Publisher's version: https://doi.org/10.1007/978-3-030-89022-3_28

Ontological Unpacking as Explanation: The Case of the Viral Conceptual Model

Giancarlo Guizzardi[1,2], Anna Bernasconi[3], Oscar Pastor[4], and Veda C. Storey[5]

[1] Free University of Bozen-Bolzano, Italy, gguizzardi@unibz.it
[2] University of Twente, Netherlands, g.guizzardi@utwente.nl
[3] Politecnico di Milano, Milan, Italy, anna.bernasconi@polimi.it
[4] Universitat Politècnica de València, Valencia, Spain, opastor@dsic.upv.es
[5] Georgia State University, Atlanta, Georgia 30302, USA, VStorey@gsu.edu

Abstract. Inspired by the need to understand the genomic aspects of COVID-19, the Viral Conceptual Model captures and represents the sequencing of viruses. Although the model has already been successfully used, it should have a strong ontological foundation to ensure that it can be consistently applied and expanded. We apply an ontological analysis of the Viral Conceptual Model, using OntoUML, to unpack and identify its core components. The analysis illustrates the feasibility of bringing ontological clarity to complex models. The process of revealing the *ontological semantics* of a data structuring model provides a fundamental type of *explanation* for symbolic models, including conceptual models.

Keywords: Viral Conceptual Model (VCM) · ontological analysis · OntoUML · virus genomics · COVID-19

1 Introduction

Since the scientific breakthrough of the sequencing of the human genome, efforts have been made to model this sequencing so it can be effectively used. Efforts to apply conceptual modeling to describe genomics databases began in the early 2000s [15]. Pastor et al. proposed the Conceptual Schema of Human Genome [14,16]; Bernasconi et al. introduced the Genomic Conceptual Model [5]. The Viral Conceptual Model (VCM) [4] was motivated by interest in representing the genomic aspects of COVID-19 so they can be shared among the scientific and medical communities that are attempting to understand SARS-CoV-2 and similar pathogens. The model is being used for designing: an integrated data warehouse and search system [8]; a linked Phenotype Data Dictionary [3]; a knowledge base of variant impacts [1]; and a visualization engine [6]. To facilitate a further, robust, application of VCM to projects with other viruses and pathogens, there needs to be some way to assess the quality of the model.

A conceptual model provides an *information structuring function* for the application domain. VCM focuses on genomic data structuring for the purposes of characterizing genomic sequences. However, a conceptual model should also be able to provide *conceptual clarification* and *unambiguous communication*. We

call this the *ontological function* of a conceptual model. The model should reconstruct the exact *intended conceptualization* (set of possible interpretations) and be explicit and transparent with respect to its *ontological semantics*. Revealing the ontological semantics of an information artifact is a fundamental type of *explanation* for symbolic models (including conceptual models). *Ontological unpacking* refers to a process of ontological analysis that reveals the *ontological conceptual model* behind an information structuring conceptual model.

2 The Viral Conceptual Model (VCM)

The Viral Conceptual Model (VCM) [4] is organized into four perspectives and centered around the notion of a virus genome SEQUENCE. A viral sequence can be either DNA or RNA. In both cases, sequences are composed of nucleotides, i.e., guanine (G), adenine (A), cytosine (C), and thymine (T)—replaced by uracil (U) in RNA.

For the **technical perspective**, sequences are derived from one experiment of a given type (EXPERIMENT TYPE entity) in which the biological material is analysed with a given *Sequencing Technology* platform (e.g., Illumina Miseq), which provides a *Coverage* and an *Assembly Method*, collecting algorithms applied to obtain the final sequence.

For the **biological perspective**, each sequence belongs to a specific VIRUS, described by a complex taxonomy, which is, in turn, represented by attributes: *Species Name* (e.g., Severe acute respiratory syndrome coronavirus 2), comparable forms in the *Equivalent List* (e.g., 2019-nCoV, COVID-19, SARS-CoV-2, SARS2), within a *Genus* (e.g., Betacoronavirus), *Sub-family* (e.g., Orthocoronavirinae), and finally *Family* (e.g., Coronaviridae). A virus species corresponds to a specific *Molecule Type* (e.g., genomic RNA, viral cRNA, unassigned DNA), which has either double or single-stranded structure. Each strand is positive or negative. A biological sample (tissue) is extracted from an organism, that hosted the virus for a certain amount of time. This is represented by the HOST SAMPLE entity. The host (of given *Age* and *Gender*) also belongs to a *Species*. The sample is extracted on a specific *Collection Date*, from a specific host tissue (e.g., nasopharyngeal or oropharyngeal swab, lung), at a specific location identified by the quadruple: *Originating Lab*, *Region*, *Country*, *Geo Group* (e.g., continent).

From an **organizational perspective**, SEQUENCING PROJECT is a project in which a particular sequencing activity is carried out. Each sequence is connected to a number of studies, represented by a research publication (with *Authors*, *Title*, *Journal*, *Publication Date*). When a study is not available, only the *Sequencing Lab* and *Submission Date* are provided, along with the *Database Source* where the sequence is deposited.

The **analytical perspective** addresses the secondary analyses of genomic sequences. ANNOTATIONS include subsequences representing segments (characterized by *Start* and *Stop* coordinates) of the original sequence with a particular *Feature Type* (e.g., gene, peptide, coding DNA region, or untranslated region), the recognized *Gene Name* to which it belongs (e.g., gene "E"), and the *Prod-*

uct it produces (e.g., Spike protein, nsp2 protein, RNA-dependent RNA polymerase, membrane glycoprotein, envelope protein). Annotations whose *Feature Type* is coding region (CDS) also have an associated *Aminoacid Sequence*. The *Nucleotide Variant* entity contains subsequences of the main sequence that differ from the reference sequence of the same virus species. They can be defined with a *Start* position coordinate for an arbitrary *Length*, a specific variant *Type* (insertion, deletion, single-nucleotide polymorphism or others), and an alternative sequence of nucleotides (*Alt Sequence*). A similar role is given to the *Aminoacid Variant* entity, containing subsequences of proteins (i.e., only of a subset of all annotations) that differ from the reference amino acid sequence of the virus species. These also have a start position, a length, a variant type, and an alternative sequence of amino acid residues.

3 OntoUML

OntoUML is a language whose meta-model complies with the ontological distinctions and axiomatization of a theoretically well-grounded foundational ontology, UFO (Unified Foundational Ontology) [13,12]. Stereotypes reflect the correspondence between the OntoUML profile and UFO ontological categories.

We start by focusing on the category of endurants (roughly objects), and to object *Kinds* (the genuine fundamental types of objects that exist according to a particular conceptualization of the given domain). All objects belong necessarily to exactly one kind (e.g, Person, Virus). There can, however, be other static subdivisions (or subtypes) of a kind. These are naturally termed *Subkinds*. For example, the kind 'Organization' can be specialized into the subkinds 'University' and 'Funding Agency'. Object kinds and subkinds represent essential properties of endurants. These include: *Phases* (e.g., 'being a living person' captures a cluster of contingent *intrinsic* properties of a person, or 'being a puppy' captures a cluster of contingent *intrinsic* properties of a dog) and *Roles* (for example, 'being a husband' captures a cluster of contingent *relational* properties of a man participating in a marriage). Kinds, Subkinds, Phases, and Roles are categories of object *Sortals*. A sortal is either a kind (e.g., 'Person') or a specialization of a kind (e.g., 'Student', 'Teenager', 'Woman').

Relators represent clusters of relational properties that "hang together" by a nexus (provided by a relator kind). Relators (e.g., marriages, enrollments, presidential mandates, citizenships) are the truth-makers of relational propositions. Relations (as classes of n-tuples) can be completely derived from relators. Objects typically participate in relators playing certain "roles".

In general, types that represent properties shared by entities of multiple kinds are termed *Non-Sortals*. Besides *RoleMixin*, another type of non-sortal in UFO is *Category*. A category represents necessary properties that are shared by entities of multiple kinds (e.g., the category 'Physical Object' represents properties of all kinds of entities that have masses, spatial extensions, etc.). In contrast to rolemixins, categories are static and *Relationally Independent* Non-Sortals.

Objects can be *collectives*, i.e., plural entities that aggregate parts (members), all of which play the same role with respect to the whole, or *functional complexes*, i.e., entities whose parts (called components) play different functional roles w.r.t. the whole. Finally, objects can also be *quantities*, i.e., portions of matter whose parts belong to the same type as the whole. Besides endurants, OntoUML has perdurants (occurrents, events) [2]. Events can bear their own properties, be decomposed, and have their types falling into taxonomies. However, events only exist in the past and, thus, are immutable. There are two categories of events: event kinds (stereotyped as «event») and event subkinds. Between endurants and events, we have a relation of participation, but events can also bring existence (i.e., *create*) objects. Finally, OntoUML embeds a theory of multi-level modeling and higher-order types [10]. Higher-order types (represented by the stereotype «type») are types whose instances are themselves types. A relation of *instantiation* connects individuals to these higher-order types.

4 Unpacking the VCM in OntoUML

We reconstruct the original conceptualization underlying VCM using ontological analysis associated with OntoUML[1]. The result of this analysis is captured in a series of modules comprising an integrated model[2].

Virus Infection. A Virus is a Biological Organism that can group, forming Virus Collectives. A Viral Infection is a bundle of relational dispositions connecting a Virus Collective and a Virus Host. The latter can be either a Living Host, which is a *role mixin* played by an Animal that could be, e.g., of Bat or Person kind, or an In Vitro Host, which is a role played by a Cell Line (itself a complex system of cells). The dispositions composing a Viral Infection can be manifested as Viral Disease events in living hosts, whereas cell lines, once infected, clearly show a Viral Response, but this is not considered a disease in full [7]. Viruses are instances of Virus Species, which are associated with Viral Disease Types. All Biological Organisms are instances of Biological Species, which can be characterized by a *name*, a *genus*, a *sub-family*, and a *family*[3]. If a Biological Organism happens to be a Virus then it instantiates a particular type of Biological Species called a Virus Species. Likewise, if a Biological Organism happens to be an Animal then it instantiates a particular type of Biological Species called an

[1] The only aspect of the VCM not presented here is the analysis the virus sequence publication. This is due to space constraints.

[2] In OntoUML models, we employ the generally-accepted color coding scheme: light red is used for types whose instances are objects, green when instances are relators, yellow when instances are events, and purple when instances are higher-order types.

[3] By using the multi-modeling support in UFO/OntoUML, one could proceed further and explicitly capture the relations of *subordination* between, for example, species and genus, namely, that a type *is subordinate to* another iff the instances of the former are specializations of instances of the latter [9]. For space reasons, we do not elaborate further.

Animal Species. Therefore, in the model of Figure 1, we have, for example, the instantiation relation between Animal and Animal Species that *redefines* the association ends of the more general relation between Biological Organism and Biological Species[4].

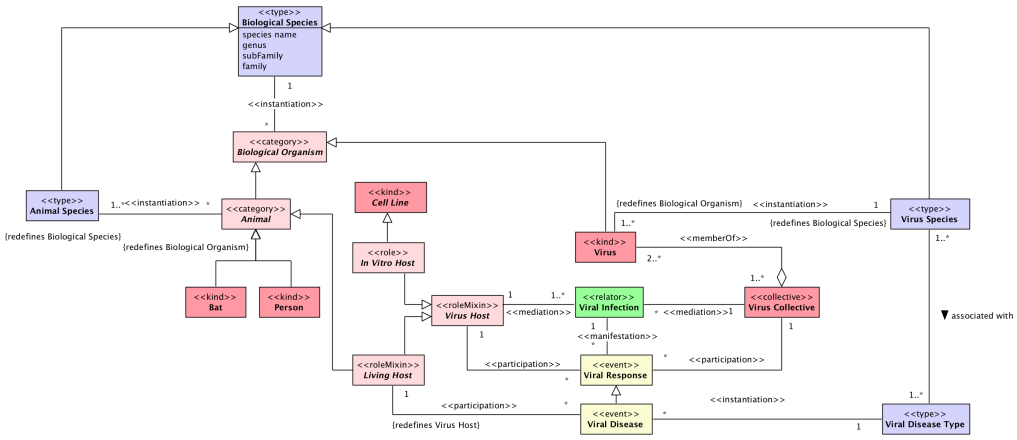

Fig. 1. A Virus Infection

Tissue Sampling. A Tissue Sampling is an event occurring at a particular *collection date*, in which a Sampling Laboratory and an Animal participate, and in which a Biological Tissue is extracted. A Biological Tissue is a portion of tissue of a given a Biological Tissue Type. If it is a Virus Infected Tissue then that Animal is a Living Host within the scope of a Viral Infection. Thus, a Virus Infected Tissue is *historically dependent* on the existence of a (current or previous) Viral Infection involving that Animal as a Virus Host[5]. A Virus Infected Tissue is then a sample of viruses of a given Virus Species (the type present in that Viral Infection). Finally, a Sampling Laboratory is a (processual) role played by a Research Laboratory (a particular type of Organization) in a Tissue Sampling event. A Research Laboratory is located in a given *country*, *region*, and *geo-group* (e.g., continent).

Virus Sequencing. A Sample Sequencing is an event characterized by a Sequencing Platform, a Sequencing Laboratory, and a Virus Infected Tissue. A Sample Sequencing event creates the collective of Virus Raw Data (comprising many thousands of Reads) from a sampled Virus Infected

[4] See [11] for an in depth discussion about the semantics of redefinition.

[5] Since the Viral Infection is existentially dependent on that particular Animal, then we can infer that the Virus Infected Tissue is also *historically dependent* on that Animal.

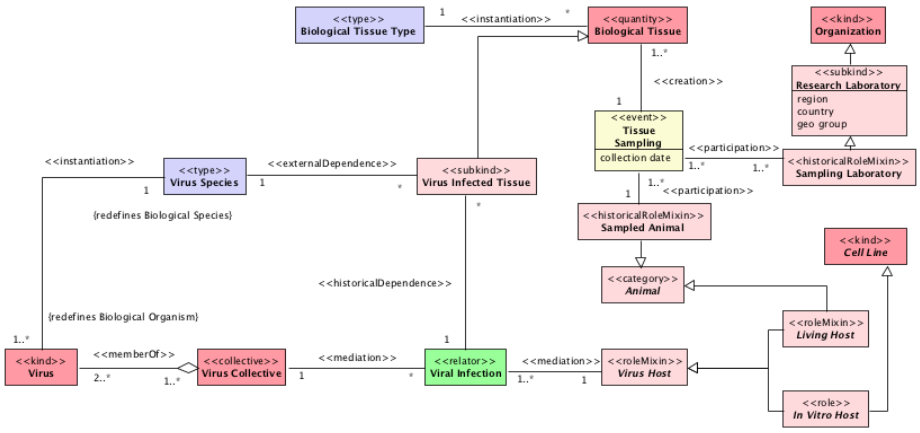

Fig. 2. Tissue Sampling

Tissue. The Virus Raw Data then participates to a Genome Assembly event performed by an Assembling Laboratory, and instantiating a particular Genome Assembly Method. This event produces a Full Consensus Sequencei i.e., a complete data record that represents the real virus sequence[6]. A Full Consensus Sequence is composed of Nucleotide Subsequences which, in turn, is composed of Nucleotides. A Virus Species is associated with molecules of a given Molecule Type (see discussion in paragraph 'Virus Sequence Annotation'). So, a Virus Sequence (of a virus of that species) instantiates exactly molecules of that type (again, see 'Virus Sequence Annotation'). The events of Tissue Sampling, Sample Sequencing, and Genome Assembly compose a super-event termed a Virus Sequencing event. Since a Virus Infected Tissue is created by a Tissue Sampling event, we can infer that a Sample Sequencing event must always be temporally preceded by a Tissue Sampling event. Analogously, since a Virus Raw Data collective is created by a Sample Sequencing event, a Genome Assembly event must always be temporally preceded by a Sample Sequencing event. Given the transitivity of these temporal precedence relations, we can infer that a Virus Sequencing event is composed of these sub-events that must occur in this particular temporal order.

[6] Full Consensus Sequence is a type of complete Virus Sequence record. Given the purpose of VCM, these entities are information/representation entities and not the actual chemical structures. In reality, a Full Consensus Sequence is a representation of a virus sequence type that is instantiated by actual complex chemical structures (actual sequences of nucleotides). Making this distinction explicit is beyond the scope of this paper.

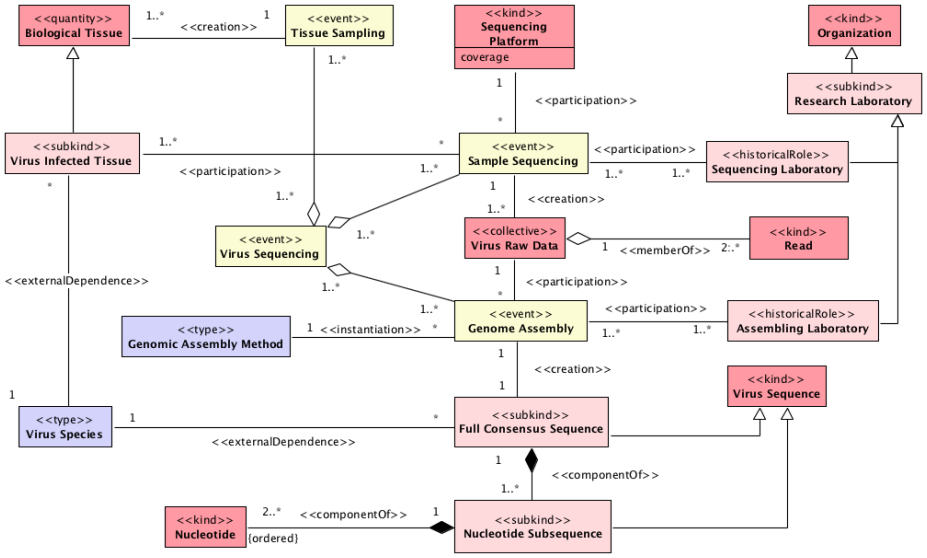

Fig. 3. Virus Sequencing

Virus Sequence Annotation. A Virus Sequence instantiates a Molecule Type[7]. A Virus Sequence is ultimately a sequence of Nucleotides. A Full Consensus Sequence is a Virus Sequence composed of a number of Nucleotide Subsequences. A Nucleotide Subsequence, thus, provides further structure to the Nucleotides composing the Full Consensus Sequences. Nucleotides are of certain types. In DNA these are Adenine (A), Cytosine (C), Guanine (G), and Thymine (T), whereas in RNA there is Uracil (U) in place of Thymine.

Particularly relevant Nucleotide Subsequences are Codons (sequences of three Nucleotides responsible for coding given Amino Acid Types) and Coding Regions (aggregations of Codons responsible for coding particular Protein types). A Coding Region participates in Translation events[8], which produce amino acid sequences (i.e., Proteins composed of a particular sequence of Amino Acids). There 20 known subtypes of Amino Acids of which we only show four examples: Leucine, Arginine, Threonine, and Proline. The type of Protein created by a Translation event can be derived from the involved Coding Region. This is because a Coding Region is composed of a

[7] Molecule Types can be double-stranded or single-stranded. These, in turn, can be positive- or negative-stranded. We restrict ourselves to single-stranded molecule types because the VCM focuses on single-stranded RNA viruses.

[8] Once more, Virus Sequences and their proper parts are information objects. Therefore, a Translation event here is not the actual biochemical event involving the real-world counterpart of these entities, but an information processing event that generates Protein representations from Coding Region representations.

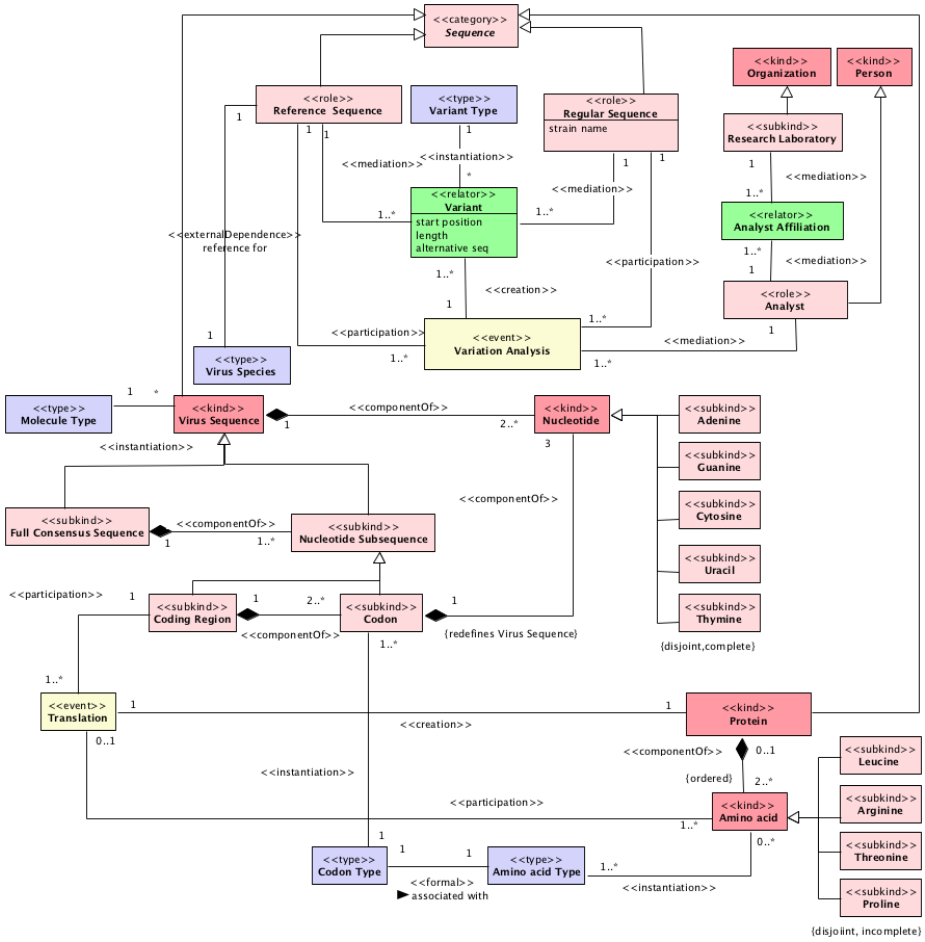

Fig. 4. Virus Sequence Annotation

particular sequence of Codons, each of which is of a Codon Type, determining a particular Amino Acid Type. So, the Protein created by that event from a particular Coding Region will be composed of a sequence of Amino Acids instantiating exactly those types determined by the Codons composing that Coding Region.

Virus Sequences and Proteins are subkinds of Sequence. A Sequence has two possible roles: Reference Sequence (unique for a Virus Species) and Regular Sequence. These two entities are mediated by the relator Variant, which is of a given Variant Type (e.g., insertion, deletion, etc.). This relator is created during a Variation Analysis event, to which one reference and one regular sequence participate. The event of Variation Analysis

is mediated by Analysts, who have an Analyst Affiliation to a Research Laboratory, which is another Organization.

5 Discussion

The Viral Conceptual Model facilitates effective and efficient data management process in a complex and challenging domain. However, focusing on data structure features alone can mask the complexity of a domain when concepts have a rich semantic structure. Our ontological unpacking provided a detailed analysis of the complex notions that are hidden within VCM and essential for acquiring a deep understanding of the domain. The ontological unpacking supports the conceptual aspect of VCM, and facilitates a detailed data manipulation process for the integration of diverse genomic data at a high level of detail. The ontological unpacking leads to the following advantages.

Making connections explicit. VCM collapses the feature types for very different biological entities (genes, peptides, etc.), which could lead to confusion. For example, genes could, in principle, have nucleotide variants; however, if they are included in the Annotation entity, they can only be connected to amino acid variants. Our unpacking makes tangible the connection between amino acid variants and nucleotide variants, now called a Translation event.

Unpacking process information. VCM compacts the technological information within Experiment Type, which could result in confusion for users who observe the sequence creation event and its sequential steps. When understood in ontological terms, it clarifies that sequence characteristics depend on the sequencing technology, which impacts both a sequencing and an assembling event.

Disambiguating collapsed concepts. The *Sequencing Lab* attribute of Sequencing Project incorrectly collapses the concepts of the sequencing laboratory and submitting laboratory, even if they are different entities. Originally, the two different interpretations were derived from two distinct data sources, then integrated using a single VCM attribute. However, this process overlooks sampling, sequencing/analysis, and submission, as instead captured by the unpacking. Further details are omitted for space reasons.

Clarifying underspecified aspects. VCM includes all the information about the host sample in one single entity, thus generating a denormalized structure that is conceptually incorrect. The event of data sampling which involves actors such as a healthcare worker, an individual being tested, and the testing facility, are underspecified, even though the event occurs at a precise point in time and space. Consider the evolution of SARS-CoV-2 variants: when a claim is made that a specific lineage is becoming prevalent in a geographical area, it is important to correctly interpret the spatial information. Sequences are assigned a "location", which could describe *where*: i) human hosts *live*; ii) phials were *sampled/stored*; iii) sequence information was *extracted* into files; or iv) information was *deposited* to public databases.

6 Conclusion

This research analyzed the Viral Conceptual Model, which was developed for sequencing viruses. Using OntoUML, we conducted an ontological analysis to unpack and identify the core components of the Viral Conceptual Model. The results demonstrate the effectiveness of this conceptual model and the usefulness of unpacking a complicated model to improve its clarity, beyond data structuring. Future research could perform similar unpacking to assesses and explain the clarity of other conceptual models developed for biological and additional complex domains.

Acknowledgements. A. Bernasconi is supported by ERC Advanced Grant 693174 GeCo; V. Storey by J. Mack Robinson College of Business, Georgia State University; G. Guizzardi by the NeXON Project (UNIBZ).

References

1. Al Khalaf, R., et al.: CoV2K: a Knowledge Base of SARS-CoV-2 Variant Impacts. In: Int. Conf. on Research Challenges in Information Science. pp. 274–282. Springer (2021)
2. Almeida, J.P.A., et al.: Events as entities in ontology-driven conceptual modeling. In: Int. Conf. on Conceptual Modeling. pp. 469–483. Springer (2019)
3. Bernasconi, A., et al.: A review on viral data sources and search systems for perspective mitigation of COVID-19. Briefings in Bioinformatics (2021)
4. Bernasconi, A., et al.: Empowering virus sequence research through conceptual modeling. In: Int. Conf. on Conceptual Modeling. pp. 388–402. Springer (2020)
5. Bernasconi, A., et al.: Conceptual modeling for genomics: building an integrated repository of open data. In: Int. Conf. on Conceptual Modeling. Springer (2017)
6. Bernasconi, A., et al.: VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants. Nucleic Acids Research (2021)
7. Boyd, K.M.: Disease, illness, sickness, health, healing and wholeness: exploring some elusive concepts. Medical Humanities **26**(1), 9–17 (2000)
8. Canakoglu, A., et al.: ViruSurf: an integrated database to investigate viral sequences. Nucleic acids research **49**(D1), D817–D824 (2021)
9. Carvalho, V.A., et al.: Multi-level ontology-based conceptual modeling. Data & Knowledge Engineering **109**, 3–24 (2017)
10. Carvalho, V.A., et al.: Using a well-founded multi-level theory to support the analysis and representation of the powertype pattern in conceptual modeling. In: Int. Conf. on Advanced Inf. Systems Engineering. pp. 309–324. Springer (2016)
11. Costal, D., et al.: Formal semantics and ontological analysis for understanding subsetting, specialization and redefinition of associations in uml. In: ER (2011)
12. Guizzardi, G.: Ontological foundations for structural conceptual models. CTIT, Centre for Telematics and Information Technology (2005)
13. Guizzardi, G., et al.: Towards ontological foundations for conceptual modeling: The Unified Foundational Ontology (UFO) story. Appl Ontol **10**(3-4), 259–271 (2015)
14. Pastor, O., et al.: Enforcing conceptual modeling to improve the understanding of human genome. In: Proc. (RCIS) 2010. pp. 85–92. IEEE (2010)
15. Paton, N.W., et al.: Conceptual modelling of genomic information. Bioinformatics **16**(6), 548–557 (2000)
16. Román, J.F.R., et al.: Applying conceptual modeling to better understand the human genome. In: Int. Conf. on Conceptual Modeling. Springer (2016)