Journal of Data Mining and Digital Humanities

Prosopographical data analysis. Application to the Angevin officers (XIII–XV centuries)

Anne Tchounikine[1], **Maryvonne Miquel**[1], **Thierry Pécout**[2], **Jean-Luc Bonnaud**[3]

[1]LIRIS-CNRS UMR 5205, INSA-Université de Lyon, Lyon, France
[2]UMR LEM-CERCOR, Université Jean Monnet, Saint Etienne, France
[3]Université de Moncton, Canada

Abstract

The EUROPANGE project, involving both medievalists and computer scientists, aims to study the emergence of a corps of administrators in the Angevin controlled territories in the XIII–XV centuries. Our project attempts to analyze the officers' careers, shared relation networks and strategies based on the study of individual biographies. In this paper, we describe methods and tools designed to analyze these prosopographical data. These include OLAP analyzes and network analyzes associated with cartographic and chronological visualization tools.

I CONTEXT

This research work was conducted within the framework of the EUROPANGE project funded by the ANR (French National Research Agency) and the Ecole Française de Rome's five-year program. It emanates from a team that links computer scientists and medievalists studying the rhythms and elaboration methods of political communities. The thirteenth to fifteenth century constitute an ideal period to observe the development of organisms, speech methods and corps of administrators, within the context of the birth of national and princely states. For this project, we examine the emergence of a particular political and societal environment through the constitution of a corps of administrators (the officers), with their relation networks, with jurisdiction within the political space of the Angevins' possession: Anjou, Maine, Provence, Lorraine, Southern Italy and Sicily, Piedmont, Lombardy and Tuscany, Hungary, Poland, Morea, and the Balkans. This spatially disjoined and temporally variable nature of the area administered by the officers, required new capabilities within the political bodies to rally and administer, creating a common political discourse, which is of special interest today in the context of current reflections on the construction of Europe. During the last two decades, historians have rethought the role of the individual as an actor ([2, 10, 5, 9]). According to Pierre-Marie Delpu ([3]), a prosopography could be minimally defined as a collective study which extracts the common characteristics of a group of historical actors based on the systematic observation of their lives and individual careers. This approach, coupled with quantitative and statistical methods and advanced display methods has shown that prosopography is the most accurate way to determine the sociological aspects of a particular group of people through its members' individual careers ([14, 8, 1]).Our project attempts to reconstitute the officers' careers, shared relation networks and strategies.

ISSN 2416-5999, an open-access journal

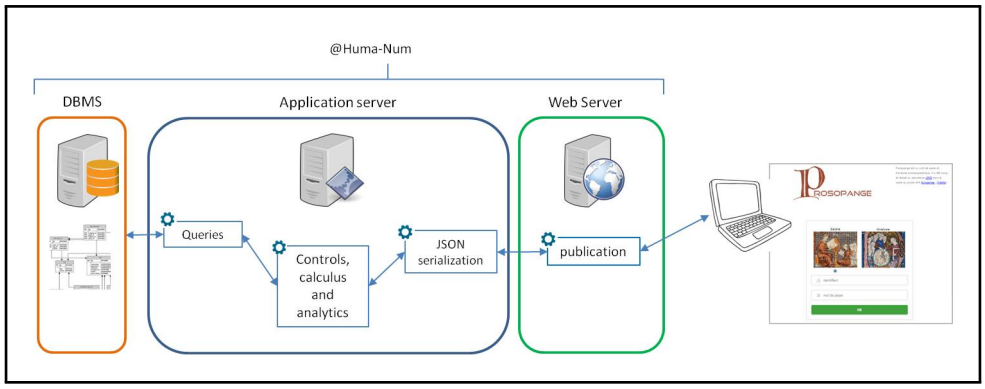


Figure 1: Architecture of the Prosopange software suite.

II CONTRIBUTIONS

The data to be gathered and analyzed include all the biographical details of the officers and their acquaintances, extracted from the inventory of archival and iconographical sources (Angevin chancery records, deeds, epigraphic sources). Our contribution hinges on 3 elements:

- a collaborative application for the capture and rendering of the data in the form of prosopographical pages made up of the consolidated, enhanced, structured and formatted data ([12]),
- an analytics application to build and survey multi-criteria populations,
- a prosopographical database used by the above applications.

All these elements make up the software suite Prosopange which is presently deployed on the French CNRS TGIR Huma-Num site. This software suite (see Figure 1) is a 3-tier client-server application. The data tier is made up of a database that allows structuring the set of data that characterizes an officer. The users who extract from and analyze the bibliographical sources usually enter this data directly, but data can be also found that are automatically computerized from the entered data. The processing tier includes algorithms for data creation and searching within the database, various controls and computation and analysis algorithms. Finally, the client tier is made up of the user Web interface equipped with data acquisition processing, and requesting and displaying tools. In this paper, we will mainly focus on the prosopographical data analytics application. This application involves multidimensional analyzes and relation network analyzes connected to temporal and cartographical representations of data. These analyzes are carried out on selected officers' populations by means of a multi-criteria filter relying upon the historical period, the concerned territories and/or individual traits (religion, gender, locations . . .).

III OLAP CUBE AND ANALYSIS

Leonardus Afflicto de Scallis, native of del Giustizierato di Terra di Lavoro Citra (Kingdom of Sicily), professor juris civilis holder since September 24th 1372 appointed in 1373 by Johanna I as judex major et secundarum appelationum in the Counties of Provence et Forcalquier. He is referred to as nobilis and vir in the documents after 1378.; these are the kinds of data that are available in the database.

It soon became apparent that the historians involved in the project as research partners had multiple expectations in terms of data analysis. This was especially true for the number of themes interesting them: some are interested in the movement of the officers, their impact and influence areas; others focus on the territory political organization, appointment games and duty

stations; or the sociology of the constituted corps, officers' places of origin, their acquaintances and their honorific or their qualifying academic degrees. The methods and the choice of analyzed populations are themselves also diverse: quantitative mass analyzes, temporal or spatial transversal analyzes, and analyzes of an identified group of officers, faction, trade and even a single individual. In order to satisfy all these requirements and be open to future requirements, we propose a flexible user interface that offers the researcher the opportunity to build his/her specific analysis dynamically.

Data warehousing and OLAP (On Line Analytical Processing) are database technologies specially designed for data analysis and decision making ([6, 7, 13]). They allow interactive multidimensional data analysis and are proving to provide interesting solutions. OLAP data modeling is guided by analysis needs. It relies on multidimensionnal models that are built around the major analysis subject (called facts) and the various axes which might be used to elaborate an analysis (called dimensions). The OLAP engine stores detailled data and can materialize all or selected aggregates so as to allow fast access to summarized data. At the front-end user interface level, OLAP technologies provide operators and tools to interactively explore the multidimensional data, supporting in the meantime the iterative nature of the analysis process, and allowing the analysts to navigate across data at different levels of detail without the use of query languages.

OLAP presupposes that the various axes which might be used to elaborate an analysis be defined in advance. As far as historians are concerned, the challenge was to identify, define and create these various analysis axes. For the computer scientist, the problem has been to adapt the OLAP concepts and operators to the specificities of prosopographical data and historians needs. The OLAP analyses focus on a unique fact: the officers. Officers have to be enumerated and identified (i.e. listed) according to various analysis axes. These axes allow us to describe the officer, his/her office or offices, his/her title or titles, (academic degrees, honorific). Within the multidimensional model, 3 dimension categories are defined:

- dimensions related to individuals: place of origin, gender, place of residence, religion, various dates (birth, residence...)...
- dimensions related to the public offices hold by individuals: office type, place of work, place of duty, appointment, exercise dates...
- dimensions related to the individual titles: honorific, academic degrees, dates of granted titles...

Concerning the OLAP analysis, the user dynamically chooses among these 3 categories, the dimensions he is willing to include in his multidimensional table, with the option for combining categories: for example, ("place of origin" x "religion") or ("place of origin" x "religion" x "types of offices"). Aggregate measures can include a distinct count of officers or offices and the calculation of the duration of offices and of associations.

Unfortunately all these dimensions have many irregularities: The fact-dimension associations often are multivalued, for example an officer may occupy several offices simultaneously or sequentially, a single office can have several places of duty. The database filled levels also are varying, because the collected information within the manuscript sources is very often incomplete or vague. Moreover, many dimensions are not orthogonal, for example the place of duty dimension depends on the office dimension, and many are also unbalanced (see Figure 2, a dimension extract type of office in the cross-table). In addition, it is impossible to build, a priori, a multidimensional schema and cube as should be done in a classical decision-support application because of the unpredictability of the historian requests and the evolution of dimension

Algorithm 1 Construction of a cube

cube(in : population of officers, list of dimensions; out : cube)
foreach o in population :
  foreach dimension d_i related to individuals :
    $E_{d_i} = \{\}$ *//coordinates on d_i axis*
    foreach j member of d_i at level l_i in o properties :
    // j can be religion of o, place of origin, gender etc.
      $E_{d_i} = E_{d_i} \bigcup \{ j \}$ *// add j to coordinates*
      $E_{d_i} = E_{d_i} \bigcup parents(j, d_i)$ *// add all ancestors of j in d_i hierarchy*
      $E_{d_i} = E_{d_i} \bigcup nr(j, d_i, l)$ *// add empty child for each incomplete lower level*
    $E_{individuals} = E_{d_1} \times E_{d_2} \times ...$
  foreach c_i in o. offices :
    foreach dimension d_i related to public offices : *// public office is a multivalued dimension*
      $E_{d_i} = \{\}$
      foreach data j corresponding to d_i at level l_i:
      // j can be office type, place of work, ...
        $E_{d_i} = E_{d_i} \bigcup \{ j \} \bigcup parents(j, d_i) \bigcup nr(j, d_i, l)$
    $E_{c_i} = E_{d_1} \times E_{d_2} \times ...$
  $E_{offices} = \cup E_{c_i}$
  foreach q_i in o. titles :
    // same as offices
foreach coordinate in $E_{individuals} \times E_{offices} \times E_{titles}$:
    add or update a point p in the cube and compute associated measure : p.measure.add(o)
    // the measure contains the list of concerned officers

members which are progressively discovered as documentary sources are analyzed . Therefore, the traditional use of an OLAP server is not possible. Our algorithm to compute the hypercube takes as an input every officer (see Algorithm 1) and is calculated on the fly at request time on the server side using the database data. Every officer's data are transformed in a point in the multidimensional target space and the set of aggregate points is updated. For example in a "place of origine" x "type of office" cube, data on Leonardus Afflicto will generate the following points in the multidimensional space : ("Terra di Lavoro", "judex"), ("Terra di Lavoro", "principal"), ("Terra di Lavoro", "justice"), ("Terra di Lavoro", "central office")...("Sicily", "judex"),("Sicily", "principal") and so on considering the dimension hierarchical instances...

Dynamic of dimensions

For each regular balanced dimension hierarchy, historians predefine the levels. For example, for the spatial dimension used in places of origin, residence, work, and duty, the predefined hierarchical schema is : $location < subdivision < political\ space < territory < all$; concerning the public office types, an unbalanced parent-child tree hierarchy is used. The members of balanced or unbalanced dimensions are built as and when database insertions are performed and gradually build up a set of nomenclature data. The choices of the dimensions and nomenclature constitutions have been performed according to both the needs expressed by the historians and the computer scientists' need to have a repository for comparative analyzes. Hence the hierarchical spatial dimension arises from a consensus on a territorial division. However, the choice related to member incorporation according to evolving needs, allows us to take into account new places during the acquisition of the officers and thus to enhanced the hierarchy.

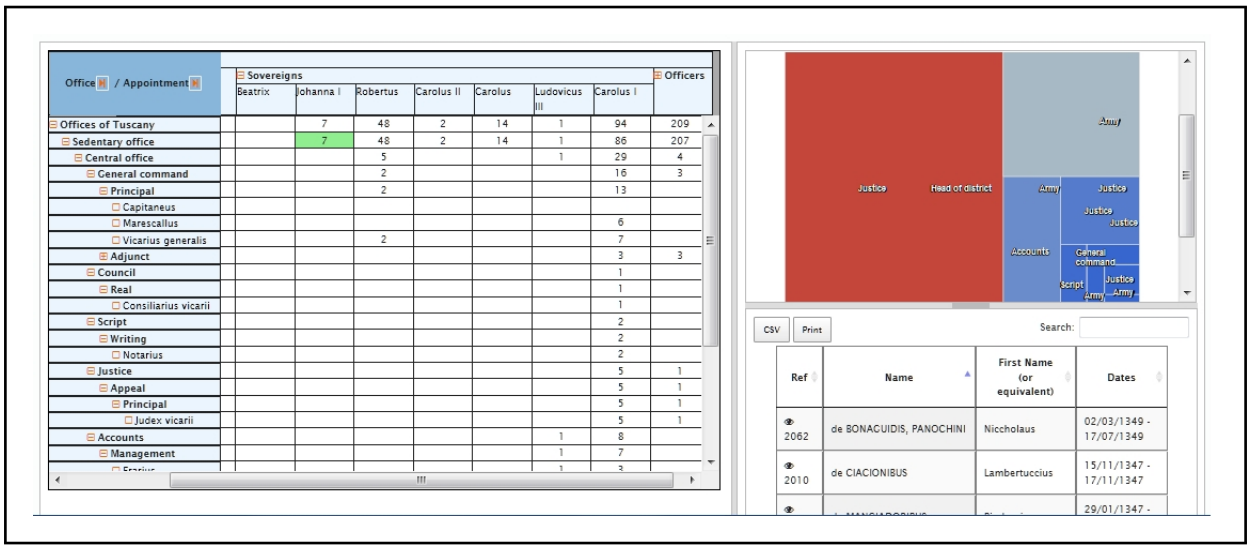


Figure 2: OLAP Analysis of the cube: “type of office” x “appointment”

Incomplete and imprecise data

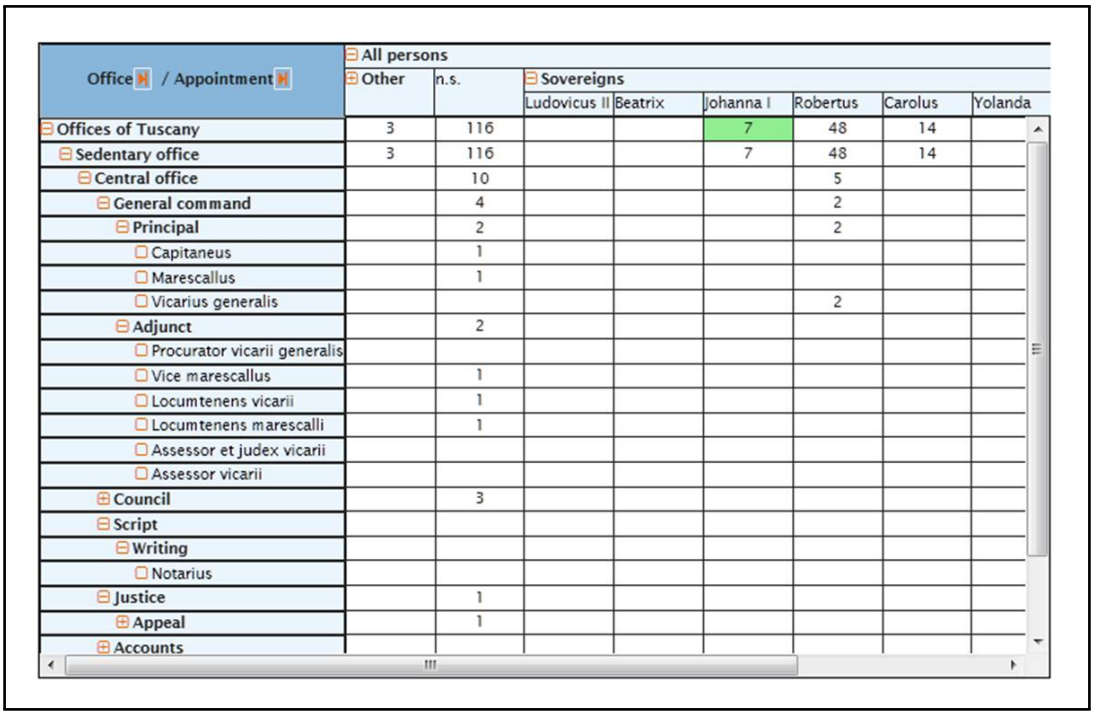
The strategy implemented by the algorithm 1 allows us to accurately generate known or entered data at various levels. For example, an officer whose place of origin is only known at the subdivision level will be obviously booked into the subdivision aggregates and the upper ones (political spaces, territories, all). At every detailed level of this dimension, this inaccuracy will be translated by means of a dynamical inclusion of an empty child member in order to provide a correct counting. The multivalued data are also taken into account. An officer who held public offices of “judex” and “notarius” types will occur in both the “Justice” and “Writting” aggregates and in the upper level aggregates (Script, Central offices...). Date entries especially are frequently imprecise or incomplete: thus, in the GUI for data entering, we allow the entry of dates with full or partial yyyy-mm-dd format but also interval, alternative or prefixed date (post, circa, ante). These dates are processed in order to be mapped into minimum and maximum timestamps that are used for chronological sorts and duration calculus.

Cube materialization and OLAP navigation

Unlike the conventional OLAP approach, multidimensional space is here dynamically built as the scan of the officers' associated data is progressing. The hypercube is thus fully materialized; the aggregate set is calculated and then sent to the client. This choice of both full materialization of the cube and officer table scan is made possible due to the small amount of data (˜7000 officers) and it allows insuring good performances during the OLAP navigation. On the client side, we implemented the set of classical OLAP operators ([4, 11]):

- sort: to sort dimension members (i.e. row and column headers within the table),
- drill-down: to move to a hierarchy finer level (i.e. spread the offspring of a row or column header),
- roll-up: to move to a hierarchy coarser level (i.e. aggregate the offspring of a row or column header),
- slice: to remove a member of the cube (i.e. remove a header and its offspring within the rows and columns of the table),
- drill-through: to make a drill-down then a slice on a row and column header).

The results are provided in the form of tables, charts, choropleth maps. Figures 2 and 3 show the result of the analysis of the number of officers by type of public office and persons or types of



Office / Appointment	All persons							
	Other	n.s.	Sovereigns					
			Ludovicus II	Beatrix	Johanna I	Robertus	Carolus	Yolanda
Offices of Tuscany	3	116			7	48	14	
Sedentary office	3	116			7	48	14	
Central office		10				5		
General command		4				2		
Principal		2				2		
Capitaneus		1						
Marescallus		1						
Vicarius generalis						2		
Adjunct		2						
Procurator vicarii generalis								
Vice marescallus		1						
Locumtenens vicarii		1						
Locumtenens marescalli		1						
Assessor et judex vicarii								
Assessor vicarii								
Council		3						
Script								
Writing								
Notarius								
Justice		1						
Appeal		1						
Accounts								

Figure 3: Details of the cross-table shown figure 2.

persons who performed their appointment. Here, the public office hierarchy is extended till the detailed level for the "General Command" type and the person type hierarchy is deployed for the "Sovereign" type. A click on a table cell highlights the cell in green and allows getting the list of the concerned officers and for each of them his prosopographical page can be displayed.

IV RELATION NETWORK ANALYSIS

One of the Europange project goals is to analyze interpersonal relationships such as family, profession and friendship. This information describing relationships that connect an officer to other individuals in the database are implemented by 4-tuples (source of relation officer, connected target individual, relation type, connection date), for example ("L. Afficto de Scallis", "sovereign Johanna I", "fidelis", "06/07/1371"). The connections are typed and oriented, and eventually associated to a reciprocal link during the acquisition. The connection type is selected among an unbalanced tree hierarchy (see Figure 5, left frame).

Starting from this information, 3 types of graph are built:

- an inter-personal graph in which individuals are vertices and directed edges are made up with acquired and computed relationships both labeled with the first and last relation reference dates and weighted according to the relationship's known or computed duration (first and last date mention),
- an officer-public office graph in which individuals and offices are vertices and which edges represent the office holding that are both labeled by means of assignment dates and weighted by the holding computed duration (first and last date mention),
- a graph for residence coincidence or assignment coincidence which vertices are officers and which edges connect the officers having resided or being assigned at the same location, same dates and weighted according to the coincidence computed duration (number of years)(see algorithm 2).

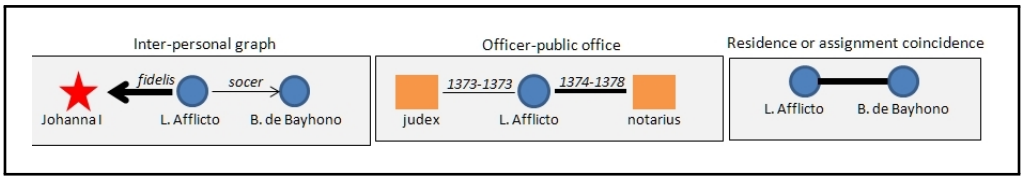


Figure 4: Types of graphs.

Algorithm 2 Construction of the assignment coincidence matrix

assignmentCoincidence(**in** : **population of officers**; **out** : **cube**)
myCube = cube(population of officers , {‘‘place of duty’’, ‘‘exercise date’’})
myCube = roll−up(myCube, {‘‘place of duty’’.’’subdivision’’, ‘‘exercise date’’.’’year’’})
foreach point(subdivision , year) in myCube:
  get point.measure *// the corresponding measure, contains the list of concerned officers*
  create or increment corresponding cell in the matrix

These different graphs are calculated on the server side thanks to suited OLAP cubes with duration of links aggregated measure. They may be displayed on the client side under the form of relation networks or coincidence matrix. Customization tools allow the historians to undertake structural analyzes and to interpret and interrogate these graphs. Thus, the user can select the relation types to be displayed or colored in the graph; the user can also make the graph dynamically evolve on a chosen period of time or search for and highlight vertices according to a label.

V EGO-CENTERED ANALYSIS

The aim of this analysis is to enrich the traditional tools of the prosopographical method that very often provide only a descriptive vision of the careers and relationships without any true

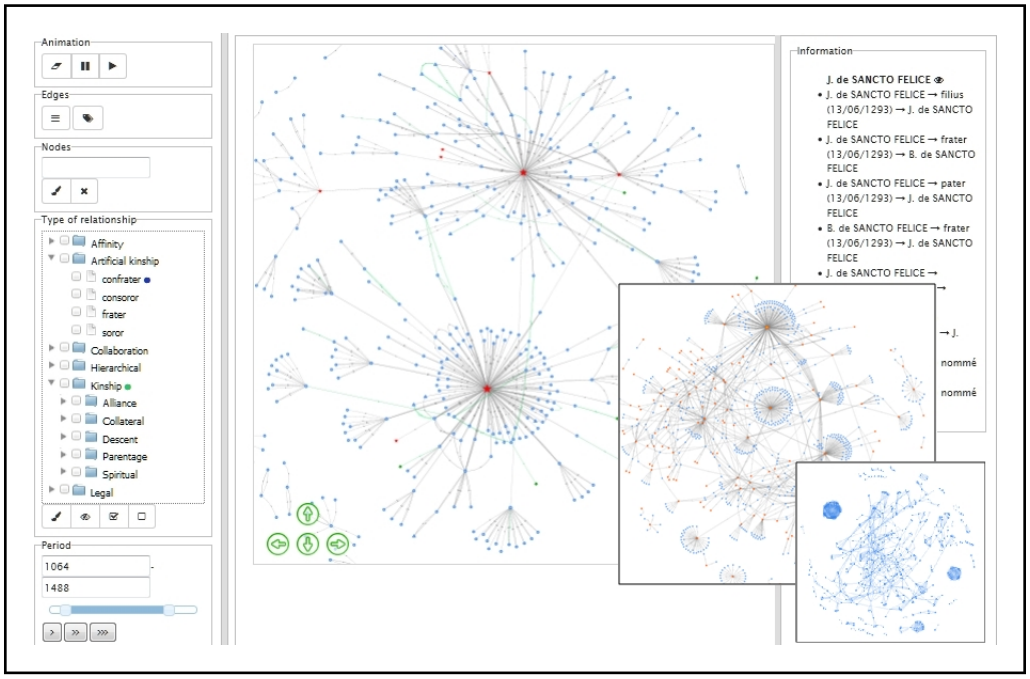


Figure 5: Visualization of an inter-personal network with family relation coloring and parametrization of the period, plus officer-public office and coincidence graphs

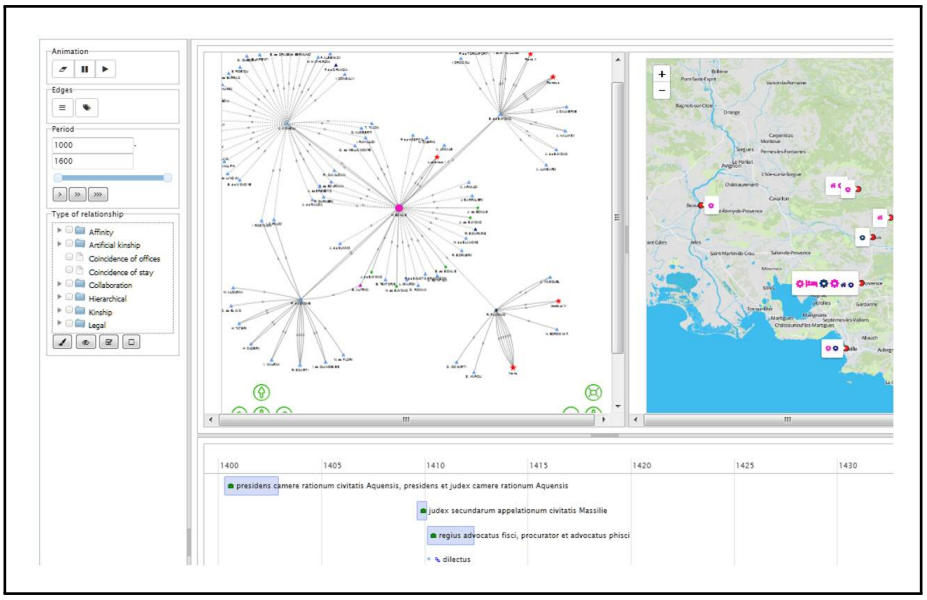


Figure 6: Visualization of an ego-centered network with cartography of officers selected in the graph and chronological record of the ego.

understanding of their formation. The ego-centered relation network investigation allows comparing the social capital of several people. It requires scaling-up (from the study of an entire group to the study of specific cases) and imposes a dynamic vision. A static vision of the relations between people is artificial since it does not take into account the role of time, a significant problem for historians. In an officer study, one of the main issues is to look into both the constitution of a homogenous milieu of officers and analyze how it was constructed. One possible perspective is to consider how newcomers are incorporated into their new environment. Varying forms of data can be examined and compared to understand and measure immigrant integration: migratory flows (origin of the officer, place of residence, place of duty); marriages as marker of the integration in the host society and as an influence on the conferring of noble title and position; professional networks; nomination and promotion; the individual or family policies of sale or purchase of goods, donations to local places of worship,chosen burial sites...

To complete this analysis, a graph centered on the central individual (called “ego”) is built which aggregates the inter-personal graphs, the coincidence graphs of the ego and the graphs of the officers connected to him. The number of iterations and depth of the graph is set by the user. A cartography of the officers present in the graph (origin of the officer, place of residence, duty place) and the ego’s chronology made up with his various associated dates (dates of the various offices, academic degree dates of graduation, link dates, . . .) are associated and displayed with the graph. Tools for time-based animation synchronized to the different components (network, cartography and chronology) allow highlighting the momentum of the network establishment and the career paths (see Figure 6).

VI CONCLUSIONS AND OUTLOOK

The Prosopange software suite and its database are currently deployed on the digital humanity Huma-Num site. About 60 medievalist researchers (including French, Italian, Hungarian, Pole, German . . .) daily feed this site with data gathered thanks to the analysis of manuscript archival documents (currently more than 7000 officers covering a period from 1210 until 1539). The database is already the largest one in the field in terms of the number of officers and of data

points per officer, and can be used for bibliographic reviews by every researcher who submits an account request. The definition of the different classifications composing the dimensions, the public offices, the relations and the positions, required a significant effort to pool, name, synthesize and reach a consensus encompassing the terms used, their classification and their prioritization. This work is fully unprecedented and amounts to a result in itself. The nomenclature of places, in turn, lays the groundwork for an administrative cartography. The use of the database and the analysis tools reveal the relevance of a collaborative approach: the historians enrich their knowledge of the officers' careers thanks to the pooling of data collected by the researchers who are experts in different political spaces; the analysis reveals the links between individuals that transcend the territories and allows them to contextualize and to compare the corps organization in the various spaces and over time. Prosopange allowed accumulating a huge amount of unpublished tools and materials. The processing of all these resources is just starting and it already allows supplying various research topics related to, for example, the public office typology, recruitment profiles, the analysis of periods during which parties are formed during severe political tensions, or further more to the study on the relation between administrative standards and the office practice. From a computer science point of view, the work is ongoing in two directions. A first perspective will be to extend the analysis tools particularly to the dynamics and the navigation in the networks and to the regularities and singularities in the officer career paths. A second challenge is the consideration of the highly variable quality and quantity of the data that are an ongoing problem when medieval documentary sources have to be processed. In fact, the bibliography of certain officers could be informative, very precise while the hitherto known information concerning other officers could be piecemeal or questionable. In this context, any quantitative analysis has to be treated with caution and should be coupled with metrics by measuring the accuracy and the significance.

VII ACKNOWLEDGMENT

This Europange project was made possible by the financial support from the ANR (French National Research Agency) and from the Ecole Française de Rome's five-year program.

References

[1] Charles Bouveyron, Laurent Jegou, Yacine Jernite, Stéphane Lamassé, Pierre Latouche, and Patrick Rivera. The Random Subgraph Model for the Analysis of an Ecclesiastical Network in Merovingian Gaul. *Annals Of Applied Statistics*, 8(1):377–405, 2014.

[2] John Bradley and Harold Short. Texts into databases: The evolving field of new-style prosopography. *Literary and Linguistic Computing*, 20(Suppl):3–24, 2005.

[3] Pierre-Marie Delpu. La prosopographie, une ressource pour l'histoire sociale. *Hypothèses*, (18):263–274, December 2015.

[4] Jim Gray, Adam Bosworth, Andrew Layman, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. In *ICDE*, pages 152–159. IEEE Computer Society, 1996.

[5] Heloise. Projet heloise, european network on digital academic history, http://heloise.hypotheses.org/.

[6] Bill Inmon. *Building the Data Warehouse*. Wiley, 4th edition, 2005.

[7] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley, Indianapolis, IN, 3 edition, 2013.

[8] Nizar Messai and Thomas Devogele. Visualisation de données de prosopographie pour la reconstruction de carrières de personnages et de réseaux socio-professionnels. In *14 ème conférence internationale sur l'extraction et la gestion des connaissances*, pages 557–560, Rennes, France, January 2014.

[9] Parisiense. Projet studium parisiense, http://lamop-vs3.univ-paris1.fr/studium/.

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

[11] M. Rafanelli, editor. *Multidimensional Databases: Problems and Solutions, Hershey, PA - USA, 2003*. Idea Group Inc., 2003.

[12] Anne Tchounikine, Maryvonne Miquel, and Thierry Pécout. Modélisation de données pour une base de données prosopographique. In *Les officiers et la chose publique dans les territoires angevins (XIIIe-XVe sicle) Vers une culture politique ?*, Colloque international de Saint-Etienne, Université Jean Monnet, 17-19 novembre 2016 2016.
[13] Alejandro Vaisman and Esteban Zimányi. *Data Warehouse Systems: Design and Implementation*. Springer, Heidelberg, 2014.
[14] Claire Zalc and Claire Lemercier. *Méthodes quantitatives pour l'historien*. coll. Repères, 2008.