Chapter 3
Modeling Life: A Conceptual Schema-centric Approach to Understand the Genome

Óscar Pastor López, Ana León Palacio, José Fabián Reyes Román and Juan Carlos Casamayor

Abstract Programs are historically the basic notion in Software Engineering, which represents the final artifact to be executed in a machine. These programs have been created by humans, using a silicon-based code, whose final components use a binary code represented by 0s and 1s. If we look at life as a program with a DNA-based genetic code with a final representation that uses four essential units (A, C, G and T), one challenging question emerges. *Can we establish a correspondence between life – from a genomic perspective – and programs – from a Software Engineering perspective?* This paper assumes a positive answer to this question and shows how genomic can benefit from Information Systems Engineering by applying conceptual modeling to determine those relevant data that life represents in order to manage them accordingly, with special emphasis in the health domain. The main contributions focus on i) a concrete materialization of a Conceptual Schema of the Human Genome, ii) the need of having a method to provide a methodological guidance concerning genome data management, and iii) the importance of assessing data quality for all generated data that are going to be used in critical domains such as health and Precision Medicine.

Óscar Pastor López
PROS Research Center, Universitat Politècnica de València, e-mail: opastor@pros.upv.es

Ana León Palacio
PROS Research Center, Universitat Politècnica de València, e-mail: aleon@pros.upv.es

José Fabián Reyes Román
PROS Research Center, Universitat Politècnica de València, e-mail: jreyes@pros.upv.es
Dept. of Engineering Sciences, Universidad Central del Este (UCE)

Juan Carlos Casamayor
Departamento de Sistemas Informáticos y Computación (DSIC), Universitat Politècnica de València, e-mail: jcarlos@dsic.upv.es

J. Cabot et al. (eds.), *Conceptual Modeling Perspectives*,
https://doi.org/10.1007/978-3-319-67271-7_3

3.1 Introduction

The goal of this paper is to show how Conceptual Modeling provides a sound background to face the challenging problem of understanding the genome. The ideas and results presented here are inspired in the keynote presented by Prof. Antoni Olivé in CAiSE 2005 in Porto, Portugal [1]. He introduced a Conceptual Schema-centric Software Development approach, intended to make true a precise model-driven development solution: in modern *Information Systems Engineering* (ISE), what it should be true is that *"the conceptual model is the code"*, instead of the conventional perspective based on the fact that *"the code is the model"*. Prof. Olivé was also explaining in his work how the term *"conceptual model"* was too frequently misused, substituting the correct term of *"conceptual schema"*, the one to be used when referring to concrete instantiations of a conceptual modeling exercise.

This inspiring keynote was reflecting very well our work in the last decade around building conceptual schema compilers and providing a software process where conceptual modeling and model transformations (*from requirements to code*) conform the strategy to be followed.

But when looking for new challenges where our experience in conceptual modeling could be effectively applied, one particular domain came to our mind: modeling life. How to face the challenge of modeling life by understanding the genome became the problem to be solved. We show in this paper how conceptual modeling can provide the required techniques to manage adequately the huge amount of data that the genome-related working context continuously generate.

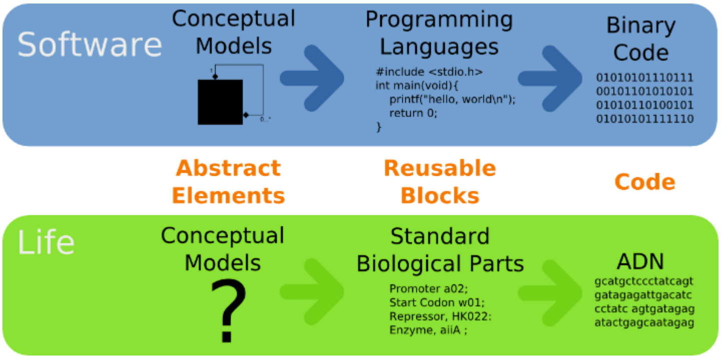
The reality is that understanding life as we know it on our planet can probably be considered the biggest challenge of our century. However, how to face the problem of understanding life from an ISE perspective is a complex question. *Can ISE help us to achieve the goal of understanding life?* Answering this question becomes a relevant issue that particularly affects how modern Precision Medicine can reach our society, changing and improving medicine, as we historically know it.

As said before, in our previous work we have been using CM to explain how we, humans, generate programs, that in their final form are constituted by silicon-based binary code. These programs are the written representation of CM that abstractly represent a relevant part of the real work we are interested in. We based our use of a conceptual modeling-based approach in the definition of conceptual modeling proposed by Prof. Olivé in his outstanding book on conceptual modeling [8]. In a few words, we assume from his work that conceptual modeling refers to the activity that elicits and describes the general knowledge that a particular IS needs to know. In this paper this particular information systems is the *"genome"*. Its main objective is to obtain that description, represented in which it is called a CS. Accordingly, a conceptual schema of the genome constitutes a significant result that will be presented later.

We also assume that conceptual schemas are written in languages called conceptual modeling languages. Additionally, in our perspective of a sound software process that covers all the conceptual modeling steps that go from requirements modeling (at the earliest software production process steps) to application code generation (as the final result of such a precisely-defined software process), conceptual modeling is an important part of that requirements engineering task, the first and most important phase in the development of an IS.

What is then the connection between conceptual modeling and life? Why did we move to the fascinating working domain of modeling life by facing the challenging problem of understanding the genome?

We applied an attractive and similar metaphor to achieve our desired clear understanding of life. In this case the programs are living beings whose genetic code includes the instructions that explain life as we perceive it. Instead of having the ISE materialization in the form of a binary executable code, in this case we have what we could call a quaternary executable code, based on four letters (A, C, G, T) that represent the four nucleotides that form the basic components of this *"carbon-based"* executable code (see the lower part of Fig. 3.1).



```
#include <stdio.h>
int main(void){
    printf("hello, world\n");
    return 0;
}
```

```
Promoter a02;
Start Codon w01;
Repressor, HK022:
Enzyme, aiiA ;
```

Fig. 3.1 From conceptual models to code: a SE-perspective and a life understanding perspective.

If we want to develop this idea, one immediate question that arises is: What is then the *"programming language of life" that would allow us to understand and manage life as we understand and manage ISE-based programs?* We are perfectly aware of the magnitude of the challenge that arises from this question. But at the same time, we are aware that the race to face this challenge has not only started but is proceeding at full speed.

In this context, *what is the role of conceptual modeling? Why a conceptual schema-centric approach intended to understand the genome?* Let us answer these questions by introducing a bottom-up perspective, instead of the conventional top-down approach that is normally used in ISE.

By top-down we mean that if the *"conceptual schema is the code"*, what we could call conventional ISE must create code from models. We start with the conceptual schema, we convert it into the final application code.

Considering life, we face a different situation. We have now living beings that can be seen as individual *"programs"*. We perceive them as running programs. But in this case, we don't know the models that these programs exactly represent. The problem is similar to trying to understand the meaning of a program just looking at its binary code, just analyzing how it executes. This is what we call a “bottom-up” perspective. Analyzing individuals, collecting data about their genomes, we should be able to infer relevant information, we should be able to identify relevant conceptual patterns. To do it, it is essential to understand the nature of the data to be managed, and to understand its structure, including basic entities and relationships among them.

Considering the complexity of the problem, and although DNA is the basis of all life as we know it on the Earth, we focus here on the human genome, where rapid progress is being made specially in the context of PM (also previously known as *Personalized Medicine*). It is in this context that we want to focus our work, and where we want to report the experience accumulated in the last years in three main areas:

1. How essential it is to have a *Conceptual Schema of the Human Genome* (CSHG) for structuring the huge amount of data and knowledge that day after day are generated in the genomic domain. A CSHG will then be introduced.
2. The need of having a method to provide methodological guidance concerning genome data management, including the crucial phases of i) valid data sources “search and selection”; ii) identification of the valid data in those selected data sources; iii) database load process; and iv) subsequent data management platform oriented to an efficient data interpretation and exploitation. A method so-called SILE (for the name of the four relevant phases of *“Search, Identification, Load, Exploitation / Interpretation”*) is going to be presented.
3. The importance of assessing Data Quality (DQ) when a big data problem is faced, as occurs when all the generated data are to be used in practical settings as critical as PM.

The conceptual thread of our book chapter is going to follow these aspects. What we want to indicate with the selected title of this paper is how important a conceptual schema-centric approach is to draw a parallel between ISE and genomics. By considering live beings a particular kind of programs whose (*genomic*) code is started to be known, a challenging needs emerge precisely: to design the conceptual schemas that must lead to the relevant genome knowledge discovery. In our work we are not simply applying one essential ISE technique (conceptual modeling) to a complex domain (human genomics). We go much further: what we want to show is how conceptual modeling and genomics can share a same picture (as Fig.3.1 represents), and particularly, how genomics can benefit from ISE by applying conceptual modeling to determine those relevant data that life represents in order to manage them accordingly, with especial emphasis in the health domain.

The structure of this chapter follows the presented ideas. A concrete materialization of a CSHG is introduced in section 3.2, explaining our experience in its evolutionary and continuous design. It conforms a solid information system core intended to correctly manage genome data. This is followed in section 3.3 by the presentation of a methodological background the SILE methodology designed to characterize a sound conceptual schema-centric genome data management process. This section ends discussing a final essential issue: what is to be done to assess the quality of the data used in the PM clinical context, guided by the CSHG and based on the use of the SILE method. Finally, our conclusions and intentions for future work close the chapter.

3.2 Conceptual Schema of the Human Genome (CSHG)

It is widely accepted that applying conceptual models [8] facilitates the understanding of complex domains (such as *genomics*). In our case we used this approach to define a model representing the characteristics and behavior of the human genome [12, 16].

Through the application of CM, a wide range of benefits are obtained, which have a positive impact on the creation of Information Systems *-based on clear and precise structures-*. For example, conceptual modeling allows to represent more precisely the relevant concepts of the studied domain. A fundamental task before beginning the process of creating a conceptual schema is the analysis of the problem domain (in our case, *genomics*). Working together with teams specialized in *Software Engineering* (SE) and genomics (i.e., *biologists*, *geneticists*, etc.) allowed us to start designing the representation of the domain, giving as a result a *"Conceptual Schema of the Human Genome (CSHG)"*.

The main objective of this CSHG is to improve the treatment and integration of genomic data in order to enhance and guarantee PM. The CSHG has been adapted according to the new discoveries made in the domain. This evolution is necessary because the genomic domain produces large amounts of information in constant change and growth. For this reason, conceptual modeling has a great advantage in representing this domain because it facilitates the integration of new knowledge in the model and provides a positive support to the knowledge on which PM is based.

Understanding the human genome is a great challenge because it requires the development of (complex) data abstraction tasks. Only in this way we can get the relevant data to be included in the conceptual representation. The first version of the CSHG (v1) was the result of a series of meetings with experts in the domain, this version focuses on the analysis of individual genes, their mutations, and their phenotypic aspects (see details in [12, 10]). Next, the classification of that first version is presented in three main views:

- *Genome view:* responsible for modeling individual human genomes (Fig. 3.2).
- *Gene-Mutation view:* used to model knowledge about genes, their structure and their allelic variants.
- *Transcription view*: intented to model the basic components of the transcription process and the synthesis of proteins (which is what we know as "*gene expression*").

After finishing version 1, we started the task of evaluating the capacity of the model to deal with the actual data manipulated in the bioinformatics domain. At the moment of putting into practice this initial version of the CSHG, it was necessary

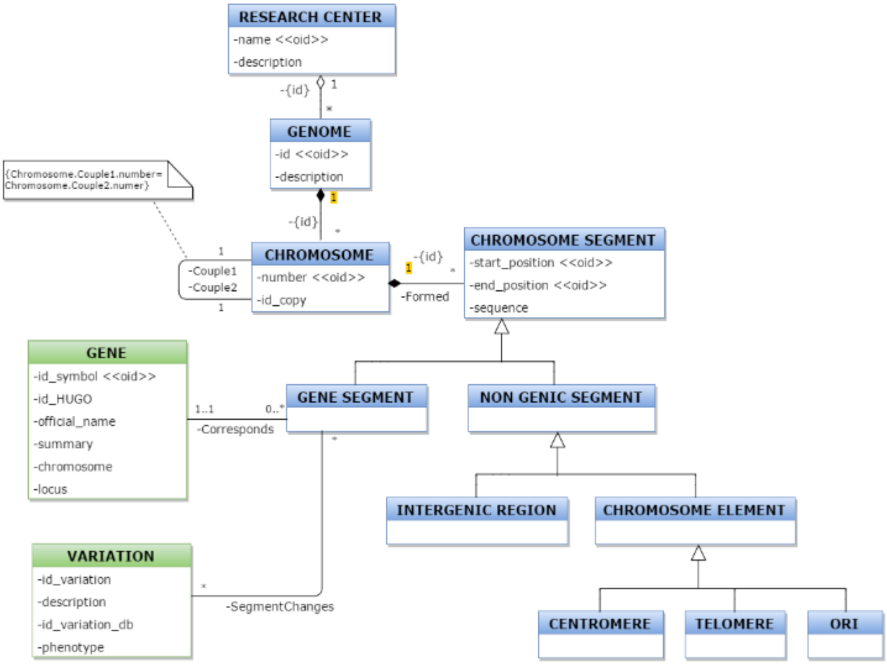


Fig. 3.2 Genome View (v1).

to generate a new version (CSHG v2), which changes its central axis and goes from representing a *"gene-centered"* vision to a vision centered on the concept of *"chromosome"*. This change of vision in the model represents the main difference with respect to previous versions of the model.

In this change of perspective, we identified a series of questions to address:

1. We were not sure about the suitability of mixing a Genome view related to the storage of individual genomes – the so-called Genome view in v1 – with a more theoretical, structural Genomic view related to the Genome configuration and characterization as a whole (the so-called Gene-Mutation and Transcription view).
2. Concerning the core concept of gene, it is not always feasible to describe DNA structure in terms of genes as basic constructs. We concluded that the most suitable structure is suing chromosome elements as the basic building blocks.
3. More relevant concepts were needed, for instance, the concept of SNPs.
4. We detected the need for extending the first version with more significant genome-related information. To go from genotype to phenotype in a complete, sound way, we needed the specification of the pathway description perspective.

The development of these four ideas are explained in detail in the following works [12], [16], and make up the so-called version 2 of the model (CSHG v2). This version of the model is organized into five main parts (called "*views*") (Fig. 3.3):

- *Structural view:* basic elements of the DNA sequence.
- *Transcription view:* components involved in going from DNA to the diversity of RNAs.
- *Variation view:* to characterize changes in the sequence of reference that have functional implications in how the genome expresses.
- *Pathways view:* intended to enrich the conceptual model with information about metabolic pathways to join genome components that participate in pathways with phenotype expressions.
- *Bibliography and data bank view:* to assess the source of any information in order to pinpoint the data source.

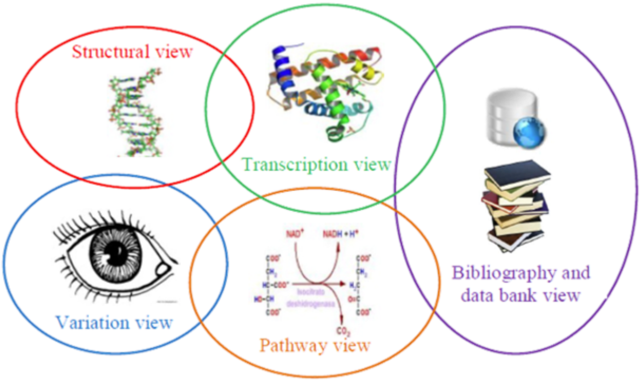


Fig. 3.3 Views of the CSHG version 2.

The CSHG v2 has been the basis for validating the management of genomic data related to diseases of genetic origin (i.e., *breast cancer* [3], *alcohol sensitivity* [13], *neuroblastoma* [4], among others). Currently, the next version of the CSHG (v3) has been developed. This new version aims to integrate all the relevant information on haplotypes [14, 15]. To do it, an extension of the *"variation view"* was done in order to manage the information related to *allelic/genotypic frequencies*, and *populations*. This conceptual schema is definitely intended to be able to generate the required number of versions in order to incorporate the genomic knowledge that continues to emerge in the domain.

The design of the CSHG is essential for the development of a *Genomic Information System* (GeIS) that guarantees the quality of the stored information. This approach facilitates a conceptual modeling perspective to provide a clear and open structure ready to be adapted to new changes, which in practical terms improves the reliability of the information and generates an accurate framework for a genomic diagnosis.

3.3 SILE Methodology and Data Quality

By defining a CM of the genomic domain we assure that data will be gathered under a single and comprehensive information perspective. The right interpretation of data is key to the PM, because it may affect *health decision making, research* and *clinical practice*. For this reason, the IS must be loaded with accurate, structured, relevant and consistent information. But this is not a trivial task. Thousands of biological databases have been developed over the last two decades, and they have widely varying content, resources, infrastructures and quality. The search and identification of relevant genomic information has become a time consuming process, highly dependent on the knowledge and experience of the researcher. Discussions with experts in the field highlighted that there is not any protocol or systematic method to search and identify relevant information. This behaviour leads to problems such as loss of relevant information resources and the collection of non-standardized data. In order to assure that as many relevant data repositories as possible are taken into account, and the data gathered are accurate and have a high quality level, the process must be performed in a systematic way. The addition of specific quality controls on each stage of the process assures an effective load of information in the database that represents the IS, and it improves the value of further analysis and exploitation.

According to this approach, the SILE (Search-Identification-Load-Exploitation) methodology has been developed in order to systematize the search and identification process of genomic information, which is loaded, analyzed and exploited by an IS that is based on the CSHG. Currently, SILE is being used by a group of researchers in an academic context, who search for genomic variations related to a set of diseases with a high social impact such as: *Alzheimer, Neuroblastoma* and *Lung Cancer*. In the next section a brief explanation of the main levels of the SILE Methodology is going to be made. Next, a first approach to the Data Quality Framework (DQF) used to complement the methodology will be presented.

3.3.1 SILE Methodology

The SILE methodology goal is to efficiently populate a *Human Genome Database* (HGDB), corresponding to the CSHG, with sound and high-quality information. But, *where can relevant information be found? Which data is significant to be loaded in the database?* And finally, *is this information of enough quality to offer an advantage to PM over traditional medicine?*

Through a methodology as SILE, as well as a proper quality framework specific for genomic data, those previous questions are precisely answered and quality errors are solved or considerably reduced.

This methodology is a four-level approach where each level provides information used as input to the next one: Search, Identification, Load and Exploitation. Next, a brief description of each level is provided.

3.3.1.1 Search (S)

Scientific sources (e.g. articles, databases) are thoroughly analyzed in order to determine the optimal ones to obtain information from. In the Search level the context of the information which is going to be searched needs to be defined (i.e., *a particular disease*). Once the context is delimited, the search must be focused on the available databases which can provide detailed information about the topic we are interested in.

Due to the huge amount of available repositories in the genomic domain, the use of biological databases catalogues is very useful to perform the search. These catalogues provide a complete list of data sources, grouped by category or topic, as well as a brief description of their content and links to the information home page. The most important ones are the catalogues provided by the *Nucleic Acid Research Journal* (NAR) [17] and the *Human Genome Variation Society* (HGVS)[1].

3.3.1.2 Identification (I)

The first step in the Identification level is to determine which information characterizes the domain of interest, according to the Conceptual Schema which describes it. As an example, Table 3.1 shows the information needed to represent a variation.

Table 3.1 HGDB Variation information

Attribute	Description
DESCRIPTION	Variation description.
DB_VARIATION_ID	Identifier provided by the data source where the information was extracted from.
CLINICALLY_IMPORTANT	Clinical importance of the variation related to a phenotype.
OTHER_IDENTIFIERS	Other possible identifier as for example HGVS expressions.
ASSOCIATED_GENES	Genes affected by the variations.
OMIM	Identifier provided by OMIM [9].
SPECIALIZATION_TYPE	Type of variation
FLANKING_RIGHT	Sequence made by 20 nucleotides on the right of the variation.
FLANKING_LEFT	Sequence made by 20 nucleotides on the left of the variation.
ALN_QUALITY	Alignment quality of the variation inside the gene.
POSITION	Position where the variation is located inside the chromosome.
INS_SEQUENCE	Sequence of inserted nucleotides.
INS_REPETITION	How many times the inserted sequence is repeated.
DEL_BASES	Number of nucleotides deleted.

Once the needed information is clear, the next step is to find out which part of the information is provided by each database selected in the previous level and how it can be extracted.

[1] HGVS Databases catalogue: `http://www.hgvs.org/content/databases-tools`

3.3.1.3 Load (L)

During the load phase the interesting information that as been previously identified will be extracted from each database and, after a transformation process, it will be used to populate the HGDB. To perform this tasks an ETL framework is used:

- The first step is to *"Extract"* (E) the information of interest from the databases, using the mechanisms they provide for such task (reports, FTP sites, APIs, etc.).
- The second step is to determine if the extracted information needs to be *"Transformed"* (T) to fit the format and the rules established by the HGDB and the CSHG.
- The final step is to *"Load"* (L) the information into the HGDB.

3.3.1.4 Exploitation (E)

The exploitation level concerns to extract knowledge from data. The data exploitation system might be able to guide experts through complex scenarios that take into account multiple types of data. In this level, the quality controls applied at the previous levels take an important value since the conclusions obtained in the extraction of knowledge depend on them. The requirements in this level are:

- *Data Discovery:* Users need to explore data by conducting ad-hoc queries with specific information goals in mind.
- *Data Visualization:* Users need ways to represent the data, identify patterns in the data and even more, to explore the most accurate interactive representation associated to those patterns.
- *Data Analysis:* Users need to analyze and understanding the relationships between the data in order to draw conclusions and inferring new relevant information.

In our case, the information stored in the HGDB is analyzed by a proper tool developed specifically for the use of variation data, called *"VarSearch"* [18]. This tool analyses the information obtained from a patient sample and determines if there are variations associated to a certain disease, according to the data stored in the database. Although the automated analysis is useful to determine the potential variation-diseases associations, additional *collaborative* and *interactive* mechanisms to explore and visualize the information are needed. Currently, a research is under-way to determine and integrate such mechanisms into *VarSearch*. The main idea is to enhance the data exploitation by easing the user-data interaction through intuitive user interfaces for non-technical users [5].

In summary, the SILE Methodology provides a framework to systematize the searching process and the population of IS developed to manage data in complex domains. This method helps to structure data collected from different public repositories, and the data-to-knowledge process becomes more efficient and more comprehensive. This methodological guidance is essential to assess an effective and efficient conceptual schema-centric genome data management environment.

3.3.2 Data Quality

Types of genomic databases range from huge data warehouses containing millions of unreviewed raw sequences to high-reviewed databases manually curated by experts in the field. Quality needs to be evaluated because these databases may affect *health decision making, research* and *clinical practice* as we mentioned before. Next, a summary of some common issues which can be found in genomic databases are briefly presented. Afterwards a first approach about the data quality framework that is proposed to be applied together with the methodology, will be explained.

3.3.2.1 Data Quality Issues

Due to its complexity and heterogeneity, genomic databases present issues related to the quality of the information that they store. These issues can be classified according to a set of six basic data quality dimensions proposed by Askham, which can be applied to genomic databases [2]:

- *Accuracy*: Data correspond to real-world values and are correct. Accuracy errors mainly affect genomic data warehouses, where DNA sequences are submitted to the database by researchers and not reviewed by external experts. Common errors are sequence conflicts, misspellings, taxonomical or curation errors.
- *Completeness*: The extent to which data is not missing and all necessary data values are represented. Primary non-curated databases have a low level of completeness while those reviewed and curated by experts are usually fairly complete.
- *Reliability:* The extent to which data is regarded as true and credible. To get proper conclusions from a study in the genomic domain, the information used must be well supported by published research results. Besides, manually curated databases are much more reliable than non-curated or automatically curated ones, due to the expert's efforts to verify the existence and correctness of assertion criteria.
- *Consistency*: Data must be consistent between systems and represented in the same format. Information extracted from one data source is not enough to reach proper and meaningful conclusions. This means that diverse data sources must be checked and integrated. But an obvious problem is faced if each one presents its information in its own format (i.e., flat files, XML, HTML, etc.) and uses specific nomenclature based on its own need (for example to determine the type of the variations). Besides, colloquial designations for genes or mutations are used so broadly that many scientists are probably unaware that they are non-standard [7]. Consistency problems lead to a highly time-consuming process of normalization to represent the information under a single normalized model.

- *Uniqueness*: The database won't have redundant data or duplicate records. The number of entries in genomic databases has grown enormously in the last few decades, but this growth was accompanied by higher redundancy. This has become a noteworthy problem and some strategies have been developed to try to minimize it. For example, UniProtKB has developed an algorithm called Proteome Redundancy Detector [11]. When it was applied to their data warehouse (TrEMBL) for the first time, 46.9 million entries were removed from its database [11].
- *Currency*: This dimension can be defined as the extent to which data is sufficiently up-to-date. Genomic domain evolves quickly and information can get obsolete in a relatively short period of time so, this dimension is one of the most important ones to be assessed.

The issues presented are the most common ones that Bioinformaticians, Geneticists and researchers have to face in their daily work. For more specific information and examples see [6]. To reduce its impact, the SILE methodology is enriched with a DQF, which ensures that the information that is collected will have the quality required by the task to be accomplished.

3.3.2.2 Data Quality Framework

In order to assure the quality of the information to be loaded in the database, a set of quality controls needs to be applied in all the three first levels of the methodology: Search, Identification and Load. In Fig. 3.4, the entire process of the methodology can be shown.

The quality controls are based on the six major data quality dimensions presented in the previous section: accuracy, completeness, uniqueness, consistency, reliability and currency:

- *Search level (S)*: The most important dimension to be checked in this level is currency. Currency problems are closely related to accuracy and completeness issues. Examples of parameters used to assess the currency of the genomic information are i) the assembly used to represent a variation ii) the version of the database and iii) the specific last update of each registry. It is very common the use of external identifiers (IDs) to enrich the information provided by the database. When external IDs are being managed, it must be assured that they are currently valid and the links to the databases are working correctly. Situations where the source that is associated to the identifier changes and the link to the involved information becomes obsolete, must be avoided.
- *Identification level (I)*: The dimension checked in this level is reliability. Once the interesting information it is identified among the databases selected in the previous level, the next step is to identify which data has enough quality to be loaded into de HGDB. The minimum criteria to check the reliability of the information depends on the context where it belongs. For example in the case of variations associated to a certain disease the minimum reliability criteria considered are:

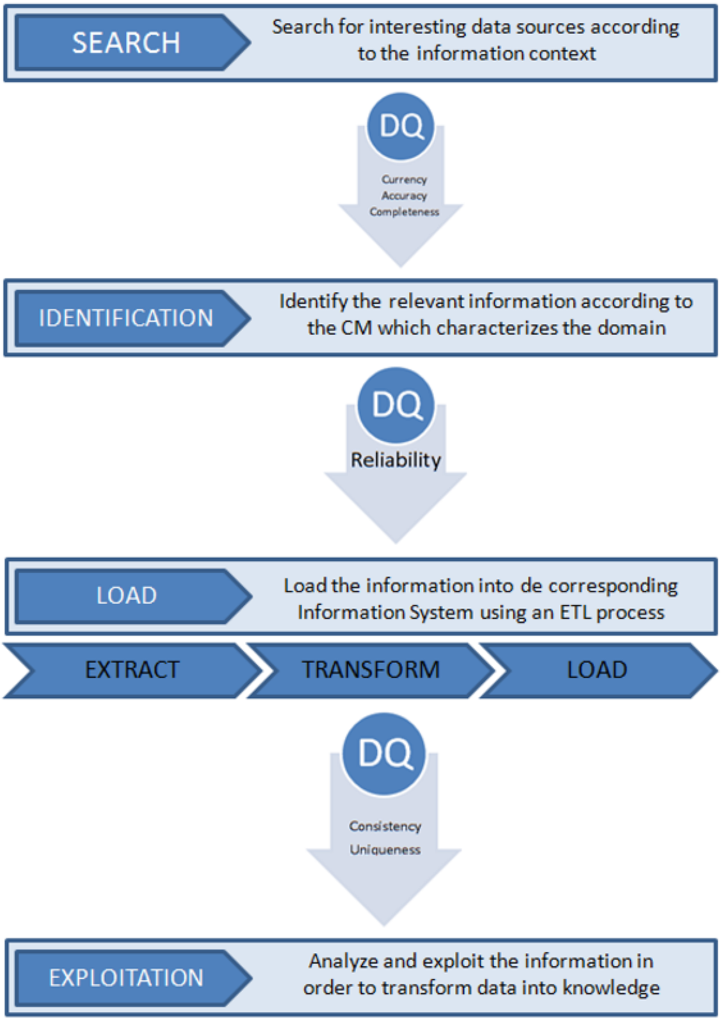

SEARCH
Search for interesting data sources according to the information context
DQ
Currency
Accuracy
Completeness
IDENTIFICATION
Identify the relevant information according to the CM which characterizes the domain
DQ
Reliability
LOAD
Load the information into de corresponding Information System using an ETL process
EXTRACT
TRANSFORM
LOAD
DQ
Consistency
Uniqueness
EXPLOITATION
Analyze and exploit the information in order to transform data into knowledge

Fig. 3.4 SILE methodology and Data Quality Assessment Process.

1. *The clinical significance* of the variation must be clearly defined.
2. There must be *enough assertion criteria* provided for the relationship between the variation and the phenotype associated to it.

Additional quality criteria can be defined, such as the number of publications supporting the evidence and their impact factor, the number and relevance of authors and research institution, statistic metrics such as *p-value, odds ratio*, etc.

- *Load level (L)*: In this level, the information identified in the data sources is going to be extracted and loaded in the HGDB. During the extraction process, information from different databases is going to be collected and merged. One of the main problems of biological databases is the lack of use of proper standards to represent the information, so the integration becomes a no trivial process. Two of the main quality problems which can appear in this level are related to the existence of redundant information (uniqueness issues) and inconsistencies in the representation of the information (consistency issues):

- Consistency: The set of semantic rules can be determined by i) looking at the allowed values; ii) looking at mandatory values (Primary Keys or not nullable values); iii) looking at the type of value the fields should have (*integers, strings, booleans*, etc) which is provided by the HGDB; and iv) looking at the integrity constraints which involve attributes of more than one table (speaking from a relational point of view) or more than one group of attributes.
- Uniqueness is defined as the absence of redundant data or duplicate records. When information from different databases is merged, it is important to identify and remove all redundant records and to assure that the information of those representing the same variation is similar and correct.

With the addition of a corresponding precise set of DQ Metrics, the methodology assures that the information is of high quality (*current, reliable, consistent* and *accurate*).

3.4 Conclusions and Future Work

Precision Medicine is going to change the way in which we have historically understood medicine. The new practical context associated with it requires a sound working environment, and the correct application of the adequate Information Software Engineering (ISE) practices. We assume that Conceptual Modeling together with Data Quality Assessment techniques are the basic strategy to design and develop the required sound and efficient Genomic Information Systems (GeIS), which will assure that both diagnosis and adequate treatment selection are fully reliable.

The paper also highlights the need of having a methodological background designed to characterize a sound conceptual schema-centric genome data management process. Following this need, the SILE methodology has been proposed as a concrete solution.

Future research work will focus on the development, improvement and assessment of all these statements, in order to face the challenge of understanding life from an Information Software Engineering (ISE) perspective, inspired by the Conceptual Schema-centric approach introduced by Prof. Olivé in this research career.

Acknowledgements The authors are grateful to the members of the PROS Center Genome group for fruitful discussions. This work has been supported by the Ministry of Higher Education, Science and Technology (MESCyT) of the Dominican Republic and also has the support of Generalitat Valenciana through project IDEO (PROMETEOII/2014/039), the Research and Development Aid Program (PAID-01-16) of the UPV under FPI Grant 2137. and Spanish Ministry of Science and Innovation through project DataME (ref: TIN2016-80811-P).

References

1. Olivé, A.: Conceptual Schema-Centric Development: A Grand Challenge for Information Systems Research. In: Pastor O., Falcão e Cunha J. (eds) Advanced Information Systems Engineering. CAiSE 2005. Lecture Notes in Computer Science, vol 3520. Springer, Berlin, Heidelberg (2005)
2. Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G., Schwarzenbach, J.: The Six Primary Dimensions for Data Quality Assessment. In: DAMA UK Working Group (2013) Available via White Papers. `http://bit.ly/2qcimnf`
3. Burriel Coll, V., Pastor, O.: Conceptual Schema of Breast Cancer: The background to design an efficient information system to manage data from diagnosis and treatment of breast cancer patients. In IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 432–435. IEEE (2014)
4. Burriel Coll, V., Reyes Román, J.F., Heredia Casanoves, A., Iñiguez-Jarrín, C., León Palacio, A.: GeIS based on Conceptual Models for the Risk Assessment of Neuroblastoma. In IEEE Eleventh International Conference on Research Challenges in Information Science (RCIS), pp. 1–2 (2017)
5. Iñiguez, C.: A conceptual modelling-based approach to generate data value through the end-user interactions: A case study in the genomics domain. PoEM Doctoral Consortium, pp. 14-21 (2016)
6. León, A., Reyes, J.F., Burriel, V., Valverde, F.: Data Quality problems when integrating genomic information. In 3rd. Workshop Quality of Models and Models of Quality (QMMQ 2016) in conjunction with the 35th International Conference on Conceptual Modeling (ER2016). Springer International Publishing pp. 173-182 (2016)
7. Ogino, S., Gulley, M.L., den Dunnen, J.T., Wilson, R.B., and the Association for Molecular Pathology Training and Education Committee: Standard Mutation Nomenclature in Molecular Diagnostics: Practical and Educational Challenges. The Journal of molecular diagnostics (2016) doi:10.2353/jmoldx.2007.060081.
8. Olivé, A.: Conceptual modeling of information systems. Springer-Verlag Berlin Heidelberg pp. 1-445 (2007)
9. Online Mendelian Inheritance in Man Homepage. Available via OMIM. `https://www.omim.org/`
10. Pastor, O., Reyes Román, J.F., Valverde, F.: Conceptual Schema of the Human Genome (CSHG). Universitat Politècnica de València (2016) Available via RiuNet. `http://hdl.handle.net/10251/67297`
11. Reducing proteome redundancy (2016) Available via UniProt. `http://www.uniprot.org/help/proteome_redundancy`
12. Reyes Román, J.F., Pastor, O., Casamayor J.C., Valverde F.: Applying Conceptual Modeling to Better Understand the Human Genome. In The 35th International Conference on Conceptual Modeling, Springer International Publishing, pp. 1-9 (2016)
13. Reyes Román, J.F., Pastor, O. Use of GeIS for Early Diagnosis of Alcohol Sensitivity. In Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016), pp. 284–289 (2016)
14. Reyes Román, J.F., Pastor, O., Valverde, F., Roldán, D.: Including haplotypes treatment in a Genomic Information Systems Management. In Ibero-American Conference on Software Engineering, pp. 11–24 (2016)
15. Reyes Román, J.F., Pastor, O., Valverde, F., Roldán, D.: How to deal with Haplotype data?: An Extension to the Conceptual Schema of the Human Genome. Universitat Politècnica de València (2016). Available via RiuNet. `https://riunet.upv.es/handle/10251/82704`
16. Reyes Román, J.F., León Palacio, A., Pastor López, O.: Software Engineering and Genomics: The Two Sides of the Same Coin?. In Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017), pp. 301–307 (2017)

17. The 24th annual Nucleic Acids Research database issue. Available via Oxford Academic. `http://www.oxfordjournals.org/our_journals/nar/database/c/`
18. Reyes Román, J.F., Iñiguez-Jarrín, C., Pastor López, O.: GenesLove.Me: A Model-basedWeb-application for Direct-to-consumer Genetic Tests. In Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017), pp. 133–143 (2017)