






Data and Conceptual Model Synchronization in Data-Intensive Domains: The Human Genome Case

Floris Emanuel^{1,2}(✉) , Verónica Burriel² , and Oscar Pastor¹ 

¹ Centro de Investigación en Métodos de Producción de Software,
Universitat Politècnica de València, Valencia, Spain
florisldn@gmail.com, opastor@pros.upv.es

² Department of Information and Computing Sciences,
Utrecht University, Utrecht, The Netherlands
v.burriel@uu.nl

Abstract. Context and Motivation: With the increasing quantity and versatility of data in data-intensive domains, designing information systems, to effectively process the relevant information is becoming increasingly challenging. Conceptual modeling could tackle such challenges in numerous manners as a preliminary phase in the software development process. But assessing data and model synchronization becomes an issue in domains where data are heterogeneous, have a diverse provenance and are subject to continuous change. **Question/problem:** The problem is how to determine and demonstrate the ability of a conceptual schema to represent the concepts and the data in the particular data-intensive domain. **Principal Ideas/Results:** A validation approach has been designed for the Conceptual Schema of the Human Genome by investigating the particular issues in the genetic domain and systematically connecting constituents of this conceptual schema with potential instances in samples of genome-related data. As a result, this approach provided us accurate insight in terms of attribute resemblance, completeness, structure and shortcomings. **Contribution:** This work demonstrates how the strategy of conceptualizing a data-intensive domain and then validating that concept by reconnecting this with the attributes of the real world data domain, can be generalized. Conceptual modeling has a limited resistance to the evolution of data, which is the next problem to face.

Keywords: Information systems · Genome · Conceptual modeling · Validation

1 Introduction

Software systems are becoming more complex due to the evolution of related techniques, as a result of the high expectations from our advancing society.

To align with this trend, software and data models must be integrated with models of the application domain, which would be technical, organizational, people centered or a mixture thereof [8]. In this work, we are especially interested in data-intensive domains (DIDs), domains where the data perspective is the most relevant one, and where applications must deal with large quantities of (more or less) structured data, using basic operations to extract valuable information. Hence, both the domain and related data should be understood.

Conceptual modeling (CM) could be applied to provide a solution to the problem context by eliciting and describing the general knowledge a particular information system needs to know [7], which tackles the complexity of managing the data of a particular domain. CM can also synergize work between experts of different domains (multiple stakeholders) and capture and structure existing knowledge [6]. Specifically testing (validate) the general (conceptualized) rules that have been created from a collection of specific observations in the problem domain (context) could tackle big data management as an overall approach. In this paper, we demonstrate this idea by introducing a validation approach to the established Conceptual Schema of the Human Genome (CSHG¹), which has been developed at PROS Research Center, Valencia.

Understanding the Human Genome is probably the biggest challenge of this century. Its main direct implication is already affecting the development of 'precision medicine' (PM), an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle [3]. PM is a practise of the interdisciplinary field of biomedical informatics. Complex genome data has exploded since the first complete read in 2003 [2,4]. The CSHG tackles the lack of the formalization of the genome concept by capturing existing knowledge by providing structure, allowing integration from numerous data sources and provides structure for the design of a data management tool. The problem is to validate the adequacy and correctness of the CSHG by assessing how the data available in existing genome data sources comply with the structure of the model, in order to demonstrate its significance. Without validation, its adequacy can only be assumed.

According to Design Science Methodology by Roël Wieringa [11], this issue can be classified a knowledge question with an underlying design problem. Because in this case we are validating the CSHG, our validation approach is the artefact that has been designed. Therefore, the question is to what extend our validation approach is able to manage the model synchronization in this context. In the next section we introduce the design of the validation approach, in the Treatment Validation section its performance is discussed. In the Conclusion and Future work we generalize this particular results to discuss the overall approach to tackle big data management, which is the overarching problem context.

¹ <https://www.dropbox.com/s/y06ov4kl6dmdgqg/CSHG.pdf?dl=0>.

2 Validation Artefact Design

To design this artefact to get a useful insight into the state of the CSHG, we initially analyzed the problem domain concerning the complexity of genome-related data and its related issues. Subsequently, we investigated the issues specific to corresponding the CSHG to genome-related data, rather than investigating methods to validate conceptual models in general, because our research purpose is to connect the data-world to the modeling-world. For the matching procedure itself, general approaches were investigated from information systems related literature. In Fig. 1, the validation artefact is depicted as a process deliverable diagram (PDD) to concisely depict what has been designed in this work. The PDD has been created on the basis of the guidelines by I. Weerd and S. Brinkkemper [10].

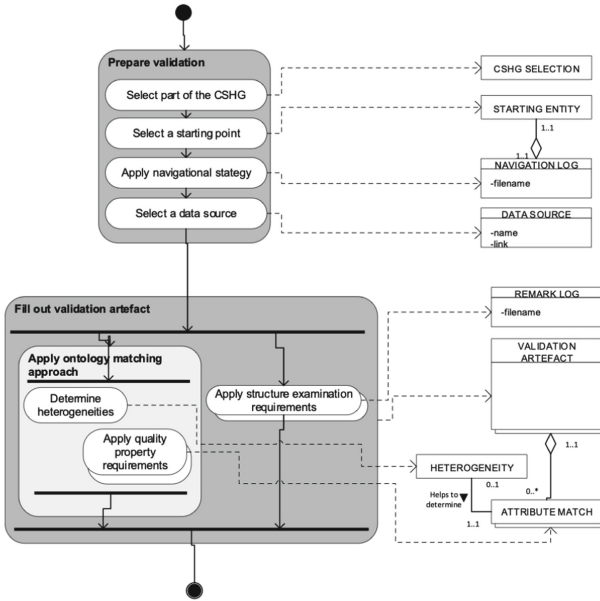


Fig. 1. A process-deliverable diagram (PDD) of the designed validation artefact.

The artefact contains two main activities, namely *preparing the validation* and *carrying out the validation*.

Prepare Validation. The preparation entails the process before establishing connections between the CSHG and the real-world data that it should represent. First, a preferred sample of the CSHG is selected. Subsequently, an entity, preferably central or less complex is chosen as a starting point. An entity could be a 'variation' in the DNA. An example attribute could be whether this is

‘benign’. Both the model and the related databases start by describing more central themes of the genome, gradually moving to more specific topics. Therefore, we adopt an ‘inside out’ strategy as a principle, combined by following the paths of the cross references between repositories, and create a *log* to keep track of the movements.

Fill Out Validation Artefact. To report findings, we created a table component to be filled out for every entity, and every repository. We store all validated entities vertically and all different sources for that same component horizontally in a spreadsheet for a structured execution. For every attribute it is determined whether it corresponds to parts of the required information from a given data record.

In order to avoid the need to justify each correspondence that is made, ‘minimum requirements’ are applied during each correspondence attempt. Justifying the abstract properties of a valid correspondence can effectively contribute to the repeatability and validity of this work. The heterogeneity amongst sources means that we will often deal with different ontologies. Therefore, we adopted an ontology matching approach by Euzenat and Svaiko [1]. This is an abstract lifecycle that guides the ontology matching workflow from analysis of the problem domain to the creation of a fitting matching approach, on the basis of an assembly of existing techniques. They state ‘no one size fits all and every case is unique’. Therefore, we adopted this approach, and used only what is relevant in our case. Along these properties, we use only two requirements for now, which is that the *data type* of a required attribute and its potential instance should always be the same. Secondly, An attribute of the CSHG and a potential instance in an external data source should always be *conceptually* the same, regardless of *terminological heterogeneity* (different definition). We think identifying and acknowledging such heterogeneities is critical to making the correspondences because ignoring them results in numerous unwanted and avoidable false positives and false negatives.

As we closely analyze the CSHG, and the data, we list discrepancies concerning how the attributes are grouped together in a *remark log*. Subsequently, we assess these with literature.

3 Validating the ‘Validation Artefact’

The validation artefact is applied to the CSHG context on the basis of genome related data sources. The results are divided into relevant general perspectives.

3.1 Results

Regarding the CSHG, our goal was to determine and demonstrate its essential ability to support the management of genomic data. Table 1 summarizes the 62 attributes of the CSHG that were validated at the hand of multiple different data

Table 1. Summary statistics of the validation component sheet (A view only copy of the correspondence recording data can be found here: https://www.dropbox.com/s/gu4893inqky0pds/View_CSHG_Validation_Record.xlsx?dl=0)

Description	Result
Total correspondence attempts	118
Number of attributes validated	62
Average number of sources per attribute	2
Ratio of present correspondences/total	78.8%
Ratio of non-present correspondences	18.6%
Ratio of undefinable correspondences	2.5%

sources per attribute, resulting in 118 correspondence attempts. Hence, regarding the correspondence between the CSHG and real-world genome related data, no perfect compliance was observed. The resemblance was however a promising 80%, taking into account *overfitting*, *derivable values* and *intended heterogeneities* in the conceptual schema (making 100% impossible). There were also no colliding flows of information under the current structure, solely discrepancies. However, discussion is always possible. Such as that there is additional information thinkable, and there were also underrepresented parts, meaning we found no real-world representation of these attributes in the data sources used. This way we demonstrate proficiency of the CHSG while remaining critical. These useful results demonstrate the success of the validation approach. This was also confirmed after presenting the results to the model authors and information systems experts from our academic network. The model authors were not actively involved in the project but were of help in understanding the problem domain, validate the results and to improve the method from a scientific point of view.

3.2 Discussion

We could generalize the essence of these two approaches and see them as subsequent or perhaps simultaneous phases. In phase one, a solution (the CSHG) is proposed by conceptualizing the problem domain, in order to provide structure for a complex concept that encompasses all kinds of related data. In subsequent phase, this is tested by using this data to examine its fit to the solution. While the artefact was initially tuned to the problems within this domain and scope, we hypothesize that this generic strategy could be useful to problem contexts of a similar nature.

This could be a problem domain where big data driven decision support is required. In such a complex DID, the related available data has properties like ‘messy’, ‘heterogeneous’, ‘vast’, ‘growing number of stakeholders and parameters’. An example of this could be the housing market, a highly dynamic environment with all kinds of constant changes. What makes this environment complex are the varying regulatory influences which cannot be withstood by conventional

house pricing models. Also, the definition of regulation is lacking due to the complexity and potential interaction of different regulatory aspects [9]. Furthermore, there are numerous heterogeneous data sources for housing data and for all kinds of environmental factors that influence the market.

4 Conclusion and Future Work

In this work, we proposed and tested a solution that should tackle the problems related to validating a conceptual schema on the basis of its real-world data that it should represent, in an exploratory manner. Some problems were specific to the domain of the management of genome related data, for which our solution has been created. In a DID where complex data is constantly evolving and valuable information is spread across heterogeneous data sources in a vast lake of data, conceptual modeling has not only been a potential solution to the data-related challenges. It also tackled the issue of the constant evolution of genome fundamental knowledge, which formed an extra challenge to IS design.

We demonstrated that this validation artefact is effective, and that the CSHG can improve through the application of this validation method; we propose it can be generalized and applied to other environments with similar properties, or used for subsequent work either in- or outside the genome domain. As the domain rapidly evolves, its related data does so as well. Where CM is a first step [5], future work should focus on how an information system could deal with data evolution.

References

1. Euzenat, J., Shvaiko, P., et al.: *Ontology Matching*, vol. 18. Springer, Heidelberg (2007). <https://doi.org/10.1007/978-3-540-49612-0>
2. Collins, F.S., Morgan, M., Patrinos, A.: The Human Genome Project: lessons from large-scale biology. *Science* **300**(5617), 286–290 (2003)
3. Larry Jameson, J., Longo, D.L.: Precision medicine-personalized, problematic, and promising. *Obstetr. Gynecol. Survey* **70**(10), 612–614 (2015)
4. Mukherjee, S., et al.: Genomes OnLine Database (GOLD) v. 6: data updates and feature enhancements. *Nucleic Acids Res.* D446–D456 (2016)
5. Pastor, O., Levin, A.M., Casamayor, J.C., Celma, M., Eraso, L.E., Villanueva, M.J., Perez-Alonso, M.: Enforcing conceptual modeling to improve the understanding of human genome. In: 2010 Fourth International Conference on Research Challenges in Information Science (RCIS) (2010), pp. 85–92. IEEE (2010)
6. Pastor, O., et al.: Conceptual modeling of human genome: integration challenges. In: Düsterhöft, A., Klettke, M., Schewe, K.-D. (eds.) *Conceptual Modelling and Its Theoretical Foundations*. LNCS, vol. 7260, pp. 231–250. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28279-9_17
7. Fabián Reyes Román, J.: *Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano*. PhD thesis (2018)

8. Sølvsberg, A.: On models of concepts and data. In: Düsterhöft, A., Klettke, M., Schewe, K.-D. (eds.) *Conceptual Modelling and Its Theoretical Foundations*. LNCS, vol. 7260, pp. 190–196. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28279-9_14
9. Tu, Q., de Haan, J., Boelhouwer, P.: The mismatch between conventional house price modeling and regulated markets: insights from The Netherlands. *J. Housing Built Environ.* **32**(3), 599–619 (2017)
10. van de Weerd, I., Brinkkemper, S.: Meta-modeling for situational analysis and design methods. In: *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*, pp. 35–54. IGI Global (2009)
11. Wieringa, R.J.: *Design Science Methodology for Information Systems and Software Engineering*. Springer, Heidelberg (2014). <https://doi.org/10.1007/978-3-662-43839-8>