An Initial Empirical Assessment of an Ontological Model of the Human Genome

Alberto García S.[1,⋆], Anna Bernasconi[1,2,⋆], Giancarlo Guizzardi[3,4], Oscar Pastor[1], Veda C. Storey[5], and Mireia Costa[1]

[1] Universitat Politècnica de València, Valencia, Spain
{algarsi3,opastor,micossan}@pros.upv.es, abernas@upvnet.upv.es
[2] Politecnico di Milano, Milan, Italy anna.bernasconi@polimi.it
[3] Free University of Bozen-Bolzano, Bolzano, Italy Giancarlo.Guizzardi@unibz.it
[4] University of Twente, Twente, The Netherlands g.guizzardi@utwente.nl
[5] Georgia State University, Atlanta, Georgia, USA vstorey@gsu.edu

Abstract. Conceptual modeling is used to model application domains for which an information system is needed. One of the most complex domains to which conceptual modeling has been applied is that of the human genome. Due to its complexity, its understanding is often left to domain experts. Conceptual models represent genomics-related concepts, with various purposes, including domain clarification or data structures design for facilitating data integration. However, traditional conceptual models, which might be expressed, for example, with UML, may not be appropriate for properly explaining such a complex domain, thus requiring an additional layer to ground the model on well-accepted ontological foundations. To achieve this result, an "ontological unpacking" method has been proposed that uses OntoUML as a visual formalism. In this research, we carry out an empirical study to compare the two mentioned representations. The study involved a small group of participants, who responded to a set of questions by reading either a UML model or its related OntoUML unpacked version; the results enabled us to assess their understanding of the domain. We aim to initiate a practical evaluation framework to assess the effectiveness, efficiency and user beliefs of models derived by ontologically unpacking traditional conceptual models. The results of the analysis provide the basis for a broader assessment.

Keywords: Empirical evaluation · Ontological Unpacking · Conceptual Model · Human Genome

1 Introduction

Genomic science is a complex interdisciplinary domain, whose understanding is so far accessible only to researchers with a strong background in biology and genetics. Its interpretation becomes problematic also because there has been a lack of effort to translate its mechanisms into modeling languages that are more

[⋆] A.G.S. and A.B. should be regarded as joint first authors.

easily understandable by computer scientists. Computer science traditionally employs modeling languages such as UML, i.e., a standard graphical modeling language that allows designers to create conceptual models. Instead, OntoUML is an ontologically well-founded language for Ontology-driven Conceptual Modeling, built as a UML extension based on the Unified Foundational Ontology [8]. OntoUML supports modelers in systematically making ontologically consistent representation choices, and thus, making explicit the ontological nature of the elements represented.

We previously created a method of ontological analysis that reveals the ontological foundation of the information represented in a conceptual model. The method, called 'ontological unpacking', allows the modeler to unfold and explain previously existing UML models, transforming them into a corresponding OntoUML version. A previous effort has been successfully performed on the viral sequences domain [9] in order to improve semantic interoperability.

We postulate that ontologically unpacked models provide a clear and understandable representation of a complex domain, such as genomics, even though it might require considerable effort to learn OntoUML, which is necessary to perform the unpacking. To investigate, we here conduct an initial experiment where students without previous biological knowledge are given competency questions regarding a conceptual model, using either a UML model or its corresponding OntoUML model, obtained as a result of an ontological unpacking procedure. The experiment is carried out using a portion of a conceptual schema of the human genome as a complex domain [4,5]. Specifically, we consider the part describing human metabolic pathways. The original schema was conceived using UML; in our recent work [3] we performed an ontological analysis exercise, producing the corresponding OntoUML version. These two models are object of the experiment thereon described. We formulate a set of research questions aimed at understanding if OntoUML delivers better quality models than UML. Our research questions can be translated into formal metrics according to ISO 25000 (i.e., effectiveness, efficiency and user beliefs).

Ontology driven conceptual modeling has previously been compared to traditional conceptual modeling in [16]. Here, we do not compare different languages or paradigms of modeling; instead, we compare the capability of different models to completely and unambiguously represent a domain, serving the intended purpose of explaining that domain to a non-expert user working in it for the first time. Empirical studies of conceptual modeling applications have been performed on tools [7] also related to genomics [2], measuring the understandability of modeling artifacts [11,12]. This paper describes our initial experiment and discusses a number of lessons learned. Future work will include further experimentation and statistical analysis to assess the use of ontological unpacking.

2 Background

The first Conceptual Schema of the Human Genome was proposed in 2011 [14] by the Research Center on Software Production Methods (PROS) at the Polytechnic

University of Valencia. Since then, several extensions have been produced [15,4]. The current version of the schema is a map of concepts and relationships grouped into different genomic knowledge modules, called the Conceptual Schema of the Genome v3 (CSG) [5]. Its UML schema includes five modules, describing the structure of the human genome, protein synthesis, changes in the sequence referring to a reference sequence, information and sources related to the elements of the conceptual schema, and human metabolic pathways. Prior work on ontological unpacking [3] of the schema focused on a relevant portion of the last view; that is, on metabolic pathways. We sought to understand the impact of ensuring ontological clarity in the concepts employed.

The original UML pathway schema presents 19 entity classes, with six generalizations, two aggregations, one self-relation, and three normal relations. We also have one integrity constraint. The unpacked OntoUML schema has 26 entities, of which 17 are from UFO-A [10] (including kinds and subkinds, collectives and categories, phase/roleMixins) and 9 from UFO-B [1], including 5 events and 4 historicaRolelMixins.

Different relationships include 11 generalizations, three aggregations, one composition, and four other regular ones, covering several relationship stereotypes; namely, «creation», «termination», «memberOf», «participational», and «historicalDependence». Here we do not explain OntoUML stereotypes but we refer the interested readers to [8].

3 Methodology

We carried out an empirical assessment by designing a study to answer three fundamental research questions:

RQ1: *Do subjects benefit from a better understanding of a complex domain with OntoUML rather than with UML?*
RQ2: *Do subjects answer competency questions faster with OntoUML than with UML?*
RQ3: *Do subjects have more positive beliefs after using OntoUML rather than UML?*

The experimental design is based on the one described in [16], thus divided into four steps, namely: variable development, subject selection, experimental design type, and instrumentation.

Variable Development. Based upon our research questions, our independent variable is the modeling language, with two possible treatments, UML and OntoUML. The dependent variable is the quality of the models, observed using three dimensions, captured with different metrics:

- *Effectiveness*: measured through the percentage of correct true/false answers given to competency questions.
- *Efficiency*: measured through the time to answer to competency questions.
- *User beliefs* – divided into the three sub-dimensions *perceived usefulness* (PU), *perceived ease of use* (PEOU), and intention to use (ITU): measured through 1-to-5 Likert scale questionnaires.

Subject Selection. The experiment was performed at the Polytechnic University of Valencia in two classes of the fourth year of the Computer Science curriculum. There were 20 participants aged 22 to 30. Given the small sample size, this can be considered as a quasi-experiment. Only the previous IT background of the participants was taken into account, since none of the subjects declared any previous experience with biological/genomic topics.

Experimental Design Type. Participants all together filled a demography survey, received an introduction to the two involved modeling languages and were tested on them. Then they were divided into four groups. Figure 1 shows the experimental setup given to each of them. We created two questionnaires, Q1 and Q2, that are equivalent in terms of difficulty and coverage of modeling constructs/topics. Participants were asked to answer Q1 and Q2 in different orders, using alternatively: 1) the original UML-based CSG pathway view, or 2) the ontologically unpacked OntoUML-based CSG pathway view. In this way, all the four possibilities (questionnaire Q1 or Q2 answered with UML or OntoUML schemata) were covered, overcoming possible biases due to the order of issuing. We performed random assignment of Q1 and Q2 to the subjects.

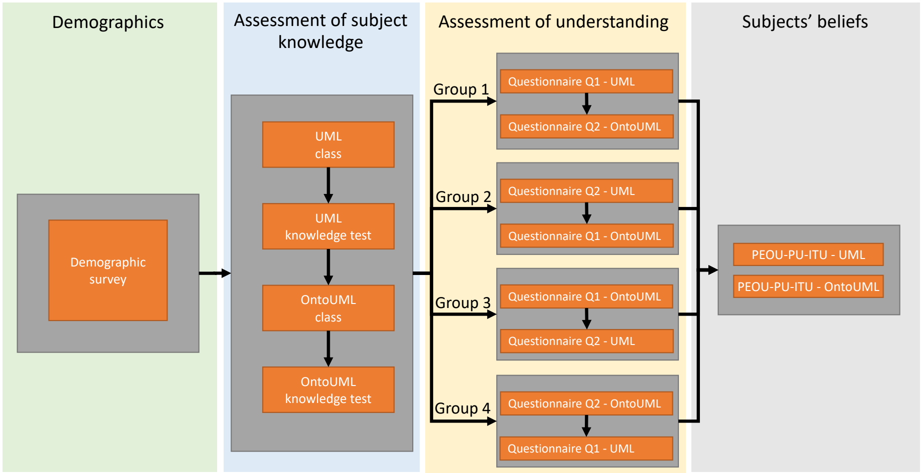

Fig. 1. Experimental design overview. [PEOU: Perceived ease of use; PU: Perceived usefulness; ITU: Intention to Use]

Instrumentation. The experiment was conducted using the PoliFormat platform (`https://poliformat.upv.es/`), on which the students had a personal login access. All the materials used for introducing the topics to the classes and to assess their understanding are provided as supplementary material on a Zenodo repository [6]. In the remainder of the text, we refer to specific handouts using the same acronyms as in the repository.

Demographics. As shown in the first block of Figure 1, the demographic survey consisted of eight questions (see file DS) aimed to grasp a more complete picture of the participants group to better interpret the results.

Table 1. Questionnaire Competency Questions

Quest.	Group	ID	Competency Questions
Q1	Entities	1	Polymers are composed of other polymers.
		2	The internal structure of any polymers is homogeneous.
		3	The internal structure of basic biological entities and polymers is the same.
	Events	4	Processes are limited in time.
		5	Pathways must be composed of other pathways.
		6	A process can be decomposed into other events.
	Interact.	7	Every biological entity must participate in at least one process.
		8	Biological entities can take part in pathways.
		9	A protein can take the roles of input, output, and regulator in the same process
Q2	Entities	10	Some polymers are composed of nucleotides.
		11	Every enzyme is a polymer.
		12	Some basic biological entities can be polymers also.
	Events	13	Every event must have a preceding event.
		14	Pathways can be composed of other pathways.
		15	Events occur in a specific time interval.
	Interact.	16	Biological entities can be created and destroyed as a result of a process.
		17	Biological entities can participate in multiple processes.
		18	A protein can take the role of input in different processes.

Assessment of subject knowledge. As shown in the second block of Figure 1, training on both UML and OntoUML was offered to all study participants (see files SK1 and SK2), aiming to eliminate all possible differences due to the background knowledge of the participants. Each training session lasted 45 minutes and was supported by a slides presentation, including theory and practical examples taken from domains not related to the one in the experiment. After each training session, participants answered to a questionnaire testing their understanding of example models (see files SK3 and SK4).

Assessment of understanding. As shown in the third block of Figure 1, the participants were divided into four groups, which answered the same sets of questions by using two different models: 1) the original UML model (see file UA5); 2) ontologically unpacked OntoUML model (see file UA6). The questions were provided by Biology expert collaborators, as they deemed them relevant for the domain. In this way, questions were guaranteed to be independent with respect to the models. After processing the statements provided by experts, we composed three groups of questions, respectively targeting Entities, Events, or Interactions between entities and events. This pre-processing allowed the composition of questionnaires Q1 and Q2 in a balanced way with respect to the models' interpretation challenges. Table 1 shows the sets of questions divided by questionnaire number and by group. The different versions of the questionnaires for the 4 groups are available in handouts UA1–UA4. A pilot run was performed before the experiment with expert collaborators to ensure that the task had the appropriate level of complexity.

Subjects' beliefs. As shown in the fourth block of Figure 1, all participants finally answered two surveys of 16 questions to assess their beliefs, in terms of PU (8 questions), PEOU (6 questions), and ITU (2 questions) according to Method Adoption Model (MAM, [13]), measured using a Likert scale.

4 Results

Results obtained by running the experiment with the selected participants follow.

Demographics. The involved students declared a Grade Point Average of about 8/10. Their working experience was very heterogeneous. Most participants had <1 year work experience, whereas two had worked longer. Their demographic survey results revealed that the participants: i) knew and had previously used UML; ii) had no previous experience with OntoUML; and iii) were not knowledgeable in the observed domain.

Assessment of subject's knowledge. Eighteen students successfully passed the two UML and OntoUML tests with > 75% correct answers in both tests. One participant did not obtain the sufficient threshold in OntoUML (62.5%); another did not pass either tests (50% and 57.14%). These two students received an additional class and were dedicated additional time for answering any questions they had. We ensured that sufficient understanding was reached before proceeding to the next stages.

Effectiveness of treatment. Figure 2A shows, for each question, the percentage of correct answers received by participants using either UML or OntoUML. Questions are grouped by category. It can be observed that: i) Entities-related questions were answered correctly by 68.33% of participants using UML and by 76.67% using OntoUML; ii) Events-related questions were answered correctly respectively by 56.67% (UML) and 83.33% (OntoUML); and iii) Interactions-related questions were answered correctly respectively by 58.33% (UML) and 56.67% (OntoUML).

Efficiency of treatment. Figure 2B shows, for each question, the working mean times (measured in seconds) spent by the participants to provide answers, using either UML or OntoUML. Questions are grouped by category. Questions answered with the OntoUML model took longer than questions answered with UML. Specifically: i) questions related to Entities and Interactions required subjects approximately 30 seconds longer to answer; ii) the difference decreases to approximately 20 seconds for Event-related questions; iii) times required to answer UML-based questions showed a higher variability than those for OntoUML-based questions. For example, the time required to answer Entities-related questions in UML ranged from 63 to 89 seconds (26 seconds difference) and from 95 to 109 seconds (14 seconds difference) in OntoUML.

Subjects' beliefs. Figure 3A shows, for each question of the MAM (grouped by sub-dimension of the user belief), the partition of respondents who strongly disagreed, disagreed, was neutral, agreed, or strongly agreed with the provided statement. The same structure is used in Figure 3B for OntoUML. From the results, it could be observed that: i) subjects perceived that UML is much easier to use (the average in PEOU questions with UML scored 0.83 more than with OntoUML); ii) subjects perceived that UML is more useful (difference of PU averages 0.38); iii) if subjects had to choose which language to use in genomics, they would prefer UML by a substantial margin (difference of ITU averages 0.9).

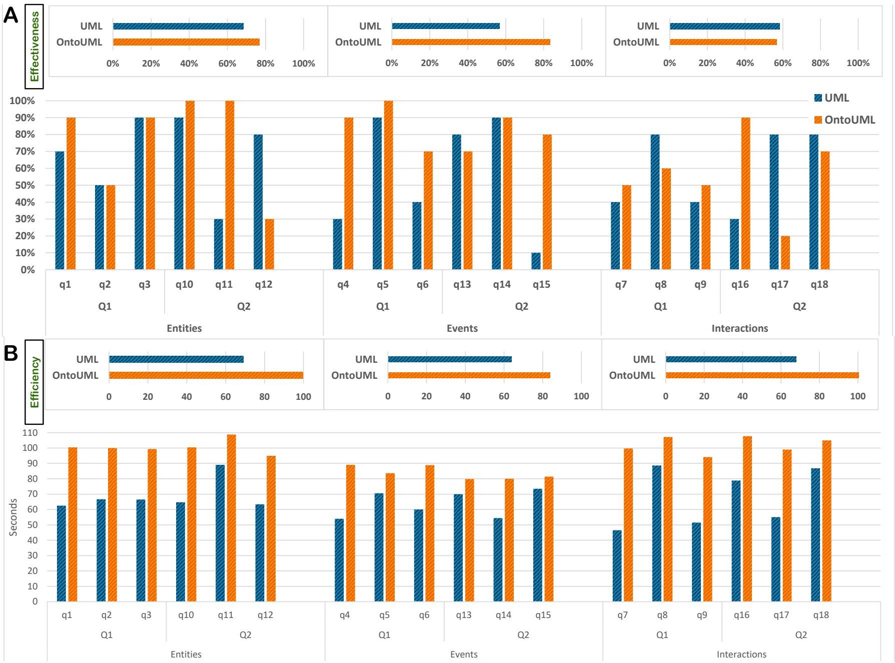

Fig. 2. Panel A: Barplot of correct answers % given by participants grouped by question number and organized by group. Panel B: Barplot of seconds employed to answer questions, grouped by question number and organized by group.

OntoUML results regarding user beliefs generally scored worse than UML results. The single OntoUML questions regarding perceived usefulness (PU) were answered with more positive scores (i.e., (strongly) agree) than negative scores (i.e., (strongly) disagree). Since two subjects reported previous experience with OntoUML, their assessments were analyzed separately. One rated OntoUML with almost 2 point less than UML on each metric. The other one, rated OntoUML higher in PU (0.375 points more) and ITU (1 point more).

5 Discussion

This research is a preliminary evaluation of the ontological unpacking method, aimed at comparing the ability to understand a complex domain through an ontologically unpacked (OntoUML) model, rather than from its corresponding traditional conceptual (UML) model. With respect to the Effectiveness assessment, the main findings can be summarized as follows. Entity-related questions were answered more successfully with OntoUML; this could be due to the fact that UFO-A contains stereotypes that helped clarifying important principles (such as rigidity). Events-related questions were also answered more successfully with OntoUML, showing an even more apparent difference; this suggests that

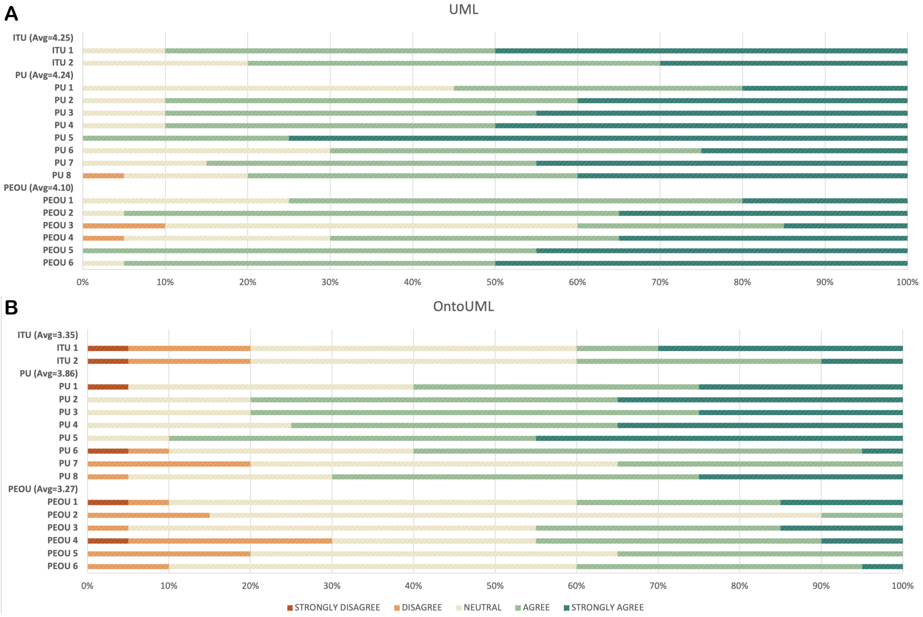

Fig. 3. The horizontal stacked bar plots represents subjects' beliefs regarding the use of UML (Panel A) and OntoUML (Panel B) in terms of intention to use (ITU), perceived usefullness (PU), and perceived ease of use (PEOU).

the ontological foundation of events presented in the UFO-B fragment may have helped participants to capture relevant details regarding event-related information. Questions related to the Interaction between events and entities were instead answered more successfully with UML (2% difference). Further comments can be formulated by analyzing the results of specific questions:

- *Temporality of events.* One of the main purposes of Conceptual Modeling is making implicit concepts explicit. From a biological perspective, it is clear that events are limited in time. However, in the UML version of the considered model, the temporal limitations of a process are implicit. From the ontological unpacking method, which also commits to UFO-B, such information was extracted and explicitly represented by means of the «event» stereotype. Results of questions Q4 (OntoUML: 90% of participants answered correctly, UML: 30%) and Q15 (OntoUML: 80%, UML: 10%), which grasped this aspect, were thus expected.
- *Mereology of events.* The UML version of the considered model provides a simple explanation of the participation of entities in the processes. OntoUML, instead, provides a more complex and detailed explanation. Note that processes in genomics are considered chemical compounds that can be divided further. Question Q6 highlighted that the UML model left the individual participation of chemical compounds in reactions implicit. As a re-

sult, the question was answered with a higher score using OntoUML (70%) instead of UML (40%). However, representing how chemical reactions are decomposed into smaller events (which capture the individual participation of each chemical compound) increased the overall complexity. This may have confused respondents of Q17 (which concerns participation in multiple processes), who ended up scoring 20% with OntoUML and 80% with UML.
- *The rigidity principle.* A significant difference is observed in Q16 (OntoUML 90% vs. UML 30%), possibly due to the capability of OntoUML to express the «phase» stereotype, exploiting the principle of rigidity [8]. This principle makes explicit the fact that chemical compounds and biological-related substances are created and destroyed as a result of chemical reactions.

Thus for *RQ1*, we can conclude that OntoUML was more effective in conveying the genomics domain to the study participants, even if for some elements, the simplicity of the UML representation still achieved the desired outcome.

Regarding the Efficiency assessment, the initial expectation suggested that a complex domain explained through a more complete and explicit model would translate into shorter answering times. However, this preliminary study suggested that OntoUML required instead more time to participants, in order for them to be able to answer questions based on it. Thus, *RQ2* receives a negative answer. A possible explanation is that OntoUML is more complex and participants had a very limited experience with it.

Regarding the User beliefs assessment, UML was more appreciated, probably due to the fact that OntoUML is more complex and was new to the participants who lacked any experience with the language. ITU opinions regarding OntoUML are strongly related to the results obtained for PEOU, because subjects will be reluctant to use a language whose learning barriers are higher than those of simpler alternatives. To answer *RQ3* considering the overall user beliefs, we can conclude that participants were hesitant to learn and use a novel modeling language, especially a complex one, in a short amount of time. However, the results indicate that performances, in terms of effectiveness, were better using OntoUML, although participants were not fully aware of this.

Previous OntoUML experience delivers better results. The two subjects with previous experience scored perfect results. Stronger opinions were revealed in the two subjects with previous experience. One had more negative beliefs; the other had better opinions than the average.

6 Conclusion

Conceptual modeling has been applied to complex domains, such as the human genome. In this paper, we describe an initial experiment for evaluating an 'ontological unpacking' method. The results showed that the participants' use of OntoUML, as needed for the ontological unpacking, achieved more correct responses than UML, although they took longer to respond. The experiment results revealed lesser intention to use, perceived ease of use, and perceived usefulness of OntoUML on the part of the participants. Based on these preliminary results, we plan to design broader experiments, that will include larger groups of participants, more heterogeneous subjects (in terms of age and background), and hypothesis testing based on the three research questions described here.

Acknowledgements. This work was supported by ACIF/2021/117, INBIO2021/AP2021-05, MICIN/AEI/10.13039/ 501100011033, and INNEST /2021/57 grants.

References

1. 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)
2. Bernasconi, A., et al.: Exploiting conceptual modeling for searching genomic metadata: a quantitative and qualitative empirical study. In: Int. Conf. on Conceptual Modeling. pp. 83–94. Springer (2019)
3. García S., A., et al.: An ontological characterization of a conceptual model of the human genome. In: Int. Conf. on Advanced Information Systems Engineering (CAiISE) Forum (2022)
4. García S., A., et al.: Towards the understanding of the human genome: a holistic conceptual modeling approach. IEEE Access **8**, 197111–197123 (2020)
5. García S., A., et al.: A conceptual model-based approach to improve the representation and management of omics data in precision medicine. IEEE Access **9**, 154071–154085 (2021)
6. García S., A., et al.: UML vs OntoUML analysis results [Data set] (Jun 2022). https://doi.org/10.5281/zenodo.6616114
7. Gray, T., et al.: Empirical evaluation of a new demo modelling tool that facilitates model transformations. In: Int. Conf. on Conceptual Modeling. pp. 189–199. Springer (2020)
8. Guizzardi, G.: Ontological foundations for structural conceptual models. CTIT, Centre for Telematics and Information Technology (2005)
9. Guizzardi, G., et al.: Ontological unpacking as explanation: The case of the viral conceptual model. In: Int. Conf. on Conceptual Modeling. pp. 356–366. Springer (2021)
10. Guizzardi, G., et al.: Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story. Appl Ontol **10**(3-4), 259–271 (2015)
11. Jošt, G., et al.: An empirical investigation of intuitive understandability of process diagrams. Computer Standards & Interfaces **48**, 90–111 (2016)
12. Liaskos, S., et al.: Experimental practices for measuring the intuitive comprehensibility of modeling constructs: an example design. In: Int. Conf. on Conceptual Modeling. pp. 231–241. Springer (2020)
13. Moody, D.L., et al.: Evaluating the quality of information models: empirical testing of a conceptual model quality framework. In: 25th Int. Conf. on Software Engineering, 2003. Proceedings. pp. 295–305. IEEE (2003)
14. Pastor, O., et al.: Model-based engineering applied to the interpretation of the human genome. In: The Evolution of Conceptual Modeling, pp. 306–330. Springer (2011)
15. Reyes Román, J.F., et al.: Applying conceptual modeling to better understand the human genome. In: Int. Conf. on Conceptual Modeling. pp. 404–412. Springer (2016)
16. Verdonck, M., et al.: Comparing traditional conceptual modeling with ontology-driven conceptual modeling: An empirical study. Inf Syst **81**, 92–103 (2019)