Modeling Genealogical Domain
An Open Problem

Joan Campanyà Artés[1], Jordi Conesa Caralt[2] and Enric Mayol[3]
[1]*Universitat Politècnica de Catalunya, Barcelona Tech, Barcelona, Spain*
[2]*Estudis d'Informàtica, Multimèdia i Telecomunicació, Universitat Oberta de Catalunya, Barcelona, Spain*
[3]*ESSI Department, Universitat Politècnica de Catalunya, Barcelona Tech, Barcelona, Spain*

Keywords: Conceptual Modeling, Genealogy, Ontologies.

Abstract: The automated processing, storing and knowledge inference of genealogical data presents several difficulties. Roughly eighteen years ago, the FamilySearch organization published GEDCOM, a new standard file format to allow genealogy software and tools to exchange genealogical data. Five years later, the GENTECH Data Modeling Project, proposed a new genealogical logic data model to support research in genealogy and to allow data inter-exchange between genealogy software. Despite being initial reference models, they still have some limitations to adapt to different cultural and social environments. Additionally, sharing genealogical data between systems is difficult since, even though they are syntactical reference models, they may have semantic mismatches. Today, we have not a common and unified proposal as a standard recognized genealogical model. In this way, in this paper we propose to consider the ontology paradigm to extend expressiveness of concepts and relationships in such standards.

1 INTRODUCTION

Genealogy, is a discipline of social sciences and history that study family composition, origin and evolution. In recent years we have seen an increasing popularity of genealogical services and specific software to build familiar pedigree. However, the absence of an accepted reference model as a universal standard to represent genealogical information makes it difficult to share and reuse such data between people.

Designed as local databases or websites, most of these applications offer importing and exporting functionalities using file formats widely accepted, like the GEDCOM[1] specification. However, given the extendability of this specification, some applications add proprietary extensions to GEDCOM to consider these cultural and historical diversity, not always recognized by others.

Another difficulty appears trying to identify equivalent, complementary or inconsistency records that refer to the same ancestor. We know that personal names and places can change over time, even if at only syntactical level. Applications based on the relational model lack of appropriate mechanisms to recognize variations of nominal attributes, making it difficult to merge equivalent instances. This situation is aggravated when the database is scarce or incomplete.

The integration between genealogical information systems could be achieved by means of schema mappings among their databases and a common reference model. Unfortunately, this model does not still exist today. So, in this paper we present a preliminary proposal to address such unmet need.

The paper is structured as follows. Section 2 gives a short overview of the most well-known genealogical reference models. In Section 3 we introduce our reference model proposal, and Section 4 describes its main features and contributions. In Section 5 we identify open problems and main challenges.

2 GENEALOGICAL DATA STANDARDS: STATE OF ART

The most emblematic compilation project of genealogical data on a large scale and worldwide has been and continues to be carried out by

[1] GEDCOM, http://homepages.rootsweb.ancestry.com/%7Epmcbride/gedcom/

Campanyà Artés J., Conesa Caralt J. and Mayol E..
Modeling Genealogical Domain - An Open Problem.
DOI: 10.5220/0004135002020207
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 202-207
ISBN: 978-989-8565-30-3

FamilySearch[2] organization. We have identified three groups of proposals, those based on GEDCOM, those on GENTECH and those following an Ontological approach.

The first version of GEDCOM appeared on 1984. GEDCOM was just a paper specification designed to allow genealogy programs to exchange genealogical data. Therefore, it is not a data model nor a genealogical application, but it is a format to support data interoperability. GEDCOM is basically oriented to a lineage-linked data model based on families and individuals. This contrasts with evidence models, where data is structured to reflect the discovered and supporting evidences. GEDCOM files are plain text similar to of markup languages. However, it has the following disadvantages: **difficult evolution** (due to its proprietary format), **family-centered** (as an individual is identified by the family it belongs, and not by the identity of their parents), **ambiguity** (since the current specification does not set limits on its hierarchical structure and its not clearly defined in which levels or identity tags where to put some data), **lack of source references** (even though sources may be informed, there is not an specific tracking method for data connected to the research process, and it is difficult to make source verification or to reuse such sources easily) and **inconsistencies** (due to data duplication).

Some extensions to the GEDCOM format had been proposed. On 2010, the "*Build a Better GEDCOM Project"*[3] was initiated to develop an international standard to store and transfer genealogical data. An adhoc committee of *BetterGEDCOM* wiki members established in 2012 the *Family History Information Standards Organization* (*FHISO*)[4]. The objective is to solve interoperability issues independently of technology platforms, genealogy products or services. At the same year, FamilySearch organization outlined a major new project called *GEDCOM X*. The proposal was a new format based on a XML language. It defines new data formats to permit traceability of sources and genealogical records. It also offers support for sharing and linking data online.

In 1995, a different genealogical model was proposed: the *GENTECH Data Modeling Project* (Mitchell 2003) by the Federation of Genealogical Societies (*FGS*, USA). Its main objective was to define a genealogical data model to model the genealogy research process and to facilitate data exchange among genealogists. Although it was just a conceptual model, it became a reference for many other implementations. This model was not so worldwide accepted. Genealogical data is defined using certain structured collections, roles and attributes. Such attributes are hardly typed, introducing some strictness and limiting its adaptation to different contexts. Moreover, the model assumes its implementation on relational databases, an unnecessary limitation for other information technologies or database models.

More recently, a new approach was initiated by (Zandhuis 2005). He proposes to use ontologies in the genealogical data treatment to take advantage of the Semantic Web in the data distribution and knowledge extraction. Genealogical data was modeled with OWL/RDF, but did not develop much beyond that the class structure and properties necessary to complete a genealogical model.

In the same direction, (Campbell 2006) proposes the creation of an open network data, scalable, extensible, based on open standards and understandable by machines. In essence it was a network of servers updated and maintained by local genealogical organizations. In order to enable automated semantic interpretation, genealogical data is fragmented in the form of subject-predicate-object sentences, in OWL-RDF files. Interconnections between nodes were fixed by equivalence relations between entities, task in which collaborate intelligent agents.

Another interesting contribution is (Woodbury 2010). Its purpose is to show the feasibility of an information system, based on individual and event information, to automatically load of unstructured genealogical data and to infer new hidden knowledge. Textual data is analyzed using ontological patterns and regular expressions. Data is stored using OWL/RDF files. This proposal use a set of tags that define SWRL rules and integrity constraints. However, this model reflects only a fraction of the complexity of the domain.

3 OUR PROPOSAL

The main difficulties when sharing data between genealogical applications can be grouped into four types. (1) Syntactic variants: the frequent existence of syntactic variations of names of individuals and locations make it difficult their proper recognition and management. (2) Structural heterogeneity: the social structure and evolution of the family, roles of its individuals depend on temporal and cultural

[2] FamilySearch, http://www.familysearch.org/

[3] BetterGEDCOM wiki, http://bettergedcom.wikispaces.com/

[4] FHISO, http://fhiso.org/

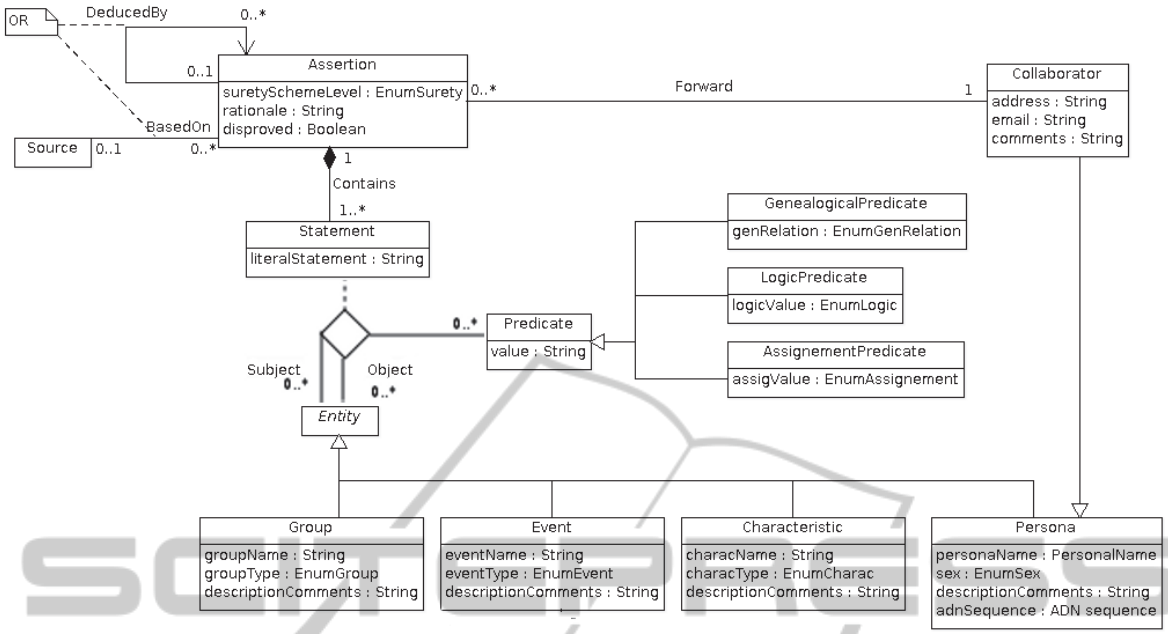

Figure 1: Modeling genealogical assertions.

contexts, so they must he interpreted correctly. (3) Semantic heterogeneity: merging and integrating genealogical data from different sources require take into account different concept interpretation and their inference rules. (4) Data quality: genealogical data may contain transcription errors, incompleteness and/or not be proved. All this characteristics are inherent to the genealogy domain and they difficult genealogical data management and sharing.

Ontology paradigm may be appropriate to solve some of these problems, since it handles semantic concepts rather than syntactic keywords in information retrieval systems. Therefore, genealogy data repositories content may be described regardless of their syntactic representation, focusing on its semantic integration.

Our proposal consists of two independent models: the *Projects* and the *Assertions* models. The first one focus on the genealogical research process. It describes projects, goals, tasks, collaborators, resources and it keep track of the document sources of genealogical information. The second one, the *Assertions* model, contains the proper genealogical information extracted or inferred from these sources. Due to space limitations, this paper only deals with the second one.

The concepts of the assertion model (Figure 1) refer to people, places, dates, events, characteristics and groups. So its design should facilitate addressing the aforesaid four groups of problems. The core is the Assertion class. Assertions may be deduced from other assertions, or may be provided by a collaborator and linked to one source. In both cases, they have a annotation of genealogical interest, and refer to one or more statements. These statements refer to entities as people, relationships among them, events, groups, personal characteristics, etc., but in a implicit way. In order to enable automation with computers, we need to make explicit this knowledge, as discussed in the next section.

Statement class records concepts and their relationships as atomic triples, in the form of <subject, predicate, object>, which is similar to the structure formats used in the semantic web: basically RDF, *RDF Schema* and OWL. In our model *Statement* is a associative class of the ternary relationship between two instances of *Entity* class (subject and object), and a *Predicate* instance. An *Event* usually occurs at known times and places. The *Place* class do not specialize to specific categories to permit the maximum adaptability to different cultural and geographical contexts (Figure 2).

4 GOING FORWARD: OUR CONTRIBUTIONS

Most part of commercial genealogical information systems and applications implement data persistence upon relational databases, following the *closed*

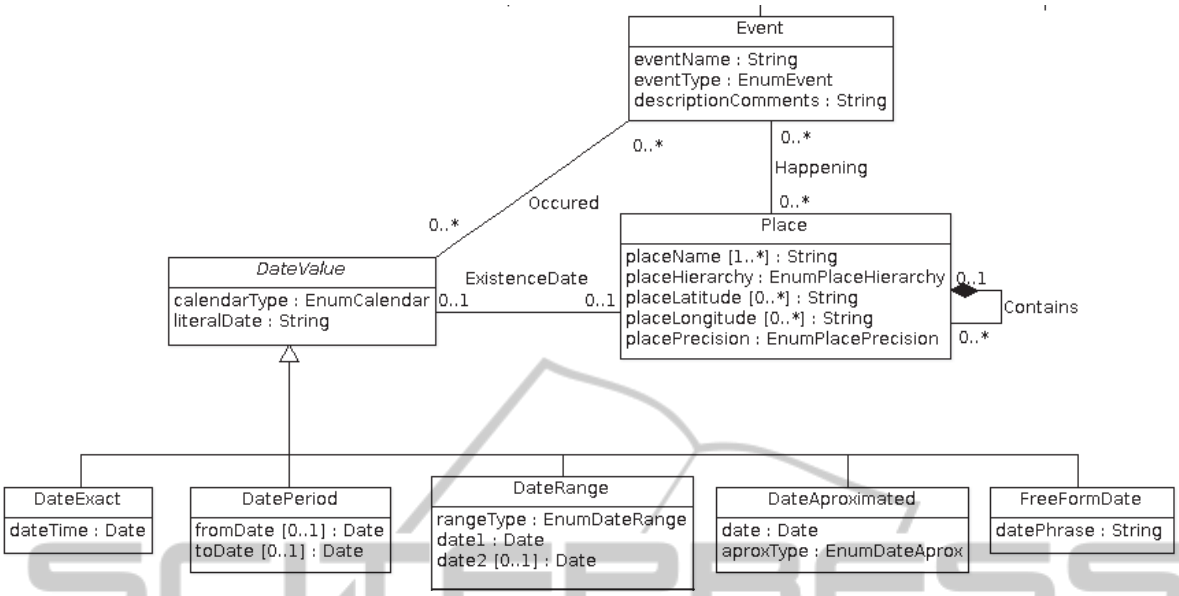

Figure 2: Relationships between Event, Place and Date.

world assumption (CWA). In this sense, only facts or assertions stored in the database are true, and therefore, any other fact not stored as a database tuple is false or non-existent. This assumption limits exploration of possible related data when information is incomplete or imprecise, which is quite common in genealogy. A similar situation may appears when integrating repositories that have data in common but without information that allow to filter duplicates. Our conceptual model releases entities and their attributes from restrictive types, facilitating integration to diverse contexts and casuistry. We are interested in the semantic value of attributes and roles, not in the explicit record syntax or types. Adopting the ontological paradigm, we can transform from implicit to explicit semantic knowledge, in a way to reaching a *open world assumption (OWA)*.

To permit interaction between different systems requires explicit semantic interoperability. To obtain a context independent model we need to address two goals. (1) The adoption of a specific vocabulary, identifiable by its IRI (International Resource Identifier) and namespace, as enumeration types. Importantly, these types could be particularized on different implementations of the model. (2) The use of ontologies (general or specific) that cover this vocabularies, to enable a constructive information merge process between information systems.

To differentiate the *Assertions* conceptual model of the ontology from where the facts come, we preferred rename the latter as *Facts* ontology. A second ontology, *PersonaEvents,* is automatically populated using the information of the *Facts* ontology that deals with a given person. This describes the specific events and relationships (*property,* on a ontology terms) of *Persona*, which is particularly of interest in genealogy. To illustrate this, we can materialize the implicit *Statement* relationships in *properties* between instances of *Person* and *Entity*, like *hasParent*, *hasWife*, *wasBorn*, *professedReligion*, etc.

So between these two ontologies, *PersonaEvents* and *Facts*, there is redundant information. This implies that, to avoid inconsistencies, all changes made in any of them should be reflected in the other. Another remarkable aspect is that instances are referenced by its IRI, avoiding the disadvantages arising from different contextual interpretation. However this has a cost, losing in *PersonaEvents* ontology any reference to information sources, and that's why it's necessary to maintain their link with statements in the *Facts* ontology.

5 NEW CHALLENGES

One of the reasons why we have chosen ontologies to represent knowledge is because the amount of data we have may evolve constantly and because some inferences can be done even when not all the data is known. The open world assumption allows us to add new knowledge incrementally and

dynamically. The only condition is that the new information cannot contradict the information of the existing knowledge base, assuring that all inferences made previously are still valid.

The ability to share information is also our objective. The current situation is characterized by an increasing number of private applications and a lack of open and recognized standards. In addition, there are an increasing number of semantic web services that provide access to data repositories. It would be desirable to agree on some specifications that provide unambiguous descriptions of their services and their mappings in a common ontology domain.

A second line of research is to consider issues related to database distribution. In this context, instances identification is a major challenge, as it is to discover duplicates (when the same instance appear in two places) or combining multiple overlapping data that refers to the same instance. To deduce equivalence between genealogical instances we must consider not only lexical coincidence or proximity of key attributes (name, date and place of birth or death) but also known kinship with others, as portions of their family tree (parents, siblings, spouse,…). Furthermore, record linkage still remains a complex problem. Different methods for automation of data linkage and for reducing manual processes have been proposed, most based on techniques from artificial intelligence. Research, despite being limited to particular environments, are promising and satisfactory enough in the validation tests performed. Neural networks (Pixton 2006), bayesian probability models (Larsen, 2005) and metric-based machine learning algorithms (Ivie, 2007) can provide the tools we need to simplify the task.

The third challenge should allow us to build the knowledge base from basic statements. As we have seen in Section 4, the base of our model lies in elementary semantic units inspired by the first-order logic, the triples <subject, predicate, object>. These triples formalize the essence of what is known and what can be said. Unfortunately, using such elemental assertions to express knowledge make undecidable the processes that would allow to infer new knowledge. However, the computational complexity problems that involve the use of first-order logic are well known. With our two related ontologies, *Facts* and *PersonaEvents*, this drawback can be fixed, as the inferences of interest would be over the second, obtained as a reduction from *Facts*. However, with this operation we can reduce to one direct Person-Entity relationship which originally may have required several statements.

To complete the challenges, we must mention problems about decidability and computational complexity. Regarding our proposal, we have chosen to reconcile description logics (DLs), which form the basis for OWL, and rule languages, while maintaining decidability:

- Using *Semantic Web Rule Language* (SWRL) rules (Horrocks 2004), but by taking certain precautions, such as restricting its applicability to certain subset of data. These rules, known as DL-safe as combination with OWL-DL, leads to decidable systems and, more importantly, computable in polynomial time. We will make reference to some published studies that propose specific solutions (Hirankitti 2011, Mei 2005, Motik 2004).

- The latest *OWL 2 Web Ontology Language Recommendation*, informally OWL 2 (Motik 2009), expands the options for integrating certain kind of rules in OWL, thereby maintaining decidability. SROIQ rules can provide interesting features.

6 CONCLUSIONS

For many years, genealogical data used by the vast majority of computer applications has been shared using the data transfer format created by GEDCOM. The problem arises when we want to integrate the information collected by different users. Despite the availability of data exchange formats widely accepted, recognition of family ties between those resources are difficult and requires some expert assistance.

In this paper we proposed a genealogical model that aims to be flexible enough to adapt to social, cultural, geographical or temporal variability. The ontological paradigm and its deployment on last years, offers a variety of experiences and practical tools competent to represent semantic information of concepts relevant to the genealogical model. These ontological tools, together with the proposed semantic definitions, can provide solutions about real problems that appear when integrating different resources, such as data inconsistencies or recognition of equivalences.

Finally, the automatic processing of information is possible only after transforming implicit knowledge from source statements to explicit semantic concepts In this way, ontologies,

OWL axioms and SWRL rules provide powerful languages understandable by computers. Future work must be done in order to achieve a reliable data processing with minimal need from expert supervision.

REFERENCES

Campbell, Hilton, 2006. Enabling the Distributed Family Tree, in Department of Computer Science - *Brigham Young University.*

Family History Department, The Church of Jesus Christ of Latter-day Saints, 2000. *GEDCOM XML Specification - Release* 6.0 (Draft)

Hirankitti, Visit and Xuan, Trang Mai, 2011. A Meta-reasoning Approach for Reasoning with SWRL Ontologies, in *Proceedings of the International MultiConference of Engineers and Computer Scientists* 2011 Vol I, IMECS 2011, March 16-18, 2011, Hong Kong

Horrocks, Ian et al., 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, in W3C Member Submission

Ivie, S. et al., 2007. A Metric-Based Machine Learning Approach to Genealogical Record Linkage, Department of Computer Science (Brigham Young University)

Larsen, Michael, 2005. Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory, *Department of Statistics* (Iowa State Univeristy)

Mei, Jing (Peking University) and Paslaru Bontas, Elena (Freie Universität Berlin), 2005. Reasoning Paradigms for SWRL-enabled Ontologies

Mitchell, Stanley, 2003. Gentech-GDM Reference Model (http://freepages.history.rootsweb.com/~mitchellsharp/gdmref/gdmref-01.pdf)

Motik, Boris et al., 2004. Query Answering for OWL-DL with Rules

Motik, Boris et al., 2009. OWL 2 Web Ontology Language – Profiles. *W3C Recommendation* 27 October 2009 (http://www.w3.org/TR/2009/REC-owl2-profiles/)

Pixton, Burdette, 2006. Improving record linkage through pedigrees, *Department of Physical and Mathematical Sciences* (Brigham Young University)

Woodbury, Charla, 2010. Automatic extraction from and reasoning about genealogical records. *Brigham Young University* (BYU, Utah – EEUU).

Zandhuis, Ivo, 2005. Towards a Genealogical Ontology for the Semantic Web, in *Association for History and Computing Conference*, (Amsterdam, 14-17 September 2005)