Projecte de recerca bàsica o aplicada
PAC3 — Tercera Prova d'avaluació continuada

Cognoms: *CAMPANYÀ ARTÉS*
Nom: *JOAN*

- Per a dubtes i aclariments sobre l'enunciat, adreceu-vos al consultor responsable de la vostra aula.
- Cal lliurar la solució en un fitxer OpenOffice.org, PDF o RTF fent servir una de les plantilles lliurades conjuntament amb aquest enunciat. Adjunteu el fitxer a un missatge adreçat a **l'espai d'avaluació de l'aula virtual**.
- El fitxer ha de tenir l'extensió *.odt* (OpenOffice.org), *.pdf* (PDF) o *.rtf* (RTF), segons el format en què feu el lliurament.
- La data límit de lliurament és el **22 de juny** (a les 24 hores).

Respostes

Per aquest lliurament final s'ha optat pel format d'article en llengua anglesa. Es considera que el model conceptual del domini genealògic desenvolupat és prou innovador i té prou qualitat com per poder pensar en la possibilitat de donar-lo a conèixer públicament.

Tanmateix, i després d'haver-ho consultat amb els co-directors d'aquest treball, en Jordi Conesa (professor de la UOC) i l'Enric Mayol (del Departament d'Enginyeria de Serveis i Sistemes d'Informació, ESSI, de la UPC), s'ha postposat la proposta de publicació per així poder contrastar opinions d'altres experts de l'àmbit tractat. Així, tanmateix, es podrien millorar alguns punts del redactat que poden resultar un tant confusos.

A genealogical conceptual model using ontologies

Abstract

Roughly fifteen years ago, the Church of Jesus Christ of Latter-day Saints published a new proposed standard file format. They call it GEDCOM. It was designed to allow different genealogy programs to exchange data.
Five years later, in may 2000, appeared the GENTECH Data Modeling Project, with the support of the Federation of Genealogical Societies (FGS) and other American genealogical societies. They attempted to define a genealogical logic data model to facilitate data exchange between different genealogical programs. Although

genealogists deal with an enormous variety of data sources, one of the central concepts of this data model was that all genealogical data could be broken down into a series of short, formal genealogical statements. It was something more versatile than only export/import data records on a predefined fields. This project was finally absorbed in 2004 by the National Genealogical Society (NGS).

Despite being a genealogical reference in many applications, these models have serious drawbacks to adapt to different cultural and social environments. At the present time we have no formal proposal for a recognized standard to represent the family domain.

Here we propose an alternative conceptual model, largely inherited from aforementioned models. The design is intended to overcome their limitations. However, its major innovation lies in applying the ontological paradigm when modeling statements and entities.

1. Introduction

Science pedigree does attract a growing number of citizens, mainly in developed countries. In a way, this is facilitated by the dissemination and popularization of computer applications for processing and data storage. The absence of an accepted conceptual representation model as a universal standard (we must clarify, as we shall see, that we will not consider GEDCOM as a conceptual model, but as a file format for data exchange) makes it difficult to share and reuse data between independent research groups. Some models have been postulated previously to facilitate sharing data between applications, but the truth is that their own limitations prevent them from adapting to the diversity of social, cultural and historical. As we shall see in subsequent paragraphs, at the present time there is not available a genealogical conceptual model adaptable to a variety of contexts and accepted by the community.

Thus, there is an outstanding need and therefore a promising research field. A prerequisite for any proposed standard is to provide great flexibility to adapt to social, cultural, geographical or over time, playing for it with the semantic meaning of the terminology used in each case as defined in the ontological associated. With this paper we propose a genealogical model that aims to be comprehensive and independent of its subsequent implementation.

1.1. The genealogical domain

Genealogical science aims at obtaining and classifying information associated with people who existed or still exist and their interrelationships between their family and others groups.

The most interesting subject is to discover the genetic ties between individuals, the named family tree. The key relationships are marriage and their direct descendants. But we find more complex relationships, even with this base simple set of relationships: multiple marriages, adultery, half-siblings, adoption, etc. Likewise, there are a undetermined number of personal data of individuals that could be important to register, depending of cultural or historical context: religious or civil events, social status, properties, education, etc.

Moreover, we can have incoherent data from different sources, such as different dates for certain events, conflicting names for the same person, insufficient information or just vague approximations, etc. This is a logical consequence when acquiring data from different documentation sources, ancient and in many cases carried out with objectives other than those pursued in further investigation.

The most emblematic compilation project of genealogical data on a large scale has been and continues to be carried out by *The Church of Jesus Christ of Latter-day Saints* with their *FamilySearch* project. There was a time when it seemed that no European documentary file, civil or ecclesiastical, could escape being microfilmed and scrupulously analyzed by their researchers. Inevitably there was a reaction and put limits on access to those files. Global powers cause distrust, and getting information that in some cases can be considered private or confidential, could result in violation of privacy rights.

By contrast, models designed to represent family genealogies have been widely accepted, especially as a result of greater interest to know the profile of their own ancestors by sectors of the population with certain cultural level. This is a global phenomenon, perhaps more pronounced in Old Europe, including countries of former Eastern bloc. Thus emerged a large number of applications, some proprietary software and other free. Some of them are web applications, so that with one user ID can access the database that stores the data set, which thus can be shared among multiple users.

Designed as local databases, most of these applications offer options for import-export data using file formats widely accepted, like the GEDCOM standard [1] discussed in Section 2. But as we will see, this can't guarantee the integrity of the information. Thus, the main drawback of this approach is that disjoint genealogical trees are constructed without the possibility of sharing data in cases where there really common nodes.

1.2. Modeling genealogy is a complex problem

Most actual tools allow import and export data from/to xml files, which facilitates compatibility with other systems, particularly with GEDCOM. Even someones have extended their options to offer classes and attributes in more expressive semantics. However, in all cases are applications aimed at amateur genealogist without any pretensions that provide an intuitive and with few demands on the operating system software.

The main problem we face when trying to reuse a data repository based on different models is the difficulty of avoiding the appearance of inconsistencies in the case that there were differences between records, even if at only syntactic level. And we can confuse a different semantic meaning that can be assigned to terms syntactically equivalent. This problem is often enlarged when the pedigree information is incomplete. On the other hand, does not seem realistic trying to impose universal models, following the example of GEDCOM, suitable for wide range of contexts in which they should adapt.

This has been traditionally solved by making an equivalence mapping between two data structures. But what happens when new applications emerge and aim to model the same entities? We find that the number of equivalences is multiplied exponentially. We see that the mapping is difficult to generalize when you need to share concepts and categories of data between several systems. Obviously, a solution would be to agree on a common reference model and establish respective equivalent relations over him. Unfortunately, this model doesn't exist today. In this paper we present a proposal for this unmet need.

1.3. Using ontological paradigm is an option

We can achieve greater expressiveness in a model by introducing greater number of structured components, but the price is sacrificing flexibility and adaptability in different contexts. It would be the case of a tool providing entities on a wide range of relationships of kinship, typology of events, personal characteristics of individuals, etc. This type of detailed structure obviously can fully define the semantic meaning of each entity while facilitating the extraction of information, but it would be difficult to reuse.

Moreover, the sleuthing skills over knowledge inference that currently offer intelligent applications are far from being able to rely exclusively on ontology on the task of integrating genealogies. Perhaps in the future can have tools (software and hardware) capable of processing raw data without giving any kind of structural information, entering documentary and automatically obtaining instances and their corresponding properties, with reference to a global ontology that includes all types of knowledge (names, places, dates, people, events, relationships, etc.). But this is still far from the current possibilities.

An intermediate option is to provide a basic structured information, following a commonly accepted model. The classes defined for this model should be flexible enough to represent different cultural situations and contexts. This relaxation would result in having, for example, a Event class with instances capable of reflect a baptism en certain cases or a transaction of property in others. Furthermore, should allow that this baptism could have different implications depending of cultures or historical contexts. To reconcile determinism with the necessary flexibility of concepts we will use specific ontologies associated with the entities of the model, containing the necessary equivalent relations. In section 9 we provide some links to ontologies that can serve as a reference.

Defined syntactic rules and semantic restrictions could obtain accurate automated data analysis and capability to deduce new knowledge. We must also consider that with the ontology we can establish rules that facilitate integrity check data from different sources. This action would allow us to discard data in some cases inaccurate (rule out a birth date declared within a range from another source if we have exact date) or of doubtful reliability (a direct source prevails over other cross-referenced with which contradicts). Ultimately, even in circumstances where it may remain about the veracity between data contradictory or incompatible, we can choose to postpone the decision and keep them all in the system, warning the user of these circumstances at the time of generating extraction of information.

The heart of the ontological paradigm is to identify entities and concepts from its URI and use thesauri and ontology to realize the relationship. This also gives a great flexibility to adapt in different contexts and local variations, which lies in the possibility of refining dynamic properties, equivalents and restrictions of the concepts described by the ontology used.

1.4. About this paper

The paper is structured as follows. Section 2 gives a short overview about the two major genealogical reference models, GEDCOM and GENTECH [2]. In our project we adapted some of their design solutions and components.
Section 3 introduces the great opportunities of ontological paradigm applied to the genealogies. First by linking the different predefined enumerated types using their URI,

and secondly harnessing the ability to integrate concepts in order to automate the generation of information.
Section 4 presents the model developed from its core components, and Section 5 describes the main features of the enumerated types listed in the model. In Section 6 you find some examples of data structure and genealogical instances using the proposed model.

2. Previous genealogical models and limitations

2.1. GEDCOM

GEDCOM, an acronym for **GE**nealogical **D**ata **COM**munication, is a de facto specification for exchanging genealogical data between different genealogy software. It meets the specific needs of *The Church of Jesus Christ of Latter-day Saints*. Then, it's not a data model or genealogical application: its purpose is to allow the exchange of data between genealogical programs without having to manually re-enter all the data.

GEDCOM files are plain text containing genealogical information about individuals, and meta data linking these records together. Its architecture is thus subject to these premises and the technological context of when it was designed, in the early 90s. Data is structured so that the elements in its structure can contain other and so on. It utilizes numbers to indicate the hierarchy and tags to indicate individual pieces of information within the file. This hierarchical structure is reminiscent of markup languages, even if earlier. Its specifications have been and continue to refer to many other models that have emerged later. However, under the current prism, it has several disadvantages:

- Being a proprietary format is not easy to propose improvements to its evolution, which otherwise do not exist
- Family-centered. An individual is identified by the family it belonged to, and not by the identity of their parents
- The current 5.5 specification don't set limits on their hierarchical structure and its not clearly defined where to put data, so that we meet with potential incompatibilities between different implementations of the standard (particularly with developer-supplied tag extensions)
- There is no support for data connected to the research process
- Inconsistencies may occur due to data duplication. For instance, it allows children to be related belonging to households, while for each person can set the parental family to which he belonged. Both reports should be consistent.
- Insufficient flexibility and little descriptive adaptability to adapt different cultures structure in some relevant data, such as names of persons, the identification of geographical locations or other additional characteristics.

2.2. GENTECH

This project, as free space of cooperation between researchers, was short-lived: the first specification appeared in May 2000 and early 2004, the development group was disbanded and the project ended up being absorbed by the U.S. National Genealogical Society.

Although it was just a conceptual model without any pretensions in the development of applications, it finished becoming a reference for many other implementations. On the one hand, providing solutions to complex problems, such as the identification of sites

regardless of the historical moment and place names associated as well as in structuring the assertions as relations between entities. On the other hand, introducing the ability to represent all kinds of useful information related to individuals, linking them with the *Characteristic* class, without forgetting the general conceptualization of the *Group* class, which naturally allows for family ties, ethnic or social group , or any property that share certain individuals and want to formalize. Also, the model offers flexibility to represent the sources of information (*Source* class), making them capable to adapt with any kind of documents, records, files of all types and for different cultural contexts.

Without questioning the interesting contributions of this model, the truth is that it also has disadvantages to their widespread acceptance as a standard. The key may be to maintain an overly layered structure types and entities with predefined roles. This introduces strictness to the model and, consequently, became difficult to adapt to different contexts.

Moreover, at the time the model was raised thinking in its implementation on relational databases. In the past ten years we have seen the emergence of successful ontology-based technologies that allow extract knowledge from stored data related to their semantic content. The search engines are a good sign.

3. The ontological paradigm

Aware of the limitations of current genealogical applications, an alternative model should consider other assumptions. And more important, it should enable that their different implementations particularize some characteristics (for the development of software, linguistic or applied in different cultural contexts, etc.) are not an impediment to effective and automated information integration. It should be borne in mind that this should be achieved without the need to establish equivalence or mapping between its instances. For this purpose, the first steep to define and associate entities with a set of ontologies that identifies concepts semantically related.

3.1. Data and properties representation requirements

Ideally, a knowledge base built under the ontological paradigm and shared by different actors credited, would be able to reference the different entities as resources uniquely identified by its URI. Also, to avoid getting into local and cultural particularities, the model should be flexible and extensible enough as the characteristics of the data may vary according to cultural or temporal contexts (names, dates, places, etc. can adopt different format or values). Moreover, we must foresee its natural evolution in the future.

Thus, our challenge is to build knowledge creating relationships across vocabulary from different applications. To obtain a context independent model we have two tools:

- The adoption of a specific vocabulary, identifiable by its URI and namespace, that here takes the form of different enumeration types (see Section 5). Importantly, these types could be particularized on different implementations of the model.
- The use of ontology (general or specific) that cover the previous vocabularies, so as to enable a constructive merge information between information systems.

3.2. Integrating and building new knowledge

Structuring genealogical information using a ontological model, apart from the advantages of being able to refer to resources identifiable (by URI), opens the door to future automated data processing by computers. Herein lies the enormous potential of a model based on this paradigm. It should be stressed that we are not talking about theoretical approaches: the semantic web is a reality in many applications. The same technology that are currently using Internet search engines could be used to build an information system able to respond to the needs arising from cultural, space or over time.

Consider for a moment in a major project, as would the development of a complete genealogy of a country of old Europe. For the case we would be thinking of Catalonia, a state of Spain. The stages of the process could be summarized in:

- First, we would like to feed our information system with data from existing genealogical databases. And we expect to do this with a automated process. Probably the sources programs can export data to GEDCOM file, which would facilitate the task. At this point we we are creating new instances of our model from the data transferred.
- Then we should integrate these different data sources into a single body of information. We must compare data from different submitters and with possible common points of connection. In some cases the equivalences and relationships between instances can be derived directly, but in other cases we can find dates that may appear inconsistent, such as different birth dates for the same person. In many cases the use of ontological model can automatically generate an appropriate assertion, for example by introducing uncertainty with the conjunction "OR". In other circumstances will require further monitoring.
- Finally, at this point that have a large integrated database, we might consider the deduction of new knowledge, as would the generation of a pedigree up or down or to establish the possible relations between two identities.

However, it is the addition of the reasoner that makes the ontology useful. For example, if we know that John is the father of Mary, we would expect a “yes” if we query whether John is the parent of Mary. The parent relationship is not asserted, but we know from our ontology that *fatherOf* is a sub-property of *parentOf*. If “John *fatherOf* Mary” is true, then “John *parentOf* Mary” is also true.

4. An ontological assertion oriented conceptual model

Here we present the model developed from its core components. We can analyze it as two partially independent sub-models, projects and assertions. The first one enables us to organize the tasks, recognize assigned employees and keep track of the sources of genealogical information. But the knowledge base is in the assertions that we take from the documentary sources. That is why the title of this section is “An ontological assertion oriented conceptual model”.

Assertions are inferred from the records kept on all types of documents and various sources. The concepts will usually refer to people, places, dates, events, characteristics or groups. Additionally, for these entities, the model should allow to reflect the changes introduced by direct factors such as space and time, with implications for the evolution of language and naming.

We must clarify that our research has focused on designing a conceptual model, the entities and relationships that reflect the genealogical domain. The effective implementation of this model presupposes the use of a wide range of ontologies, as

raised in Section 3, for whose development will be a specific project. However, we must clarify that we don't have a single ontology domain associated with, but a comprehensive and extensible set of specific ones reporting concepts and properties to the entities we have in the model. It's this flexibility that allows easy adaptation of the model to new contexts and enhance the capacity of semantic processing in their implementations.

4.1. Modeling projects and partners

Projects can satisfy different needs, some will be internal to the organization (defined search strategies, data processing and information structure, classification of archives and repositories, etc.) and others perhaps research on their own documentary. Moreover, as a dynamic entity, projects progress trough different phases, identified by our attribute *status*. To better monitoring and planning, the ongoing projects have relationships with interval dates (start and end), planned and actual.

When a new project is created, the first step is to assign collaborators who will be responsible for different tasks. A collaborator may be assigned to one or more roles, each of which may be restricted to a limited time period in the context of the project. And every project bases on specific objectives, each of which may require carrying out different activities. These activities can be either: research, data source extraction, integration of databases, etc. Thus, generic *Activity* class can specialize, if desired, in many other ways as desired (research, organizational, computerization, microfilm activity, etc.).

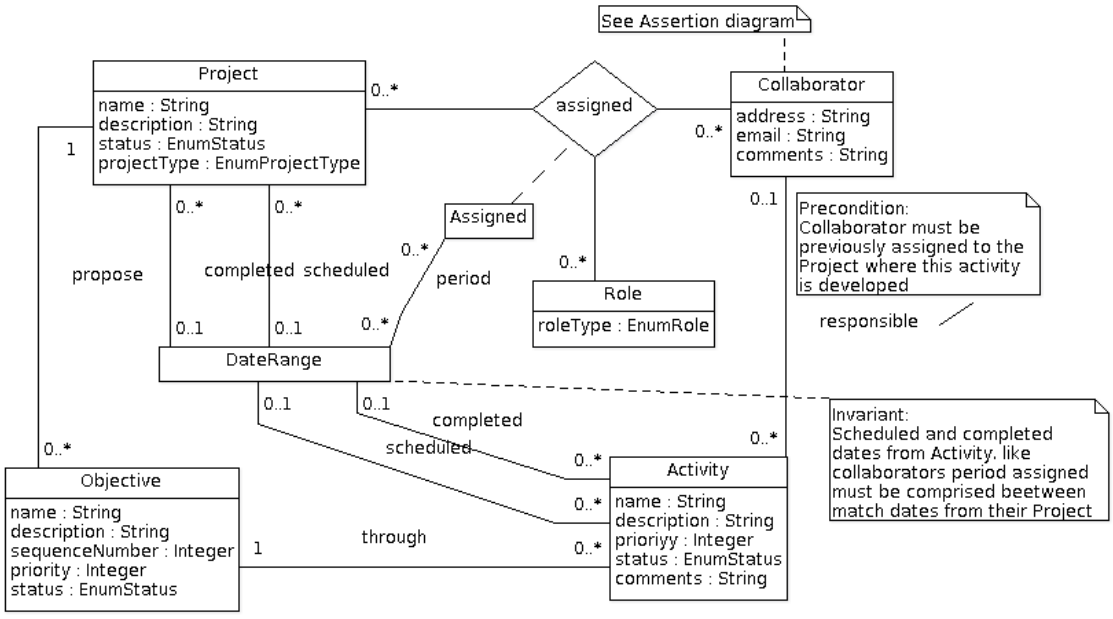


Figure 1: Project structure and collaborators

4.2. Modeling *Source* and *Repository*

Repository instances contains information about the place where data is found, typically library, public or private archives, etc. Inside them we can find several *Source,* collection of data useful for genealogical research such as a historical documents, official register, books, electronic database, etc.

Search activities will target repositories (for example when you want to make an inventory of its contents) or to specific information sources, perhaps available in one or more repositories. The different and complex hierarchical structure of sources, the plural cultural contexts, sorting methods, etc. make it virtually impossible to foresee all possible cases. Thus, it was decided to design a single class that allows recursively modeling the fact that certain resources can in turn contain other ones of a different nature.

As the GENTECH model, we offer the source linking with a sphere of jurisdiction (a relationship with a *Place* instance which identifies the place of the jurisdiction of the *Source*) and with a subject *Place* instance of the facts.

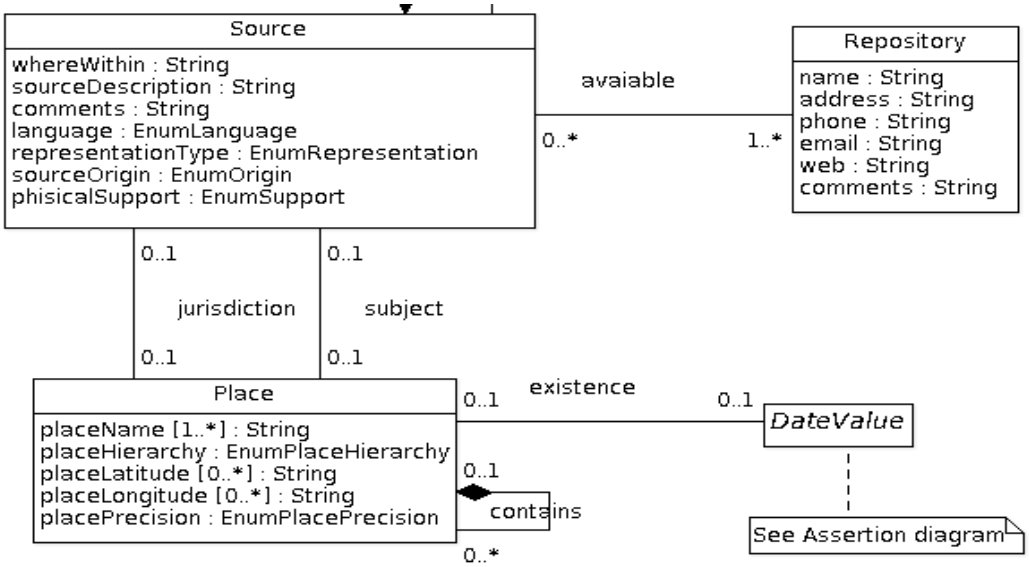


Figure 2: Repositories and sources

4.3. Modeling *Assertion, Statement* and *Collaborator*

Assertion instances contents a single annotation of genealogical interest taken from a particular source (usually the lowest level within the recursive sources hierarchy), or it can be deduced from one or more predefined assertions.

A *Assertion* will consist of one or more statement. Although genealogists deal with an enormous variety of data sources, one of the central concepts of this data model is that all genealogical data can be broken down into a series of short, formal genealogical statements. These statements may refer to people, relationships among them, events, groups and collectives, personal characteristics, etc.

Statement class records concepts and their relationships as atomic triples, in the form of <subject, predicate, object>. Can be seen that this structure is similar than the proposed formats for the description of knowledge in the semantic web: the Resource Description Framework (RDF) / RDF Schema and the Web Ontology Language (OWL) [7]. This is not accidental and reflects our interest in facilitating the subsequent implementation in these languages: RDF was created to be based on semantics, so it should be easier to make inferences, with the help of an RDF processor.

In RDF triples, the subject corresponds to some class in the ontology while the object can either correspond to a class or to a basic literal data type. In the first case the

predicate represents a relation between classes in the ontology and is called an object property, while in the second case it represents an attribute of the class corresponding to the subject in the triple and is called a datatype property. In our model we will not use this second possibility, using instead the attributes required by the adequate subclass of *Entity*.

Also in *Statement* we make certain concessions on the RDF standard: according it, the nature of *Predicate*, which should correspond to an RDF *Resource*. In our model, keeping this option, we left open the possibility of using a literal when not needed more precision.

Collaborator in information search tasks are the responsible for submit *Assertion* from the sources analyzed. It is also possible in some cases, to propose *Assertion* that arise as a deduction from a subset of other *Assertion*. In other cases, a collaborator can be a computer engineer, a project coordinator, etc.

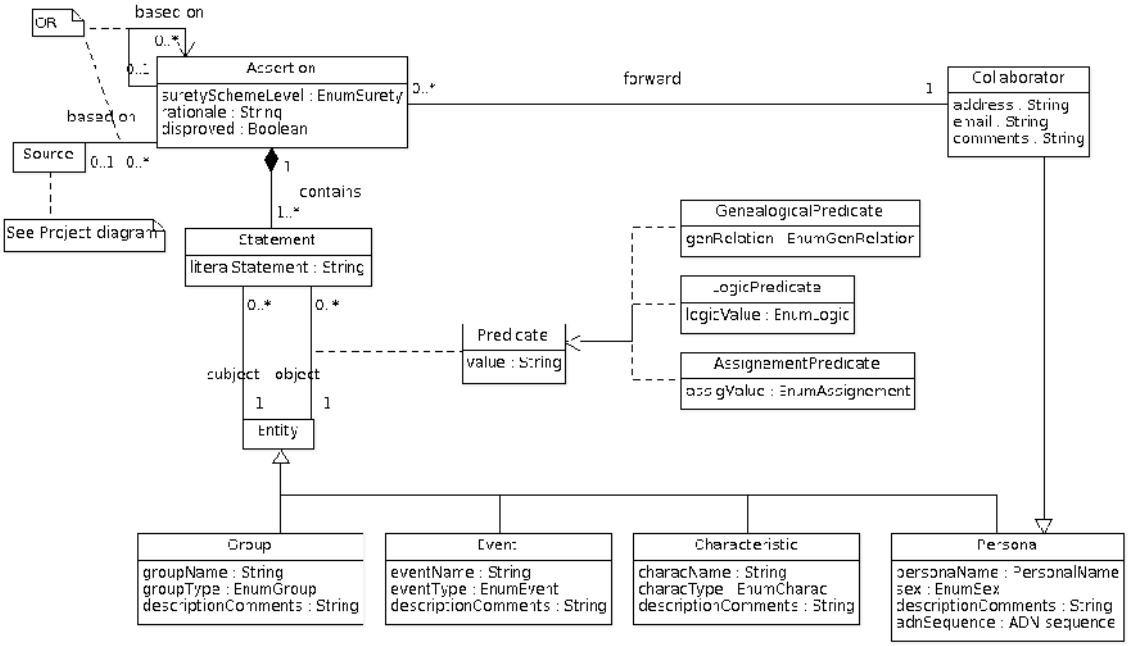


Figure 3: Modeling genealogical assertions

4.4. Modeling *Resource*

One of the most important contributions of the ontological model is the option to identify the concepts and entities for their URI, universal resource identifier. If we add the ability to relate concepts that provide an appropriate combination of ontology (genealogical, general, geographical ,...) we can begin to think about developing systems capable of identify semantically concepts.

In the proposed model, all classes describing concepts and real world entities are specializations of the *Resource* class. We will not have direct instances from this class, and it always will be through one of its specializations. The attributes of this super-class (URI and namespace) may be null in instances that are not defined this information.

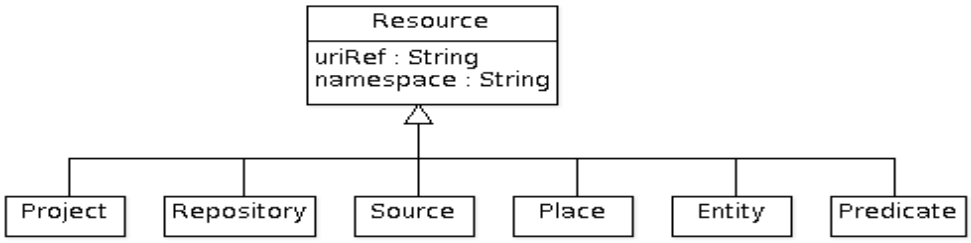


Figure 4: Resource and their subclasses

4.5. Modeling *Event*, *Place* and *Date*

A *Event*, specialization of *Entity* class (section 4.3), usually occurs at times and known places. Moreover, the class *Place* is designed with the maximum adaptability to different cultural and geographical contexts. Our proposal avoids defining specific categories or levels associated with the scope of "Place" (both a city and district are instances of *Place*), given the relativity of its meaning. Instead, we give the class *Place* the possibility of a reflexive relationship that allows each object associated with lower-level geographic entities. Thus, a Country contains several states, a *State* can contain *County*, this ones *City*, etc. An enumerated attribute, which can be particularized and extensible for each particular context, define levels and rates provided.

DateValue is another class of general utility. It and its specializations (*DateExact*, *DatePeriod*, etc.) inherited from Gedcom. With this design we can accommodate all types of time references. In addition, we added a enumerate attribute to identify the calendar of the original data base (essential with ancient documents).

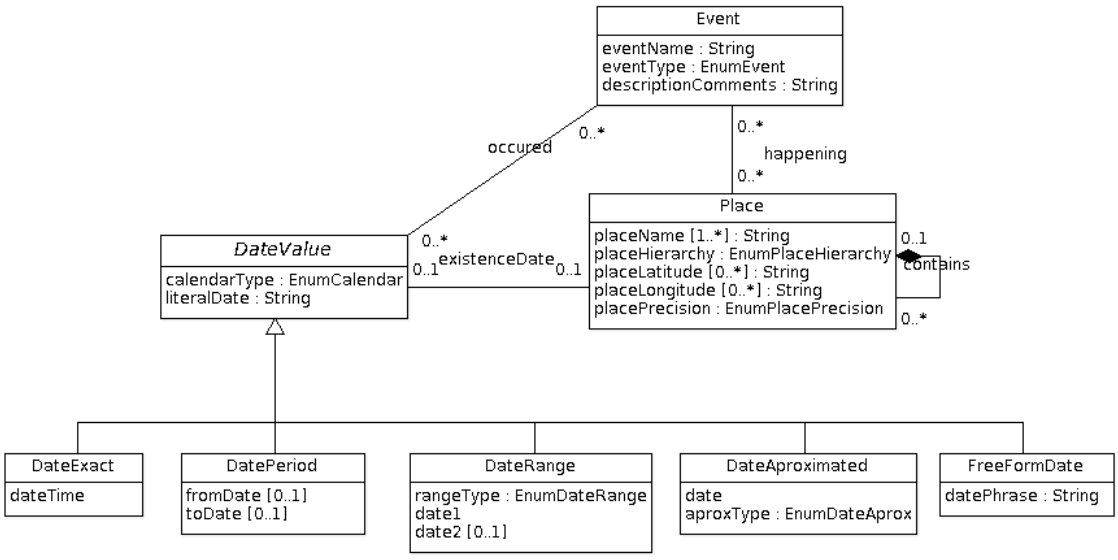


Figure 5: Relationships between Event, Place and Date

4.6. *Integrity constraints*

In the diagram are reflected certain integrity constraints, but not all. This is because the purpose of keeping the design simple and legible. Some of the restrictions that we have in mind are:

- A collaborator that submit certain *Assertion* should be assigned the necessary role in some project related to the *Source*. This is necessary to keep track of the origin of the information available in the system.
- *DateRange* and *DatePeriod* types should verify that the first date must be previous or at most equal to the second.
- Obviously, as the genealogist facts are always in the past, we can't have dates after the current
- We can only propose *Assertion* in projects where their related activity has the *status* attribute in the state of "*in progress*". Clearly, consistency must meet the respective *status* attributes available for the classes *Project*, *Objective* y *Activity* (we can't have a activity *in progress* on a project *finished*, for example).
- All *Activity* aimed at extracting data from a *Source* must necessarily be associated with the *Repository* where it comes from this source. This does not mean that we can have similar sources (may be copies) in other repositories.

5. Ontological specific terminology: Enumerations

Here we describe the main features of the types listed in the model. These descriptions are not intended to be exhaustive, since it may depend on specific cultural contexts. We must also think that in a different language implementations they end up using special terminology.

EnumProjectType

Allow classify different type of project: organizational, computerization, sources extraction, research

EnumStatus

Actual progress: proposed, approved, rejected, updated, in progress, canceled, finished

EnumRole

Assigned to project collaborators: direction, coordination, translator, documentary, submitter, validador, etc.

EnumRepresentation

Source type: archive, book, chapter, document, page, section, part, etc.

EnumOrigin

Source qualification: original, reprint, issue, copy, certificate, form, testimony

EnumSupport

Source phisical support: audio, digital, handwritten, inscription, microfilm, printed, etc.

EnumLanguage

Identifies the language of the tag in which it is defined. Legal values are language codes as defined by ISO 639, a combination of such language codes and country codes (ISO 3166-1) as "en-US", or IANA LANGUAGE CODES [14]. Frequently, documents can be written in older versions of certain language. In these cases, please indicate it in the comments attribute of the instance.

EnumPlaceHierarchy

Location hierarchy: country, state, region, department, town, municipality, etc.

EnumPlacePrecision

Sometimes we have only relative references: precise, nearby, around, aside, etc.

EnumSurety

Source quality: unreiable, questionable, secondary evidence, primary evidence

EnumGenRelation

Family relations: parent, father, mother, biological child, ancestor, parentage, etc. For extensive relations we can use the Protègè ontology [15]

EnumLogic

Logic vocabulary: true, false, null, exists, or, and, not, etc. For extensive relations we can use the fipaFOL.owl ontology [16]

EnumAssignement

Assignment vocabulary: equivalent to, is a, different from, requires, has value, etc.

EnumSex

Male, female, unknown

EnumNamePiece

Name piece depends of cultural and geographical environment. Usually we have: name, nickname, religious name, surname_1, surname_2, middle name, etc.

EnumGroup

The list of different group of people and interests are extensive. Some samples: parents family, great family, religious, ethnic, professional, political, army, organization, congregation,nobleness, society, etc.

EnumEvent

Events in people live are extensive: birth, marriage, death, burial, adoption, baptism, blessing, divorce, emigration, graduation, naturalization, ordination, residence, retirement, etc.

EnumCharac

This is a very open enumeration type, because classify people characteristic. For example: education, language, nationality, physical feature, affection, work, title, religion, age, etc.

EnumCalendar

When analyzing old documents wee need to specify the source calendar: french, gregorian julian, islamic, precolombin, etc.

EnumDateRange

Specifies this date: before, after, between

EnumDateAprox

Specifies this date: about, calculated, estimated

6. Instantiating the model. Examples

Here we propose some examples of how we can instantiate a *Statement* (part of *Assertion*) from documentary sources. *Statement* class, as noted in Section 4, conforms the atomic triples, in the form of <subject, predicate, object>. Do not focus on the inferences made, although in some cases may seem insufficiently justified: we can assume that the ontologies used make the necessary correspondences (eg, if two people marry automatically constitute a family).

Fact (Assertion or Statement)	subject	predicate	object
Paul was carrier	Paul Class: Persona	was Class: Predicate	carrier Class: Characteristic
Martí Fusté was the same person than Martí Fuster	Martí Fusté Class: Persona	equivalent Class: AssignementPredicate	Martí Fuster Class: Persona
Martí Fuster married Maria Pons in 13/4/1760	Family Fuster Pons Class: Group	formalize Class: Predicate	marriage in 13/4/1760 Class: Event
	Family Fuster Pons Class: Group	husband Class: GenealogicalPredicate	Martí Fuster Class: Persona
	Family Fuster Pons Class: Group	wife Class: GenealogicalPredicate	Maria Pons Class: Persona
Núria, Fuster Pons couple's daughter, was born in 1834	Family Fuster Pons Class: Group	daughter Class: GenealogicalPredicate	Nuria Fuster Pons Class: Persona
	Nuria Fuster Pons Class: Persona	was Class: Predicate	born in 1834 Class: Event
Núria, being the heiress, inherited the family home (Mas Fuster) in 1879	Family Fuster Pons Class: Group	heiress Class: GenealogicalPredicate	Nuria Fuster Pons Class: Persona
	Nuria Fuster Pons Class: Persona	she Class: Predicate	inherited Mas Fuster in 1879 Class: Event
	inherited Mas Fuster Class: Event	implies ownership of Class: Predicate	Mas Fuster Class: Resource
Charles died of tuberculosis on 10/2/1813 in Montcada	Charles Class: Persona	he Class: Predicate	died on 10/2/1813 in Montcada Class: Event
	Charles Class: Persona	suffered from Class: Predicate	tuberculosis Class: Characteristic

7. Related Works

In our model we propose to use ontology to reconcile data from different contexts or from different computer applications. In the last years interesting proposals have been published in the field of semantic web related to family studies. To cite some examples, on the one hand we have the prototype by Charla Woodbury and David W. Embley, from BYU Computer Science Department [3], in their article "*Family History Research on the Semantic Web: Building a Semantic Prototype for Danish Research*". Here they identify a set of entities for which they should involve high-level ontology to provide semantic content in the associated fields in their databases. Thus, the attributes name, date, place, relationship, occupation, record_type and source can be interpreted correctly regardless of the linguistic context used in their description.

To allow that our model is applicable in different contexts in which we find even with different vocabularies and terminologies, we mention in this paper the need to reconcile different ontologies using mechanisms of equivalence between concepts. In this regard, an interesting approach that gives an idea how to semantically combine content from different ontological, is the contribution of Dejing Dou, Drew McDermott and Peishen Qi in "*Ontology Translation on the Semantic Web*" [4]. They propose a online system (as OntoMerge, ontology merging and automated reasoning), where the merge of two related ontologies is obtained by taking the union of the concepts and the axioms defining them, using XML namespaces to avoid name clashes. They focus on formal inference from facts expressed in one ontology to facts expressed in another

With the aim of integrating genealogical content in the semantic web, we have the work done by Ivo Zandhuis "*Towards a Genealogical Ontology for the Semantic Web*" [5]. The model presented here proposes to develop two ontologies, *srcont* and *genont,* specific to set the person (the subject of all genealogical study) and sources of information, respectively. This work has largely been an important reference in the construction of our model, for its clarity and simplicity.

We would also like to mention the interesting work done by Christoffer Owe, GenXML [6]. Inspired by Gentech initiative, it aims to offer an alternative to GEDCOM for exchanging genealogical data. Taking the form of OWL-RDF files [7], gets to resolve many of the drawbacks and inconsistencies that could occur between structured data under that model.

Finally, we mention the great number of domain-specific family ontologies now available online. To cite some examples, we have in *FOAF* [8] a structured namespace RDF model designed for individuals and groups. Also it is worth mention *Relationship* [10], which provides an extensive vocabulary for describing relationships between people. At the same line, *The Event Ontology* [11] also provides very full descriptions in RDF to model people, events, time and location.

8. Conclusions

For many years data modeling of genealogical used by the vast majority of computer applications has been dominated by the data transfer format created by Gedcom. In a way, computer applications in this field seemed to haven't crossed the limits of family relatives: data storage and building personal family trees. 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 is not automatic and requires the assistance of the expert. And yet this does not guarantee that inconsistencies arise between the repositories at the time of merging the information, due to the use of specific terminology, different contexts, and so on.

The automatic processing of information is possible only if we complement data with a semantic type related information that is understandable for machines. We refer, of course, the use of ontology as RDF and OWL standards that allow us to give an understandable format for computers. Also, giving instances of the family domain (people, events, groups, places, etc.) with unique identifier, a predefined namespace and and a URI data, we provide additional contextual information that allows subsequent automatic inference of knowledge.

The proposed model aims to be comprehensive and independent of its subsequent implementation. It provides great flexibility to adapt to social, cultural, geographical or over time, playing for it with the semantic meaning of the terminology used in each

case as defined in the ontological associated. It also solves the main problems identified in integrating genealogical data such as data inconsistencies, recognition of equivalence or new relationships, enabling the coexistence of different data for the same attribute to the point of allowing conflicting data by assigning a confidence level for each one of them.

Our ultimate goal was to build an integrated conceptual model, with enough flexibility to adapt different cultural, geographical or historic contexts, designed under the ontological paradigm. Also, we consider the events as the base of the information system, with the difference that these will be represented, in its atomic form, with the triplet <Subject, Predicate, Object>. In the background, the idea is to enable a future evolution towards an information system capable to infer the implicit knowledge from atomic data recorded.

9. References

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

[2] ***Gentech-GDM Reference Model***
Author: Stanley Mitchell
Publication: (2003)
http://freepages.history.rootsweb.com/~mitchellsharp/gdmref/gdmref-01.pdf

[3] ***Family History Research on the Semantic Web: Building a Semantic Prototype for Danish Research***
Author: Charla Woodbury and David W. Embley
Publication: BYU Computer Science Department (July 2005)

[4] ***Ontology Translation on the Semantic Web***
Author: Dejing Dou, Drew McDermott and Peishen Qi
Publication: Yale Computer Science Department (2003)

[5] ***Towards a Genealogical Ontology for the Semantic Web***
Author: Ivo Zandhuis
Publication: Association for History and Computing Conference, (Amsterdam, 14-17 September 2005)
http://www.zandhuis.nl/sw/genealogy/genont.pdf

[6] GenXML
Author: Christoffer Owe
Publication: (2007)
http://cosoft.org/genxml/

[7] World Wide Web Consortium -w3c
http://www.w3.org/
http://www.w3.org/TR/rdf-concepts/
http://www.w3.org/TR/owl-ref/

[8] FOAF
Author: Dan Brickley, Libby Miller
Publication: (December 2009)
http://xmlns.com/foaf/spec/

[9] Ontology Design Patterns.org (ODP)
http://ontologydesignpatterns.org/

[10] Relationship
Author: Ian Davis i David Galbraith
Publication: (February 2010)
http://vocab.org/relationship/
http://vocab.org/bio/0.1/

[11] The Event Ontology
Author: Yves Raimond i Samer Abdallah
Publication: Quenn Mary, University of London (2007)
http://motools.sf.net/event/event.html

[12] Semantic Web.org
http://semanticweb.org

[13] Swoogle
http://swoogle.umbc.edu/

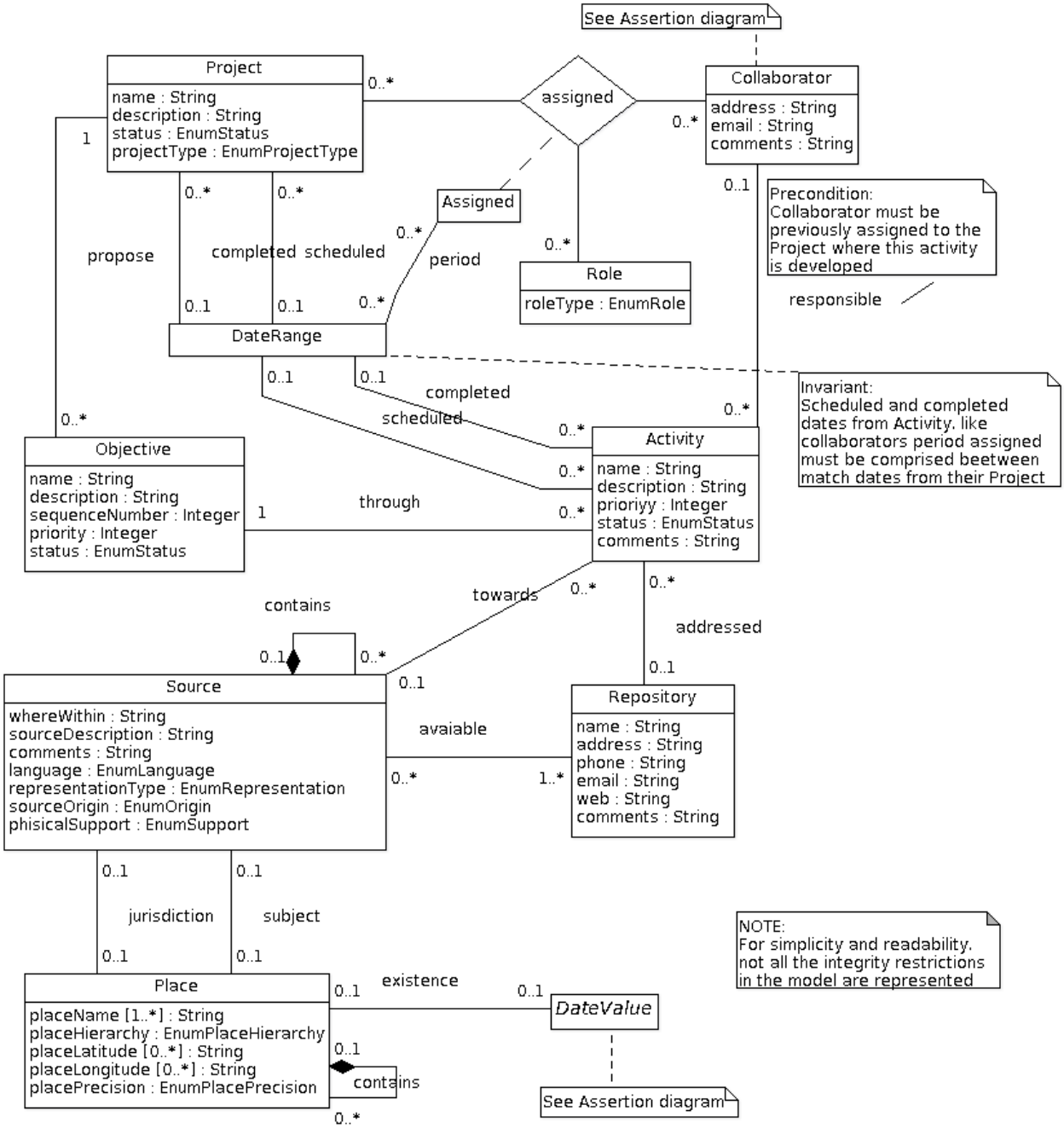
[14] IANA LANGUAGE CODES
http://www.iana.org/assignments/language-subtag-registry

[15] Protègè family ontology
http://protege.cim3.net/file/pub/ontologies/family.swrl.owl/family.swrl.owl

[16] fipaFOL.owl ontology, Universitat Politecnica de Catalunya and Ecole Polytechnique Federale de Lausanne
http://www.lsi.upc.es/~ffernandez/resource/ontology/fipaFOL.owl

APENDIX. CONCEPTUAL MODEL

UML Diagram: “Projects”



UML Diagram: “Assertion”