Conceptual modeling of prosopographic databases integrating quality dimensions

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

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, France

2 Institut Mines-Télécom Business School, France

3 ESSEC Business School, France

4 Lab. PIREH, University of Paris I, France

*Corresponding author: Cédric du Mouza dumouza@cnam.fr

Abstract
Prosopographical databases, which allow the study of social groups through their bibliography, become a paramount for a significant number of historians. Computerization has enabled intensive and large-scale exploitation of these databases. The modeling of these prosopographical 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. To illustrate the genericity of our conceptual model, we performed a mapping process of the latter to the models underlying three different existing prosopographical projects. Finally, an informed argument allowed us to validate our conceptual model.

keywords
Conceptual modelling; prosopographical database; quality; uncertainty; mapping process

I INTRODUCTION

Prosopography is a research method for studying a social group by comparing the biographical (sequence of) data describing its members. It aims at understanding how the groups operates, without neglecting the singular behaviors. Prosopography is based on a precise, documented investigation of each person in the population under consideration. Historians combine a methodology and an advanced erudition to collect all the traces that will constitute the record of each individual. This method of investigation is implemented for all historical periods.

The word "prosopographia" appears in the 16th century. Historians resort to prosopography to answer research questions such as:

1. Is there a link between disciplines (arts, medicine, canon law, theological law) and geographical origin of scholars involved in these disciplinary fields?
2. Is there a link between university degrees and geographical origin?
3. What is the nature and quantity of contentious cases in which Parisian academics are involved in the thirteenth and fourteenth centuries?

Since the end of the nineteenth century, researchers in the Humanities rely largely on the use of index cards allowing them to describe and analyze their objects of study. These files are not intended to receive all the information available related to the objects of study. They are built

to highlight some characteristics selected by the researcher to meet her requirements. For instance, a research methodology in prosopography consists of the reconstruction of biographical elements. This reconstruction requires aggregating and combining many documents in order to build an artefact required for conducting a prosopographical study.

Quantitative analysis and computer science profoundly transformed historian's methodology towards the end of the twentieth century. Many academic journals have been interested in this digital transformation and published articles and special issues on this theme. Many historians have even proposed specific software developments. [Schreibman et al. 2004] describe the advent of digital humanities and illustrates the broad spectrum of the many disciplines. The latter "have gone beyond simply wishing to preserve humanistic artifacts to represent and manipulate them, to reveal properties and traits not evident when the artifact was in its native form". Of particular interest is the contribution that databases have introduced to prosopography regarding the identification of the sources of information. More precisely, four main features of databases led to their large use by researchers in digital humanities: computing capacity, complex storage, organization of information, and availability of data for other researchers. Using only paper cards, it remains 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. Similarly, it was not easy to manage contradictory information with index cards since 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 verified by multiple sources, many data are missing, inaccurate, or appearing in sources known to have low credibility. This uncertain context has a significant impact on the formulation of hypotheses that historians attempt to verify using index cards. Using databases, historians can query, more easily, prosopographical data to determine, for instance, the degree of reliability and precision of the information. For example, in the STUDIUM PARISIENSE database [Genet et al., 2016], information about students' curriculum of the medieval period is stored. The information about curriculum, from the baccalaureate to the higher grades, may be complete, incomplete, or purely hypothetical, depending on the student's status ('Student', 'Graduate' and 'Master'). Therefore, historians can achieve a better level of trust when querying databases such as STUDIUM PARISIENSE. In this context, the problems related to information uncertainty and quality arise with renewed preponderance and a stronger acuity. Web techniques enable the efficient manipulation of voluminous information, but the problems related to the management of contradictory information and of low quality have not been resolved. In the field of prosopography, it remains judicious and essential to use databases that are more structured than the web and that integrate the management of this uncertainty at the conceptual level. Big data architectures can be of help if we are able to represent the information and the associated uncertainty in a structured way. The need for conceptual modeling of prosopographical data, its uncertainty and its quality, is still present.

In this vein, all prosopographical databases, whatever their underlying logical model, have been implemented on an ad hoc basis, sometimes without a conceptualization phase. In some projects, conceptual models have been developed but they lack genericity. Like all databases, significant maintenance effort is necessary, requiring upgrading of the underlying models. It is to face these limitations that we propose a generic conceptual model. The latter describes in a general and enriched way the information contained in a prosopographical database. We illustrate its properties by mapping it successively to 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. Section II is devoted to a state-of-the-art of the digitization of prosopographical databases and the management of the quality of historical data. In Section III, we present our generic conceptual model for prosopographical data which encompasses uncertainty and quality management. In Section IV we illustrate the genericity of our model by describing the mapping process aligning our generic model with the main concepts underlying three different prosopographical databases. Section V describes the validation process of the generic model. Section VI concludes the paper and presents some future work.

II STATE OF THE ART

Developing a generic conceptual model of a prosopographical database requires a good knowledge and understanding of four aspects: the role of computer science in prosopographical databases, the main concepts used in prosopographical databases, the quality management issues related to historical data, and conceptual models for prosopographical projects. These four aspects are the subject of the state of the art presented below.

2.1 Prosopographical databases and computer science

The use of prosopographical databases has become widespread among researchers in history since the 1970'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 was the first to highlight the contribution of computer science as a tool for prosopographical research [Werner, 1977]. The increasing volume of recorded data makes their exploitation (analysis and cross-referencing of data) extremely time-consuming. Using a database approach has emerged as one of the solutions to this volumetry problem. Significant examples of prosopographical databases include COEL [Keats-Rohan, 1998], PASE [Bradley and Short, 2005], ASFE [Brizzi, 2014], RAG [Schwinges, 2015], PADU-A [Gallo, 2018], and STUDIUM PARISIENSE [Genet et al., 2016].

Moving from a collection of paper cards to databases first involves designing a data model. Among the existing data models, we 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]. This structured representation enables to perform efficient search queries crossing a limited number of tables. Semi structured models include STUDIUM PARISIENSE which is based on XML files [Genet et al., 2016]. 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 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 [Graham and Ruffini, 2007; Verbruggen, 2007; Jackson, 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 [Tuominen, 2016] which is an event-based,

person and role-centric model for representing the activities a person has participated in during his life. [Tchounikine et al., 2018] propose OLAP analyzes and network analyzes associated with cartographic and chronological visualization tools to study the careers and the shared relation networks.

2.2 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 [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 [Gregersen and Jensen, 1999] as well as international standards [GIT-Schema, 2002] consider two types of data: "instant" and "interval". [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 offers several time types including instant, interval, multiInstant, and multiInterval types [Moisuc et al., 2012]. A variety of approaches have been proposed to represent temporal information in RDF [Manola et al., 2004] and in OWL [McGuinness and Van Harmelen, 2004]. Recently, [Ogawa et al., 2020] introduce a new model where a person is not described as a single entity, but as a collection of contextual entities each of them corresponding to a temporal aspect of a person.

Uncertainty is defined as "a general concept that reflects our lack of sureness about something or some-one" [National Research 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]. [Edmond, 2019] enumerates implicit characteristics of uncertainty in historical sources and the role it plays in historical interpretation. In the URREF ontology, uncertainty encompasses a variety of aspects including ambiguity, incompleteness, vagueness, randomness, and inconsistency [da Costa et al., 2012]. 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.

2.3 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. Numerous contributions have categorized quality issues, as well as metrics to measure the extent of these issues, methods, and tools to improve it [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 (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, e.g., 641, 5 August,
2. Assertions with a lower fine granularity, when data are available but less precise, e.g., a bishop is dead in 667,
3. Incomplete assertions where some information is missing for accurate identification, e.g., 5 August,
4. Uncertain assertions with an absolute specification of uncertainty, e.g., reader in theology between 1305 and 1312.
5. Uncertain assertions with a relative specification of uncertainty, e.g., reader in theology probably from 1305.
6. Assertions referring to other assertions containing temporal properties, e.g., a bishop is recorded as dying in 667,
7. Assertions with unknown or missing information, e.g., Hymn Adoro Te, commonly referred to by its Latin title: date unknown.

[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 spatio-temporal data modeling in a historical setting. It then needs to be quantified. An example of such quantification has been implemented using fuzzy logic [Martin-Rodilla et al., 2019].

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.

2.4 Conceptual models for prosopographical projects

Several conceptual models representing prosopographical data have been proposed. The Factoid model represents the most common conceptual approach. PASE [Pasin and Bradley, 2015] is an example of a factoid model. In the factoid model, a source S believes that the fact F can be stated about the entity E. This can be coupled with a set of temporal information T. It can express a relation between the entity (subject) E and other entities (objects) O1, O2, etc. Factoid types correspond to traits, states, and events. A factoid is not an absolute assertion. Factoids can be contradictory with each other. Finally, some degree of reliability can be associated with factoids.

CHARM is a reference model for cultural heritage, based on ConML [Gonzalez-Perez, 2018]. It is first dedicated to the representation of valuable entities, valorizations, and representations (the three main classes of CHARM). However, CHARM is not directly usable for prosopography, even if the factoid model could be reconstructed with CHARM classes, which, in turn, include subjectivity and vagueness.

Conceptual modeling of prosopographical projects also calls on ontologies. The Factoid Prosopography Ontology (FPO), ontology of the PASE project, is based on OWL / RDFS. The CIDOC-CRM ontology is not based on factoids but on temporal entities which generalize the concept of event [Doerr et al., 2020]. It represents information relating to cultural heritage. To our knowledge, these ontologies do not consider vagueness of information.

As it can be seen, there is not a single conceptual data model that encompasses the different viewpoints of prosopographical researchers, including vagueness of information and quality management.

In this paper, we present a conceptual model that gathers and makes more generic the information contained in different prosopographical databases, namely the concepts of People, Factoids, Places and Sources. It also incorporates a representation of uncertainty. A main advantage of our approach is to represent explicitly the measures of uncertainty, confidence, time, and precision attributes attached to all prosopographical concepts.

III CONCEPTUAL MODELING OF PROSOPOGRAPHICAL DATABASES

The model proposed in this article, and presented in Figure 1, 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. The first version of this model was presented in [Akoka et al., 2019]. This conceptual model must be validated using 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.

- 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 aiming at generalizing the event renders 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.

- The dimensions of all prosopographical concepts including factoids are associated with 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 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 vary over time (their name, their boundaries, etc).
- 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.
- 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.
- 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].
- 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 instance, when two documents give a different information related to the date, e.g., 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 a 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.

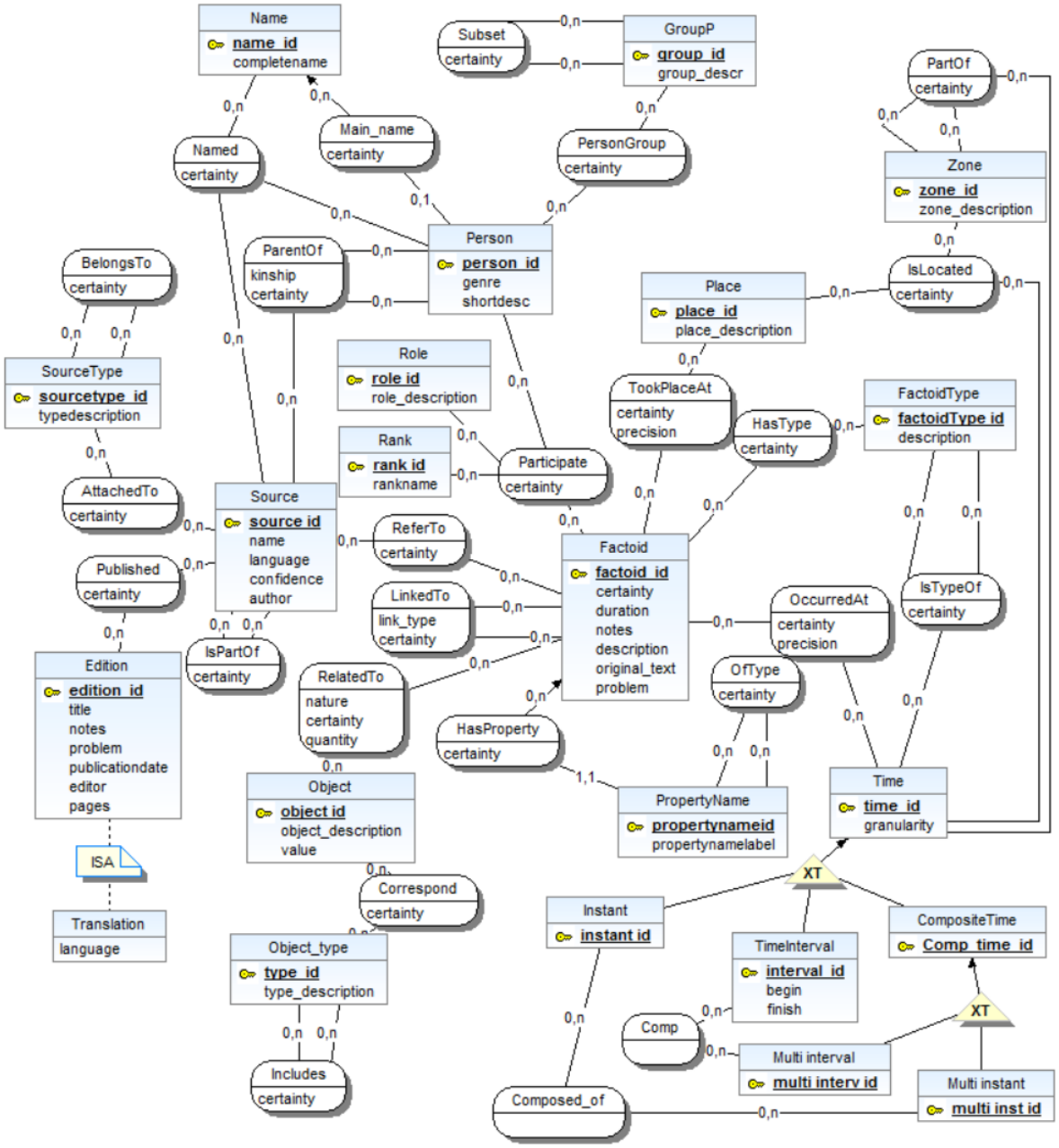

Figure 1. Our generic conceptual model for a prosopographical database.

This generic model makes it possible to cover the information contained in PASE (except for traces), in STUDIUM PARISIENSE and in PADU-A. It extends the model proposed in [Akoka et al., 2019] to better fit the historians needs, e.g., the fact that a source can be part of another source and to store information related to the different editions if any, that a factoid may have properties related to a property type, that a person may have a rank in his/her

participation to a factoid, etc. Moreover, a variant of this model in which we split the factoid into different concepts is described in [Akoka et al., 2020b], leading to a data model associated with a process model inspired from the information fusion process. We propose below to illustrate the genericity of this model by mapping it to three different prosopographical database models.

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. These three projects are representative of three classical approaches in prosopography (see [Akoka et al., 2020a] for an overview of the main prosopographical models, including ontological approaches). They rely respectively on a relational database based on a factoid model, an XML database relying on a source-oriented approach, and a relational database integrating different data sources with a person-based approach.

The mapping process is based on an alignment mechanism such as the ones implemented in all schema matching tools [Shvaiko and Euzenat, 2005]. 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 prosopographical 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, we compared 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 prosopographical project are identified and constitute the too specific part of the project which could not be considered in the generic model.

We briefly describe below the application of this process to the three prosopographical projects.

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
Object	object id	Possession	possessionkey
Object	object description	Possession	description
ObjectType	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.

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.

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	corresponds to the rank of the person
Origin	corresponds to the Origin event which takes place in a location
Bachelier ès arts (Paris) 1460	corresponds to the diplomation event with a location and a date

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

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 at the schools and the University of Paris since the appearance of the cathedral school at the end of the eleventh 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). It should be noted that 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

[1] http://www.pase.ac.uk/ (available on 2021, Feb. 18th)
[2] http://studium.univ-paris1.fr/ (available on 2021, Feb. 18th)

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 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.

Finally, Table 4 represents the mapping between the PADU-A[3] concepts and the ones we proposed in our generic model. The Prosopographical-Access-Database of University-Agenda project (PADU-A) intends to put the bases of a prosopographical 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 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 precisely defined within 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.

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, PADU-A, etc.), which would result from the adaptation of this generic model to their research needs. In addition,

[3] https://www.dissgea.unipd.it/padu-prosopographical-access-database-university-agenda-verso-una-banca-dati-di-studenti-e-docenti (available on 2021, Feb. 18th)

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 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 way 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 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 us 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 is concerned with the logic of interaction between models which is not relevant in our context where we have only one model. The layout design is achieved using 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 PARISIENSE, then we generated the relational schema and finally, we executed the following two queries [Akoka et al., 2019]:

- The first one compares 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.), in managing linguistic terms with vagueness interpretation, and solving the onomastics issues.
- The second query looks for more complex career patterns and considers the source reliability (estimated by historians): *Who are the Italians from the XIVth 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 deal with the reliability of the sources when evaluating a query. It also shows how to take into consideration 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 PARISIENSE 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 is 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 to quickly realize many tedious treatments. This digitization of prosopographical data has led to the emergence of many data models. In this article, we propose a generic data model allowing to describe most of the existing prosopographical databases. Moreover, we provide a way to enrich existing prosopographical databases by integrating several quality concepts such as certainty, credibility, consistency, and completeness. We illustrate the genericity of our conceptual model by performing a mapping process of our generic model to the models underlying three different existing prosopographical projects, namely PASE, STUDIUM PARISIENSE and PADU-A. Finally, an informed argument allowed us to validate our conceptual model.

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. Another future research will consist in enriching our model with the properties of ontologies. In this vein, we plan to rely on FPO to produce an ontology-based generic representation of prosopographical databases. Finally, another future research will consist in applying our generic model to prosopography of other historical groups since the three databases under consideration in this paper relate to the Middle-Age.

References

Akoka J., Comyn-Wattiau I. and du Mouza C. (a) Conception de bases de données prosopographiques en histoire – un état de l'art. *Revue Ouverte d'Ingénierie des Systèmes d'Information.* 2020;1(3).

Akoka J., Comyn-Wattiau I., Lamassé S. and du Mouza C. Modeling Historical Social Networks Databases. *International Conference on System Sciences (HICSS)*, 2019;1:10.

Akoka J., Comyn-Wattiau I., Lamassé S. and du Mouza, C. (b) Contribution of Conceptual Modeling to Enhancing Historians' Intuition-Application to Prosopography. *International Conference on Conceptual Modeling* 2020;164-173. Springer, Cham.

Allen J.F. Maintaining Knowledge about Temporal Intervals. *Commun. ACM*. 1983;26(11):832-843.

Barabucci G. and Zingoni J. PROSO: prosopographic records. *International Workshop on Collaborative Annotations in Shared Environment, DH-CASE@DocEng* 2013;3:1-3:7.

Barroso Júnior J.S., Pimentel M., Nunes V. and Cappelli C. Design Science Research to Design a Conceptual Model About Prosopographic Information Related to Politicians. *IV Brazilian Symposium on Information Systems (SBSI).* 2019;24:1-24:8.

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

Bol P.K. GIS, Prosopography and History. *Annals of GIS.* 2012;18(1):3-15.

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

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

da Costa P.C.G., Blackmond Laskey K., Blasch E., and Jousselme A-L. Towards unbiased evaluation of uncertainty reasoning: The URREF ontology. *Intl. Conf. on Information Fusion*, 2012;2301–2308.

Doerr M., Light R. and Hiebel G. Implementing the CIDOC Conceptual Reference Model in RDF. http://cidoc-crm.org/versions-of-the-cidoc-crm, Available on 2021, Feb. 18.

Edmond J. Strategies and Recommendations for the Management of Uncertainty in Research Tools and Environments for Digital History. *Informatics*. 2019; 6(3):36.

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

Genet J-P., Idabal H., Kouamé T., Lamassé S., Priol C., and Tournieroux A. General Introduction to the Studium Project. *Medieval Prosopography*, 2016 (31):156–172.

GENTECH. Genealogical data model: A comprehensive data model for genealogical research and analysis. http://xml.coverpages.org/GENTECH-DataModelV11.pdf, 2011, Available on 2021, Feb. 18.

GIT-Schema. Geographic information -Temporal schema, ISO 19108:2002. https://www.iso.org/obp/ui#iso:std:iso:19123:ed-1:v1:fr:sec:B, 2002, Available on 2021, Feb. 18.

Gonzalez-Perez C. *Information Modelling for Archaeology and Anthropology. Software Engineering Principles for Cultural Heritage*. Springer 2018.

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

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

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

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

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

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

Lindland O.I. and Krogstie J. Validating conceptual models by transformational prototyping. Proc. Intl Conf. on Advanced Information Systems Engineering (CAiSE), Springer Lecture Notes in Computer Science 685, 1993;165–183.

Liu Y., Mcgrath R.E., Wang S., Pietrowicz M., Futrelle J. and Myers J. Towards A Spatiotemporal Event-Oriented Ontology. In *Microsoft eScience Workshop*, 2008.

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

Martin-Rodilla P., Pereira-Fariña M. and Gonzalez-Perez C. Qualifying and quantifying uncertainty in digital humanities: a fuzzy-logic approach. *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2019;788-794.

Matousek K., Falc M. and Kouba Z. Extending Temporal Ontology with Uncertain Historical Time. *Computing and Informatics*, 2007;26(3):239–254.

McGuinness D.L. and Van Harmelen F. Owl web ontology language overview, w3c recommendation, 2004.
Moisuc B., Dia Miron A., Villanova-Oliver M. and Gensel J. Spatiotemporal Knowledge Representation in AROM-ST. *Innovative Software Development in GIS*, 2012;91–119.
National Research Council. *Risk analysis and uncertainty in flood damage reduction studies*. National academy press, 2000.
Ogawa J., Nakamura S., Ohmukai I. and Nagasaki K. Creating a New Semantic Model for Ancient Greco-Roman Prosopography - Toward a Contextual and Historical Description of the Prosopographical Data. *Intl Conf. of the Alliance of Digital Humanities Organizations (DH)*, 2020.
Pasin M. and Bradley J. Factoid-based prosopography and computer ontologies: towards an integrated approach. *Digital Scholarship in the Humanities*, 2015;30(1):86–97.
Pfeiffer D. and Niehaves B. Evaluation of Conceptual Models - A Structuralist Approach. *European Conference on Information Systems (ECIS), Information Systems in a Rapidly Changing Economy*, 2005;459–470.
Plewe B. The Nature of Uncertainty in Historical Geographic Information. *Trans. GIS*, 2002;6(4):431–456.
Prat N., Comyn-Wattiau I. and Akoka J. A Taxonomy of Evaluation Methods for Information Systems Artifacts. *J. Manag. Inf. Syst.*, 2015;32(3):229–267.
Ranade S., Bell M. Traces through Time: a Case-study of Applying Statistical Methods to Refine Algorithms for Linking Biographical Data. *Intl. Conf. on Biographical Data in a Digital World*, 2015;24–32.
Rittgen P. Quality and perceived usefulness of process models. *ACM Symposium on Applied Computing (SAC)*, ACM.2010;65–72.
Roblot T.K. and Link S. Cardinality Constraints with Probabilistic Intervals. Intl. Conf. on Conceptual Modeling (ER), 2017;251–265.
Schreibman S., Siemens R. and Unsworth J. The digital humanities and humanities computing: An introduction. *A companion to digital humanities*, 2004, xxiii-xxvii.
Schütte R. and Rotthowe T. The Guidelines of Modeling - An Approach to Enhance the Quality in Information Models. *Intl Conf. on Conceptual Modeling (ER)*, Springer Lecture Notes in Computer Science 1507, 1998;240–254.
Schwinges R.C. Das Repertorium Academicum Germanicum (RAG). Ein digitales Forschungsvorhaben zur Geschichte der Gelehrten des alten Reiches (1250-1550). In Jahrbuch für Universitätsgeschichte, 2015;215–232.
Shanks G.G., Tansley E. and Weber R. Using ontology to validate conceptual models. *Commun. ACM*, 2003;46(10):85–89.
Shaw R. and Larson R.R. Event Representation in Temporal and Geographic Context. Europ. Conf. on Research and Advanced Technology for Digital Libraries (ECDL), 2008;415–418.
Shvaiko P. and Euzenat J. A survey of schema-based matching approaches. *Journal on data semantics IV*. Springer, Berlin, Heidelberg, 2005;146-171.
Tchounikine A., Miquel M., Pécout T. and Bonnaud J-L. Prosopographical data analysis. Application to the Angevin officers (XIII-XV centuries). *Journal of Data Mining & Digital Humanities (JDMDH)*, 2018.
Tuominen J. Emlo prosopographical data model: Towards a biographical conceptual reference model. Technical report, Cost Action IS1310, Reassembling the Republic of Letters, Aalto University, 2016.
Verbruggen C. Combining Social Network Analysis and Prosopography. In *Prosopography Approaches and Applications. A Handbook*, pages 579–601. Linacre College, 2007.
Werner K-F. Problèmes de l'Exploitation des Documents Textuels Concernant les Noms et les Personnes du Monde Latin (IIIe-XIIe siècles). *Informatique et Histoire Médiévale*, pages 205–212, 1977.
Westermann U. and Jain R. Toward a Common Event Model for Multimedia Applications. *IEEE MultiMedia*, 2007;14(1):19–29.