HAL
open science

Conceptual modeling of prosopographic databases integrating quality dimensions

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

To cite this version:

Jacky Akoka, Isabelle Comyn-Wattiau, Stéphane Lamassé, Cédric Du Mouza. Conceptual modeling of prosopographic databases integrating quality dimensions. Journal of Data Mining and Digital Humanities, In press. hal-01966374v3

HAL Id: hal-01966374

https://hal.science/hal-01966374v3

Submitted on 18 Sep 2020 (v3), last revised 6 May 2021 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal of Data Mining and Digital Humanities

Conceptual Modeling of prosopographical Databases Integrating Quality Dimensions

Jacky Akoka[1,2], Isabelle Comyn-Wattiau[3], Stéphane Lamassé[4], Cédric du Mouza[1]

[1]Lab. CEDRIC, CNAM, Paris
[2]Institut Mines-Télécom Business School, Paris
[3]ESSEC Business School, Paris
[4]Lab. PIREH, University of Paris I, Paris

Abstract

Prosopographical databases, which allow the study of social groups through their bibliography, are used today by a significant number of historians. Computerization has allowed intensive and large scale exploitation of these databases. The modeling of these proposopographic databases has given rise to several data models. An important problem is to ensure a level of quality of the stored information. In this article, we propose a generic data model allowing to describe most of the existing prosopographical databases and to enrich them by integrating several quality concepts such as certainty, credibility, consistency, and completeness. The first two criteria (certainty, credibility) are stored in the database whereas the last two criteria are derived.

Keywords

Conceptual modeling, prosopographicalal database, quality

I. INTRODUCTION

Prosopography is a research method for studying a social group by comparing the biographical itineraries of each of its members. Its aim is to understand how the groups operate, without neglecting the singular trajectories. Prosopography is based on a precise, documented investigation of each individual in the determined population. In history, it is thanks to a methodology and an advanced erudition that all the traces that will constitute the record of each person are collected. All historical periods use this method of investigation. The word "prosopographia" appears in the 16th century. This research method is used by historians to answer research questions such as *"Is there a link between disciplines (arts, medicine, canon law, theological law) and geographical origin and within these disciplinary fields, is there a link between grades and geographical origin?"* or *"What is the nature and quantity of contentious cases in which Parisian academics are involved in the thirteenth and fourteenth centuries?"*.

Prosopography is a research method for studying a social group by comparing the biographical itineraries of each of its members. Its aim is to understand how the groups operate, without neglecting the singular trajectories. Prosopography is based on a precise, documented investigation of each individual in the determined population. In history, it is thanks to a methodology and an advanced erudition that all the traces that will constitute the record of each person are collected. All historical periods use this method of investigation. The word "prosopographia" appears in the 16th century. This research method is used by historians to answer research questions such as *"Is there a link between disciplines (arts, medicine, canon law, theological law) and geographical origin and within these disciplinary fields, is there a link between grades*

and geographical origin?" or *"What is the nature and quantity of contentious cases in which Parisian academics are involved in the thirteenth and fourteenth centuries?"*.

Quantitative analysis and computer science profoundly transformed its methodology in the 20th century. Many periodicals have been interested in this aspect and published articles and special issues on this theme. Many historians have even proposed dedicated software developments. We can particularly mention the contribution of database systems that allowed, for example, to address very concretely the "sourcing" of information. Using only paper cards, it is difficult to index every fact constituting a person?s career with the document that allowed to establish it, while this is no longer the case with databases. In the same way, it is not easy to manage contradictory information. However, it is possible that two different documents provide contradictory information on an individual. Additionally, another characteristic inherent in historical data is their unequal quality. While some data are accurate and proven by multiple sources, many data are missing, inaccurate, or appearing in sources known to have low credibility. This has a significant impact on the formulation of hypotheses that historians attempt to verify. So they can query the prosopographical database to determine for instance *What is the degree of reliability and precision of the curriculum (curriculum known from the baccalaureate to the higher grades in a complete, incomplete, or purely hypothetical way) according to the status of the Parisian masters and students ('Student', 'Graduate' and 'Master')*.

Nowadays, the problem arises with another acuity in humanities since the Web becomes also a way of study, as evidenced by the project *Traces through Time: Prosopography in practice across Big Data* [Mark Bell, 2015]. In this article we propose a conceptual model to describe in a general and enriched way the information contained in a database of prosopographical data. We then study how this model can be instantiated with the PASE database [Bradley and Short, 2005], STUDIUM PARISIENSE [Genet et al., 2016] and PADU-A database [Gallo, 2018].

Our article is structured as follows. After a state-of-the-art about the digitization of prosopographicalal databases and the management of the quality for historical data in Section II, we present our generic conceptual model for prosopographicalal data which encompasses temporal and quality management in Section III. In Section IV we illustrate the genericity of our model by describing the different mappings for the concepts of three different prosopographicalal databases to our generic model. Section VI concludes the paper and presents some future work.

II STATE OF THE ART

prosopographical databases and computer science

The use of prosopographical databases has become widespread among researchers in history since the 70's, transforming much of their research approach [Keats-Rohan, 2000]. Although this phenomenon coincided with the rise of computer science, both sciences have evolved without interaction for a long time, despite the visionary approach of Karl Ferdinand Werner who first in 1977, with his PROL project [Werner, 1977], realized the contribution of computer science as a tool for prosopography researchers. The increasing volume of recorded data makes their exploitation (the analysis and the cross-referencing of data) extremely time-consuming. Using a database approach has emerged as one of the solutions to this volumetry problem, for example in COEL [Keats-Rohan, 1998], PASE [Bradley and Short, 2005], ASFE [Brizzi, 2014], RAG [Schwinges, 2015], PADU-A [Gallo, 2018] or STUDIUM PARISIENSE [Genet et al., 2016] projects.

Moving from a collection of paper cards to databases first involves thinking about a data model. Among the proposed data models, we will distinguish relational models, semi-structured

models, and network models. The first proposals for prosopographical databases relied on the relational model [Keats-Rohan, 1998, Bradley and Short, 2005]. In [Tchounikine et al., 2018], the authors propose OLAP analyzes and network analyzes associated with cartographic and chronological visualization tools to analyze the careers and the shared relation networks. Recent work [Bol, 2012] propose the use of geographic information systems, supported by relational databases, in order to detect for example spatial patterns.

This structured representation enables to perform efficient search queries crossing a limited number of tables. Semi-structured representation, in addition to its contribution to semantics, allows to limit join operations by exploiting the tree structure. It allows thus multivalued attributes and the integration of (semi-)structured objects within a (semi-)structured object. It is therefore adapted to the prosopographical databases where an element "person" can be composed of the elements "production", "diploma", etc., being themselves structured elements. The STUDIUM PARISIENSE [Genet et al., 2016] and PROSO [Barabucci and Zingoni, 2013] projects are two examples of such a choice of representation. If the semi-structured model allows structurally to represent links between people / objects / places / facts, it makes it difficult to query more complex links between elements.

For this reason, recent works apply the "social networks" type of representation for example [Graham and Ruffini, 2007, Verbruggen, 2007, Jackson, 2016, 2017]. This approach allows the search of data to discover links between people / objects / places / facts, or recurring patterns. Recent approaches also rely on linked data representation like [Tuominen, 2016] which is an event-based, person and role-centric model for representing the activities a person has participated in during his life.

Main concepts of prosopographical databases

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. Generally, the event-based approach is used to model life stories of a group of persons Westermann and Jain [2007]. The latter can take different roles. Events are linked to other events, persons, places, time periods, and documents. Shaw and Larson [2008] distinguishes different types of events, supporting both discrete and continuous events, and expressing various temporal aspects of events. Several ontologies describing events have been proposed, see Liu et al. [2008].

Representing temporal data is a problem facing historians. Time can be the source of vagueness and/or uncertainty. Temporal database research, see Gregersen and Jensen [1999], considers two types of data: "instant" and "interval" GIT-Schema [2002]. Allen Allen [1983] proposes a time model based on time intervals. The GENTECH model GENTECH [2011] supports the creation of conflicting temporal relationships expressing different points of view. The time model in AROM-ST Moisuc et al. [2012] offers several time types including instant, interval, multiInstant, and multiInterval types. A variety of approaches have been proposed to represent temporal information in RDF Manola et al. [2004] and OWL McGuinness and Van Harmelen [2004].

Uncertainty is defined as "a general concept that reflects our lack of sureness about something or some-one", see Council [2000]. 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 Li et al. [2013]. In the URREF ontology da Costa et al. [2012], 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. In Barroso et al. [2019], the authors rely on the Design Science Research, which directs the construction of an artifact in a given context, whose theoretical conjectures are based on the search and production of knowledge. This approach allows to contribute to the knowledge base, and to deliver reliable and relevant information about the life of a politician. 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. Roblot and Link [2017]. Pasin and Bradley [2015] presents HiCO, an ontology which aims to outline relevant issues related to the workflow for stating, and formalizing, authoritative assertions about context information. It particularly focuses on the different interpretations of a cultural object which highly depend on this context.

Quality management of historical data.
One of the important issues of databases in general, and prosopographical databases in particular, is the quality of the information stored. Data quality is a field of research in itself. Numerous contributions have categorized quality issues, as well as metrics to measure the extent of these issues and methods and tools to improve it (see a large survey in [Batini and Scannapieco, 2016]). The latter refers to ISO standard data quality dimensions (ISO/IEC 25012 :2008). It encompasses fifteen quality dimensions including completeness, consistency, credibility, and precision (see Table 1).

Data Quality characteristic	Definition
Completeness	subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use
Consistency	the degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use
Credibility	the degree to which data has attributes that are regarded as true and believable by users in a specific context of use
Precision	the degree to which data has attributes that are exact or that provide discrimination in a specific context of use

Table 1: ISO standard data quality dimensions (an excerpt)

For reasons of space, we focus our state of the art on the precision factor, which is only one aspect but it seems to be particularly relevant in the context of social science.

[Matousek et al., 2007] propose the following categorization of imprecise temporal assertions:

1. Accurate assertions where all data is available and where maximum accuracy is reached,
2. Assertions with a lower fine granularity, when data are available but less precise,
3. Incomplete assertions where some information is missing for accurate identification,
4. Uncertain assertions with an absolute specification of uncertainty,
5. Uncertain assertions with a relative specification of uncertainty,
6. Assertions referring to other assertions containing temporal properties,
7. Assertions with unknown or missing information.

Plewe [2002] proposes a model on the nature of uncertainty, specifically for thematic, spatial and temporal representation of geo-historical phenomena. The goal is to provide a framework for spatiotemporal data modeling in a historical setting.

To the best of our knowledge, there is no prosopographical database incorporating the representation of uncertain information at the model level. Some systems, such as STUDIUM PARISIENSE, insert marks (mainly the question mark, or natural language) to alert the user about the uncertain nature of the information. However, this home-made representation does not allow the evaluation of the certainty associated with the corresponding information.

A main advantage of our approach is to represent explicitly the measures of uncertainty, confidence, time, and precision attributes attached to all prosopographical concepts. The model is presented and described in the next section.

III CONCEPTUAL MODELING OF PROSOPOGRAPHICAL DATABASES

The model proposed in this article, and presented in Figure 1, (4) is a conceptual model that includes Source, Person, Place, Time, Factoid (or Event or Fact or Assertion, or State or Trait). It also includes uncertainty to deal with contradictory sources and represent the reliability of information.

This conceptual model must be validated with historians' queries such as:

- Who studied canon law in Paris at the same time as Petru de Quercu and then got an ecclesiastic position?
- Who are the Italian living in the fourteenth or fifteenth century who studied a PhD degree in Bologna after studies in Paris?

This model has the advantage of being generic. It puts together and makes more generic the information contained in different prosopographical databases, namely the concepts of persons, factoids, places, and sources. It also incorporates a broad representation of uncertainty. We summarize in the following the different contributions of our proposal.

1. The notion of **factoid** is taken in a broad sense. It includes the factoids of certain prosopographical representations, but also all the facts that characterize individuals. It is a piece of information that becomes accepted as a fact even though it is not actually true. It also can be considered as an invented fact believed to be true because it appears in print. It may represent a state, a trait, an event. Factoids may be linked together. Persons play roles in factoids that can belong to categories (FactoidTypes) which are specific to research projects. For example a publication is also a factoid. This choice to generalize the event makes the model compact without losing the wealth of information that can be represented. However, it led us to define the factoid with a larger number of dimensions. For example, 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, etc.

2. The dimensions of all prosopographical concepts including factoids are associated to **hierarchical** repositories. For example, places, sources, people and factoids are generalized to one or more levels. Factoids are grouped into types of factoids, like in PASE where confession is a factoid of Christian piety, itself a religious act. This

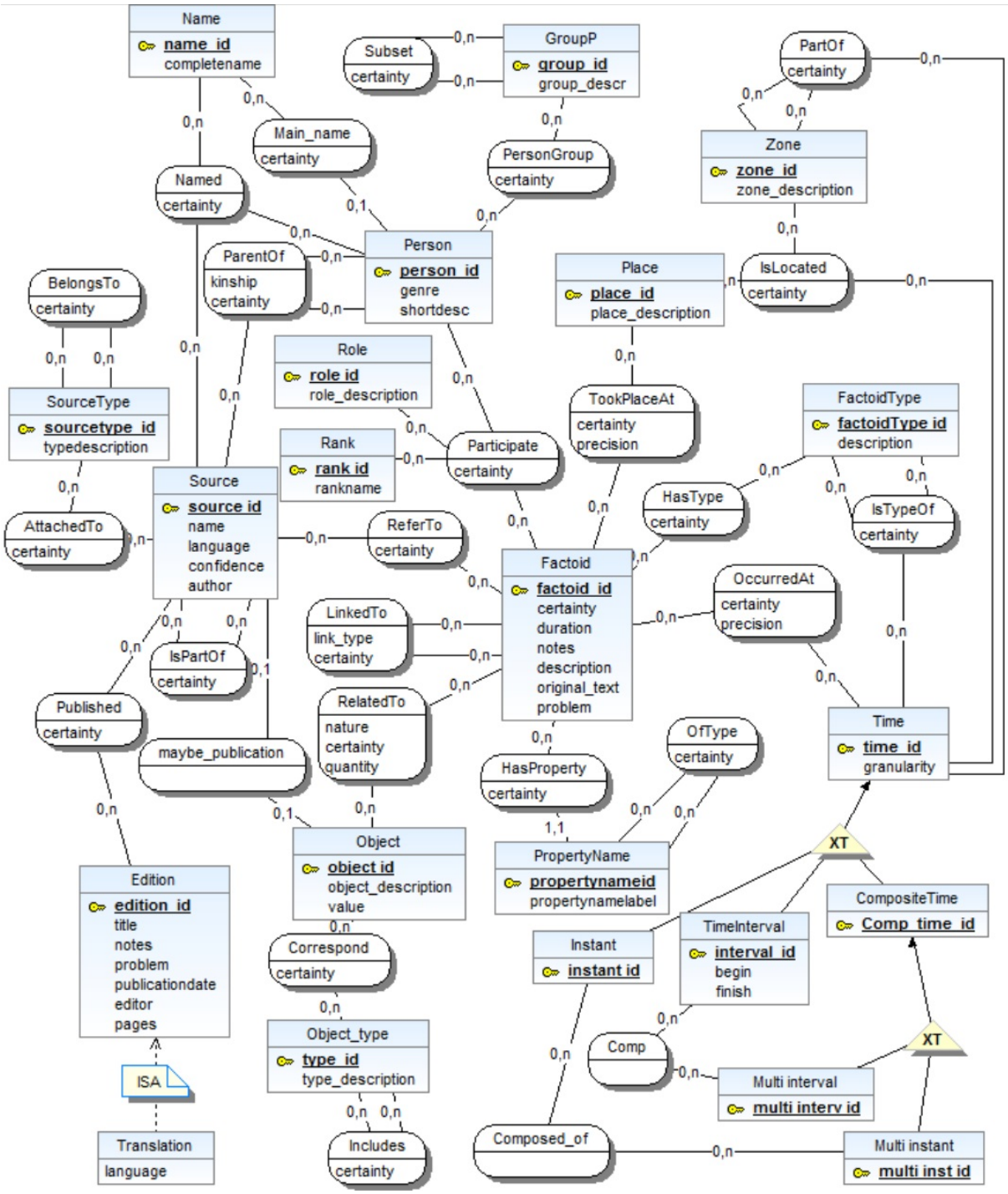

Figure 1: Our generic conceptual model for a prosopographical database

aggregation mechanism incorporates time as a dimension since this categorization may also vary over time. Time characterizes every factoid. Factoid types may also depend on time. Geographic elements also varies over time (their name, their boundaries, etc).

3. Depending on the area targeted by the prosopographical database, the **names of individuals** may be known imprecisely. So our model includes a representation of several names , since People may be known by different names. Each one is 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). Every known potential name is associated with the person with a measure of the certainty, if it is available. The representation of different names of people allows to have several names with a certainty associated with each.

4. Some relationships between concepts are typed, in the sense that a **Type** attribute describes them. For example, the attribute *nature* between the factoid and the object makes it clear that, during a barter event, an object is assigned, and an object is granted in exchange. This *nature* attribute can take the value "dowry" at a wedding. Between factoids, information "link_type" allows to define a set of dependencies between factoids as "precedes", "causes", etc. The role of a person in a factoid is also a type that has been represented in the form of an entity to the extent that the same person can sometimes play multiple roles in the same factoid.

5. The representation of **time** integrates discrete time (a date), continuous time (an interval) and their composition (several potential dates, or several possible intervals, or several cumulative intervals, for example "he was present from 1492 to 1500 then from 1503 to 1508"). It is adapted from AROM-ST model [Moisuc et al., 2012].

6. Finally, it integrates the management of uncertain information into three forms: a degree of **certainty**, **confidence** and **precision**. 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 range [0,1]. **Confidence** is a shared feature of information as measured by a degree between 0 and 1. In this model, we have restricted its use to the characterization of sources of information, as this is the main information available. Historians rely on many sources and their experience allows them to associate to each source a confidence that results from this experience. An example of uncertainty is, for example, when two documents give a different information related to the date for example using terminus ante quem or post quem. All documents concerning Johannes Vitalis allow us to say that his activity is between 1380 and 1395. He is known as Franciscan, a beggar order. We know that he was a bachelor, a graduate in theology. He is quoted as a Doctor of Theology in a request for forgiveness between September 8 and 11, 1390 of another Dominican brother Johannes Nicolai. So we can think that he got his rank before this moment. Then we find him at the trial of Jean Blanchard and in the convocation of the students in theology for the trial where he is quoted as a Dominican, which is probably a mistake. **Precision** is a representation of approximate information. For example, accuracy may be relative to the location of an event. The values it can take in this case are: *near, around, not far from, a few kilometers from, etc*. When it characterizes the moment when an event takes place, it can take the values of: *around, before, well before, shortly after, etc.*

This generic model makes it possible to cover the information contained in PASE (except for traces), in STUDIUM PARISIENSE and in PADU-A.

IV MAPPING TO PASE, STUDIUM PARISIENSE AND PADU-A

The aim of this section is to illustrate the genericity of our model. To this end, we performed a mapping process from our model to the models underlying three different existing prosopographical projects, namely PASE, STUDIUM PARISIENSE and PADU-A.

The mapping process is based on an alignment mechanism such as the ones implemented in all schema matching tools [survey trouvé]. However, given the limited size of the models under

consideration and the relative complexity of the objects described in these models, we applied manually this alignment mechanism. For each of the three mappings, the following steps were applied :

- Comparing two by two the central objects of the two models, namely person, place, time, source, etc. For each pair, similar properties were matched, specific properties (present only in the three prosopographic projects) were identified.
- The other homonymous objects in both models were compared in the same way.
- By browsing the model, the remaining objects were then compared in pairs to detect synonyms, for example Object in the generic model and Possession in PASE. For each pair thus detected, similar properties are matched and specific properties are identified.
- Finally, the remaining objects in the prosopographic project are identified and constitute the too specific part of the project which could not be taken into account in the generic model. We briefly describe below the application of this process to the three prosopographical projects.

The *Prosopography of Anglo-Saxon England (*PASE*)* [1] is a database which aims to provide structured information relating to all the recorded inhabitants of England from the late sixth to the late eleventh century. It is based on a systematic examination of the available written sources for the period, including chronicles, saints' Lives, charters, libri vitae, inscriptions, Domesday Book and coins, etc. PASE is based in the Department of History and the Centre for Computing in the Humanities, at King's College, London, and in the Department of Anglo-Saxon, Norse, and Celtic, at the University of Cambridge.

Table 2 presents the comparison of our model with PASE. The first two columns designate the entity or relationship in our model and the associated property. The last two designate the table and the corresponding column in PASE. For example, the groups of people in our model correspond to the types represented in the table *alfactoidpersontype*. This effort to match two models allowed us to verify that our model incorporates all the information from PASE. Moreover, the addition of certain dimensions to factoids improves the representation of the information. For example, the *OBJECT* entity that allows structuring the description of certain factoids (graduation, marriage, etc.) avoids the description in natural language of unstructured fields, more difficult to exploit by queries.

In the same way, Table 3 compares some STUDIUM PARISIENSE topics and their alternative representation in our model. The STUDIUM PARISIENSE database is an online database that has been developed by the LAMOP laboratory [2]. It includes the students and teachers of the schools and the University of Paris since the appearance of the cathedral school at the end of the XIth century until 1500. Each individual is described by a structured sheet which gives all the known biographical information (origin, university curriculum, ecclesiastical career, place of residence, writings (more than 10% of the individuals are authors)). Currently STUDIUM PARISIENSE consists of 15,000 records - some are brief, but others represent nearly 100 printed pages, 7500 of which are online, and in the future there should be more than 40,000. We made the comparison between our model and that of STUDIUM PARISIENSE. Thus, the variants of the name that STUDIUM PARISIENSE allows are represented, in our model, by the relation `Named` between persons and names. The activity period of STUDIUM PARISIENSE is represented by a factoid of type `Activity` with a start date and an end date. The median of activity is an information calculated from these dates. The `status` of a person in STUDIUM PARISIENSE is their

[1] http://www.pase.ac.uk/

[2] http://lamop-vs3.univ-paris1.fr/studium/

Object	Property	PASE object	PASE property
GroupP	group id	alfactoidpersontype	alfactoidpersontypekey
GroupP	group title	alfactoidpersontype	alfactoidpersontype
Name	name id	Person	headname
Name	complete name	Person	descriptionname
Zone	zone id	allocation	allocationkey
Zone	zone description	allocation	allocation
SourceType	source type id	alsourcetype	alsourcetypekey
SourceType	typedescription	alsourcetype	alsourcetype
Person	person id	Person	personkey
Person	genre	AlGender	AlGenderAbrv
Person	shortdesc	alfactoidpersonrank	alfactoidpersonrank
Place	place id	factoidlocation	factoidlocationkey
Place	place description	factoidlocation	alplace
Objet	object id	Possession	possessionkey
Objet	object description	Possession	description
ObjetType	type description	alpossessiontype	alpossessiontype
Source	source id	Source	sourcekey
Source	source name	Source	sourcetitle
Source	author	Source	author
Source	language	alLanguage	allanguage
Source	confidence	Archivequality	archivequalityname
SourceType	typedescription	Source	description
Edition	edition id	Editioninfo	editioninfokey
Edition	title	Editioninfo	articletitle
Edition	editor	Editioninfo	editor
Factoid	factoid id	Factoid	factoidkey
Factoid	description	Factoid	shortdesc

Table 2: Extract of the mapping between our model and PASE

role in our model. The information *Bachelor es arts (Paris) 1460* in STUDIUM PARISIENSE corresponds to a *graduation* factoid taking place in Paris in 1460.

STUDIUM PARISIENSE field	its representation in our model
Name variants	are linked to the corresponding person by the `named` relationship
Activity period	represented by the `Activity` event with a start date and an end date
Activity medium	computed
Status	it is the rank of the person
Origin	it is the `Origin` event which takes place in a location
Bachelier ès arts (Paris) 1460	it is the diplomation event with a location and a date

Table 3: Examples of mapping between our model and STUDIUM PARISIENSE

Finally Table 4 represents the mapping between the PADU-A concepts and the ones we proposed in our generic model. The prosopographicalal-Access-Database of University-Agenda

project (PADU-A) [3] intends to put the bases of a prosopographicalal data bank in order to make available the data related to the students and teachers from the first two centuries of the Padua University (1222-1405). The work starts from the sources published in press and completes with the contribution of other unpublished works. It aims essentially at being a useful tool for historians investigating specific questioning fields related to backgrounds, careers and disciplinary areas of students and teachers.

We observe that several concepts with a spatial and temporal information are mapped in our model to the `Factoid` entity. PADU-A database also manages the onomastics through the `Individui` relationship associated to the `AttNomi` relationship. These two concepts are covered by our `Person` and `Name` entities along with the `Main_name` relationship. The PRODUZINT table which stores all the information about the production (written or not) of a student or a teacher corresponds to the OBJECT entity associated to FACTOID which represents the event of production. The nature of the production can be precised thanks to the OBJECT_TYPE entity.

PADU-A relation	Representation in our generic model
INDIVIDUI	PERSON entity along with the NAME entity and the relationship MAIN_NAME
ATTNOMI	NAME entity and the relationship NAMED
ATTQUALIFICHE	GROUPP entity with the recursive relationship SUBSET
TITOLIUNIV	FACTOID entity associated to the FACTOIDTYPE entity with description value set to academic degree, and associated to INSTANT entity for graduation date
ATTPOSUNIV	FACTOID entity associated to the FACTOIDTYPE entity with description value set to academic position, and associated to TIMEINTERVAL
ATTASSOCIAZIONI	GROUPP entity with the recursive relationship SUBSET
ORIGINE	FACTOID entity for the birth event associated to the PLACE which is connected to the ZONE entity and its recursive PARTOF relationship
FAMIGLIA	PERSON entity along with its recursive relationship PARENTOF with its *kinship* attribute
RESIDENZA	FACTOID entity for the "reside" event associated to the PLACE which is connected to the ZONE entity and its recursive PARTOF relationship
ALTREPERSONE	PERSON entity
SOURCE	SOURCE entity associated to the SOURCETYPE entity
PRODUZINT	OBJECT entity associated to FACTOID corresponding to the production
BIBLIOGRAFIA	OBJECT entity associated to FACTOID corresponding to the writing and to OBJECT_TYPE entity to written work
EVENTI	FACTOID entity associated to FACTOIDTYPE, PLACE and TIME entities

Table 4: Extract of the mapping between our model and PADU-A

[3] https://www.dissgea.unipd.it/padu-prosopographicalal-access-database-university-agenda-verso-una-banca-dati-di-studenti-e-docenti

Our approach has the advantage of offering a generic model for all these databases, which makes it possible to pool development and maintenance efforts. Thus, the different communities of historians would each have their specific base (PASE, STUDIUM PARISIENSE, PBW, etc.), which would result from the adaptation of this generic model to their research needs. In addition, the management of uncertain information allows a query of better quality, associating each answer with certainty. Moreover, our generic model may work as a pivot model making possible interoperability of the various existing bases.

V EVALUATION OF THE APPROACH

The main contribution described in this paper is the generic conceptual model. The previous section made it possible to show the genericity of the model in the sense that it could be put in mapped to the underlying models of three prosopographical projects. There are many approaches and criteria used to perform the evaluation of a conceptual model. Shanks et al. [2003] proposes four criteria that such models must meet: accuracy, completeness, conflict free, and no redundancy. Validation approaches mainly include test with transactions and review with users. They also mention that many rules have been proposed but they are not generalizable since they highly depend on the context and the objective of the conceptualization effort. Rittgen [2010] lists many criteria using the framework of Lindland and Krogstie [1993] which differentiates between syntactic, semantic, pragmatic, and social qualities. Prat et al. [2015] considers that a design science approach generally results in a set of artifacts constituting a system and propose to validate the properties of this system. Pfeiffer and Niehaves [2005] mentions Guidelines of Modeling (GoM) as a list of requirements that models must meet: construction adequacy, language adequacy, economic efficiency, clarity, comparability, and systematic design. These guidelines were first described in Schütte and Rotthowe [1998]. We propose to use their framework to check the quality of our generic model.

GoM differentiates between necessary principles (construction adequacy, language adequacy, economic efficiency) and supplementary principles (systematic design, comparability, clarity).

Principle of *construction adequacy* seeks to achieve consensus about the problem definition and about the model representation. Consensus was achieved by the team of researchers, which is composed of analysts (two conceptual modeling researchers, one computer scientist) and a historian playing the role of user.

The second necessary principle is *language adequacy*. It includes language correctness as well as its suitability. We built our generic model using a modeling tool, which ensures that the resulting model conforms to the underlying meta-model, in terms of consistency and completeness. Consistency results from the fact that the tool constrains the representation in terms of concepts. Completeness is achieved when the model is saved, which includes the validation of structural properties, such as the obligation to associate a name with each concept, an identifier with each entity, entities participating in each relationship, etc.

The third necessary principle is *economic efficiency*. The generic model was designed with the objective of pooling the design effort for reuse in several prosopographical projects. In addition, conceptual modeling, based on a semantically rich language, makes it possible to reduce the cost of subsequent modifications. Its purpose is to obtain, very early in the database design process, a means of validating the coverage of user needs by the to-be system, upstream of any implementation.

The principle of *systematic design* is relevant in the context of multi-model design, measuring in particular inter-model consistency, which is not our purpose.

The principle of *comparability* aims at the semantic comparison of two models. The mapping described in the previous section consisted of the systematic comparison of our model with three models of prosopographical projects. It was made difficult by the unavailability of their corresponding conceptual models. Therefore, we had to deduce the concepts from their logical models. Let?s note that the comparability can be made more difficult when the size of the models is important, which is not our case.

Finally, the principle of *clarity* is broken down into three properties: hierarchy, layout design and filtering. The hierarchy concerns the logic of interaction between models which is not relevant in our context where we have only one model. The layout design is achieved through the use of the modeling tool. Finally, the filtering capacity obtained thanks to the model was verified by the following process. We first customized the generic model to the context of Studium, then we generated the relational schema and finally, we executed the following two queries:

- In the first one, we compare two careers as follows: Who studied canon law in Paris at the same time than Petru de Quercu and got an ecclesiastic position after? 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 second query looks for more complex career patterns and considers the source reliability (estimated by historians): Who are the Italian from the XIV or XVth century who studied a PhD degree in Bologna after studies in France, according to sources with a reliability greater than 0.85? This query illustrates how we consider the reliability of the sources when evaluating a query and how the hierarchy of locations or of factoid types (here for diploma which is a subtype of curriculum).

This informed argument allows us to validate our conceptual model. It offers richer semantics for prosopographical historians. The prototype developed to test its applicability to Studium has shown its usefulness. It can also be used as a pivot model between prosopographical projects. Finally, it meets the requirements of the Guidelines of Modeling (GoM) which are one of the reference approaches for the evaluation of conceptual models.

VI CONCLUSION

prosopographical databases are an indispensable tool for many history researchers who have turned their attention to computers in order to quickly realize many tedious treatments. This digitization of prosopographical data has led to the emergence of many data models. This article proposes a generic conceptual model covering the concepts and relationships between concepts present in different models (we have seen that this model generalizes and enriches those of PASE, STUDIUM PARISIENSE and PADU-A for example), but it is distinguished by its representation of data quality, such as uncertainty, completeness, reliability, represented by the attributes certainty, confidence, and precision. Our future research will consist in validating the model by confronting it to other references in the field of prosopographical databases. It will also include checking its applicability by transforming it into a logical and physical model (relational, graph or document for example). This article has put forward the representation of uncertainty, enriching the possibilities offered by prosopographical databases. Future research will be dedicated to the definition of different modes for aggregating these representations of the uncertain.

References

James F. Allen. Maintaining Knowledge about Temporal Intervals. *Commun. ACM*, 26(11):832–843, 1983.

Gioele Barabucci and Jacopo Zingoni. PROSO: prosopographic records. In *Proc. Intl Work. on Collaborative Annotations in Shared Environment, DH-CASE@DocEng*, pages 3:1–3:7, 2013.

José S. Barroso, Júnior, Mariano Pimentel, Vanessa Nunes, and Claudia Cappelli. Design Science Research to Design a Conceptual Model About Prosopographic Information Related to Politicians. In *Proc. XV Brazilian Symp. on Information Systems (SBSI)*, pages 24:1–24:8, 2019.

Carlo Batini and Monica Scannapieco. *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer, 2016.

Peter K. Bol. GIS, Prosopography and History. *Annals of GIS*, 18(1):3–15, 2012.

J Bradley and Harold Short. Texts into Databases: The Evolving Field of New-style Prosopography. *Literary and Linguistic Computing*, 20(Suppl 1):3–24, 2005.

Gian Paolo Brizzi. Asfe: une Base de Données pour Trois Projets. In *Eur. Work. on Historical Academic Databases*, 2014.

National Research Council. Risk analysis and uncertainty in flood damage reduction studies. national academy press, 2000.

Paulo Cesar G. da Costa, Kathryn Blackmond Laskey, Erik Blasch, and Anne-Laure Jousselme. Towards unbiased evaluation of uncertainty reasoning: The URREF ontology. In *Proc. Intl. Conf. on Information Fusion FUSION*, pages 2301–2308, 2012.

Donato Gallo. Padu-a: Prosopographical-access-database of university-agenda. Technical report, University of Padova, 2018.

Jean-Philippe Genet, Hicham Idabal, Thierry Kouamé, Stéphane Lamassé, Claire Priol, and Anne Tournieroux. General Introduction to the Studium Project. *Medieval Prosopography*, (31):156–172, 2016.

GENTECH. Genealogical data model: A comprehen-sive data model for genealogical research and analysis. `http://xml.coverpages.org/GENTECH-DataModelV11.pdf`, 2011.

GIT-Schema. Geographic information-Temporal schema, ISO 19108 :2002 . `https://www.iso.org/standard/26013.html`, 2002.

Shawn Graham and Giovanni Ruffini. Network Analysis and Greco-Roman Prosopography. In *Prosopography Approaches and Applications. A Handbook.*, pages 325–336. K.S.B. Keats-Rohan, (ed.), 2007.

Heidi Gregersen and Christian S. Jensen. Temporal Entity-Relationship Models - A Survey. *IEEE Trans. Knowl. Data Eng.*, 11(3):464–497, 1999.

Cornell Jackson. Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland. *Digital Scholarship in the Humanities*, 32:fqv070, 02 2016.

Cornell Jackson. Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland. *Digital Scholarship in the Humanities*, 32(2):336–343, 2017.

K. S. B. Keats-Rohan. Prosopography and Computing: a Marriage Made in Heaven? *History and Computing*, 12: 1–12, 2000.

K.S.B. Keats-Rohan. Historical Text Archives and Prosopography: the COEL Database System. *History & Computing*, 10(1-2-3):57–72, 1998.

Yiping Li, Jianwen Chen, and Ling Feng. Dealing with Uncertainty: A Survey of Theories and Practices. *IEEE Trans. Knowl. Data Eng.*, 25(11):2463–2482, 2013.

Odd Ivar Lindland and John Krogstie. Validating conceptual models by transformational prototyping. In *Proc. Intl Conf. on Advanced Information Systems Engineering (CAiSE)*, volume 685 of *Lecture Notes in Computer Science*, pages 165–183. Springer, 1993.

Yong Liu, Robert E Mcgrath, Shaowen Wang, Mary Pietrowicz, Joe Futrelle, and James Myers. Towards A Spatiotemporal Event-Oriented Ontology. In *Microsoft eScience Workshop*, 2008.

F. Manola, E. Miller, and B. McBride. Rdf primer, w3c recommendation, uazuay.edu.ec, 2004.

Sonia Ranade Mark Bell. Traces through Time: a Case-study of Applying Statistical Methods to Refine Algorithms for Linking Biographical Data . In *Proc. Intl. Conf. on Biographical Data in a Digital World*, pages 24–32, 2015.

Kamil Matousek, Matrin Falc, and Zdenek Kouba. Extending Temporal Ontology with Uncertain Historical Time. *Computing and Informatics*, 26(3):239–254, 2007.

D.L. McGuinness and F. Van Harmelen. Owl web ontol-ogy language overview, w3c recommendation, 2004.

Bogdan Moisuc, Alina Dia Miron, Marlène Villanova-Oliver, and Jérôme Gensel. Spatiotemporal Knowledge Representation in AROM-ST. In *Innovative Software Development in GIS*, pages 91–119, 2012.

Michele Pasin and John Bradley. Factoid-based prosopography and computer ontologies: towards an integrated approach. *Digital Scholarship in the Humanities (DSH)*, 30(1):86–97, 2015.

Daniel Pfeiffer and Björn Niehaves. Evaluation of Conceptual Models - A Structuralist Approach. In Dieter Bartmann, Federico Rajola, Jannis Kallinikos, David E. Avison, Robert Winter, Phillip Ein-Dor, Jörg Becker, Freimut Bodendorf, and Christof Weinhardt, editors, *Proc. Europ. Conf. on Information Systems, Information Systems in a Rapidly Changing Economy, (ECIS)*, pages 459–470, 2005.

Brandon Plewe. The Nature of Uncertainty in Historical Geographic Information. *Trans. GIS*, 6(4):431–456, 2002.

Nicolas Prat, Isabelle Comyn-Wattiau, and Jacky Akoka. A Taxonomy of Evaluation Methods for Information

Systems Artifacts. *J. Manag. Inf. Syst.*, 32(3):229–267, 2015.

Peter Rittgen. Quality and perceived usefulness of process models. In *Proc. ACM Symposium on Applied Computing (SAC)*, pages 65–72. ACM, 2010.

Tania Katell Roblot and Sebastian Link. Cardinality Constraints with Probabilistic Intervals. In *Proc. Intl. Conf. on Conceptual Modeling ER*, pages 251–265, 2017.

Reinhard Schütte and Thomas Rotthowe. The Guidelines of Modeling - An Approach to Enhance the Quality in Information Models. In Tok Wang Ling, Sudha Ram, and Mong-Li Lee, editors, *Proc. Intl Conf. on Conceptual Modeling (ER*, volume 1507 of *Lecture Notes in Computer Science*, pages 240–254. Springer, 1998.

Rainer C. Schwinges. Das Repertorium Academicum Germanicum (RAG). Ein digitales Forschungsvorhaben zur Geschichte der Gelehrten des alten Reiches (1250-1550). In *Jahrbuch für Universitätsgeschichte*, pages 215–232, 2015.

Graeme G. Shanks, Elizabeth Tansley, and Ron Weber. Using ontology to validate conceptual models. *Commun. ACM*, 46(10):85–89, 2003.

Ryan Shaw and Ray R. Larson. Event Representation in Temporal and Geographic Context. In *Proc. Europ. Conf. on Research and Advanced Technology for Digital Libraries ECDL*, pages 415–418, 2008.

Anne Tchounikine, Maryvonne Miquel, Thierry Pécout, and Jean-Luc Bonnaud. Prosopographical data analysis. Application to the Angevin officers (XIII-XV centuries). *Journal of Data Mining & Digital Humanities (JDMDH)*, 2018, 2018.

Jouni Tuominen. Emlo prosopographical data model: Towards a biographical conceptual reference model. Technical report, Cost Action IS1310, Reassembling the Republic of Letters, Aalto University, 2016.

Christophe Verbruggen. Combining Social Network Analysis and Prosopography. In *Prosopography Approaches and Applications. A Handbook*, pages 579–601. Linacre College, 2007.

Karl-Ferdinand Werner. Problèmes de l'Exploitation des Documents Textuels Concernant les Noms et les Personnes du Monde Latin (IIIe-XIIe siÃ¨cles). In *Informatique et Histoire Médiévale*, pages 205–212, 1977.

Utz Westermann and Ramesh Jain. Toward a Common Event Model for Multimedia Applications. *IEEE MultiMedia*, 14(1):19–29, 2007.