Providing a Realist Perspective on the eyeGENE Database System

Werner Ceusters[1]
[1]New York State Center of Excellence in Bioinformatics & Life Sciences, Buffalo, NY, USA

Abstract

One of the achievements of the eyeGENE Network is a repository of DNA samples of patients with inherited eye diseases and an associated database that tracks key elements of phenotype and genotype information for each patient. Although its database structure serves its direct research needs, eyeGENE has set a goal of enhancing the structure to become increasingly well integrated with medical information standards over time. This goal should be achieved by ensuring semantic interoperability with other information systems but without adopting the incoherencies and inconsistencies found in available biomedical standards. Therefore, eyeGENE's current pragmatic perspective with focus on data and information should shift to a realism-based perspective that includes also the portion of reality described, and the competing opinions that clinicians may hold about it. An analysis of eyeGENE's database structure and user interfaces suggests that such a transition is possible indeed.

Introduction

The eyeGENE database is a repository of genotype and phenotype information of patients with inherited eye diseases collected through the National Ophthalmic Disease Genotyping Network, an initiative launched by the National Eye Institute in 2006.[1] The database design used the innovate approach of defining the structure of phenotype information by means of metadata, so that new diagnoses and clinical finding questions could be added or modified by the eyeGENE administrator at any time. The goal was to allow collection of a large number of samples with a minimal data entry burden to the clinician and genetics testing labs, and to provide an easy overview of key data for a researcher who may wish to study details of an attached eye image or otherwise study the patient's data in more depth. To avoid this system becoming yet another information silo, eyeGENE set a further goal of integrating the eyeGENE data with applicable medical information standards over time. The goal of evolving to meet emerging medical information standards is to provide an additional layer of benefits in more easily collecting, sharing and analyzing data in the future.

As a first step, an extensive study was performed on existing and emerging standards relevant to clinical research data, including the identification of gaps and overlaps.[2] This study revealed that this goal is confounded by deficiencies in many standards pertinent to clinical data registration, which suffer from reductionist views on reality which are constrained by what can be seen through the lenses of either information systems[3] or terminologies and ontologies that adhere to what is called 'concept representation'.[4] Without appropriate remediation, semantic interoperability between systems adhering to such standards will be on a less than fully logically sound foundation and will suffer limitations over time.

Objectives

As witnessed by the success of the OBO-Foundry a growing number of scholars adheres to a realist view on reality and to an implementation along these lines both in ontologies[5] and information systems.[6] The goal of the study reported on here was (1) to understand the type of view embedded in the eyeGENE database and (2) in case this view would differ from the realist one, to propose a migration path towards the latter.

Material and methods

We studied the available documentation about eyeGENE's core medical information, including parts of its information model and user interfaces. We looked at some of the clinical questions (and corresponding possible-answer sets) that are asked when data are entered in the system, as well as to system generated reports about lab procedures performed on genes. We did not have access to a data-dictionary with data-definitions and corresponding business rules.

We checked in the first place for design choices in the system that would lead the information to be collected not to match with the corresponding structure of reality, the latter under the realist view consisting of:

1. first-order reality, which includes entities such as specific patients, their relatives, the disorders they are suffering from, the lab tests that have been conducted, and so forth;

2. second-order reality, including, for instance, interpretations and opinions on the side of clinicians, including hypotheses and diagnoses;
3. third-order reality, which is composed of information about first- or second-order reality, examples being entries in information systems such as the eyeGENE database.

We also checked for structural and functional issues in eyeGENE that in absence of sufficient background information for disambiguation would lead to difficulties in interpreting data once entered.

Results and discussion

We found that to meet its goal of future integration with high quality medical information systems over time, the pragmatic design approach initially followed by the eyeGENE developers should be transformed to remove current limitations of (1) conflating the three levels of reality as described above, and (2) not representing faithfully the relevant portions of reality at each level.

An example of a non-faithful representation of first-order reality is eyeGENE's treatment of the patient's demographic information: the user interface lists a number of data entry fields, amongst which the postal code, as 'required'. A motivation for including 'required fields' in data entry forms is to have data as complete as possible. Sometimes, however, as is the case here, these constraints violate what is the case in reality: many countries do not use postal codes at all. If eyeGENE's realm is not limited to patients living in the US, such constraints pose a problem as they force the user to enter fake data, or, when the latter is against the user's principles, prevent him from entering data at all. A strategy often applied is to allow for various sorts of null-values, but that changes the semantics of the data field drastically: it would then not always contain strings that denote postal codes, but strings that denote, for example, that there are no postal codes in the corresponding country, or that the postal code is not known by the user.

Another example of a required field in eyeGENE is 'gender' with the two possible values 'female' and 'male'. This might seem to be consistent with first-order reality as each human being can be expected to be either 'male' or 'female'. However, for each of the three possible interpretations of what the word 'gender' here might stand for, matters are not that obvious. *Phenotypic gender* is not either male or female in hermaphrodites, *genotypic gender* comes in many more flavors, while, finally, *administrative gender*, depending on the community in which it is defined, is based not only on scientific grounds but also on political, ethical, and even religious considerations, thereby giving rise to oddities to the effect that the different treatments of the right of gender self-identification makes it possible that the same person has a different administrative gender in Australia and in the US.[7]

The eyeGene database contains many examples not of unfaithful representation of reality but rather of undocumented reductionism. It allows, for instance, the eye fundus to be described as being normal or exhibiting any combination of four types of anomalies. By 'undocumented', we mean that it is left unspecified whether these four types are the *only* possible types in reality, or whether there are many more possibilities of a sort which are not relevant for the purposes for which eyeGENE has been designed, and therefore are not offered as additional alternatives.

An example of a conflation of first-order and second-order reality is in the registration of diagnoses. Clinicians are requested to provide the date of examination and then to select one or more types of diagnoses out of a list of 21. Based on that information and with the goal to collect further data about signs and symptoms, clinical data entry forms specific for each type of diagnosis are generated. These forms are composed out of building blocks some of which, for example to provide details about the patient's 'best corrected visual acuity', can appear in forms related to more than one diagnosis. Once data are provided in the context of one diagnosis, the same data re-appear in the form corresponding to another diagnosis. This setup, although being very pragmatic – it frees the clinician from entering the same data more than once – leads to ambiguities from an ontological perspective.

One arises from the mere fact of entering diagnoses without identifying the corresponding disorder *about which* that diagnosis is a diagnosis: disorders are first-order entities on the side of the patient while diagnoses are second-order entities on the side of, for instance, the clinician.[8] Disorders and diagnoses live totally different lives: patients may have a disorder without any diagnosis being made; clinicians may come to one diagnosis while the patient may have either two or more disorders or no disorder at all; distinct clinicians may bring forward different diagnoses for the one disorder the patient has; a clinician may change his diagnosis over time, while the disorder does not change at all, and so forth. The problem becomes obvious when more than one clinical examination form is entered: in absence of identifiers for the disorder, it is not possible to deduce formally in case a diagnosis entered on an earlier form is different from the diagnosis on a later form whether the difference is because the earlier diagnosis is revised, whether there is a second disorder involved, or, if distinct clinicians entered the forms whether there is disagreement about the correct diagnosis.

Another ambiguity, when multiple diagnoses are specified, is to what the individual clinical signs relate. Although clinical signs provide evidence in favor or against certain diagnoses, a particular clinical sign in some patient is not related to any diagnoses entertained for that patient, but rather to at least one disorder from which that patient suffers.

Recommendations

It is no surprise that the information model of the eyeGENE database exhibits the sorts of mismatches with reality just described: to our best knowledge, ***all*** information systems designed according to the state of the art in information modeling suffer from these incoherencies because of at least two misconceptions.

One is the *erroneous assumption of inherent classification* adhered to in many database design circles according to which entities can be referred to only as instances of pre-specified classes.[9] We, in contrast, defend the position that in information systems entities should exclusively be referred to by means of globally unique and singular identifiers.[6] These identifiers can then to be used in descriptions of various sorts indicating, for instance, what universals are instantiated by the entity referred to, what terms from a terminology or concept-based ontology apply to it, or how the entity relates to other entities.

The other misconception is the tyranny of the use case, what leads some to argue that '*if most people wrongly believe that crocodiles are a kind of mammal, then most people would find it easier to locate information about crocodiles if it were located in a mammals grouping, rather than where it factually belonged*'.[10 p89]

Of course, the incoherencies of the information model and business rules as compared to what is the case in reality are not relevant to the original goals for which eyeGENE has been designed. But they do become a problem when the data have to be pooled with data coming from other information systems that describe partially or in total the same domain from a different perspective and are collected for another purpose. In that case, the second system, if designed following prevailing approaches, will also contain incoherencies with respect to reality, but in different ways than eyeGENE. A comparative analysis of the underlying information models may reveal areas where they are in agreement and other places where they can not have it right both. But in absence of an external benchmark, we have no means to assess which one is right, not even when both models are in agreement because they both might have it wrong in the same way.

We argue that reality should function as that benchmark, and that realism-based ontology provides the means to reach that goal in similar ways as it is increasingly and successfully used for quality assurance in biomedical terminologies and ontologies.[5] The reason is that no portion of reality depends on the information used to describe it or on the purposes for which such information is collected. This is not to say that such information does not contribute to the evolution of reality at all. On the contrary, as soon as it is generated, that information is part of reality itself (level 3), and so is the system used to manage it. Therefore any attempt to make such system, in our case the eyeGENE database system, coherent with respect to reality, should acknowledge the priorities and objectives that have been taken into account at design time. If, for instance, through realism-based analysis one discovers a reductionist approach (e.g. the eye fundus description described earlier), it would be a bad idea to bother the users of the eyeGENE database with a more complex interface that does not bring them advantage in any way, even if it would help secondary users of the data.

The right way forward, so we argue, is by mapping the information model of eyeGENE to a domain model that itself is not reductionist in nature. Reductionist models are typically created when UML is used as this language forces reality to be viewed through the eyes of an information system, using a (partially graphic) vocabulary which is inadequate to describe reality faithfully. The HL7 RIM is the most dramatic example, dramatic because its acceptance as ISO standard gives it an unjustified aura of excellence.[11]

Note that we see no harm in using an existing information model to scope the corresponding domain model. The procedure, in the context of eyeGENE, would be to study each of its tables, data fields and associated allowed values, as well as any hard- or soft-coded business rules that restrict data-input, with the following goals: (1) to assess what (type of) entity in reality would be denoted by any data instance, (2) to represent how these entities in reality relate to each other as well as to other ontologically relevant entities that are not explicitly addressed in the information model, this being the domain model proper, and (3) to describe formally how the information model has to be interpreted in terms of the domain model. The latter can then be used to inform third party systems with which the eyeGENE database system would exchange data about the implicit restrictions in eyeGENE. It can also be used to identify issues that must be resolved in further releases.

As an example, eyeGENE's information model relates a *PatientDiagnosis* to (1) a *ClinicalEncounter* which itself is related to a *Patient* and a *PhysicianPerson* and (2) a *Diagnosis*. The eyeGENE database system limits the latter to 21 types, however, upon closer inspection, not to types of diagnoses, but to types of diseases. The domain model would tell us that there are of course many more types of diseases. The interpretation model would then contain statements clarifying this distinction. With respect to (1), the interpretation model could clarify, for instance, whether the date of the *ClinicalEncounter* is the date that the diagnosis was made, and that this by itself would not allow inferences to be made about when the disease started. To some extent, eyeGENE users can clarify such issues in free text, but this cannot be used for automated processing.

Conclusion

The eyeGENE database system is successfully in use since July 2006 and processes 35 samples per month. It is foreseen that this number will grow to 100 by end 2009. To most optimally fulfill its ambitious goal of integration with high quality medical information systems in future developments, the eyeGENE database system can become a model of fulfilling a stated objective in the NIH roadmap to '*require new ways to organize how clinical research information is recorded, new standards for clinical research protocols, modern information technology*'. One expectation, in the context of the patient profile, is that at some future time relevant phenotypic data can be automatically extracted from an electronic medical record using a standard in widespread use. At that point, a larger set of base patient data in more specific detail would be practical to collect. Realism-based ontology combined with adequate identification and reference of entities at each level of reality is one new way that can be explored to turn these data into knowledge.

Acknowledgements

With thanks to David Scheim of the National Eye Institute for many useful comments on the first version of this paper.

References

1. National Eye Institute. eyeGENE - National Ophthalmic Disease Genotyping Network. Jan 2009;http://www.nei.nih.gov/resources/eyegene.asp. Accessed January 26, 2009.

2. Rudnicki R, Ceusters W. *Emerging medical information standards as applicable to clinical research data: A study performed in the context of the project 'Exploring eyeGENE, an International Genotype / Phenotype Database, from a Bioinformatics Perspective'.* Buffalo NY: NYS Center of Excellence in Bioinformatics & Life Sciences; July 16 2008.

3. Ceusters W, Smith B. What do identifiers in HL7 identify? An essay in the ontology of identity. In: Ogawa Y, ed. *InterOntology09*. Tokyo: Keio University Press; 2009 (in press).

4. Smith B. From Concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies. *Journal of Biomedical Informatics.* 2006;39(3):288-298.

5. Smith B, Ashburner M, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*. 2007;25:1251-1255.

6. Ceusters W, Manzoor S. How to track Absolutely Everyting? In: Obrst L, Ceusters W, Janssen T, eds. *Ontologies for Intelligence*. Amsterdam: IOS Press; 2009 (in press).

7. Milton SK. Top-Level ontology: the problem with naturalism. In: Guarino N, ed. *Formal Ontology in Information Systems.* Vol 85-94. Amsterdam, The Netherlands: IOS Press; 1998.

8. Scheuermann RH, Ceusters W, Smith B. Toward an Ontological Treatment of Disease and Diagnosis. *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*: American Medical Informatics Association; 2009 (forthcoming).

9. Parsons J, Wand Y. Emancipating Instances from the Tyranny of Classes in Information Modeling. *ACM Transactions on Database Systems.* June 2000 2000;25(2):228-268.

10. Huhns MN, Stephens LM. Semantic Bridging of Independent Enterprise Ontologies. In: Kosanke K, ed. *Enterprise Inter- and Intra-Organizational Integration: Building International Consensus*. Boston, MA: Kluwer Academic Publishers; 2002:83 - 90.

11. Aerts J. Ten good reasons why an HL7-XML message is not always the best solution as a format for a CDISC standard. February 10, 2009; http://www.xml4pharma.com/HL7-XML/HL7-XML_for_CDISC_Standards.pdf.