Exploiting Conceptual Modeling for Searching Genomic Metadata: A Quantitative and Qualitative Empirical Study

Anna Bernasconi(✉), Arif Canakoglu, and Stefano Ceri

Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, Milan, Italy
{anna.bernasconi,arif.canakoglu,stefano.ceri}@polimi.it

Abstract. Providing a common data model for the metadata of several heterogeneous genomic data sources is hard, as they do not share any standard or agreed practice for metadata description. Two years ago we managed to discover a subset of common metadata present in most sources and to organize it as a smart genomic conceptual model (GCM); the model has been instrumental to our efforts in the development of a major software pipeline for data integration.

More recently, we developed a user-friendly search interface, based on a simplified version of GCM. In this paper, we report our evaluation of the effectiveness of this new user interface. Specifically, we present the results of a compendious empirical study to answer the research question: *How well is such a simple interface understood by a standard user?* The target of this study is a mixed population, composed by biologists, bioinformaticians and computer scientists.

The result of our empirical study shows that the users were successful in producing search queries starting from their natural language description, as they did it with good accuracy and small error rate. The study also shows that most users were generally satisfied; it provides indications on how to improve our search system and how to continue our effort in integration of genomic sources. We are consequently adapting the user interface, that will be soon opened to public use.

Keywords: Conceptual model · Data integration · Genomics · Next generation sequencing · Open data · Evaluation · Usability

1 Introduction

With progress of DNA sequencing technology, many international consortia are providing public, open datasets that can be used for answering research questions, from biological (e.g. what are the basic mechanisms for explaining DNA organization and gene activation) to clinical (e.g., finding gene panels that can be

G. Guizzardi et al. (Eds.): ER 2019 Workshops, LNCS 11787, pp. 83–94, 2019.
https://doi.org/10.1007/978-3-030-34146-6_8

used for effectively separate cancer patients into classes by observing their expression). In most cases, public datasets must be assembled from several sources, each providing specialized information (e.g., genome annotations, mutations, gene expression, protein bindings to DNA, and so on). Thus, researchers must be able to inspect metadata that describe experimental conditions, so as to ascertain their relevance with respect to the research question and how many instances of compatible data are available for supporting their study.

To facilitate this task, we started two years ago a large data integration project, with the ambitious objective of collecting the open source content of many important genomic sources into a single repository, with integrated and normalized metadata. The integrated repository offers to users a single data organization that can be inspected with a single search query. Before integration, metadata at the various sources were understood and translated to a standard conceptual model, designed at the start of our project, and discussed in [3].

The conceptual model drives the periodic integration process of genomic data, as it allows to recognize and periodically extract non-overlapping portions of datasets at each source; a software pipeline is used for data injection from each source to an integrated repository. For what concerns metadata in particular, the pipeline includes value normalization and enrichment steps that improve the ability to compare metadata from different sources. Currently, we have integrated experimental genomic data from Encyclopedia of DNA Elements (ENCODE), The Cancer Genome Atlas (TCGA), Roadmap Epigenomics, subsets of Gene Expression Omnibus, Cistrome, and annotations from GENCODE and RefSeq (see references in companion paper [2]); we plan to add many other sources.

For searching metadata, we provided two very different user interfaces. One interface, described in [2], is focused on explaining the inference process that we can perform on metadata in order to do data matching. Such interface is made available to expert users (and to us) to clarify the process of query and inference, with a diagrammatic representation of inference results, where all the connections are extensively shown. We soon realized that such interface is too complex for most of our generic users, who are biologists or bioinformaticians with no experience of knowledge graph matching. We then developed a user-friendly interface, which actually hides not only the inference steps, but also the complexity of the conceptual model. We translated the ER model into a much simpler denormalized structure consisting of a star with four related dimensions, which can be queried by using a structured form; the complexity of ontological inference was implemented in the user-friendly interface as just a *check-board*, by means of which the user can augment or reduce the inference process, hence the number of choices that are made available for satisfying a search query.

In principle, we were uncertain that this transformation could capture at the same time the original semantics and the user understanding. The main focus of this paper is to report our evaluation of the effectiveness of the user interface and henceforth of our data transformation. Specifically, we report the results of a compendious empirical study to answer our research question: *How well is such a simple interface understood by a standard user (i.e., students, biologists, interdisciplinary user base)? In how many cases it fulfills the data integration needs without errors?*

The target of this study is a mixed population, composed by biologists, bioinformaticians and computer scientists, who participated to the empirical study; most of them were totally unaware of conceptual modelling guidelines and, as such, well represented our target users. We are currently considering their feedback and adapting the user interface, that will be soon opened to public use.

Related Work. In the last few years, the focus of empirical studies dedicated to conceptual modelling has ranged from works on tools based on CM [10], to process mining [1] and to artifact sampling [5]. A broad study has compared traditional conceptual modeling with ontology-driven conceptual modeling [9]. Some recent works employ conceptual models to explain biological entities and their interactions [6,8], or to characterize the objects during analysis workflows [7]. Our use of conceptual modeling is aimed at data integration for the purpose of building a new resource and make it publicly available.

Paper Organization. In Sect. 2 we describe the CM and explain its reduction to four simple views, which drive the user-friendly search system; we also explain how search is performed. In Sect. 3 we illustrate how we designed our study, by first providing instructional material and then asking to provide answers to an online questionnaire, whose questions test specific aspects of our system; in Sect. 4 we discuss the study results and in Sect. 5 we conclude.

2 Genomic Conceptual Model: Original and Simplified

We report here a synthetic description of the Genomic Conceptual Model [3] and then we explain its simplification operated to support the user-friendly interface.

Genomic Conceptual Model. GCM is a star-like entity-relation model that summarizes the common organization of a limited set of concepts supported by most genomic data sources, although originally with different formats and names. In the upper part of Fig. 1 we show its sketch from [3] (this conceptual representation is also used for the advanced user interface, see [2]); with respect to the original GCM one can note some small changes, which are due our experience of use of the model. The Item represents the central entity of the schema: a single experimental (or annotation) file of genomic regions with their properties. The schema includes four dimensions (or views) that describe the biological phenomena observed in the experiment (entities Donor, BioSample and Replicate), the management aspects of the experiment (entities Project and CaseStudy), the technological process used for the production of the item (entity ExperimentType), and the extraction parameters used for internal selection and organization of items (entity Dataset). One-to-many relationships connect the various entities to the Item; two many-to-many relationships are needed for the relationships between Item and Replicate (as the same item can be used in replicated experiments) and between Item and CaseStudy (as the same item can be used in several use cases).

The GCM schema is extended by two sub-models representing, respectively, the original unstructured metadata and the semantic enrichment for specific attributes. Many attributes and their respective values discovered within sources

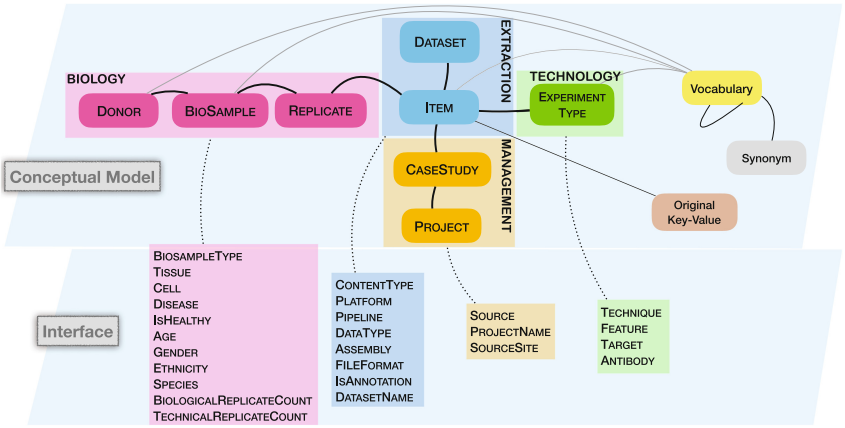

Fig. 1. Upper part: the genomic conceptual schema, which includes 8 entities connected by 5 one-to-many relationships and 2 many-to-many relationships. Lower Part: the simplified conceptual model used for our user-friendly search interface, which is based on 4 denormalized views; the figure includes the attributes selected in each view.

cannot be mapped to the same conceptual model. Thus, these metadata are directly downloaded from the original sources and transformed into key-value pairs. Moreover, as result of a normalization and enrichment phase, we associate specific values of the GCM with controlled terms (see [4]). Out of all GCM attributes, we selected ten of them as worthy of enrichment. We selected one or two preferred bio-ontologies for each attribute, and linked to each value a term from the chosen ontology, equipped with its synonyms and a small hierarchy of hypernyms and hyponyms, connected though *is_a* and *part_of* relationships.

Design of a Simplified View Supporting the User-Friendly Interface. As the start point for a user-friendly interface we opted for a drastic simplification of the model. We merged the ITEM entity with the extraction dimension and we denormalized all many-to-many relationships; denormalization was applied to items having multiple replicas and to items appearing in the same case study. We also selected some of the attributes from the entities of each dimension (26 out of 38 attributes in the current GCM) based on typical use, while the other attributes were re-inserted as key-value pairs. The bottom part of Fig. 1 illustrates the resulting schema, with the four dimensions connected to each ITEM and the 26 attributes selected from the entities of each dimension that were chosen to appear in the interface.

Items are implicitly gathered within a folder or dataset; the simplified *Extraction View* includes: DATASETNAME, denoting the folder gathering the items; CONTENTTYPE, type of genomic regions in the file (such as gene segments, introns, transcripts); PLATFORM, instrument used to sequence the raw data related to the item; PIPELINE, list of methods used for processing phases, from raw data to processed data; DATATYPE (e.g., peaks, expression quantifications, methylation levels); ASSEMBLY, reference genome such as hg19 or GRCh38;

FileFormat, standard data format of the items, dictating region schema, number and semantics of columns, and a Boolean variable IsAnnotation, indicating if the dataset includes experimental data or genome annotations. The *Biology View* is centered on the description of biological samples, i.e. material sample taken from a biological entity and used for the experiment, with information on: Tissue, a multicellular component in its natural state; Cell, denoting single cells in natural state, immortalized cell lines, or cells differentiated from specific cell types; Disease, possibly carried by the sample, with the Boolean health status (IsHealthy). Biological material is possibly provided by a donor, described by: Age (in number of days), Gender, Ethnicity, and Species. Finally, when an assay is performed multiple times on separate biological samples (or on the same one), multiple replicates of the experiment are generated. To keep track of the replication process, we store the BiologicalReplicateCount and the TechnicalReplicateCount. The *Management View* describes the project producing the item, and includes: SourceSite where the material is analyzed and the item produced (e.g., universities, biobanks, hospitals, research centers, or laboratories); ProjectName, particularly relevant in the context of cancer Fenomics (e.g., TCGA-BRCA is the study for Breast Invasive Carcinoma of The Cancer Genome Atlas); Source, the program or consortium responsible for the production of genomic items (e.g, ENCODE/TCGA/...). The *Technology View* describes the technology producing the itiem; it includes: Technique, the procedure conducted to produce the items; Feature, the specific genomic aspect studied with the experiment (such as gene expression, mutations...); then, for epigenomic experiments such as ChIP-Seq: Antibody, a protein employed against the Target proteins.

User Interface. The user interface presents to users the possibility of opting for structured search (based on the described 25 attributes) or unstructured search (based on key-value pars). In both cases, it extracts matching items; the number of matching items is dynamically provided while the user enters search values. In the case of structured search, possible matching values are shown in a drop-down list; the list is dynamically updated while the search proceeds. The search query is a conjunction over its structured and semi-structured search steps; within structured search, it is a conjunction of the search clauses which are progressively built by selecting attributes, while every selected search value provides a disjunctive option. Abstract examples of queries are shown in Fig. 2, in the next section.

3 Experiment Description

Study Rationale. For evaluating the usability and usefulness of our interface, we planned an empirical study consisting of presenting a questionnaire to a group of biologists, bioinformaticians, and computer scientists/software developers with interest in Genomics. Before being engaged with the search system, we provided users with WIKI documentation and video tutorials. We planned questions of progressive levels of difficulty; each question presents a specific research

scenario and participants are asked to use our interface for extracting items, thereby simulating the typical search task (i.e., checking that our repository stores sufficient information for addressing the needs of each scenario). After the submission of answers, we show the right answers to users, and provide explanations of each answer; we expect that during the process users can develop a better understanding and progressively master the search system. After such training, we ask the users to evaluate the overall experience and specify the degree of expertise in the domain.

Table 1. Proposed survey questions.

Q1. How many datasets do we provide from the source TCGA with assembly GRCh38?
Q2. How many items do we provide for TCGA, assembly GRCh38, in the normal (**a**)/tumoral (**b**) cases?
Q3. Which TCGA GRCh38 project among COAD (Colon adenocarcinoma), LUAD (Lung adenocarcinoma), and STAD (Stomach adenocarcinoma) has more gene expression data?
Q4. How many sources contain data annotated with the human fetal lung cell line IMR-90 (both using original spelling (**a**) and alternative syntaxes (**b**))?
Q5. How many sources contain data annotated with the tissue uterus (both using original spelling (**a**) and the broadest possible intepretation (**b**))?
Q6. In ENCODE, how many items of ChIP-Seq can you find for the histone modifications H3K4me1, H3K4me2, and H3K4me3?
Q7. Assume you want to retrieve items from the TADs source that correspond to combined replicates (i.e., they belong to at least 2 biological replicates). How many items can you find?
Q8. We would like to retrieve items of hg19 assembly from healthy brain tissue (and possibly its subparts) of male gender, up to 30 years old. How many items can you find with these characteristics in the sources ENCODE (**a**) and TCGA (**b**)?
Q9. We are interested in ovarian cancer patients at clinical Stage III and IV. Select TCGA-OV project data. Then, select pairs with the key 'clinical_patient__clinical_stage' corresponding to the stage iii and iv (e.g., stage iiia, stage iiib, ...). How many items can you retrieve?
Q10. Suppose you need to identify DNA promotorial regions bound by the MYC transcription factor that present somatic mutations in breast cancer patients. For each of the following steps, provide the number of retrieved items. First, get from ENCODE source, ChIP-seq narrowpeak data from the cell line MCF-7, regarding MYC binding sites (**a**). Second, DNA-seq data is needed from TCGA BRCA patients which encountered a new tumor occurrence (**b**). Third, genomic region annotations describing promoters locations should be retrieved from RefSeq (**c**).

Experiment Design. During the conception of the survey, we followed a number of study design principles. We attempted to lower the ambiguity of the questions and to provide some guidance to the users; we used questions that could have exact answers (i.e., numbers), to lower the possible interpretation biases; we stratified questions by complexity, to capture different levels of understanding of the interface and its structure; we diversified the challenges addressed in the questions, to overview all search possibilities encompassed by our system.

Table 2. Input features tested in the survey. Desired output column contains numbers of items (#I), datasets (#D), or sources (#S).

Group	Question	Sub-questions	Desired output	Cross-dimension attributes	Logical disjunction	Semantic enrichment	Combination original/integr.	Complete study
1	Q1	1	#D	×				
	Q2	2	#I	×				
	Q3	1	#I	×				
2	Q4	2	#S			×		
	Q5	2	#S			×		
	Q6	1	#I	×	×			
	Q7	1	#I	×				
3	Q8	2	#I	×		×		
	Q9	1	#I		×		×	
	Q10	3	#I	×			×	×

In Table 1 we show the complete list of 10 proposed questions (some of which contain two or three sub-questions). We divided the questionnaire according to three groups of questions, in order of complexity: the first provides a simple scenario with incremental addition of filters: first a source with the assembly (Q1), then selection of normal/tumor patients (Q2) and of specific disease projects (Q3); the second explores peculiar (i.e., less standard) features of the search, e.g., semantic enrichment with synonyms (Q4), ontological hierarchies (Q5), disjunction of attribute values (Q6), and aggregate attributes (Q7); the third builds three more complex cases: combination of many filters (Q8), joined use of original metadata (in key-value format) and structured metadata (Q9), composition of three selections from data sources to simulate a complete study (Q10). Figure 2 visually explains the process of attribute selection and value provisioning required by questions Q2, Q5, and Q9.

As shown in Table 2, in different questions, we tested: the ability to compose queries by combining attribute filters coming from different dimensions, the use of value filters in disjunction one with the other, the understanding of semantic enrichment options, the combined used of original metadata filters (using a key-value-based interface) with structured integrated metadata (based on the GCM). With respect to the interplay between original and structured metadata: the query interface must enable interaction with both (in the key-value pairs it is important that people can ask separately what are the key—typically defining the property associated to the item—and what are the values—associated to

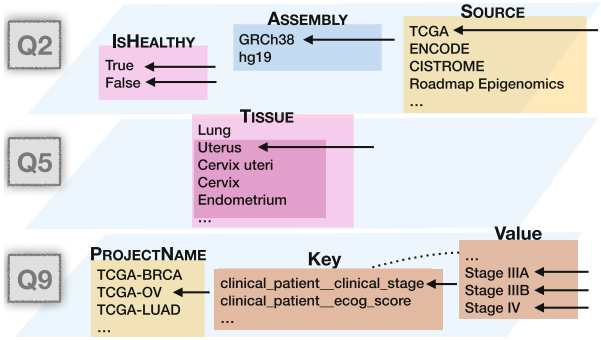

Fig. 2. Q2 describes a case in which the user selects from the IsHealthy attribute list first the value "True" and then the value "False", corresponding to two sub-questions. Then, she selects "GRCh38" among the possible values in the Assembly attribute list and "TCGA" as a Source. Q5 presents an enriched list of values for the attribute Tissue—note that "Cervix uteri" and "Cervix" are synonyms and, together with "Endometrium", they are hyponyms of uterus. For Q9, after selecting the ProjectName, the user explores keys and values through a specific interface.

the specific property). In different questions we alternatively asked to report the number of items, datasets, or sources.

Study Execution. The experiment target users were sourced from within our research group (GeCo) and from several collaborating institutions (such as Politecnico di Torino, Istituto Nazionale dei Tumori, Università di Torino, Università di Roma Tre, Istituto Italiano di Tecnologia, Radboud Universiteit Nijmegen, Freie Universität Berlin, Harvard University, Broad Institute, National University of Singapore, University of Toronto), including researchers with different backgrounds (computational and molecular biology, bioinformatics, and computer science) but also students and pure software developers with interest in Genomics. Out of about 60 invitations, we received 40 completed responses.

4 Results

We first describe how many answers were correctly provided, then how the users evaluated their experience with our system.

Correct Answers. In Table 3 we report: the required semantic level to set at the beginning of the query, the numbers of dimensions, integrated attributes and original keys involved in the query. Then we show percentages of correct answers (scores) of each specific sub-question and aggregated by group. Note that, if we consider together the performances of each group, as expected, group 1 reached a high percentage of correct answers (93.33%), group 2 a little less (75.94%), while group 3 had the worse score (68.47%). Some typical errors spotted in many answers are also reported. Question 8 had a low rate of correct answers

Table 3. Result features. Semantic levels include original values (O), synonyms and vocabulary terms (S), or the expanded option, with also hierarchical hyponyms (E).

Question	Semantic level	#dim.	#integr. attributes	#orig. keys	Scores	Group score	Typical errors
Q1	O	2	2	0	97.50%	93.33%	#items instead of datasets
Q2a	O	3	3	0	97.50%		
Q2b	O	3	3	0	92.50%		
Q3	O	3	2	0	87.50%		
Q4a	O	1	1	0	72.50%	75.94%	#items instead of sources and wrong spelling
Q4b	S	1	1	0	82.50%		#items instead of sources
Q5a	O	1	1	0	82.50%		#items instead of sources
Q5b	E	1	1	0	70.00%		#items instead of sources
Q6	O	2	3	0	67.50%		
Q7	O	2	2	0	82.50%		Wrong use of replicate count
Q8a	E	3	6	0	50.00%	68.47%	Wrong use of age selector
Q8b	E	3	5	0	52.50%		Wrong use of age selector
Q9	O	1	1	1	75.00%		
Q10a	O	4	5	0	82.50%		
Q10b	O	2	2	1	70.00%		
Q10c	O	2	3	0	85.00%		

(50% and 52.63%); we asked to retrieve the number of items in two sources for a specific assembly from a healthy tissue (using the semantic option that includes ontological hierarchy) of one gender in a restricted age range. Such question combined many elements (six data search filters, use of semantic expansion, age feature).

Overall, users replied correctly to 78.92% of the questions (grouping together the sub-questions of a same entry). Five users answered correctly to all questions. On average, it took them less than 44 min to answer all the 10 questions.

Lessons Learned. In retrospective, we made mistakes in the formulation of some of our queries. Users were confused when we asked them to count the containers (e.g. sources and datasets) instead of the data items, probably because they do not understand the notions of sources and of datasets. Distinguishing the dataset and data source storing the items probably requires a computer science background that was not present in many users. As in these cases users made the exact choices of attributes and values and just provided a wrong numerical answer, we considered their answers as valid. In one question (Q8) users did not reach a satisfying percentage, probably due to the misinterpretation of some filters.

In spite of these mistakes, our user study provided us with an important feedback. We were forced to denormalize and simplify the conceptual schema, but the logical organization of our simplified schema, centered on the item with selected attributes and organized along four dimensions, still proved to be effective; it facilitated both the training and the search interface organization. Clustering attributes along the four dimensions allowed us to explain them first collectively and then individually; users understood well their meaning and in most cases were able to translate narrative questions into the correct choice of attributes embedding the questions' semantics.

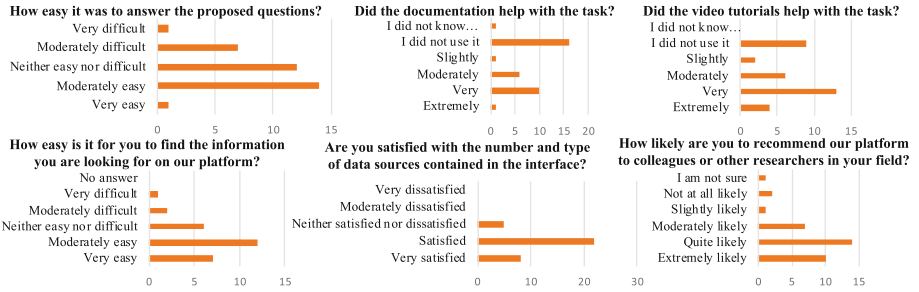

Fig. 3. Histograms showing the user's evaluations of the search system.

Qualitative Results. After filling the first part of the questionnaire, we asked users if they learned from the system and if they liked it, and to give us hints on how to proceed in our work (possibly with open suggestions to improve it). Answers to this part of the questionnaire are shown in Fig. 3.

Two thirds of users declared that answering to the proposed questions was "Moderately easy" or "Neither easy nor difficult". Most users either did not use the documentation or found it moderately/very useful, while users who watched the video tutorials were generally satisfied with them. When asked to perform a query to reach items useful to their own research, most users declared it was moderately easy. The majority was satisfied with the data sources available in the interfaces and was quite likely to recommend the platform to colleagues and researchers in the field.

Fig. 4. Histograms showing the user's expertise on genomic data analysis.

Figure 4 shows histograms on the self-assessment of users about their experience in the field of genomic data analysis. Users present expertise scores that range from "None" to "Expert". When asked about their use of platform to find data for analysis and about their need to combine inter-sources data, answers ranged from "Never" to "Daily", confirming that our users' test-set was well-assorted. Users also provided interesting suggestions, including a number of relevant sources to add to the framework and particular features that could be useful for practitioners.

5 Conclusions

Although we were forced to denormalize and simplify the conceptual schema, still its logical organization helps the users in translating a natural language question into the right choice of attributes, keys and values for querying our search interface; thus, we conclude that the most important aspects of attribute semantics are conveyed to users also in the context of the simplified conceptual schema. Some specific feedback and the observation of users' mistakes allowed us to improve the instructions for learning how to best use the search interface, as we eliminated some sources of ambiguities that could have created some of the misunderstanding. We also received important indications about missing data sources according to users' experience; this information will drive us in selecting the next sources to be integrated to our repository.

Acknowledgement. This research is funded by the ERC Advanced Grant 693174 GeCo (Data-Driven Genomic Computing), 2016–2021.

References

1. Back, C.O., Debois, S., Slaats, T.: Towards an empirical evaluation of imperative and declarative process mining. In: Woo, C., Lu, J., Li, Z., Ling, T.W., Li, G., Lee, M.L. (eds.) ER 2018. LNCS, vol. 11158, pp. 191–198. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01391-2_24
2. Bernasconi, A., Canakoglu, A., Ceri, S.: From a conceptual model to a knowledge graph for genomic datasets. In: Laender, A.H.F., Pernici, B., Lim, E., de Oliveira, J.P.M. (eds.) ER 2019, LNCS, vol. 11788, pp. 352–360, 2019. Springer, Cham (2019)
3. Bernasconi, A., Ceri, S., Campi, A., Masseroli, M.: Conceptual modeling for genomics: building an integrated repository of open data. In: Mayr, H.C., Guizzardi, G., Ma, H., Pastor, O. (eds.) ER 2017. LNCS, vol. 10650, pp. 325–339. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69904-2_26
4. Bernasconi, A., et al.: Ontology-driven metadata enrichment for genomic datasets. In: International Conference on Semantic Web Applications and Tools for Life Sciences, vol. 2275. CEUR-WS (2018)
5. Lukyananko, R., Parsons, J., Samuel, B.M.: Artifact sampling in experimental conceptual modeling research. In: Woo, C., Lu, J., Li, Z., Ling, T.W., Li, G., Lee, M.L. (eds.) ER 2018. LNCS, vol. 11158, pp. 199–205. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01391-2_25

6. Palacio, A.L., López, Ó.P., Ródenas, J.C.C.: A method to identify relevant genome data: conceptual modeling for the medicine of precision. In: Trujillo, J.C., Davis, K.C., Du, X., Li, Z., Ling, T.W., Li, G., Lee, M.L. (eds.) ER 2018. LNCS, vol. 11157, pp. 597–609. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00847-5_44
7. Rambold, G., et al.: Meta-omics data and collection objects (MOD-CO): a conceptual schema and data model for processing sample data in meta-omics research. Database **2019**, baz002 (2019)
8. Reyes Román, J.F., Pastor, Ó., Casamayor, J.C., Valverde, F.: Applying conceptual modeling to better understand the human genome. In: Comyn-Wattiau, I., Tanaka, K., Song, I.-Y., Yamamoto, S., Saeki, M. (eds.) ER 2016. LNCS, vol. 9974, pp. 404–412. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46397-1_31
9. Verdonck, M., et al.: Comparing traditional conceptual modeling with ontology-driven conceptual modeling: an empirical study. Inf. Syst. **81**, 92–103 (2019)
10. Zhang, H., Li, T., Wang, Y.: Design of an empirical study for evaluating an automatic layout tool. In: Woo, C., Lu, J., Li, Z., Ling, T.W., Li, G., Lee, M.L. (eds.) ER 2018. LNCS, vol. 11158, pp. 206–211. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01391-2_26