Dipartimento di Ingegneria e Scienza dell’Informazione

KDI Lab 2018-19

Document data:	**Reference persons:**	
15.11.2018	Thomas Cantore	198329
	Giuseppe Minardi	197754
	Davide Golzato	193553

KnowDive (internal) reports are for internal only use within the KnowDive Group. They describe preliminary or instrumental work which should not be disclosed outside the group. KnowDive reports cannot be mentioned or cited by documents which are not KnowDive reports. KnowDive reports are the result of the collaborative work of members of the KnowDive group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the KnowDive group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.

Disclaimer: our work mainly focuses on the task 3 but overlaps with task 4, we focused a bit more on the formal model, considering a possible top level grounding in order to tune and refine our informal model (section 6).

1 - Scenario Description

Research in the healthcare field is a very complex world, peopled by many different workers, employees, with different jobs and roles, but with a common purpose: trying to move a step further the human knowledge related to the nature of the human body and its health. Most of the times, this shared purpose goes beyond private interests and leads to what many people ask for nowadays: an open-source world. Being able to communicate easily between different laboratories around the globe and sharing knowledge, becomes crucial. This approach speeds-up research, leading to a better understanding of pathologies and invention of new and more affordable therapies. Open-source data is a critical issue also due to new high-throughput technologies, such as Omics technologies that produces tons of biologically interesting data, giving a huge help to shed light on very complex mechanisms. Such machines produce information about the whole genome sequence of a patient, or its transcriptome (gene expression information).

Researchers need to share knowledge easily but often have to face problems such as the high cost of the technology used. These two aspects make the role of data integration for data repurposing central. Being able to share data and results in a scaled and standardized way could represent a key answer to these problems. Also the use of previously obtained data, going beyond their use for very specific research works, could be enlighteining. Here we purpose four different personas involved in cancer research, taking advantage of omics technologies, that could benefit of such system.

Amanda Dawson, the P.I. of a cancer computational biology laboratory (i.e. a lab focused on the study of cancer from a computational perspective). Bob Giusti, a lab technician, interested in specific issues regarding Omics machines (in particular related to genomics and transcirptomics experiments). John Dorbal, a wet-lab biologist, interested in specific information that could lead his future experiments, and Susan White an oncologist that wants to have a wider sight on possible problems regarding her patients. These people represent common workers in the field, they are usually part of the same research structure, or, more interestingly, they can cooperate from different facilities.

2 – Personas description

Amanda Dawson is an affirmed academic researcher, PI of the *Laboratory of Computational and Functional Oncology*. Her research is focused in studying which is the role of genetic diversity and genomic abnormalities in how cancer evolves and progresses. The laboratory's approach to cancer studies is obviously based on computational and quantitative approaches The ultimate purpose is to discover potentially new biomarkers in cancer that could contribute towards personalized treatments for patients. Amanda works with genomic and trascriptomic data produced by her own NGS facility, as well with data provided by other related research laboratories. Additionally, she is also collaborating with several cancer wards based in UK and USA, which share with her their biopses and sequenced samples coming from real cancer patients. Considering the huge amount of data accumulated during previous experiments and other people, she's interested in developing a system capable of retrieving old datasets that match her criteria.

Data-driven hypotheses are crucial in her field, so for example she may want to see all the transcriptomes of cancer patients exposed under particular environmental conditions or from cells treated with a peculiar drug. Moreover, producing NGS data is economically burdensome for the laboratory, so every sample is precious when it comes to test hypotheses.

Bob Giusti is a 35 years old a lab technician. He works in the same university as Amanda. He’s specialised in NGS (Next Generation Sequencing) technologies. Such technologies are part of a subgroup of Omics technologies, specifically regarding sequencing issues (i.e. sequencing genome or transcriptome).

Since such technologies became of crucial interest in almost every biology laboratory, from healthcare to ecology, their maintenance as well as their usage is usually in charge of prepared specialised technicians. This is also supported by the high cost behind each usage and the complex technology behind their functioning. Hence, Bob Giusti is part of the team working in the Next Generation Sequencing core facility of the university.

He’s interested in the system, since, else than just focusing on general problems related to each time the machine works, he can work on a wider scale, having access to many sessions with specifically use, of each machine in the university. Hence, detailed reports on the overall behaviour of the machines, could help its maintenance work, preventing specific ground problems and errors and avoid laboratories to waste inefficiently finance resources.

John Dorbal is a molecular biologyst, working as post-doctoral scientist at the Laboratory for Molecular Cancer Biology. His work is wet lab based: he aims to understand the sophisticated mechanisms that control cell growth and differentiation, and the ways in which these are disrupted in cancerous cells. His research approach encompasses disciplines such as biochemistry, pharmacology, cell biology and genetics.

The ambitious aim of him and his laboratory team is to discover molecular and genetic elements that could be target of new cancer treatments. He's currently focusing his efforts on studying the molecular mechanisms that link genetical mutations in TP53 and its molecular partners with uncontrolled cell proliferation.

His deep knowledge of TP53's protein structure makes him speculate that its loss of function in cancer cells is due to nucleotide mutations in the encoding gene sequence.

He also wants to study if TP53 transcriptional levels are influenced by methylation in certain regions in the chromatine.

He needs a system capable of retrieving the nucleotide sequence of TP53 and his molecular interactors from cells coming from biopsies, in order to study if his speculations are supported by real clinical data.

Susan White is a 42 years old oncologist. She’s a doctor specialized in oncology working for a public hospital in the same city of the university. She works directly with patients, following them from their diagnosis, in general assessed by pathologists, to the best therapy.

As well as many other hospitals, hers tries always to provide the best therapies and technologies to cure diseases, especially cancer. Since, in the cancer field, research and new therapies are everyday more advanced, having access to personalized therapies based on the mutational landscape of the patients cancer (trough specific genomic and transcriptomic analysis) can be of utmost importance.

Hence, she is very interested in the system. Having access to a huge amount of information, coming from the university research lab, and providing them her patients informations in order to populate their system and improve their researches is a huge opportunity. Having access to such information can lead to a better follow-up of her patients, as well as better understanding of the disease. Cooperation with Amanda's lab can be beneficial for both: both in order to plan better therapies and make medical research.

3 – Storytelling definition

Persona	Real world action by persona	System/Demo action
Amanda Dawson	Search for patients of caucasian ethnicity	Returns all the patients associated to persons of caucasian ethnicity
Amanda Dawson	Search for the methylome of lung cancer biopses	Return all the methylomes obtained from lung cancer biopses
Amanda Dawson	Search for all RNA samples extracted with Qiagen extraction kit	Returns all RNA samples extracted with Qiagen extraction kit
Amanda Dawson	Search for genomic data which contains a SNV at position 29839495	Returns all the Genome data which have a SNV at position 29839495
Amanda Dawson	Look for all the possible SNVs within a range of positions	Selects Genome data which contains any SNV within the specified range of positions
Amanda Dawson	Search for genomes that have at least n° of total $5*10^6$ total reads	Shows all the Genome data which exceed the threshold of total_reads of $5*10^6$
Bob Giusti	Search for data generated with warning errors	Returns all Omics data with *error logs*
Bob Giusti	Search for all Omics files with a specified level of quality	Returns all the raw_omics_files that exceed a threshold of *quality_check*

Bob Giusti	Search for all Omics file obtained from all the NGS sequencer of brand *PacBio*	Returns all the raw_omics_files produced with *brand* “PacBio”
Bob Giusti	Search for Omics file produced from a specific NGS sequencer	Return all raw_omics_files produced by NGS sequencer with the specified *machine ID*
Bob Giusti	Search for Omics file produced by all the *Illumina MySeq*	Returns all raw_omics_files produced by *Illumina MySeq* machines
Bob Giusti	Search for Omics files produced by *Illumina* machines with warning errors	Returns raw_omics_files produced by *Illumina* machine with at least 1 error
John Dorbal	Search the transcription levels for a specific isoform of TP53 from a tumor biopsy	Returns the expression levels of the selected gene.
John Dorbal	Search for the SNVs occuring within TP53 genes from a cohort of tumor biopses	Returns all the SNVs within the position range of TP53 in reference genome for tumor biopses
John Dorbal	Search for methylated regions in a genome from a cancer biopsy	Return the methylome associated to a genome file
John Dorbal	Search for all the methylated regions that fall within exons	Returns all the methylated regions falling into exons
John Dorbal	Search for all the SNV classified as missense mutation in patients with prostate cancer	Returns all genomes with a SNV at the specified positions
Susan White	Susan White Search for the age and gender of patients with lung cancer	Returns all the patients age and gender with lung cancer
Susan White	Susan White Search for the patients with prostate cancer and a specific single nucleotide variant	Returns the patients with prostate cancer and a specific single nucleotide variant
Susan White	Search for patients with breast cancer recovered during between 1995 and 1996	Returns patients that have been recovered within the specified period of time
Susan White	Search for male patients with prostate cancer and older than 45 years	Returns male patients with prostate cancer with more than 45 years
Susan White	Search for the genome data of one of her patients	Return genome data for patient with a specific patient ID

4 – Generalized Queries

Persona	Generalized Query	Displayed Result
Amanda Dawson	Give me all patients such that (Person/*ethnicity* = *“Caucasian”)*	Returns a dataset with the patients' IDs along with their respective attributes
Amanda Dawson	Give me all methylome_data such that (Biopsy/*body_source* = *“Lungs”* and *Biopsy/Control* = “No”*)*	Return a dataset with all the methylomes IDs along with their respective attributes
Amanda Dawson	Give me all Transcriptome_data such that (Extracted_material/*Extraction Kit* = *“QiaGen”*)	Returns a dataset with all Transcriptome_files IDs along with their attributes
Amanda Dawson	Give me all Genome_Data such that (SNV_data/*position* = 29839495)	Returns a dataset with the Genome_data file IDs along with their attributes
Amanda Dawson	Give me all the SNVs such that (SNV_data/*position* in [x_start. x_end])	Returns a dataset with the SNP_data IDs along with their attributes
Amanda Dawson	Give me all *Genome_data* such that (Genome_data/*mapped reads* > $5*10^{6}$ reads)	Returns a dataset with the Genome_data IDs along with their attributes
Bob Giusti	Give me all *raw_omics_file* such that (Omics_file/*error_logs > 0)*	Returns a dataset with the raw_omics_file IDs along with their attributes
Bob Giusti	Give me all *raw_omics_files* such that (Raw_*quality_control* > 30%	Returns a dataset with the raw_omics_files IDs along with their attributes
Bob Giusti	Give me all *raw_omics_files* such that *(Sequencer/*brand = *“PacBio”)*	Returns a dataset with the raw_omics_files IDs along with their attributes
Bob Giusti	Give me all *raw_omics_files* with (Sequencer/*machine's ID* = ”SIM:1:FCX:1:15:6329:1045 1:N:0”)	Return a dataset with the raw_omics_files IDs along with their attributes
Bob Giusti	Give me all *raw_omics_file* such that (Sequencer/Brand = “*Illumina”)*	Return a dataset with the raw_omics_files IDs along with their attributes
Bob Giusti	Give me all *raw_omics_files* such that *(Sequencer/*Brand = “Illumina” and *raw_omics_files/*log_Errors > 0*)*	Return a dataset with the raw_omics_files IDs along with their attributes

Persona	Generalized Query	Displayed Result
John Dorbal	Give me Transcriptome_data such that (Transcriptome/Ensembl ID = “ENSTP53x”; Biopsy/*Control* = *“No”*)	A dataset with the Transcriptome_data along with their attributes
John Dorbal	Give me SNV_data such that (SNV_data/*position* in [x_start. x_end])	A dataset with the SNV_data IDS along with their attributes
John Dorbal	Give me Methylome_file such that (Biopsy/Control = “No”)	A dataset with the Methylome_data IDs along with their attributes
John Dorbal	Give me Methylome_file such that (Methylome_file/*exon* = “Yes”)	A dataset with the Methylome_data IDs along with their attributes
John Dorbal	Give me SNV_data such that (SNV_data/*variant_classifier* = “missense mutation” and Biopsy/*Body_source* = “Prostate” and Biopsy/*Control* = “No”)	A dataset with the SNV_data IDs along with their attributes
Susan White	Give me patients such that (Patient/*Gender* = “Male” and Patient/*Birthdate* = 1995)	A dataset with the person's IDs along with their attributes
Susan White	Give me patients such that (Biopsy/Control = “No” and Biopsy/Body_source = “Prostate” and SNV_data/*position* = 1983249)	A dataset with the patient's IDs along with their attributes
Susan White	Give me patients such that (Biopsy/Control = “No” and Biopsy/Body_source = “Breast” and 01/01/1995 < Patient/*recover-visit_date* <*31/12/1996*)	A dataset with the patient's IDs along with their attributes
Susan White	Give me patients such that (Patient/*Gender* = “Male” and Patient/*Birthdate* = *xx/xx/*1978 and (Biopsy/Control = “No” and Biopsy/Body_source = “Prostate”)	A dataset with the patient's IDs along with their attributes
Susan White	Give me genome_data such that *(*Patient/*patient_ID* = “PN14053”*)*	A dataset with the genome_file ID's along with their attributes

5 – Model Design

Observing our entities we noticed they they can be broken down into four macro-areas, connected to our ontology. Living being, Physical and Machine are representative of the real world, while Information object is part of the information world. These aspects are further explained in the report section dedicated to the ontology. This model was designed with real users in mind and so it will try to account for real users needs. The cohort of users to whom the model is addressed to is quite restricted , since it comprises biologists, computational biologists, technicians or oncologists. Nevertheless, since the model is intended to be used also by people working in different fields our aim was to simplify the knowledge transfer between those fields. We started from scratch by taking data from an article published by the Prof. Francesca DeMichelis. The EER model was built using those data as reference and by adding entities and attributes, in order to create a more coherent and complete model.

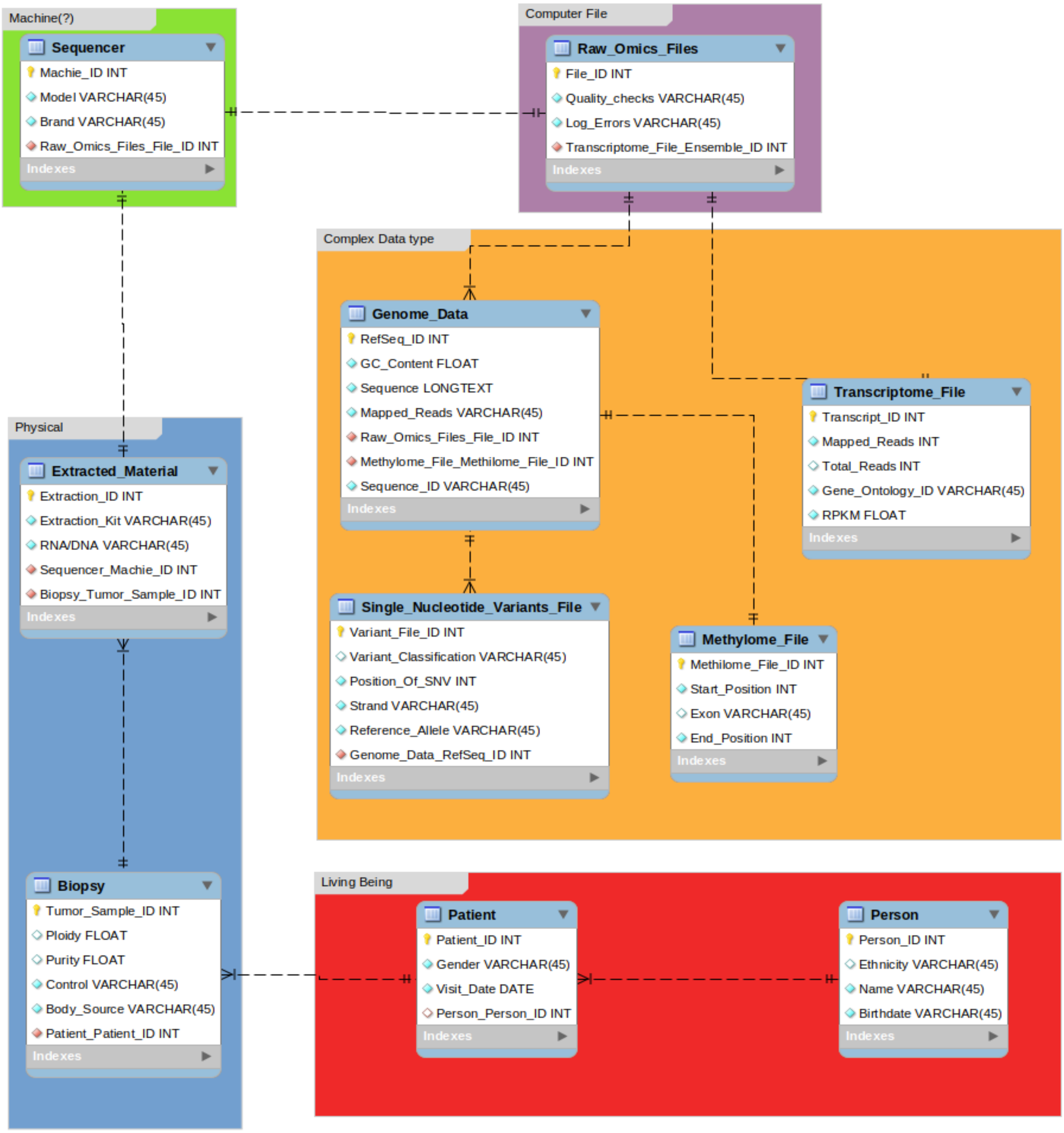


Figure 1: Relational model produced with MySQLWorkBench

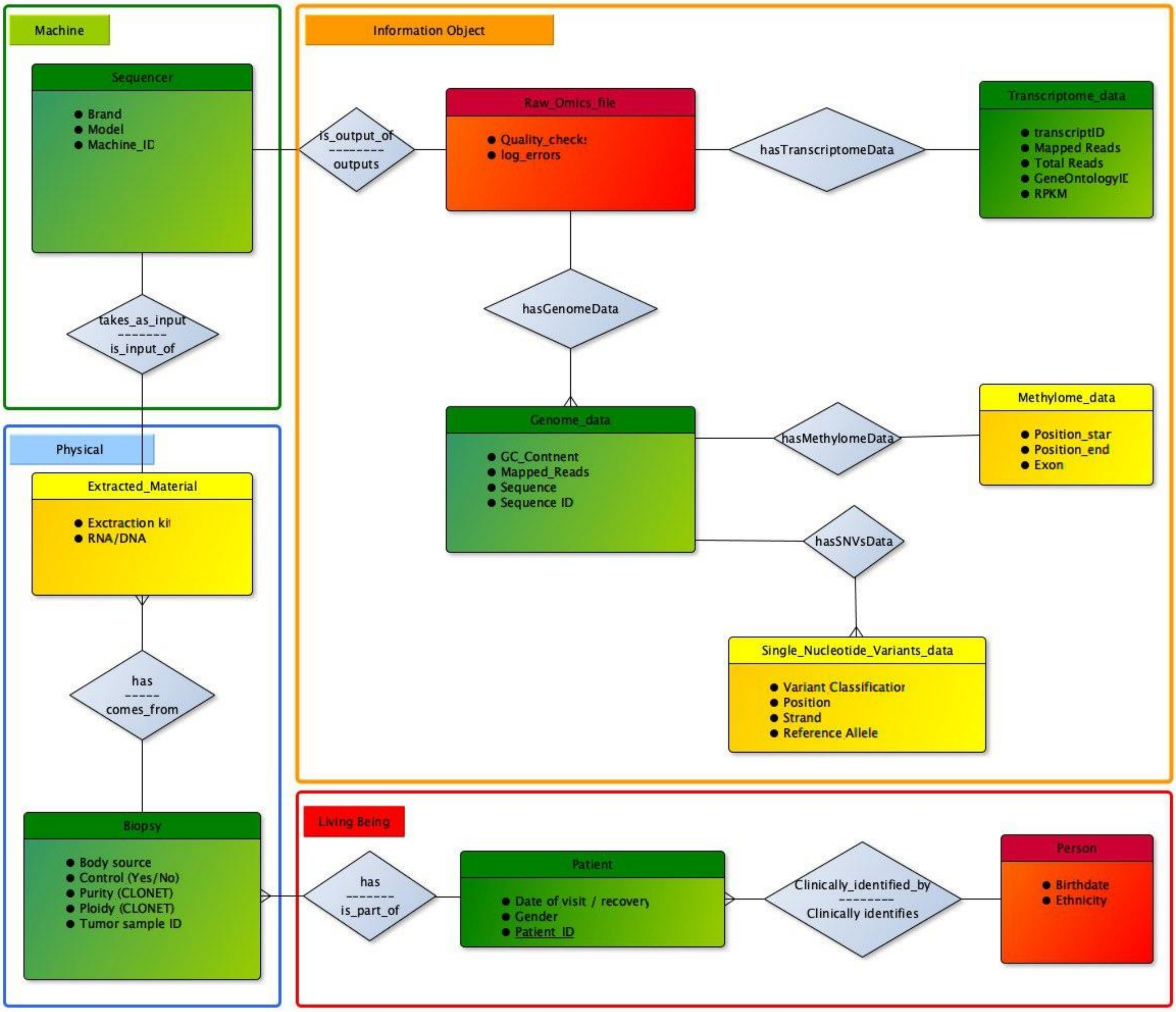


Illustration 2: ER model produced with yED live

Core Entities: Our core entities are Patient, Biopsy, Genome Data, Transcriptome Data and Sequencer. We decided to categorize these as core entities, since they can be frequent subject of query from the cohort of users we had in mind

Patient: a doctor or a researchers could be highly interested in knowing the description of the patient since therapies and hypothesis can vary depending on the gender or the age of the patient. In order to be fully evaluated, a patient must be subjected to different biopsies (*hasBiopsy:* one-to-many, *clinicallyIdentifies:* many-to-one).

- **Gender:**the gender of a patient is very important because male and female bodies, being different, affect the possible cause of the cancer (*dt:* varchar)
- **Visit Date:** the date of the visit (or recovery) is important beause can be related to the age of the patient using the birthday oh him taken from the *Person* entity. (*dt:* datetime)

Biopsy: The biopsy is the first step to the computational analysis of the patient, the provenience and the nature of it can be object of study during research. In order to be analyzed from a biops many genetic material must be extracted with an extraction kits (*isBiopsyOf:* many-to-one,*hasExtractedMaterial:* one-to-many).

- **Ploidy:** is an output value that tells how much the gene/region is multiplied in the *genome.(dt:* percentage*)*
- **Purity:** is an output value that tells the percentage of tumor/normal cells there are in the *biopsy.(dt:* purity*)*
- **Control:** is a variable that tells us if the biopsy is a normal tissue or a tumor *tissue.(dt:*boolean*)*
- **Body Source**: is a variable that indicate from which part of the body comes the *biopsy(dt*:varchar)

Genome Data: The genome data are one of the two biggest sources of information in order to do computational analysis. Each genomic data can have associated a methylome and many nucleotide variants that can explain the cause of the cancer (*IsGenomeDataOf:* one-to-one)

- **GC content:** Is the percentage of cytosine and guanine present in the genome.It can be useful to detect genomic aberrations (*dt*:percentage)
- **Mapped reads:** how many reads are mapped against the reference genome (dt:integer)
- **Sequence:** Is the sequenced genomic sequence (*dt:* string)

Transcriptome Data: it represents gene expression levels of a patient

- **Mapped Reads:** It tells us how many reads are mapped against this piece of this (*dt:* integer)
- **Total Reads:** How many reads are found in the sample (*id:*integer)
- **Gene Ontology ID:** This is the ID that can be used in order to find a gene in the Gene ontology. (*id:* varchar)

Sequencer: The sequencer is the machine that allows the physical entity to be translated into machine-readable data. A sequencer outputs omic files that contain information about the status of the machine *(take-as_input*:one-to-one, *outputs:* one-to-one***).***

- **Model:** is the model of the machine (*id*:varchar)
- **Brand**: is the brand machine (*id*:varchar)
- **Machine ID:** the ID of the specific machine (*id*:varchar)

Auxiliary Entities: we chose Single Nucleotide Variants data and Methylome data as auxiliary entity, since they add informations to the core entities

Methylome Data: represents the ensemble of methylations occuring in nucleotides for a determined genomic region

- **Start position:**defines the position of the starting nucleotide of the methylated region (int)
- **End position:**defines the position of the ending nucleotide of the methylated region (int)
- **Exon:**tells whether the methylated region occurs inside an exon (bool)

Single Nucleotide Variant Data: data containing all the nucleotide positions that differ from the reference.

- **Variant Classification:**tells which kind of SNV is according to which region it occurs in (varchar)
- **Position:**the position of the SNV mapped to the reference genome(int)
- **Strand:**the straind in which the SNA is located(bool)
- **Reference Allele:**the nucleotide expected to be found in that position for the reference genome (character)

Extracted Material:represents the concept of the physical sample obtained from the biopsy. It's the input of the NGS sequencer (*has:* many-to-one, *is_input_of:* one-to-one).

- **Extraction kit:**specifies the kit used for the extraction of the nucleic acids (varchar)
- **DNA/RNA:**tells if the sample contains RNA or DNA (varchar)

Common entities: we chose Person and as common entity types, since they belong to the common knowledge among biologists.

Person: is the entity that record univocally the person in the system, since the person can be monitored multiple times during his lifetime is clinically identified by the patient *(ClinicallyIdentifiedBy / Identifies :* one-to-many/many-to-one*)*

- **Birthdate :** is the birthdate of the person, is useful because a person can be a patient multiple times.(*dt:*datetime)
- **Ethnicity:** The Ethnicity is important since various genomic mutations affect more some populations.(*dt:*varchar)

Raw Omics Files: Are output files of the sequencer and have transcriptomic and genomic data *(isOutputOf:* one-to-one, *hasTranscriptomeData:* one-to-one, *hasGenomeData*:one-to-one).

- **Quality check:** In a percentage that express the output quality of the sequencer (*dt: percentage*)
- **Log error:**It expresses the number warning errors of the sequencer. *(dt:* int*)*

6 - Related ontology proposal and formalization attempt

In this section we will present ontologies related to our model, from which we took inspiration for a formalization attempt. As we will explain in details, trying to formalise such technical and specific information in a wider knowledge context has been no easy task. We then tried to go beyond the simple formalization attempt by top-level grounding our model: this helped us to finetune the formal model in a more precise, and, ideally*, easy-to-integrate* way.

6.1 Related ontologies

Before describing the ontologies we found more in details, two main considerations need to be stated:

Our representation of these data has two main aims. The first is to represent them in a "conceptual way": abstracting the representation of the object from the concrete experimental result. The latter is to create a system where concrete results from different experiments are easy to manage, and organise, answering to more concrete needs of our possible users.

We found three ontologies related to our model. Here follows their description. The ranking of the following list, represents a measure of consistency between our model aim and the knowledge represented by the ontology. These are listed from the less representative to the most representative.

1. Gene Ontology:

It is the most used ontology in the biological field. It is highly used in research: it tries to represent the knowledge related to the biological world from the molecular perspective. It is a communication port between molecular data and their conceptual representation. Each gene, protein, transcript founds here a representation in a three terms defined namespace: *molecular_function, biological_process and cellular_component*. All these high level entities give an overall idea of the specific molecular object. Every term has a term name (e.g. mitochondrion, glucose transport, amino acid binding) and a unique zero-padded seven digit identifier (often called the term accession or term accession number) prefixed by GO:, e.g. GO:0005125 or GO:0060092. A textual description of what the term represents, plus reference(s) to the source of the information. All new terms added to the ontology must have a definition; there remains a very small set of terms from the original ontology that lack definitions, but the vast majority of terms are defined.

This ontology is the less representative for a possible model formalization of our data. It is true that our data include also molecular entities, such as transcripts or genes, but it is not what we actually want to represent. Nevertheless, since this ontology is ubiquitously present in the biology research world, we wanted to give access to it, including under *transcriptome* entity, the attribute *"Gene_Ontology_ID"*.

2. EDAM

EDAM is a comprehensive ontology of well-established, familiar concepts that are prevalent within bioinformatics and computational biology, including types of data and data identifiers, data formats, operations and topics. EDAM is suitable for large-scale semantic annotations and categorization of diverse bioinformatics resources. It focuses mainly on bioinformatical files representation.

EDAM gives a detailed representation of omics data, but doesn't capture the overall process that we want to represent: it is very detailed about Omics technology, but there's no detailed enough connection with real world, through biopsy, patient and persona.

3. TCGA (The Cancer Genome Atlas)

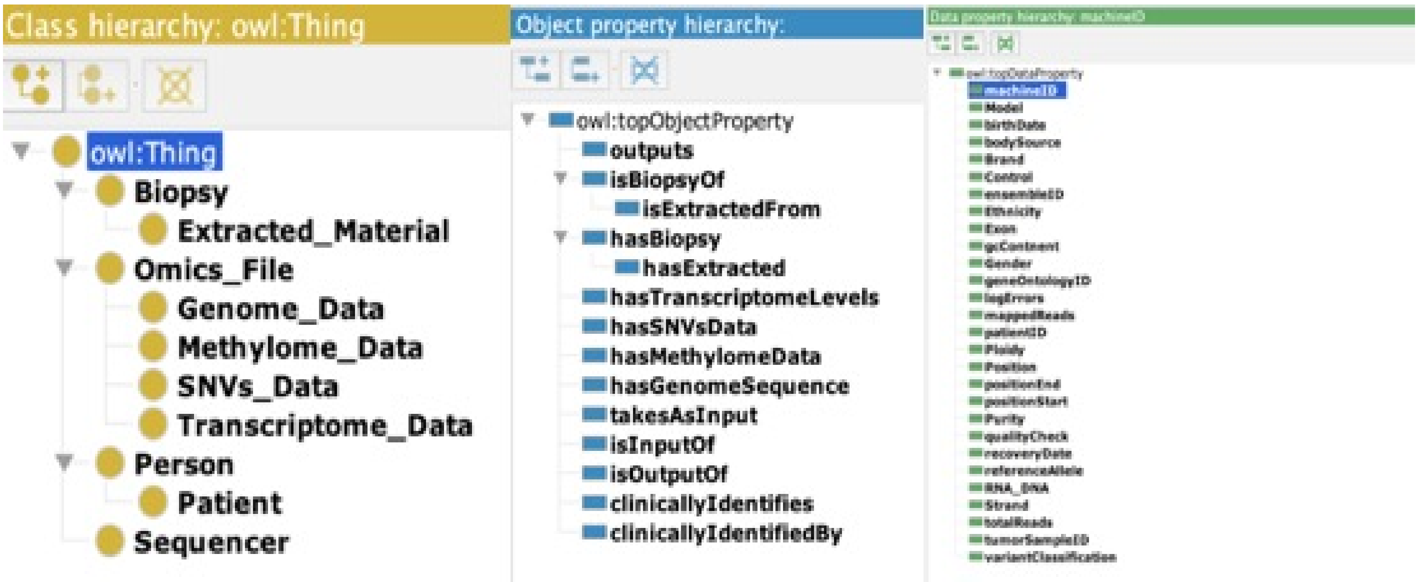
The Cancer Genome Atlas (TCGA) is one of the richest and most complete genomics datasets and was compiled to understand the molecular basis of cancers. Data collection for TCGA began in 2006 as a joint effort by the National Cancer Institute (NCI), National Human Genome Research Institute (NHGRI), the National Institutes of Health (NIH), and the U.S. Department of Health and Human Services. It has a specific ontology, specialised in cancer data representation.

TCGA ontology is very interesting for our final aim. It represents quite easily the knowledge related to the oncology world, going from genomics data to sample (biopsy) description. Indeed, it includes some of our purposed core entity types, such as genomic data and biopsy. By the way, it doesn't capture the whole set of information we want to cover, lacking of more detailed specifications regarding sequencer machine, which is for us a core entity type.

6.2 First formal model attempt

Our first attempt of formal model consisted in capturing the organisation of both TCGA and EDAM ontologies regarding the core, common and auxiliary entity types present in our EER. Here follows our resulting model:

Illustration 3: First attempt of formalization of the model with Protegé

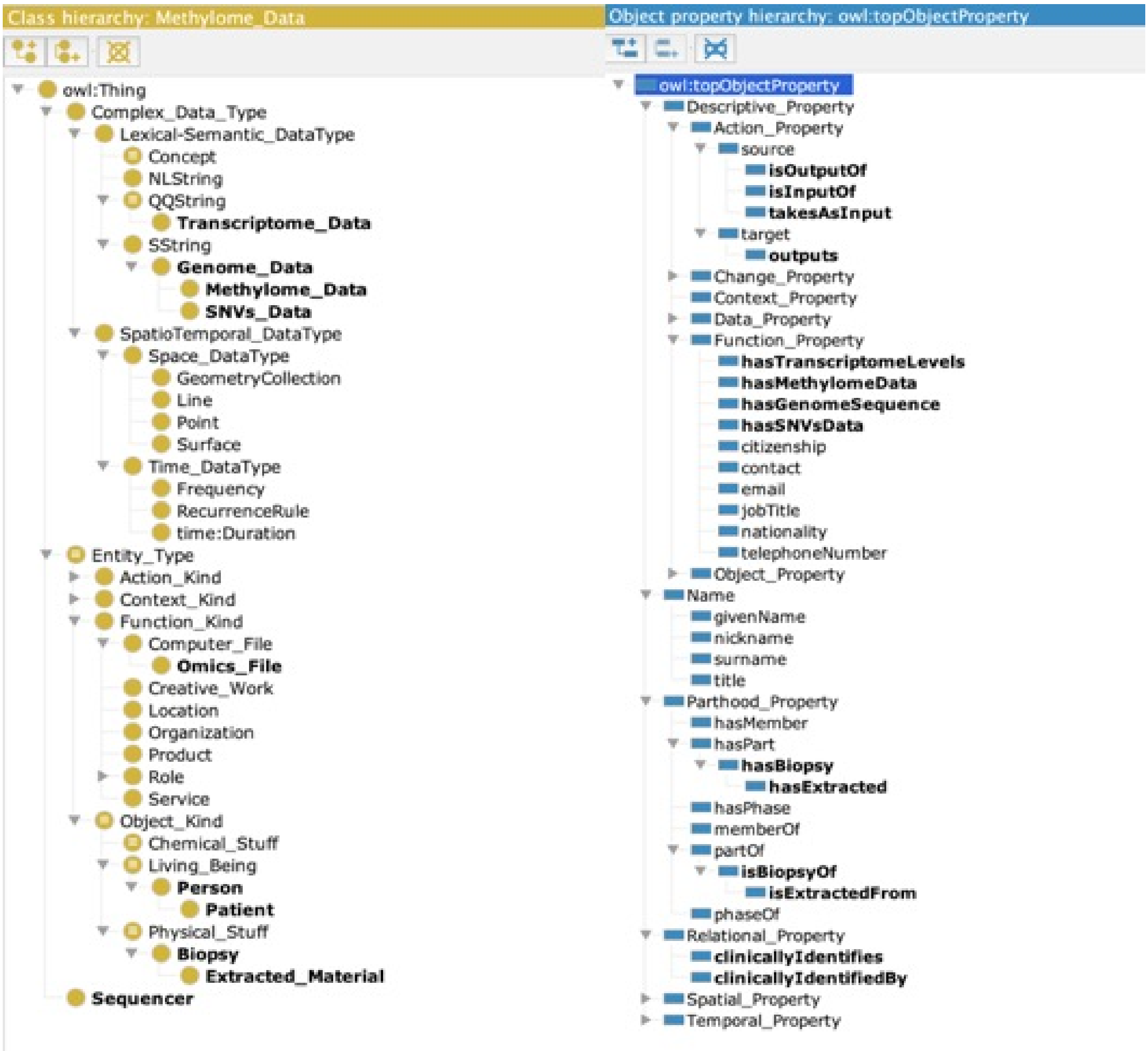


This model structure was induced by the fact that the "omics_file", "Person", "Biopsy" and "Sequencer", were the super-classes representative of the four knowledge domains previously identified in the EER model. Biopsy, is super-class representing physical objects, different from *living* *beings* which are represented by Person. The sequencer (i.e. the machine producing the data) is the connection between the physical object and with the *Information Objects "world"*.

This consideration took us to the decision of considering them as disjoint classes, connected by the specific object properties, representative of the EER relationships. Moreover, we created also inverse object properties, in order to increase the usability of the model, adding more flexibility to the queries. For each owl class we also created specific restrictions representing the cardinality of the relations reported in our EER. The reasoner didn't return consistency errors.

6.4 Top level grounding

We further proceeded with top level grounding in order to check our model consistency and structure respect to CSK (Common Sense Knowledge) top-level ontology. This procedure helped us not only in the refinement of our formal model (even if not completely defined), but, in particular, in a last clarification about our EER. Here reported are the owl classes and object properties mapped over the CSK after having imported it with protegé software.



Class hierarchy: Methylome_Data

- owl:Thing
  - Complex_Data_Type
    - Lexical-Semantic_DataType
      - Concept
      - NLString
      - QQString
        - **Transcriptome_Data**
      - SString
        - **Genome_Data**
          - **Methylome_Data**
          - **SNVs_Data**
    - SpatioTemporal_DataType
      - Space_DataType
        - GeometryCollection
        - Line
        - Point
        - Surface
      - Time_DataType
        - Frequency
        - RecurrenceRule
        - time:Duration
  - Entity_Type
    - Action_Kind
    - Context_Kind
    - Function_Kind
      - Computer_File
        - **Omics_File**
      - Creative_Work
      - Location
      - Organization
      - Product
      - Role
      - Service
    - Object_Kind
      - Chemical_Stuff
      - Living_Being
        - **Person**
          - **Patient**
      - Physical_Stuff
        - **Biopsy**
          - **Extracted_Material**
  - **Sequencer**

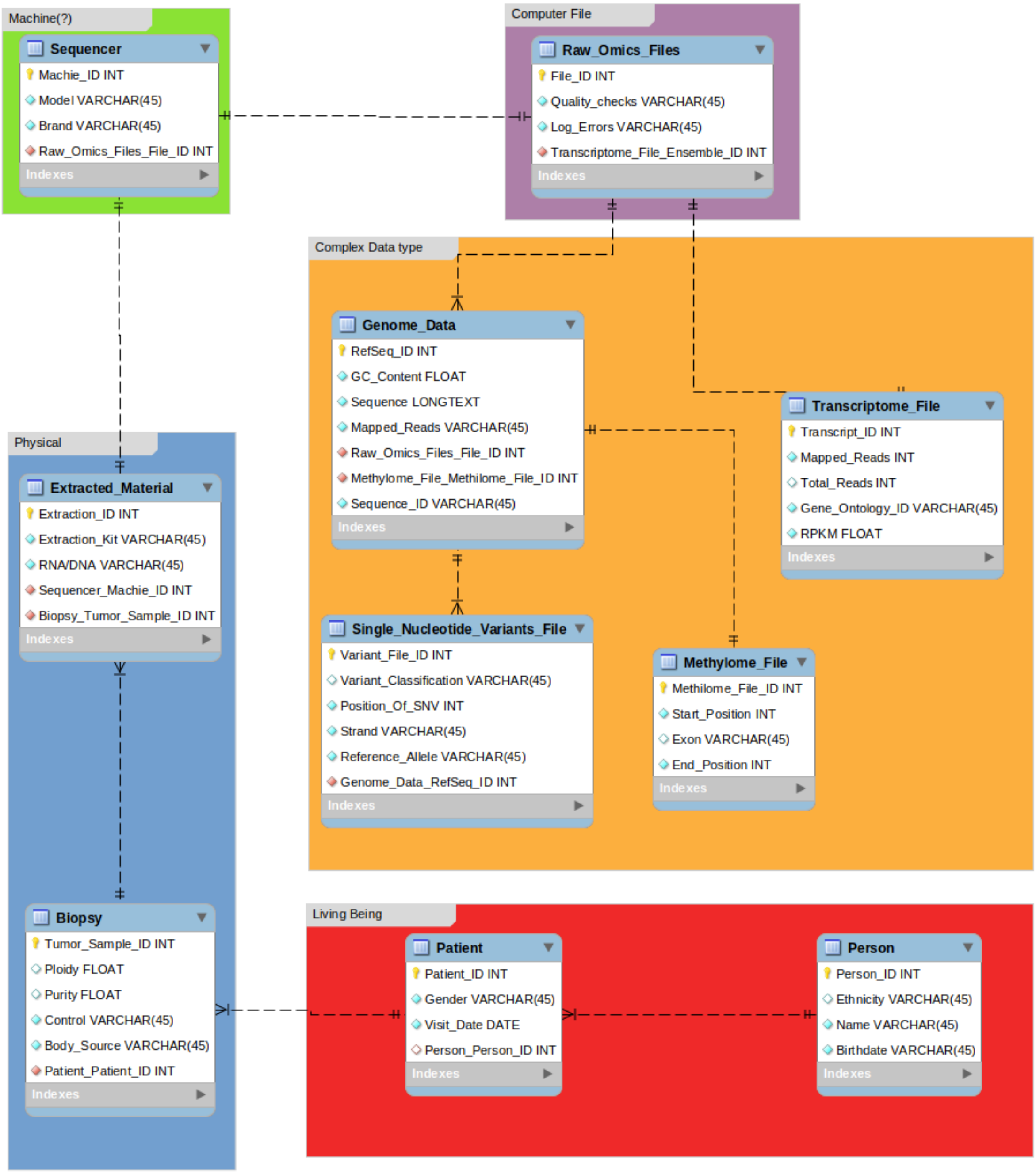
Object property hierarchy: owl:topObjectProperty

- owl:topObjectProperty
  - Descriptive_Property
    - Action_Property
      - source
        - **isOutputOf**
        - **isInputOf**
        - **takesAsInput**
      - target
        - **outputs**
    - Change_Property
    - Context_Property
    - Data_Property
    - Function_Property
      - **hasTranscriptomeLevels**
      - **hasMethylomeData**
      - **hasGenomeSequence**
      - **hasSNVsData**
      - citizenship
      - contact
      - email
      - jobTitle
      - nationality
      - telephoneNumber
    - Object_Property
  - Name
    - givenName
    - nickname
    - surname
    - title
  - Parthood_Property
    - hasMember
    - hasPart
      - **hasBiopsy**
        - **hasExtracted**
    - hasPhase
    - memberOf
    - partOf
      - **isBiopsyOf**
        - **isExtractedFrom**
    - phaseOf
  - Relational_Property
    - **clinicallyIdentifies**
    - **clinicallyIdentifiedBy**
  - Spatial_Property
  - Temporal_Property

This mapping helped us to define several aspects of our EER: first of all, we realised that *“Omics_File”* and all related data, such as *“Transcriptome_Data”*, *“SNVs_Data”*, *“Methylome_Data”* and *“Genome_Data”* should be considered disjoint, since they are sub-classes of

Complex_Data_Type, which is disjoint from Entity_Type. Indeed, these are actual “data” not single files, contained in the latter. We were not sure about the actual mapping of “Sequencer”, but still sure about considering it as disjoint from the other classes of our ontology. Here we also re-defined the correct classification for Methylome_Data and SNVs_Data which was not clearly stated in the previous formal model attempt.

From this mapping we obtained the final relational model here reported. Having defined suchconnections with the used top level ontology, it should be easier to integrate with other ontologies.



7 – Final Considerations and Open Issues

Our work focused on the building of a versatile EER (and relational model). We tried to capture the knowledge behind our starting dataset, trying to move to an higher knowledge representation level. We think that this model can be a good cross point in order to organize the knowledge that can be extracted from many datasets coming from research works.

Open issues remain the formalization of the process regarding the first data preparation that we performed informally (especially considering a data integration problem) and further tuning of the formal model. Lexical information upload remains an open issue, even if we preferred to give to use and easy to understand but precise language (hopefully), also for non-technicians.