Using Ontologies and Semantic Web Technology on a Clinical Pedigree Information System

João Miguel Santos[1], Beatriz Sousa Santos[1,2], and Leonor Teixeira[2,3]

[1] Dep. Electrónica, Telecomunicações e Informática, Universidade de Aveiro, Aveiro, Portugal
[2] Instituto de Engenharia Electrónica e Telemática, Universidade de Aveiro, Aveiro, Portugal
[3] Dep. Economia, Gestão e Engenharia Industrial, Universidade de Aveiro, Aveiro, Portugal
{miguelsantos,bss,lteixeira}@ua.pt

Abstract. Clinical family histories, in the form of clinical pedigrees, are recognized as valuable tools in the diagnostic, risk assessment and treatment of patients and their family members. The lack of adequate tools in present health information systems (HIS) is one of the factors that currently deter practitioners from making full use of these tools. In this paper we present OntoFam, an ontology-based clinical pedigree information system that can be integrated with existing HIS. We focus on the usage of ontologies and semantic web technology in the context of this information system and present a practical scenario of integration with hemo@care, a HIS designed for hemophilia care.

Keywords: ontology-based information system, clinical family history, clinical pedigree, semantic web technology.

1 Introduction

It has long been observed that certain diseases are more prevalent in some families than in others. Since ancient times, medical practitioners have been complementing patient symptoms with their respective family history to better understand disease manifestations [1]. Advances in Genetics have unveiled dozens of medical conditions that are linked to genetic factors, including common illnesses such as diabetes, Alzheimer, schizophrenia and depression, among many others [2-4]. It is therefore becoming increasingly important to record patients' clinical family histories to aid in the diagnostic, risk assessment and treatment of patients and their family members [4, 5]. Risk assessment is particularly important, as the timely detection of at-risk individuals may allow preventive medicine to delay, diminish or avoid illnesses or symptoms [6].

There are several ways to record clinical family histories, in either text form (checklists, forms, descriptive text) or graphic form (genograms, ecomaps, clinical pedigrees) [2, 7]. Clinical pedigrees are a particularly well-accepted tool for recording and presenting clinical family histories, as they are relatively easy to generate and understand – a basic hand-drawn pedigree can be constructed in minutes and, because it is a graphical representation, important information such as disease heredity

V.G. Duffy (Ed.): DHM 2014, LNCS 8529, pp. 448–459, 2014.

patterns, penetrability, mortality and risk can be quickly assessed by observers [1, 2, 8]. Standard symbols and rules are available to graphically represent family structure, patient symptoms, environmental factors, test results, genetic traits and other relevant information on a pedigree, thanks to the efforts of the Pedigree Standardization Work Group (PSWG) [9, 10]. Fig. 1 represents an example clinical pedigree using PSWG notation: circles and squares represent females and males, respectively; spouses are joined at the sides by horizontal lines from which a vertical line may depart, denoting the existence of offspring; in the case of multiple children, siblings are grouped below an horizontal line that connects to their parents' vertical offspring line; filled shapes represent medical conditions and dotted shapes denote carriers; diagonal strokes symbolize deceased individuals and arrows point to individuals that have been studied; roman numerals denote generations, and arabic numerals enumerate and distinguish individuals within a generation.

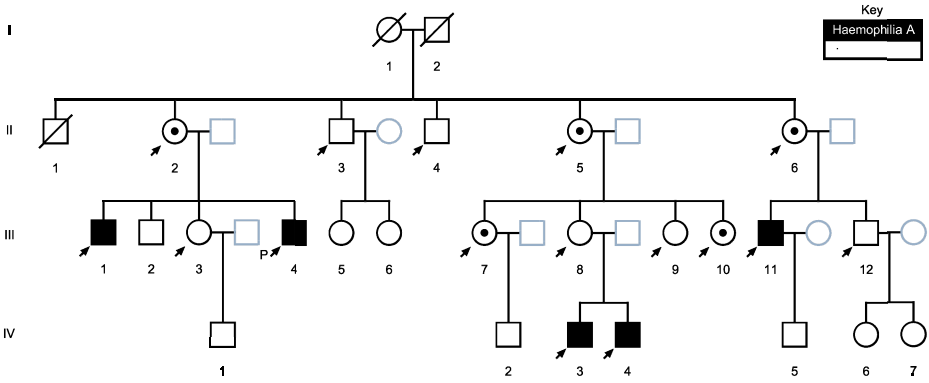

Fig. 1. A clinical pedigree using PSWG notation

Even though the usefulness of clinical pedigrees has been proven in several studies, they are still not as widely used as would be desirable [1, 6, 11]. We believe the reasons for this are twofold: on the one hand, practitioners may not be fully aware of the advantages of keeping clinical family histories of patients; on the other hand, many Healthcare Information Systems (HIS) do not provide adequate tools for keeping family history information or representing clinical pedigrees [3, 4, 11, 12]. Some practitioners solve the latter problem by resorting to external pedigree-drawing tools or generic graphic design software, but this approach results in pedigrees that are "disconnected" of current clinical information and are therefore difficult to keep up-to-date. It appears evident that the use of an information system to effectively manage clinical family histories and pedigrees, and that integrates with existing HIS, will remove many of the hurdles that currently deter practitioners from adopting and taking full advantage of clinical family histories and pedigrees.

In this paper we present OntoFam – an ontology-based information system that facilitates the creation and management of clinical pedigrees and that can be integrated with existing Health Information Systems. This is, to the best of our knowledge, the

first open system to use ontologies and semantic data to represent knowledge about clinical family histories and pedigrees.

We begin by describing the design goals and resulting system architecture in section 2. Section 3 focuses on ontologies and semantic web technology, briefly describing the relevant concepts before detailing the chosen ontology and exemplifying the resulting semantic data. The expected benefits of choosing these technologies are also presented. Section 4 summarily describes a practical application of the system and section 5 presents our conclusions and suggestions for future work.

2 Overview of System Architecture

We aimed at designing a system that could be used either in standalone form or, more importantly, in conjunction with existing Healthcare Information Systems (HIS); that stored family history information in an open and standard format which could be understood and used by other systems; and that could adapt to specific clinical areas without system redesign. As such, Integration, Openness and Extensibility played a big part in system design and technological decisions.

Fig. 2 presents a high-level view of the resulting system architecture. The system follows a typical n-tier design, where a central business layer mediates the data and user presentation layers. Optionally, the system interfaces with existing HIS to gather clinical data regarding individuals represented in pedigrees. The system uses ontologies and semantic web technology to store, validate and reason with family history data, rather than rely on more "traditional" relational database-oriented approaches.

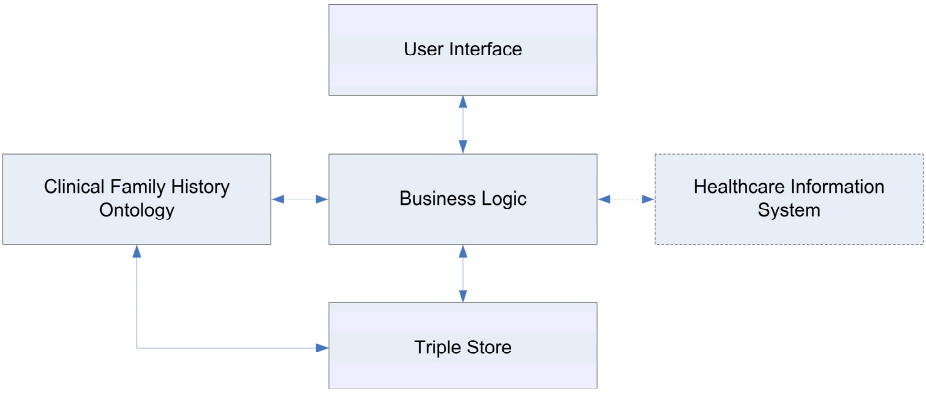

Fig. 2. High-level System Architecture

The **user interface layer** consists of a web application that uses Scalable Vector Graphics (SVG) to represent pedigrees in PSWG notation (refer back to Fig. 1 in section 1 for an example). The Madeline 2.0 Pedigree Drawing Engine [13, 14] is used internally to generate the predigree SVGs, as we have found that this engine produces the most standard-compliant clinical pedigrees, out of several open-source and freeware tools that were considered (Cranefoot, Haplopainter, Hughes RiskApps

Pedigree Module, Kinship package in R, PedHunter, PedigreeQuery, Pelican, and My Family Health Portrait). The pedigree representation is interactive – users can add and remove family members and edit their clinical information – therefore the same graphical representation is used for input and output of information. This contrasts with some pedigree-generation systems that use tables and forms to gather pedigree information and produce a graphical pedigree as a result (for example, the My Family Health Portrait tool [15]). While tables and forms may be sufficient for household consumers and simple family structures, we believe that a more direct "What You See Is What You Get" (WYSIWYG) graphical approach is preferable for practitioners' use and for representing complex families, as seen on professional systems such as Progeny Clinical [16]. We have considered the latter as a reference on the sort of interaction to achieve on our system.

The **business logic layer** coordinates data exchange between the user interface and data layers. It resides on the server-side component of the web application and essentially acts as a translator between the visual and ontological representations of clinical family histories. It uses ontologies to validate and reason with data, ensuring that it is semantically correct and inferring new information from existing data, when possible. The dotNetRDF library [17] is used internally to connect to the triple store and to process semantic data and ontologies.

The **triple store** acts as the data layer of the system. Thanks to the flexibility offered by the dotNetRDF library, many different triple stores can be used to store data, such as AllegroGraph, 4store, Fuseki, Virtuoso, Stardog or any Sesame-based store. Most of these stores are capable of understanding ontologies and can therefore infer information from stored triples, a task that the dotNetRDF library is also capable of. At present, we are experimenting with several triple stores in order to determine the best fit, but this will most likely be a choice delegated to the deployment stage, allowing an organization to choose whichever store is more appropriate for its environment.

The optional **interface to existing Health Information Systems** allows patient data to be gathered and refreshed automatically from the HIS, so that the clinical information on the pedigree is always up-do-date. Internally, the system uses a plug-in architecture to allow communication with different HIS. The task of the plug-in is to translate patient data from HIS to ontological representations. At present, we are utilizing plug-ins that import data in custom HIS formats, but are also working on a plug-in that consumes Health Level Seven International (HL7) standards, namely SECTION 1 Primary standards, which appear to be the most popular for integration and interoperability purposes [18].

3 Usage of Ontologies and Semantic Web Technology

This section focuses on OntoFam's usage of ontologies and semantic web technology. The most relevant concepts are briefly described before presenting the ontology and example semantic data used by the system. The expected benefits of choosing these technologies conclude this section.

3.1 Introduction to Ontologies and Semantic Web Technology

Originally a philosophical term related to the study of reality and existence, in computer science an ontology can be briefly described as the representation of an area of knowledge in a way that machines can understand [19]. Built on World Wide Web Consortium (W3C) standards such as eXtensible Markup Language (XML), Resource Description Framework (RDF) and Web Ontology Language (OWL), among others, ontologies use classes, attributes, relations, restrictions and rules to describe a knowledge domain in a consensual, shared and formal manner [19, 20]. These technologies were initially aimed at building the Semantic Web, Tim Berners-Lee vision of a "web of data" where the World Wide Web would no longer consist of a series of unrelated documents but rather a web of interrelated knowledge usable by humans and computers agents alike [21]. However, semantic web technology found its way into information systems not directly related to the Web, such as the system described in this paper. The BioMedical field appears to be particularly proficient in ontology construction and usage, with applications employing ontologies for knowledge and workflow management, data integration and interoperability, decision support and computer reasoning [22].

3.2 Representation of Semantic Data

As previously mentioned, an ontology defines a set of classes, properties and rules that describe a knowledge domain. Actual semantic data, that is, data that conforms to a certain ontology, is represented in *triples*: statements describing that an entity (subject) has a certain relation (predicate) with another entity (object). The ontology dictates what subjects can relate in what way to what objects. It can also further describe that relation by imposing restrictions and rules.

At the beginning of any ontology-oriented project, a choice must be made whether to build new or reuse existing ontologies. Given that interoperability is one of the goals of ontologies, and indeed of our information system, reuse of existing ontologies, when available, is advisable. A third option is to extend an existing ontology that is a "close fit" by adding the elements that are missing. For example, new classes and properties can be added, or derived from existing entities, to fill gaps and further describe the knowledge domain.

To facilitate share and reuse of ontologies, several repositories exist on the Web, with Unified Medical Language System and BioPortal being two of the largest [23]. Using these tools, we were hoping to find genealogy and clinical ontologies that we could combine to represent clinical family histories. Instead, we have found that work had already been done on an ontology to specifically represent clinical family histories [24, 25], resulting in the Family Health History Ontology (FHHO)[1]. This ontology defines 240 classes and 290 properties that allow representation of family relations and illnesses, including non-biological relations such as adoptive, foster and

[1] Available at `http://bioportal.bioontology.org/ontologies/FHHO` (accessed 2012-10-30)

emotional. It also includes rules to automatically compute 3 generations of family relations. The ontology has been instance-tested and found adequate for representing most, though not all, of the relevant information for the tested families [26]. Fig. 3 represents FHHO's conceptual model, demonstrating how persons relate to each other and to their health states in this ontology.

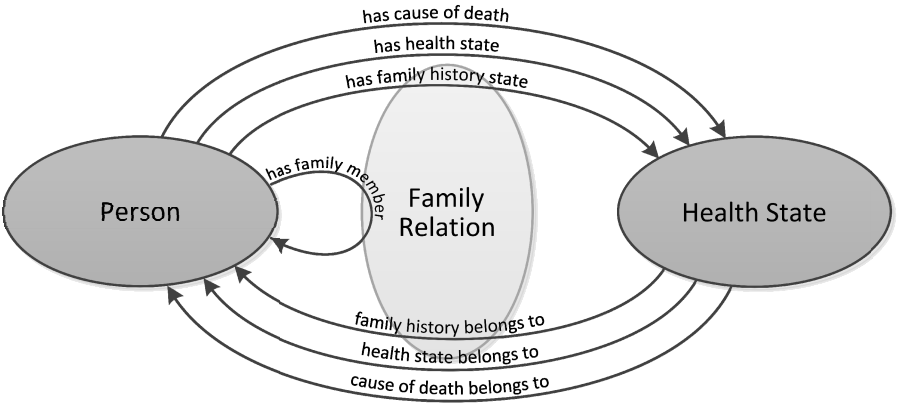

Fig. 3. Family Health History Ontology conceptual model, adapted from [25]

Given that FHHO closely matched the needs of our system, we chose to use it as the foundation for knowledge representation about clinical family histories, extending it where necessary, such as further detailing health conditions, symptoms and health risk behaviors (which we found to be lacking in completeness).

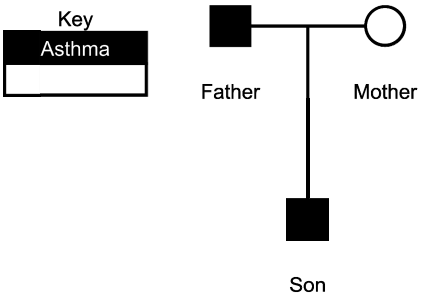

Fig. 4. Clinical pedigree representation of a simple family structure

As an example of what semantic data that follow this ontology looks like, Fig. 4 presents a simple clinical pedigree (3 individuals, two of which suffer from asthma) and Fig. 5 contains a graphical representation of the semantic data involved (simplified for brevity). Directed edges represent predicates, nodes from which edges depart represent subjects and remaining nodes represent objects. Subjects and objects which are instances of classes are represented as ellipses, while simple objects, such as plain

strings, are represented as rectangles. Fig. 6 presents the same semantic data in Turtle syntax (a terse alternative to XML-based RDF representation). The "fam:" prefix is used to shorten the full path to classes and properties and the "a" shortcut is used to define that a subject is of a certain type (i.e., is an instance of a certain class). Consecutive triples that refer to the same subject are separated by semi-colons and omit the subject from the second statement onwards.

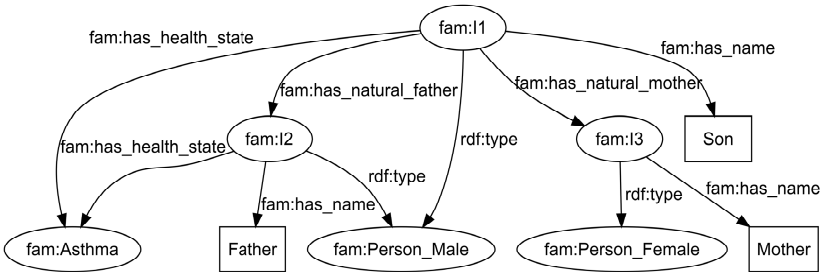

Fig. 5. Graphical representation of semantic data for the previous family structure

```
@prefix fam: <http://www.owl-ontologies.com/Ontology1172270693.owl#>.
fam:I1 a fam:Person_Male;
     fam:has_natural_father fam:I2;
     fam:has_natural_mother fam:I3;
     fam:has_name "Son";
     fam:has_health_state fam:Asthma.
fam:I2 a fam:Person_Male;
      fam:has_health_state fam:Asthma;
      fam:has_name "Father".
fam:I3 a fam:Person_Female;
      fam:has_name "Mother".
```

Fig. 6. Turtle representation of semantic data for the previous family structure

The semantic data from figures 5 and 6 is simplified for brevity. Because of inference and ontology rules, many more triples are used to fully describe this family history. For example, the fact that I1 is a Person_Male implies that it is also a Person, the class from which Person_Male derives. And since I1 has I2 for a father, that implies that I2 is_father_of I1. These inferred triples are materialized and stored in the triple store, in order to allow powerful and flexible queries to be quickly performed.

3.3 Benefits of Using Ontologies and Semantic Web Technology

The decision to design OntoFam around ontologies and semantic web technology, rather than use more traditional approaches such as relational databases, was largely based on a series of benefits that contribute towards our goal of designing an open, extensible and interoperable information system.

Interoperability: XML, RDF, OWL and other semantic web technologies are W3C standards, which means that the clinical family history data generated by our system is inherently interoperable with other systems. Data integration and interoperability are indeed two of the main reasons why ontologies are used in biomedical information systems [22]. Furthermore, because semantic data and ontologies can be represented in plain text format, recorded family histories are relatively safe from obsolescence and will be readable in the future, even if the original software used to create them is no longer available.

Flexibility: Family histories can be quite complex. While a basic ancestor tree can be organized around a simple Person class, with recursive relations for Mother and Father properties, a clinical family tree needs to handle not-so-obvious relations between individuals such as twins, adoption, fostering, interbreeding and interrupted pregnancies, among others. Furthermore, clinical information regarding individuals must also be stored. Rather than depend on relational data tuples, semantic data relies on simple subject-predicate-object triples as units of information. This simplicity offers great flexibility, as any entity characteristic or relation can ultimately be decomposed into triples. This means that complex relations that would imply intricate relational database schemas can be easily represented in a series of related triples.

Extensibility: Existing ontologies can be extended, further describing certain aspects of a knowledge domain. A basic clinical ontology describing high-level diseases and symptoms can be extended to handle more specific information, by branching existing classes into subclasses or by adding new attributes and classes. Unlike relational databases, subclassing or adding properties to an ontology does not break the existing "schema" and does not invalidate existing data [27]. This means that the information system is able to adapt to particular or evolving scenarios with little or no code changes and without raising incompatibilities with existing data.

Computer Reasoning and Inference: Ontologies include relations, restrictions and rules, which can be used to validate input, reason with data and infer new information from existing data. Family structures are a good example of where inference may be helpful, indeed FHHO includes a set of rules that allow the system to automatically infer family relations up to the third generation (for example, the fact that A is mother of B and B is mother of C implies that A is grandmother of C). More importantly, rules can be used to automatic perform risk analysis and detect at-risk individuals.

Queriability: Semantic data can be queried using SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language). This query language builds on the triple nature of semantic data to allow very flexible queries. Fig. 7 presents an example SPARQL query for obtaining the names of the persons whose great-grandparent have asthma.

```
PREFIX fam:<http://www.owl-ontologies.com/Ontology1172270693.owl>
SELECT ?name
WHERE {
   ?person a fam:Person.
   ?person fam:name ?name.
   ?person fam:has_great_grandparent ?ggp.
   ?ggp fam:has_health_state fam:Asthma.
}
```

Fig. 7. Example SPARQL query to obtain persons whose great-grandparent have asthma

4 Prototype Integration with a Health Information System

Hemo@care is a HIS built specifically for hemophilia care. It was initially deployed at the Hematology Service of Coimbra Hospital Center, in Portugal, and has since been extended to provide a nationwide registry of hemophiliac patients [28, 29].

While hemo@care stores electronic health records (EHR) that contain most of the information pertinent to the treatment and health condition of patients, the system had limited family history capabilities. Fig. 8 presents a clinical family history as presented by hemo@care prior to OntoFam integration. This representation uses background colors to distinguish hemophiliacs (pink) and carriers (blue) from non-carriers (white). Studied individuals are denoted with a red outline, known family members are represented in full color icons which contrast with unknown family members, represented in grayscale. Family members that have EHRs have their family relation written in bold, while those without EHRs are strikethrough.

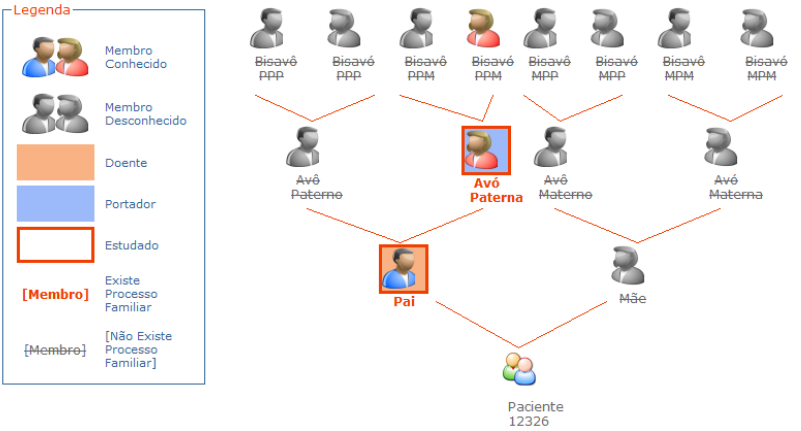

Fig. 8. Hemo@care's representation of a clinical family history (user interface in Portuguese)

The main problem with this representation is that it only includes information regarding ancestors (parents, grandparents and great-grandparents) of patients. Because of this, practitioners that wanted to study the full family history resorted to using external drawing tools to represent clinical pedigrees. As mentioned before, this practice produces "disconnected" pedigrees that are difficult to keep up-to-date.

Seeing as an integrated clinical pedigree information system would benefit users, hemo@care's team welcomed a prototype integration with OntoFam, which is currently being implemented. This involved extending the FHHO to incorporate concepts of hemophilia, as it was limited in that respect. Subclasses of Health_State were created to describe patients with Haemophilia A, B, C and asymptomatic carriers. A specific plug-in was developed to allow OntoFam to acquire patient data from hemo@care. Fig. 1, in section 1, provides an example of a clinical pedigree in PSWG notation applied to the hemophilia field, resulting from the integration of OntoFam and hemo@care. In the process of standardizing the pedigree representation, hemo@care's notation was replaced with the corresponding PSWG symbols, maintaining all relevant information on the pedigree.

Even though the hemo@care ⇔ OntoFam integration is currently at a prototype stage, the preliminary results and user feedback are promising. Practitioners are now able to build and study the complete family histories of patients in context and without resorting to external, disconnected tools.

5 Conclusions and Future Work

Several authors have identified the need for family history-aware Health Information Systems (HIS), yet presently most systems have little or no such capabilities. We believe that an open clinical pedigree information system, which can be integrated with existing HIS, may provide an immediate solution. OntoFam represents our efforts in designing and implementing such a system. We chose to design the system around ontologies and semantic web technology in order to achieve the degrees of Openness, Interoperability and Extensibility that we find necessary for its successful integration with existing HIS.

Though the system is currently at a prototype stage, integration with an existing HIS dedicated to hemophilia care yielded positive and encouraging feedback. In the near future, we expect to finalize the system implementation and begin performance and user acceptance tests to confirm the system's viability and adequacy. In the longer term, we intend to further explore inference and computer reasoning capabilities inherent to semantic web technology in order to enable automatic risk analysis and decision support in OntoFam.

Acknowledgments. This work is funded by National Funds through FCT - Foundation for Science and Technology, in the context of the project PEst-OE/EEI/UI0127/2014.

References

1. Hinton Jr., R.B.: The Family History: Reemergence of an Established Tool. Crit. Care Nurs. Clin. North. Am. 20, 149–158 (2008)
2. Bennett, R.L.: The Practical Guide to the Genetic Family History, 2nd edn. Wiley-Blackwell (2010)
3. Kmiecik, T., Sanders, D.: Integration of Genetic and Familial Data into Electronic Medical Records and Healthcare Processes (2009), http://www.surgery.northwestern.edu/docs/KmiecikSandersArticle.pdf (November 13, 2012)
4. Rich, E.C., Burke, W., Heaton, C.J., Haga, S., Pinsky, L., Short, M.P., Acheson, L.: Reconsidering the Family History in Primary Care. Journal of General Internal Medicine 19, 273–280 (2004)
5. Morales, A., Cowan, J., Dagua, J., Hershberger, R.E.: Family History: An Essential Tool for Cardiovascular Genetic Medicine. Congestive Heart Failure (Greenwich, Conn.) 14, 37–45 (2008)
6. Frezzo, T.M., Rubinstein, W.S., Dunham, D., Ormond, K.E.: The Genetic Family History as a Risk Assessment Tool in Internal Medicine. Genet. Med. 5, 84–91 (2003)
7. American College of Obstetricians and Gynecologists, "Committee Opinion No. 478: Family History as a Risk Assessment Tool," Obstet Gynecol, vol. 117, pp. 747-750 (Mar 2011)
8. Wattendorf, D.J., Hadley, D.W.: Family History: The Three-Generation Pedigree. Am. Fam. Physician 72, 441–448 (2005)
9. Bennett, R.L., Steinhaus, K.A., Uhrich, S.B., O'Sullivan, C.K., Resta, R.G., Lochner-Doyle, D., Markel, D.S., Vincent, V., Hamanishi, J.: Recommendations for Standardized Human Pedigree Nomenclature. J. Genet. Couns. 4, 267–279 (1995)
10. Bennett, R.L., French, K.S., Resta, R.G., Doyle, D.L.: Standardized Human Pedigree Nomenclature: Update and Assessment of the Recommendations of the National Society of Genetic Counselors. J. Genet. Couns. 17, 424–433 (2008)
11. Feero, W.G., Bigley, M.B., Brinner, K.M.: New Standards and Enhanced Utility for Family Health History Information in the Electronic Health Record: An Update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. Journal of the American Medical Informatics Association 15, 723–728 (2008)
12. Scheuner, M.T., de Vries, H., Kim, B., Meili, R.C., Olmstead, S.H., Teleki, S.: Are Electronic Health Records Ready for Genomic Medicine? Genet. Med. 11, 510–517 (2009)
13. Trager, E.H., Khanna, R., Marrs, A., Siden, L., Branham, K.E.H., Swaroop, A., Richards, J.E.: Madeline 2. 0 PDE: A new program for local and web-based pedigree drawing. Bioinformatics 23, 1854–1856 (2007)
14. Trager, E.H., Khanna, R., Marrs, A.: Madeline Pedigree Drawing Engine, version 2.0 rev. 99 (2011), http://eyegene.ophthy.med.umich.edu/madeline/index.php
15. U.S. Department of Health & Human Services. My Family Health Portrait (2009), http://familyhistory.hhs.gov/ (November 11, 2012)
16. Progeny Software, "Progeny Clinical", version 8.0 (2011), http://www.progenygenetics.com/clinical/
17. Vesse, R., Zettlemoyer, R.M., Ahmed, K., Moore, G., Pluskiewicz, T.: Dotnetrdf - Semantic Web, RDF and SPARQL Library for C#/.Net, version 1.0.3 (2014), http://www.dotnetrdf.org/

18. Health Level Seven International. HL7 Standards - Section 1: Primary Standards (2013), http://www.hl7.org/implement/standards/product_sectioncfm?section=1 (January 09, 2013)
19. Boulos, M.N.K., Roudsari, A.V., Carson, E.R.: Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research 8, MT124 (2002)
20. Antoniou, G., Van Harmelen, F.: A semantic web primer. MIT Press (2004)
21. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 284, 1–5 (2001)
22. Bodenreider, O.: Biomedical ontologies in action: Role in knowledge management, data integration and decision support. Yearb. Med. Inform. 47, 67–79 (2008)
23. Fung, K.W., Bodenreider, O.: Knowledge representation and ontologies. In: Richesson, R.L., Andrews, J.E. (eds.) Clinical Research Informatics, pp. 255–275. Springer, Heidelberg (2012)
24. Peace, J., Brennan, P.F.: Ontological representation of family and family history. AMIA Annu. Symp. Proc. 1072 (2007)
25. Peace, J., Brennan, P.F.: Formalizing nursing knowledge: from theories and models to ontologies. Stud. Health Technol. Inform. 146, 347–351 (2009)
26. Peace, J., Brennan, P.F.: Instance testing of the family history ontology. AMIA Annu. Symp. Proc., 1088 (2008)
27. Segaran, T., Evans, C., Taylor, J.: Programming the semantic web. O'Reilly Media, Incorporated (2009)
28. Teixeira, L., Ferreira, C., Santos, B.S.: User-centered requirements engineering in health information systems: A study in the hemophilia field. Computer Methods and Programs in Biomedicine 106, 160–174 (2012)
29. Teixeira, L., Ferreira, C., Santos, B.S., Saavedra, V.: Web-enabled registry of inherited bleeding disorders in Portugal: Conditions and perception of the patients. Haemophilia 18, 56–62 (2012)