminology as diagnosis, but represent the pathologist's findings.

6. **Tumor Grade** For now we are using PDXNet codes, but we will switch to the NCIT subhierarchy, although I think they may need some TLC. We will work with NCIT to revise these terms as a part of Monarch's ongoing collaboration with NCIT.

7. **Disease Stage; classification** T3N2M1; TNM or Non applicable (example blood cancer) Should follow Tumor Grade; classification standard Use AJCC. This will be seperated into pT,pN,PM and stage

8. **Specific markers (diagnostic linked)** Clinically relevant bio markers. Pairs of Marker:Status where status can be "positive", "negative", a percent value, or a variant and optional platform Most of the assays such as IHC are covered by the NCIT under the subhierarchy "Laboratory Procedure". That NCIT subhierachy also includes items for Receptor status (e.g., HER2/Neu positive), and these will be linked to external representations of genes/proteins by the Monarch collaboration.

9. **Treatment naive patient?** yes/no (enumeration) Yes means patient has never had neoadjuvant treatment. per TCGA

10. **Original tumor sample type** The process used to collect the sample. biopsy, surgical sample, punch NCIT has a subhierarchy of terms for biopsy and biopsy locations, that will be linked to uberon etc by the Monarch collaboration. Some terms appear to be missing, e.g., "ascites fluid", but will be added to NCIT as needed for PDXNet.

11. **Subline of** If this model is created as a subline of an existing model indicate which model it is a subline of.

12. **Subline reason** We may need to create our own mini-terminology to describe the reasons for using a subline

CHAPTER 6

Model creation module

The following table shows the recommendations from the PDX-MI manuscript.

Field	Rec	Description	Example	PDXNet
Submitter PDX ID	E	Unique ID for the PDX Model	PDX#123	CenterID:PatientID
Passage	E	Tissue Passsage. P0 tissue is from PT	P1	create PDXNet classes?
Mouse strain	E	Name of strain used for engraftment	NOD.Cg-Prkdc<scid> Il2rg<tm1Wj>l/SzJ	create PDXNet classes?
Mouse source	E	Institution supplying strain	The Jackson Laboratory	create PDXNet classes?
Mouse sex	D	Sex of mouse used for engaftment	Male	
Mouse immune system humanised?	E		yes/no	
Type of humanisation	E		CD34+hematopoietic stem cell-engrafted	
Engraftment Procedure	D		suspension	?
Engraftment Method	D			?
Engraftment Site	D		right flank	?
Mouse treatment for engraftment	D		estrogen treatment	?
Engraftment rate	D		80%	literal
Engraftment time	D		8 weeks	literal

1. **Submitter PDX ID**. This field is analogous to PatientID. We display as CenterID:PDXID and keep an internal ID that will not be shown externally to act as a primary key. For instance, PDX 123 from JAX would be shown on the PDXNet website as JAX:PDX-123

2. **Passage**. We neet to have a standard for this. Suggesting that P0 would indicate engrafted tissue is from the patient. P1 tissue comes from P0. P2 from P1 etc.

3. **Mouse Strain**. We will allow strains to be denoted according to the MGI guidelines.

4. **Mouse Source**. Institution providing the strain. An enumeration.

5 **Mouse Sex**.

6. **Strain Immune System Humanised?** Yes or No

7. **Type of Humanisation**. Description of humanisation method.

8. **Engraftmet Procedure**. Enumeration (**TODO** Need input: list of all the methods and whether any vocabulary exists?)

9. **Engraftment Method**. Enumeration (**TODO** Need input: list of all the methods and whether any vocabulary exists?)

10. **Engraftment Site**. Subcutaenous, Mammary Fat Pad, Orthotopic, etc. Where the tissue is engrafted to the mouse.

11. **Mouse Treatment for Engraftment**. estrogen treatment Enumeration (**TODO** Need input: list of all the methods and whether any vocabulary exists?)

12. **Engraftment Rate**. 80%. Would it be better to state "n of m" rather than a percentage? **TODO** Use percent or N of M?

13. **Engraftment Time**. 8 weeks Number of weeks/days PDMR uses "Estimated days from implant to 500 mm3" Both Rate and Time would benefit from noting if tissue was cryopreserved. Should this be captured as a field?

CHAPTER 7

Model quality assurance

The model quality assurance module captures information about tissue provenance and fidelity of the passaged tumor with respect to key characteristics of the patient tumor.

The following table shows the recommendations from the PDX-MI manuscript.

In terms of RDF modeling, we will add these items as properties of the Pdx Sample.

Field	Rec	Example	PDXNet
Quality control method(s)	E	histology and IHC	create PDXNet classes?
Quality control results	E		Text
Animal health status	D	SPF/SOPF	
Passage QA performed	D	P4	Integer

Table 2.4. Model creation Q/A module. Rec: Recommendation; E: essential; D:desirable.

1. **Quality control method TODO** Need to get a set of QC methods.

2. **Quality control results** Description of the results of the QA/QC method

3. **Animal health status** Pertains to the status of mouse room models are kept in.

4. **Passage QA performed** The passage or passages on which QA was performed. As models are repeately passaged QA status may change. If QA/QC is done on multiple passages multiple QA sections can be added

CHAPTER 8

Model study module

Tumors from PDX often undergo comprehensive genomic characterization and/or treatment in controlled dosing studies to define therapeutic response and resistance. PDX-MI includes desirable fields in the reporting of these studies that supplement existing guidelines for reporting on in vivo biomedical research (Meehan et al., 2017).

Field	Rec	Example	PDXNet
Study name or identifier		PDX-123P3 Pertuzmab/Trastuzumab	Needs to be unique to attach files
Treatment	D	pertuzumab in combination with trastuzumab; CHEMBL2007641	List of 1 or more generic drugs
Treatment protocol	D	trastuzumab (30 mg/kg loading dose, 15 mg/kg weekly);	Additional module to capture Drug,Dose,Route,Frequency
Treatment Response	D		RECIST Term
Passage	D	P2	Integer
Metastasis	D	Yes	Yes or No
Metastasized to	D	Liver	Uberon
Metastasis in passage	D	P3	Integer
Lag time/doubling time	D	48h	separate elements

Table 2.5. Model study module. Rec: Recommendation; E: essential; D:desirable.

1. **Study Name or identifier** (if not available could be model + passage + treatment) A way to uniquly identify the study. Human readable or unique ID

2. **Treatment** List of treatment(s) as generic terms for medications.

3. **Treatment protocol** For each treatment drug provide drug name, dosage, route and frequency.

4. **Treatment response** complete response, partial response, stable disease, progressive disease.

5 **Model Passage** Passage(s) of models used in study. Assumption is P0 modles have tissue directly from patient. P1 is engrafted with tissue from P0 etc.

> **Tumor OMICS**: This was removed. It will be populated based on types of files uploaded for Study/Model.

6,7,8 **Development of metastases in strain** We will code this as Yes/no; site as uberon; passage as enumeration

9. **Lag time/doubling time of tumor** We will code this as the number of hours.

CHAPTER 9

Demonstration of simulating cases and performing SPARQL queries

This page demonstrates how to run the demonstration query.

First we run the simulate command of PDXIntegrator to produce an RDF file with "randomized" cases. By default, 5 random cases are produced.

```
$ java -jar target/PdxIntegrator.jar simulate
```

This will produce a file called `simulatedCases.rdf` in the current working directory. This file will use the RDF/XML format, but the program will also emit the same RDF data in the Turtle format. For instance,

```
@prefix PDXNET: <http://pdxnetwork/pdxmodel#> .
@prefix NCIT:  <http://purl.obolibrary.org/obo/NCIT#> .
@prefix UBERON: <http://purl.obolibrary.org/obo/UBERON#> .

PDXNET:PAT-1  PDXNET:age_group  "0-4 years" ;
    PDXNET:consent       PDXNET:consent_YES ;
    PDXNET:ethnicity     "Sephardic" ;
    PDXNET:gender        PDXNET:female ;
    PDXNET:hasDiagnosis  NCIT:C130038 ;
    PDXNET:hasTumor      PDXNET:TUMOR-PAT-1 ;
    PDXNET:patient_id    "PAT-1" .

PDXNET:TUMOR-PAT-1  PDXNET:hasSubmitterTumorId
            "TUMOR-PAT-1" ;
    PDXNET:stage                NCIT:C19251 ;
    PDXNET:tissueOfOrigin       UBERON:35975 ;
    PDXNET:tumorCategory        NCIT:C3352 ;
    PDXNET:tumorGrade           NCIT:C121173 ;
    PDXNET:tumorHistology       NCIT:C130038 .

PDXNET:PAT-0  PDXNET:age_group  "15-19 years" ;
    PDXNET:consent       PDXNET:consent_NO ;
    PDXNET:ethnicity     "hispanic or latino" ;
    PDXNET:gender        PDXNET:male ;
```

(continues on next page)

(continued from previous page)

```
    PDXNET:hasDiagnosis  NCIT:C7326 ;
    PDXNET:hasTumor      PDXNET:TUMOR-PAT-0 ;
    PDXNET:patient_id    "PAT-0" .

PDXNET:TUMOR-PAT-0  PDXNET:hasSubmitterTumorId
            "TUMOR-PAT-0" ;
    PDXNET:stage                NCIT:C19251 ;
    PDXNET:tissueOfOrigin       UBERON:4146 ;
    PDXNET:tumorCategory        NCIT:C8509 ;
    PDXNET:tumorGrade           NCIT:C48934 ;
    PDXNET:tumorHistology       NCIT:C7326 .
```

9.1 SPARQL Queries

We will use the corresponding RDF/XML file to perform demonstration SQPARL queries. For this, we use the query command, which produces output like this.

```
PREFIX pdxnet: <http://pdxnetwork/pdxmodel_>
PREFIX ncit: <http://purl.obolibrary.org/obo/NCIT_>
SELECT ?patient_id ?consent ?diagnosis
WHERE {
  ?x pdxnet:patient_id ?patient_id .
  ?x pdxnet:consent ?consent .
  ?x pdxnet:hasDiagnosis ?diagnosis .
  }
LIMIT 5
Lock : main
Lock : main
-------------------------------------------------------
| patient_id | consent                       | diagnosis  |
=======================================================
| "PAT-846"  | pdxnet:consent_NO             | ncit:C5235 |
| "PAT-1256" | pdxnet:consent_ACADEMIC_ONLY  | ncit:C4887 |
| "PAT-127"  | pdxnet:consent_NO             | ncit:C5656 |
| "PAT-179"  | pdxnet:consent_YES            | ncit:C7811 |
| "PAT-1477" | pdxnet:consent_ACADEMIC_ONLY  | ncit:C7965 |
-------------------------------------------------------

###########  Next Query  ###########  Next Query

PREFIX pdxnet: <http://pdxnetwork/pdxmodel_>
PREFIX ncit: <http://purl.obolibrary.org/obo/NCIT_>
PREFIX uberon: <http://purl.obolibrary.org/obo/UBERON_>
SELECT ?patient_id ?currentTreatmentDrug ?diagnosis
WHERE {
  ?x pdxnet:patient_id ?patient_id .
  ?x pdxnet:currentTreatmentDrug ?currentTreatmentDrug .
  ?x pdxnet:gender pdxnet:female .
  ?x pdxnet:hasDiagnosis ?diagnosis .
  }
LIMIT 5
Lock : main
Lock : main
-------------------------------------------------------------------------
```

(continues on next page)

(continued from previous page)

```
| patient_id | currentTreatmentDrug                          | diagnosis  |
=================================================================================
| "PAT-846"  | "Goserelin[DB00014;65807-02-5]"               | ncit:C5235 |
| "PAT-1256" | "Sargramostim[DB00020;123774-72-1]"           | ncit:C4887 |
| "PAT-1477" | "Peginterferon alfa-2a[DB00008;198153-51-4]" | ncit:C7965 |
| "PAT-1770" | "Cetuximab[DB00002;205923-56-4]"              | ncit:C7061 |
| "PAT-1676" | "Cetuximab[DB00002;205923-56-4]"              | ncit:C8834 |
---------------------------------------------------------------------------------

############  Next Query  ############  Next Query

PREFIX pdxnet: <http://pdxnetwork/pdxmodel_>
PREFIX ncit: <http://purl.obolibrary.org/obo/NCIT_>
PREFIX uberon: <http://purl.obolibrary.org/obo/UBERON_>
SELECT ?patient_id ?currentTreatmentDrug ?diagnosis ?age_lowerrange ?age_upperrange
WHERE {
  ?x pdxnet:patient_id ?patient_id .
  ?x pdxnet:currentTreatmentDrug ?currentTreatmentDrug .
  ?x pdxnet:gender pdxnet:female .
  ?x pdxnet:hasDiagnosis ?diagnosis .
  ?x pdxnet:ageBinLowerRange ?age_lowerrange .
  ?x pdxnet:ageBinUpperRange ?age_upperrange .
  FILTER (?age_lowerrange > 55) .
  }
LIMIT 5
Lock : main
Lock : main
-------------------------------------------------------------------------------------
↪---------------------
| patient_id | currentTreatmentDrug                        | diagnosis   | age_
↪lowerrange | age_upperrange |
=====================================================================================================================
| "PAT-1256" | "Sargramostim[DB00020;123774-72-1]"         | ncit:C4887  | 75          ⌟
↪    | 79             |
| "PAT-1770" | "Cetuximab[DB00002;205923-56-4]"            | ncit:C7061  | 105         ⌟
↪    | 109            |
| "PAT-75"   | "Denileukin diftitox[DB00004;173146-27-5]" | ncit:C5631  | 75          ⌟
↪    | 79             |
| "PAT-1765" | "Pegfilgrastim[DB00019;208265-92-3]"        | ncit:C7964  | 80          ⌟
↪    | 84             |
| "PAT-851"  | "Leuprolide[DB00007;53714-56-0]"            | ncit:C27754 | 65          ⌟
↪    | 69             |
-------------------------------------------------------------------------------------
↪---------------------
```

9.2 Development plans

Currently, there are prototype versions of all modules but one. We will go through the entire PDX-MI ontology specification in this document : https://docs.google.com/document/d/1M81y8wbT5gegUe35RZwS92bvHLYJrVPhaFnnkECgbto/edit and will implement RDF patterns, and will test the ability to query the data with SPARQL. Once this is mature and tested, we will adapt the code to provide ETL and Q/C functionalities.

9.3 Visualization

This is a nice tool for visualizing RDF graphs: http://visgraph3.org/

CHAPTER 10

PDXIntegrator

PDXIntegrator is inteded to provide a common semantic model for PDXNet. The project is currently in a preliminary stage and we do not recommend use of any of this code in production environments. Instead this site is intended to provide information about current plans for collaborators. Please visit https://github.com/TheJacksonLaboratory/PDXintegrator to view the current code base.

To get a very first impression of this project, we recommend that you go through the Demonstration of simulating cases and performing SPARQL queries.