Modeling historical social networks databases

Jacky Akoka, Isabelle Comyn-Wattiau, Stéphane Lamassé, Cédric Du Mouza

Modeling Historical Social Networks Databases

Jacky Akoka
CEDRIC-CNAM &
IMT-TEM
jacky.akoka@lecnam.net

Isabelle Comyn-Wattiau
ESSEC Business School
wattiau@essec.edu

Stéphane Lamassé
LAMOP
Stephane.lamasse@univ-paris1.fr

Cédric du Mouza
CEDRIC-CNAM
dumouza@cnam.fr

Abstract

Historical social networks are analyzed using prosopographical methods. Prosopography is a branch of historical research that focuses on the identification of social networks that appear in historical sources. It aims to represent and to interpret historical data, sourced from texts. Conceptual modeling imparts the capability to process these large data sets. This paper outlines a conceptual approach to designing a prosopographical database encompassing uncertainty. Our contribution is threefold: i) a generic certainty-based prosopographical conceptual model; ii) two meta-models with a mapping between them; iii) an illustrative example generating a customized prosopographical relational model. Unlike past approaches, our design process helps us to integrate disparate points of view as expressed in the prosopography community. We apply our approach to the prosopographical database Studium Parisiense dedicated to members of Paris schools and university between the twelfth and sixteenth centuries. This instantiation validates the usefulness of our approach.

1. Introduction

Prosopography is a domain of digital humanities related to the inquiry into the common characteristics of a group of historical actors by means of a collective study of their lives [1]. It relies generally on a database containing information related to persons from a specific milieu defined chronologically and geographically [2]. Its purpose is to collect and analyze data describing the individual lives of the historical actors under consideration, targeting mainly their common characteristics. Historians generally study large groups of individuals poorly documented. They fill in manually for each actor a record with all information they have regarding the milestones of his/her life, the places he/she visited, the people he/she met, his/her production, etc, according to a schema they decide for their prosopographical database. The reliability and the quality of the source material (demographic, economic, administrative, religious, family archives, etc.) is crucial. Moreover, historians are confronted with the relative scarcity of source material. Representing the time and the uncertainty dimensions related to people, locations, factoids, and source material constitutes another problem. Prosopography deals with information which is often incomplete, imprecise, and contradictory. Therefore, there is a need to develop data models accommodating all types of uncertainty including the one characterizing the dating phenomena.

There exist several models representing prosopographical data. The most common model is that of the factoid [3]. Prosopography of Anglo-Saxon England (PASE) is based on a factoid model in which statements about persons, possessions and places are derived from sources [3]. Another example of project based on the factoid model is the Roman Republic [4]. Most existing digital prosopography projects use relational databases. However, to the best of our knowledge, no prosopographical project uses conceptual modeling to derive the associated relational model, which considerably limits the ability of merging or querying different prosopographical databases. The aim of this paper is to present an approach allowing us to build a generic certainty-based prosopographical conceptual model which serves as a basis for the instantiation of contextualized conceptual and relational models of prosopographical databases such as PASE and Studium Parisiense [5].

We begin in Section 2 by identifying structured elements of prosopographical models and databases. We then present and discuss our approach in Section 3. Section 4 is dedicated to the application of the approach to the Studium Parisiense prosopography. We conclude in Section 5 and present future research directions. Our contribution is threefold: i) a generic certainty-based prosopographical conceptual model; ii) two meta-models with a mapping between them; iii) an illustrative example generating a customized prosopographical relational model.

2. Prosopographical Concepts: A State of the Art

There are many prosopographical databases such as PASE, Studium Parisiense, prosopography of the Byzantine Empire, China Biographical Database Project, The Making of Charlemagne's Europe, and Paradox of Medieval Scotland [6]. Prosopography analyzes information on sets of individuals in the context of historical societies. Central to any prosopographical project are the concepts of *event*, *time*, and *uncertainty*.

Modeling life stories of a group of persons can be performed using the event-based approach [7]. In an event, a person can take different roles. Events are linked to other events, persons, places, time periods, and documents. [8] distinguishes different types of events, supporting both discrete and continuous events, and expressing various temporal aspects of events. Event times are generally specified as date ranges and have time-spans with durations. Most of the standards mentioned by [8] enable the association of events with location terms, including geographical place names. Events play the role of linking persons to places and times. Individual events can be linked to multiple documents and vice-versa. Several ontologies describing events have been proposed [9]. *Our work is inspired by the recognized and successfully used in several contexts event model [10] since it focuses on the main concepts of interest in prosopographical projects.*

Any historian faces the problem of representing temporal data. Time can be the source of vagueness and/or uncertainty. Temporal database research [11] consider two types of data: "instant" and "interval" [12]. Allen [13] proposes a time model based on time intervals. A number of temporal relationship types are based on Allen's temporal logic [14]. Very few research works offer support for modeling relative times. The GENTECH model [15] supports the creation of conflicting temporal relationships expressing different points of view. Some databases integrate data temporal aspect by relying on a temporal version of SQL (TSQL2) [16]. The time model in AROM-ST [17] offers several time types including instant, interval, multiInstant, and multiInterval types. The importance of time considerations in ontologies was initiated by the semantic web community [18]. A variety of approaches have been proposed to represent temporal information in RDF [19] and OWL [20]. Several approaches have been proposed for time modeling using the ER conceptual model [11]. *In our approach, we selected the AROM-ST model because of its generality.*

Uncertainty is defined as "a general concept that reflects our lack of sureness about something or someone" [21]. Uncertainty reflects a lack of confidence in an object, in an event or in a person. A survey about theories and practices in handling uncertainty can be found in [22]. There exist many uncertainty classifications [23]. In the URREF ontology [24], uncertainty encompasses a variety of aspects including ambiguity, incompleteness, vagueness, randomness, and inconsistency. Ambiguity arises when the information lacks complete semantics**.** Incompleteness reflects a lack of information. Vagueness arises when a situation is characterized by an incomplete knowledge of the facts and events under consideration. Randomness expresses the lack of pattern or predictability in events. Finally, inconsistency arises when two or more information cannot be true at the same time. These uncertainties may be supported by different uncertainty models or theories, such as probability theory, possibility theory, fuzzy sets, etc. [25]. A review of the literature on fuzzy conceptual modeling and databases is presented in [26, 27]. *In our approach, we use the URREF ontology which seems to be the most appropriate for representing the uncertainty that characterizes prosopographical data*.

Concepts such as event, time, and uncertainty are central to our modeling approach described in the next section.

3. Our Approach

Our methodology consists of three main steps. The first one is dedicated to building a generic certainty-based prosopographical conceptual model. Then we proceed to its customization leading to a specific prosopographical project. Finally, we automatically convert the resulting conceptual model into a prosopographical relational database.

3.1 Building a generic prosopographical conceptual model

We first proceeded to requirements gathering which encompasses the following tasks: interviewing historians, browsing through prosopographical databases, analyzing the factoid models, and studying the literature on time and uncertainty modeling. Then, we chose to capitalize on the factoid model by putting emphasis on a limited number of concepts named factoid objects, such as *Person*, *Place*, *Factoid*, and *Source*. *Time* is an important dimension too. Moreover, all the information is tainted with uncertainty. A factoid may be considered as an event taken in a broad sense including all the facts that characterize individuals. For example, a publication is also an event. The choice to generalize the event into a factoid enables a

compact model. However, it led us to define the factoid with a larger number of dimensions. For instance, the fact that an event impacts an object allows us to cover the publication written by an author, the purchase of a property, the dowry at a wedding,

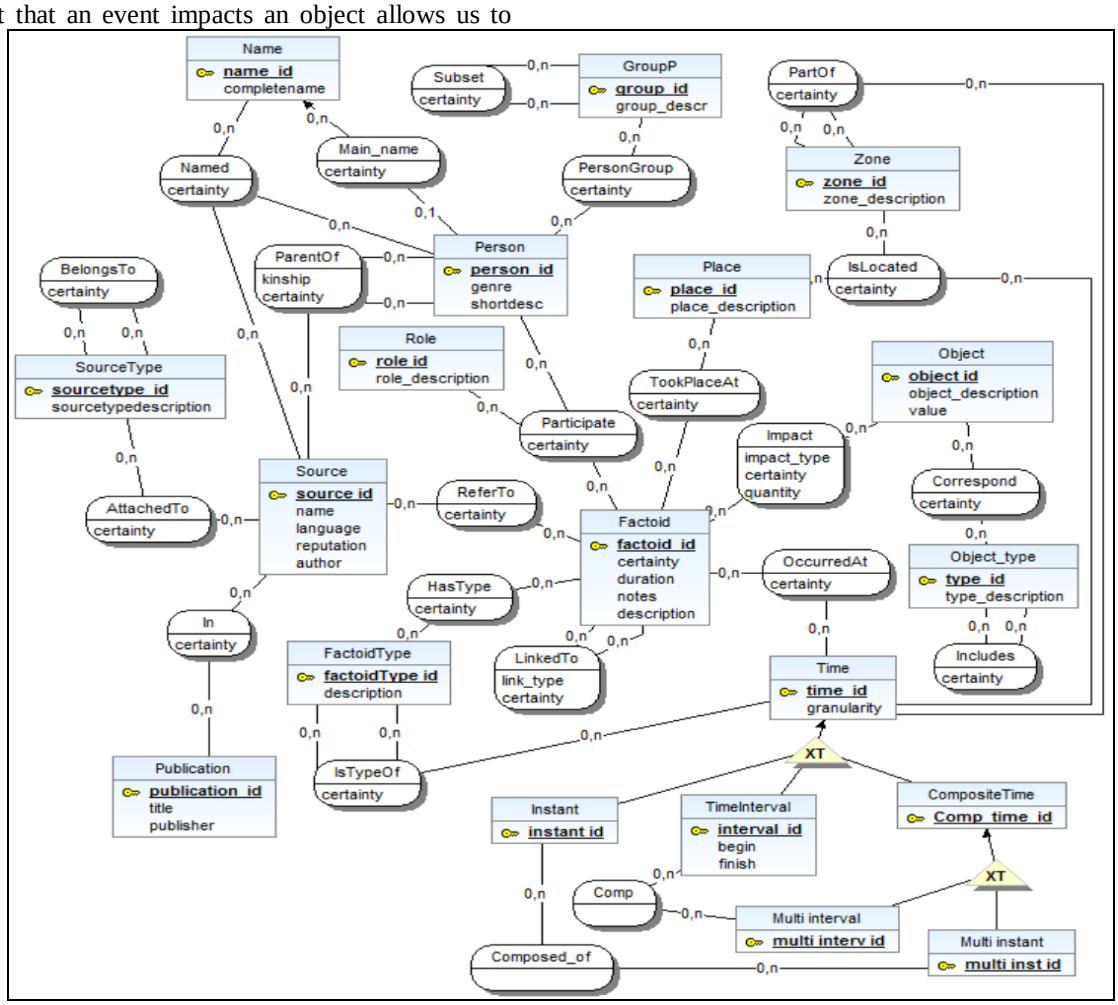

Fig. 1 The generic conceptual prosopographical model

etc. Our model also supports a multi-level hierarchy of concepts. For example, *Places*, *Sources*, *Persons*, and *Factoids* are generalized to one or more levels. The factoids are grouped recursively into types of factoids (*FactoidType*), like in PASE where the confession is an event of Christian piety, itself a religious act. The resulting generic prosopographical conceptual model is presented below (Fig. 1).

One difficulty of prosopographical research lies in onomastics, *i.e.* the need to identify the people that may be known by different names, each one associated with an uncertainty degree. People are also generally linked to groups. Our model supports the ambiguity attached to names as well as the concept of groups (*GroupP*).

Most relationships between concepts are typed. For example, the type of impact between an event and an object allows us to specify that, during a barter event, an object is transferred and an object is granted in exchange. Between factoids, the *Linkedto* relationship is used to define dependencies between events such as precedes, provokes, and so on. The *Role* of a *Person* in a *Factoid* is an entity since the same person can sometimes play more than one role in the same event.

The representation of time integrates discrete time (*Instant*), continuous time (*Interval*), and their composition. It is adapted from the AROM-ST model [17].

Finally, our generic conceptual model integrates the management of uncertain information in four forms: incomplete data leading to null values, ambiguous information due to linguistic terms (*e.g.* about, probably, not far from, etc.), vague information (membership degree, importance degree, etc.) [28], and inconsistent assertions. In our model, *certainty* is a

representation of the degree of reliability of the information to which it is attached. Generally, it takes its value in the interval [0, 1]. In certain cases, these values are: near, around, close to, in the vicinity of, not far from, a few kilometers from, etc. When it characterizes the timing of an event, *certainty* can take the values of: around, before, well before, shortly after, and so on. Moreover, *Time* and *Place* are connected. For example, Flanders belonged to France at one time but not always. This led us to introduce the ternary relationship *IsLocated* between *Place*, *Zone*, and *Time*. Similarly, factoid types are linked to the *Time* entity. We manage the contradictory assertions by associating each source with a confidence degree associated to its *reputation*.

3.2 Customizing the generic prosopographical conceptual model

The customization process to any particular prosopographical database requires the following steps:

Step 1. **Model pruning.** Each prosopographical project concentrates on some specific features, implying a pruning of all the irrelevant parts of the model. As an example, Studium Parisiense does not consider links between factoids.

Step 2. **Model refinement.** The objective of this step is to facilitate data updating and to reduce quality issues. Each prosopographical project refers to basic assumptions and authority lists. Based on them, the database designer, with the help of historians, lists the possible values for each meaningful concept, enabling a precise definition of attribute domains. As an example, the *Role* entity may be characterized by a closed set of values. In PASE, it takes a number of values such as apostate, apostle, disciple, fugitive, etc. Moreover, for each hierarchy, the database designer has to set the number of hierarchy levels, the type of hierarchy (one-to-many or many-to-many), and the list of possible values for each level. As an illustration, PASE project includes a 3-level hierarchy of factoids whose first level contains the five following categories: 1) acts of crime, law-breaking/violence, 2) legal/governmental/administrative acts and legitimate use of violence, 3) life-events/social and economic acts and relations, 4) power-taking and power-leaving, and 5) religious/ecclesiastical acts. Finally, in some cases, the customizing process encompasses the addition of specific attributes to some concepts. As an illustration, ethnicity is an important information in PASE project.

Step 3. **Temporal model management.** Depending on the timeline of the prosopographical project, we have to associate each *Time* entity (*Instant*, *Interval*, etc.) with a specific grain. In Studium Parisiense, the time grain is the year whereas, in PASE, the dates are subdivisions of centuries (early, middle, and late).

Step 4. Linguistic terms management. Prosopographical databases rely on sources containing natural language descriptions. Thus, in particular for dates and places, there is a need to check a sample of representative sources for extracting fuzzy expressions, such as: around, about, probably, etc. and mapping linguistic terms to an evaluation of their value, *i.e.*, around may take the value *less than 20 km*.

Step 5. Fuzzy attribute elicitation. For each attribute of the model, we check with historians whether this attribute is fuzzy and, as the case may be, its type of vagueness among the five categories: membership, importance, fulfillment, possibility, uncertainty, as defined by [28]. As an example, *kinship* in Studium Parisiense is sometimes fuzzy.

Step 6. **Fuzzy object elicitation.** The generic prosopographical model contains only one uncertain entity (*Factoid*) and many uncertain relationships (*BelongsTo*, *Participate*, etc.). For each certainty found in the generic model, we check whether it should be maintained, and, if so, define its type of vagueness. As an example, in Studium Parisiense, people belong to different groups: student, master, graduate, etc. This information is often uncertain.

At the end of this customization step, the conceptual model is annotated for a specific prosopographical project.

3.3 Mapping the customized conceptual model to a relational database

To carry out this step automatically, we adopted a model-driven approach. To this end, we have defined: a) an Extended Entity Relationship (EER) conceptual meta-model incorporating uncertainty, b) a relational meta-model incorporating uncertainty, and c) a set of mapping rules from conceptual to relational meta-models. Due to the presence of vagueness in the resulting conceptual model, standard mapping rules do not apply, requiring the following approach.

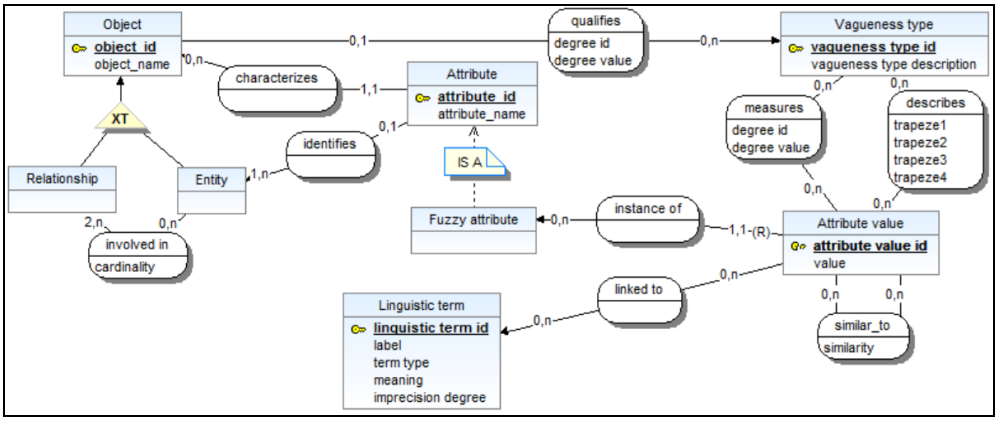

Fig. 2 The EER meta-model incorporating uncertainty

3.3.1 The EER conceptual meta-model incorporating uncertainty

The EER meta-model encompasses the standard concepts of entity and association (Fig. 2). They form a partition of the set of objects of the conceptual model. They are characterized by attributes. For entities, a subset of these attributes constitutes an identifier (to simplify, we consider here that an entity has only one identifier).

Some attributes may be fuzzy, requiring specific relational mappings. We have improved the expressiveness of the model by allowing fuzzy attributes to cover different types of vagueness across different ranges of values. Therefore, in the meta-model, we associate with *Attribute values* four relationships corresponding to four different modalities: (i) an attribute value can be linked to a linguistic term. For example, 1530 is a value of the *Year* attribute. This value can be associated with a linguistic term of the temporal type such as "around", which can have a meaning: "in an interval centered on this value" and a degree of inaccuracy: 10%, which makes it possible to calculate the interval around 1530; (ii) an attribute value may be similar to another one. For example, the kinship "elder brother" is similar to the kinship "brother" with a similarity that can be quantified; (iii) an attribute value can be qualified by a trapezoidal function. As an example, "young" is defined over the trapezoid (20, 30, 40, 50) that represents four successive x-axis values such that a medieval clergyman is undoubtedly "young" between 30 and 40 years old, possibly "young" between 20 and 30 or between 40 and 50 years old); (iv) an attribute value is defined with a degree with respect to a type of vagueness (membership, importance, possibility, etc.).

3.3.2. A relational meta-model incorporating uncertainty

Similarly, the relational meta-model contains the standard relational schema concepts: Relation, Column, Columnset (aggregates columns to define candidate keys) (Fig. 3).

Four relation subtypes are added to represent the concepts related to uncertainty: (i) *Vagueness type* that will become a relational table listing all the types of uncertainty represented in the database; (ii) the *Degree* table which, in the same way, will contain all the degrees of uncertainty or inaccuracy associated with either the objects (entities or relations) or the attributes of the prosopographical model; (iii) the *Linguistic term* table which contains the linguistic terms describing the uncertainties (probably, perhaps, not impossible, probably, etc.) or the inaccuracies (close to, around, near, etc.); (iv) the *Trapezoid_description* table contains all trapezoidal type coordinates to represent possibilistic elements.

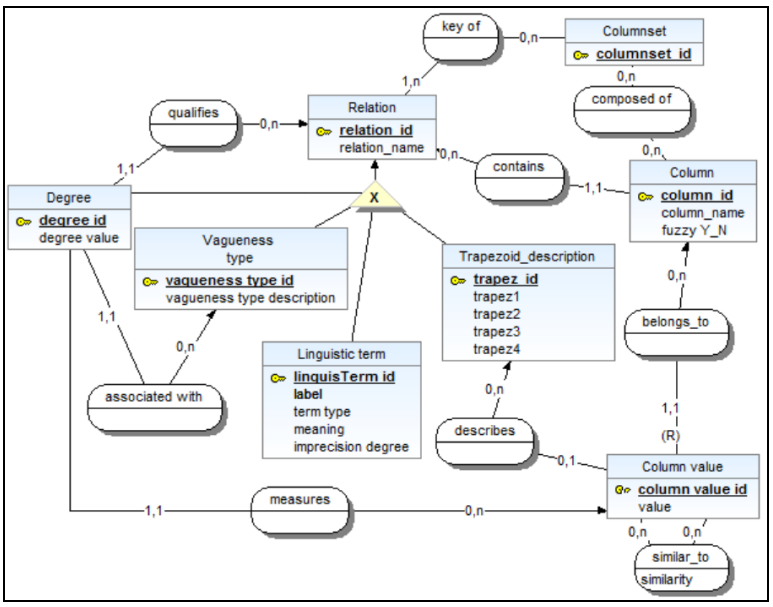

Fig. 3 The relational meta-model with uncertainty

3.3.3. A set of mapping rules from conceptual to relational meta-models

Beyond the standard rules mapping the EER model to the relational model by means of meta-models, we defined special rules dedicated to the mapping of uncertain information. In particular, in order to incorporate the vagueness in the relational database, we generate the specific tables containing the corresponding elements. Thus, if a fuzzy attribute is defined, for each of its fuzzy values represented by a trapeze, there will be a row in the table *Trapezoid_description*. Table 1 summarizes the mapping between the two meta-models at the concept level.

Conceptual meta-model		**Relational meta-model**	
Object	**Attribute**	**Object**	**Attribute**
Attribute	attribute_id, attribute_name	Column	column_id, column_name
qualifies	degree id, degree value	Degree	degree id, degree value
measures	degree id, degree value	Degree	degree id, degree value
Object	object_id, object_name	Relation	relation_id, relation_name
describes	trapeze1, trapeze2, trapeze3, trapeze4	Trapezoid description	trapez1, trapez2, trapez3, trapez4
Vagueness type	vagueness type description	Vagueness type	vagueness type description
	vagueness type id		vagueness type id
Attribute value	attribute value id, value	Column value	Column value id, value
Attribute	ISA between Fuzzy attribute and Attribute	Column	fuzzy Y_N
similar_to	similarity	similar_to	similarity
Linguistic term	linguistic term id, label, term type, meaning, imprecision degree	Linguistic term	linguisTerm id, label, term type, meaning, imprecision degree
Object	object id	Columnset	columnset_id
involved in	cardinality	Column	column_name
		Relation	relation_name

Table 1. Mapping the meta-models

An example of a rule is described below. It deals with the case where an attribute A linked to O (O may be either an entity or a relationship) contains fuzzy values defined with trapezoid functions.

```
For each O ∈ Object
   For each A ∈ Fuzzy_attribute characterizing O
     For each V ∈ Attribute_value such that V is an
instance of A
      If VT is the Vagueness_type describing V whose
attributes trapeze1 to
        trapeze4 take respectively t1 to t4 values
       Then                                    Tra-
pezoid_description=Trapezoid_description      ∪ {(id,
t1,t2,t3,t4)};
   /* the tuple (id, t1, t2, t3, t4) is inserted in
table Trapezoid_description */
            Column_value = Column_value ∪ {(vid, V,
id)};
   /* the tuple (vid, V, id) is added to table Col-
umn_value, id being the foreign key linking this
column value V to its fuzzy trapezoidal description
*/
   End If; End For; End For; End For;
```

For space reasons, we cannot provide the reader with the whole set of rules. In this section we provide a generic prosopographical model and its mapping to a relational meta-model which allows to model any prosopographical databases with uncertainty management. We illustrate in the following how our generic model can be derived to a specific prosopographical project to demonstrate its feasibility and usefulness.

4. Illustrative example: Application to Studium Parisiense

Our research is part of a project funded by the French National Research Agency and related to several prosopographical contexts. We illustrate our approach on the Studium Parisiense project aiming at creating an online biographical-bibliographic database describing members of Paris' schools and university from the twelfth century until the end of the Middle Age. The project currently totals more than 16,000 records, of which almost 9,000 are already online using an XML format.

4.1 Customizing the generic model to the context of Studium

Applying **Step 1** of our approach leads to the deletion of (i) the entities *Publication*, *CompositeTime*, *Multi_Instant*, *Multi_Interval*, (ii) the relationships *In*, *BelongsTo*, *Comp*, *Composed_of*, *LinkedTo*, *PartOf*, *Includes*, *Subset*, (iii) some attributes such as *value* (entity *Object)*, *duration* (entity *Factoid)*, *quantity* (relationship *Impact*), etc. Notice that we also added some attributes, such as *language* (entity *Object)* and *social class of origin* (entity *Person*).

The application of **Step 2** allows us to define all the authority lists. Among them, let's mention: (*i)* the list of role labels: *per se*, *author*, *grantee*, etc., (*ii)* the list of factoid types: *birth*, *death*, *activity*, *origin*, *university or studium attended*, *ecclesiastical position*, *functions with the pope*, etc. We also set the different hierarchy types and levels. As an illustration, a factoid is not defined for Studium by an N-N relationship but by a purely hierarchical set of types with only two levels. For instance, *functions with the pope* (first level) is part of *ecclesiastical position* (top level) characterizing the career of Alexander de Kininmund who was Prosecutor of Thomas de Fingask, Bishop of Caithness at the Curia in Avignon in 1348.

Step 3 allows us to define the time unit, here *year*. Even if years are provided in sources, they are often qualified by a linguistic term describing the uncertainty level.

During **Step 4**, parsing a significant sample of XML files (in which many fields are expressed in natural language), we collected a consequent list of linguistic terms representing an uncertainty level of the information mentioned (e.g. *nothing should allow us to know*, *probably*, *it is not impossible that*, *perhaps*, *unlikely*). We met the team of historians in charge of for Studium Parisiense and asked them to validate this list and to enrich it with a numeric scale.

Step 5 generates the list of fuzzy attributes which is very limited in Studium. Among them, let us mention the *kinship* attribute characterizing *ParentOf* links between *Persons*. The peculiarity of this attribute led us to build a table linking two by two all the possible values of *kinship* and to characterize the links by a similarity measure. As an example, *brother* and *elder brother* are very similar.

Finally, **Step 6** required more effort to qualify the uncertainty feature and the vagueness type of each entity and/or relationship. For instance, relationship *OccurredAt* comes often with linguistic terms listed in **Step 4.** Regarding the relationship *TookPlaceAt*, the vagueness type *uncertainty* applies whereas *importance* better qualifies the *reputation* of *Source*.

For space reasons, we cannot present the final conceptual model. As it can be seen in the different steps, the involvement of historians is crucial to the success of this customization process.

4.2 Generating the relational schema

Firing the mapping rules described above, we obtain the following customized relational schema.

Correspond (correspond_id, *object_id*, *type_id*, *degree_id*)
Degree (degree_id, degreevalue, *vaguenessType_id*, *object_id*)
Factoid (factoid_id, *degree_id*, duration, notes, description)
FactoidType (factoidType_id, description)
GroupP (group_id, group_descr)
HasType (hasType_id, *factoid_id*, *factoidType_id*, *degree_id*)
Impact (impact_id, *factoid_id*, *object_id*, impact_type, *degree_id*)
Instant (time_id, *instant_id*, granularity)
IsLocated (isLocated_id, *place_id*, *zone_id*, *degree_id*, *time_id*)
IsTypeOf(istypeof_id,*factoidType_id1*,*factoidType_id2*,*degree_id*, *time_id*)
KinshipLink (kinship1, kinship2, similarity)
LinguisticTerm(linguisTerm_id, label, term_type, meaning, imprecision degree)
Name (name_id, completename)
Named (name_id, person_id, source_id)
Object (object_id, object_description)
ObjectType (type_id, type_description)
OccurredAt(Occurredat_id, *factoid_id*, *time_id*, *degree_id*, *linguisTerm_id*)
ParentOf (parent_id,*person_id1*,*person_id2*,kinship, *degree_id*, *source_id*)
Participate (participate_id, *person_id*, *role_id*, *factoid_id*, *degree_id*)
Person(person_id,shortdesc,*main_name_id*,genre,maincompletename, *degree_id*)
PersonGroup (pg_id, *person_id*, *group_id*, *degree_id*)
Place (place_id, place_description)
ReferTo(referTo_id, *factoid_id*, *source_id*, *degree_id*)
Role (role_id, role_description)
Source (source_id, name, language, reputation, author, *type_source_id*)
SourceType(sourcetype_id, sourcetypedescription)
TimeInterval(time_id, *interval_id*, begin, finish, granularity)
TookPlaceAt(tookPlaceAt_id, *factoid_id*, *place_id*, *degree_id*)
VaguenessType (vaguenessType_id, vaguenessType_description)
Zone(zone_id, zone_description)

The following two queries shed the light on the opportunities offered by the resulting Studium database. The first query compares two individual careers as follows: *Who studied canon law in Paris at the same time as Petru de Quercu and then got an ecclesiastic position?*

This query shows how we succeed in capturing the uncertainty of the different data (factoids, places, times, etc.), and in managing linguistic terms with vagueness interpretation and the onomastics. The corresponding SQL query is:

```
SELECT P1.maincompletename,  D1.degreevalue as 'confidence
in scholarship period', D2.degreevalue as 'confidence in
study place',  D3.degreevalue as 'confidence in student
status', D4.degreevalue as 'confidence in ecclesiastic po-
sition after'
FROM Person P1, Person P2, Factoid F1, Factoid F2, Factoid
F3, FactoidType FT1, FactoidType FT2, HasType HT1, HasType
HT2, HasType HT3, Participate PA1, Participate PA2, Took-
PlaceAt TP1, TookPlaceAt TP2, OccurredAt OA1, OccurredAt
OA2, OccurredAt OA3, TimeInterval TI1, TimeInterval TI2,
TimeInterval TI3, Degree D1, Degree D2, Degree D3, Place
PL, IsLocated IL1, IsLocated IL2
WHERE P2.maincompletename='Petru de Quercu' and
P2.person_id=PA2.person_id
---- refers to main complete name because 'Petru de Quercu'
may correspond to several entries in the Person table ----
and F2.factoid_id= HT2.factoid_id and
HT2.factoidType_id=FT2.factoidType_id and
FT2.description='student in canon law' and
PA2.factoid_id=F2.factoid_id and
F2.factoid_id=TP2.factoid_id and TP1.place_id=PL.place_id
and TP2.place_id=PL.place_id and PL.description='Paris' and
P1.maincompletename != P2.maincompletename and
F1.factoid_id=HT1.factoid_id and HT1.factoidType_id=
FT2.factoidType_id and PA1.factoid_id=F1.factoid_id and
PA1.person_id= P1.person_id and F1.factoid_id=TP1.factoid_id
and OA1.time_id=TI1.time_id and OA2.time_id=TI2.time_id and
TI2.finish >= TI1.begin and TI2.begin <= TI1.finish
--- check if the two factoids are associated to overlapping
time intervals ---
and OA1.factoid_id=F1.factoid_id and
OA2.factoid_id=F2.factoid_id and
F3.factoid_id=HT3.factoid_id and HT3.factoidType_id=
FT1.factoidType_id and FT1.description='ecclesiastic posi-
tion' and F3.factoid_id= OA3.factoid_id and
OA3.time_id=TI3.time_id and TI3.begin>=TI1.finish
---- check if factoid of type 'ecclesiastic position' oc-
curred after factoid 'student in canon law' in Paris for
this person ----
and PA3.factoid_id=F3.factoid_id and
PA3.person_id=P1.person_id
and D1.degree_id=OA1.degree_id and
D4.degree_id=HT3.degree_id
and D2.degree_id=TP1.degree_id and
D3.degree_id=HT1.degree_id;
---- we consider information uncertainty degrees ----
```

The evaluation of this query on the Studium dataset returns the following results (extract).

Complete Name	Confidence in scholarship period	Confidence in study place	Confidence in student status	Confidence in ecclesiastic position after
Gerard de Manso	0.7	0.7	0.5	0.9
Nicolaus de Freauvilla	1.0	1.0	1.0	0.8
Blasius Eximini	1.0	1.0	0.5	1.0
Curatus Sancti Illari	1.0	1.0	0.5	0.5

The second query looks for more complex career patterns and takes into account sources reputation (evaluated by historians): Who are the Italian living in the fourteenth or fifteenth century who studied a PhD

degree in Bologna after studies in Paris, according to sources with a reputation greater than 0.5.

```
   SELECT P.maincompletename, I.instant as 'birthdate', PL1.place_description as 'birth-place', D1.degreevalue as 'confidence in the PhD place', D2.degreevalue as 'confidence in former studies',
(S1.reputation+S2.reputation)/2 as 'reputation of the sources'
--- global reputation of the sources for the curriculum is the average reputation of the sources which assess the PhD and the studies in Paris respectively ----
FROM Person P, Factoid F1, Factoid F2, Factoid F3, Participate Pa1, Participate Pa2, Participate Pa3, HasType HT1, HasType HT2, HasType HT3, FactoidType FT1, FactoidType FT2, FactoidType FT3, FactoidType FT4, TookPlaceAt TP1, Took-PlaceAt TP2, TookPlaceAt TP3,  Place PL1, Place PL2, Place PL3, IsLocated IL, Zone Z, OccurredAt OA1, OccurredAt OA2, OccurredAt OA3, Instant I, IsTypeOf ITO, TimeInterval TI1, TimeInterval TI2, ReferTo RT1, ReferTo RT2, Source S1, Source S2, Degree D1,Degree D2
WHERE P.person_id= Pa1.person_id and Pa1.factoid_id= F1.factoid_id and F1.factoid_id=HT1.factoid_id and HT1.factoidType_id=FT1.factoid_type and FT1.description= 'birth' and F1.factoid_id= TP1.factoid_id and TP1.place_id=PL1.place_id and PL1.place_id= IL.place_id and IL.zone_id= Z.zone_id and Z.zone_description= 'Italy'
---- Check if the zone which contains the birth-place is Italy ----
and F1.factoid_id= OA1.factoid_id and OA1.time_id= I.time_id and I.instant_id>1300 and I.instant_id <=1500 and P.person_id=Pa2.person_id and Pa2.factoid_id= F2.factoid_id and F2.factoid_id= HT2.factoid_id and HT2.factoidType_id= FT2.factoid_type_id and FT2.description= 'PhD' and F2.factoid_id= TP2.factoid_id and TP2.place_id=PL2.place_id and PL2.description= 'Bologna' and P.person_id= Pa3.person_id and Pa3.factoid_id= F3.factoid_id and F3.factoid_id= HT3.factoid_id and HT3.factoidType_id= FT3.factoid_type and ITO.factoidType_id1=FT3.factoid_type and ITO.factoidType_id2=FT4.factoid_type and FT4.description='curriculum'
---- Check if the factoid type which occurs earlier in Paris is a factoid subtype of the 'curriculum' factoid type ----
and F3.factoid_id= TP3.factoid_id and TP3.place_id = PL3.place_id and PL3.description='Paris' and OA2.factoid_id=F2.factoid_id and OA2.time_id=TI1.time_id and OA3.time_id=TI2.time_id and TI1.begin>TI2.finish and RT1.factoid_id=F2.factoid_id and S1.source_id=RT1.source_id and S1.reputation> 0.85 and RT2.factoid_id= F3.factoid_id and S2.source_id= RT2.source_id and S2.reputation >0.5
-- Check the reputation of the sources for PhD in Bologna and studies at Paris --
and D1.degree_id= TP2.degree_id and D2.degree id=TP3.degree id;
```

This query illustrates how we take into account the source reputation when evaluating a query related to the hierarchy of locations or of factoid types. The evaluation of this query on the Studium dataset returns the following results (extract).

Complete name	Birth date	Birth place	Confi-dence in the PhD place	Confi-dence in former studies	Reputa-tion of the sources
Castellanus Nicolai de Bunarellis			0.5	0.8	0.6
Faustus andrelinus	1462	For li	1.0	0.7	0.8
Bonaventura Badoer de Peraga	1332	Pa-doue	0.5	1.0	1.0
Laurentius de Bononia			0.8	1.0	0.6

Observe that the different confidence and reputation scores are set by experts who filled in the database. This illustrative example validates the ability of the generic model to be customized to a specific prosopographical project like Studium Parisiense as well as the usefulness of this representation to deal with certainty issues.

5. Conclusion and future research

This paper proposes a modelling approach to certainty-based prosopographical databases. It consists of three successive steps: building a generic certainty-based prosopographical conceptual model, customizing this generic conceptual model to a specific prosopographical project, and mapping the customized conceptual model to a relational database. Our contribution encompasses a generic conceptual model, two meta-models including uncertainty features, and a set of mapping rules. We illustrate the application of the approach with the Studium Parisiense prosopographical database and we propose two SQL queries demonstrating the ability of the approach to go beyond previous approaches. Observe that our approach is not dedicated to the sole History field but can also be deployed in other contexts using prosopographical approaches like societal studies, biology, tourism, etc.

Future research will confront our generic certainty-based conceptual model with more prosopographical projects in order to ensure its completeness. Since prosopography can be seen as a social network accommodating different uncertainty relationships of people, place, events and time periods which can be handled using probabilistic or fuzzy social networking approaches, we also plan to map our generic conceptu-

al model to a graph database in order to efficiently perform complex graph queries.

6. Acknowledgements

This research has been partly funded by a national French grant (ANR Daphne 17-CE28-0013-01.)

7. References

[1] Stone, L.: Prosopography, Daedalus, 100, 1, (1971), pp. 46-79

[2] Bulst N, Genet J-P.: Medieval lives and the historian: studies in medieval prosopography. In Int. Interdisciplinary Conf. on Medieval Prosopography (1986), pp. 1-16

[3] Pasin, M., Bradley, J.: Factoid-based prosopography and computer ontologies: towards an integrated approach, In: Digital Scholarship in the Humanities 30.1 (2013), pp. 86-97

[4] Figueira, L.,Vieira, M.: Modelling a Prosopography for the Roman Republic.
https://dh2017.adho.org/abstracts/091/091.pdf

[5] Studium Parisiense,
http://lamop-vs3.univ-paris1.fr/studium/

[6] Bradley, J., Pasin, M.: Annotation and Ontology in most Humanities research: accommodating a more informal interpretation context, NeDiMaH Ontology Workshop (2012)

[7] Westermann, U., Jain, R.: Toward a common event model for multimedia applications. IEEE Multimedia, 14(1), 19–29 (2007)

[8] Shaw, R., Larson, R.R.: Event representation in temporal and geographic context. In: Int Conf on Theory and Practice of Digital Libraries. Springer (2008)

[9] Liu, Y., McGrath, R. E., Wang, S., Pietrowicz, M., Futrelle, J., Myers, J.D.: Towards A Spatiotemporal Event-Oriented Ontology, In: Microsoft eScience Workshop (2008)

[10] Raimond, Y., Abdallah, S.: The Event Ontology, http://motools.sf.net/event/event.html (rdf)

[11] Gregersen, H., Jensen, C.S.: Temporal Entity-Relationship models-a survey. In: IEEE Transactions on knowledge and data engineering, 11 (3), 464-497 (1999)

[12] ISO 19108 :2002 - Geographic information-Temporal schema. https://www.iso.org/standard/26013.html

[13] Allen J.F.: Maintaining knowledge about temporal intervals.CACM, 26(11):832–843 (1983)

[14] Grüninger, M., Li, Z.: The Time Ontology of Allen's Interval Algebra. In: 24th Int. Symposium on Temporal Representation and Reasoning (TIME 2017)

[15] GENTECH. Genealogical data model: A comprehensive data model for genealogical research and analysis, http://xml.coverpages.org/GENTECH-DataModelV11.pdf

[16] Snodgrass, R.T.: The TSQL2 temporal query language. Vol. 330. Springer Science & Business Media, (2012)

[17] Moisuc, B., Miron, A., Villanova-Olivier, M., Gensel, J. : Spatiotemporal Knowledge Representation in AROM-ST. Innovative Soft. Development in GIS, 91-119 (2012)

[18] Bry, F., Marchiori, M.: Reasoning on the semantic web: beyond ontology languages and reasoners. In Eur. Work. on the Integration of Knowledge, Semantics and Digital Media Technology, pp. 317 – 321 (2005)

[19] Manola, F., Miller, E., McBride, B. : RDF primer, W3C recommendation, uazuay.edu.ec (2004)

[20] McGuinness, D.L., Van Harmelen, F.: OWL web ontology language overview, W3C recommendation (2004)

[21] National Research Council. Risk Analysis and Uncertainty in Flood Damage Reduction Studies. National Academy Press (2000)

[22] Li, Y., Chen, J., Feng, L.: Dealing with Uncertainty: A Survey of Theories and Practices, TKDE, 25(11), (2013)

[23] Thunnissen, D.: Uncertainty classification for the design and development of complex systems. In: 3rd Annual Predictive Methods Conference, Veros Software (2003)

[24] Costa, P.C.G., Laskey, K.B., Blasch, E., Jousselme, A.L.: Towards Unbiased Evaluation of Uncertainty Reasoning: The URREF Ontology. In Int. Conf. on Info Fusion (2012)

[25] Roblot, T.K, Link, S.: Cardinality Constraints with Probabilistic Intervals. In Conceptual Modeling ER (2017)

[26] Galindo, J., Urrutia, A., Piattini, M.: Fuzzy Databases: Modeling, Design and Implementation, Idea Group (2006)

[27] Ma, Z., Yan, L.: A Literature Overview of Fuzzy Conceptual Data Modeling, J. Information Science and Eng., vol. 26, pp. 427-441 (2010)

[28] Galindo, J., Urrutia, A., Piattini, M.: Representation of fuzzy knowledge in relational databases. In DEXA 2004.