Digital Scholarship in the Humanities, 2023, **38**, 1014–1032
https://doi.org/10.1093/llc/fqad033
Advance access publication 22 April 2023
Full Paper

OXFORD

A new approach for the construction of historical databases—NoSQL Document-oriented databases: the example of *AtlantoCracies*

Manuel Diaz-Ordoñez [1,*], **Domingo Savio Rodríguez Baena** [2], **Bartolomé Yun-Casalilla** [3]

[1]Departamento de Economía e Historia Económica, Universidad de Sevilla, Sevilla, España
[2]Departamento de Informática, Universidad Pablo de Olavide, Sevilla, España
[3]Departamento de Filosofía, Geografía e Historia, Universidad Pablo de Olavide, Sevilla, España

*Correspondence: Manuel Diaz-Ordoñez, Departamento de Economía e Historia Económica, Universidad de Sevilla, España.
E-mail: mdiazord@us.es

Abstract

This article proposes, and justifies, the use of the Document-oriented databases as a flexible, easy to use, and powerful digital tool in the field of historical research. First, the reasons that have made relational databases the predominant instrument among historians are studied, while detailing the problems involved in their use. Next, the way in which historians have tried to face these problems by using other digital tools is explained, as well as the limitations that such use entails. Through a case study—that of European aristocratic networks in early modern times—it is shown, however, that Document-oriented databases, present notable advantages and have greater explanatory power for the historian's work. Thanks to their flexibility, they are better adapted to the often-unpredictable nature of historical sources without diminishing their ease of use or their analytical potential.

1 Introduction

Today, databases are a fairly common methodological resource in historical research. The DataBase Management Systems (DBMS) most used by historians are usually commercial ones such as Microsoft ACCESS. They also use open-source software such as LibreOffice Base or OpenOffice Base. However, when it comes to large projects, more powerful software such as Microsoft SQL Server, Oracle Database, IBM DB2, MySQL, MariaDB, or PostgreSQL are frequent. This work tries to explain why current historians have a clear preference for Structure Query Language (SQL) databases, also known as relational databases, and to propose the use of Document-oriented NoSQL databases in substitution. It is proposed here that the reason for the preferential use of SQL databases is due to the evident theoretical and technical deficiencies in the historians' computer training, as well as due to the fact that they feel more comfortable in front of the manuscript than typing in a computer (Bodenhamer, 2008). It should also be considered that historians may have difficulties communicating with computer engineers, with whom they work in multidisciplinary research teams. This fact is especially important when it comes to explaining their methodological needs, which can lead to the technician not recommending the most suitable resource to satisfy them. It is contended here that, as a consequence, the historian frequently has a major problem derived from the very characteristics of the database technology, due to the enormous rigidity of the structures and definitions of the Relational Database Management Systems (RDBMS), which are the most used as resources in current historical research. In order to develop this argument, the architectures, database engines, and data models that have been used in historical research to overcome the limits of relational systems (RDF and Key-Value models, graph-oriented databases, etc.) will also be analyzed. Following this analysis, the problems with relational databases for our project, *AtlantoCracies*, are exposed. The multi-dimensionality and complexity of historical research demand database solutions that provide more flexibility when changes are introduced during the

development of the research. Finally, an example of Document-oriented NoSQL (DoNoSQL) database is proposed, which, applied to the study of aristocratic networks in early modern times, demonstrates the advantages of this type of database. In summary, we argue that the semi-structured nature of the DoNoSQL databases allows the researcher greater freedom, since the historian does not need to define a rigid scheme at the beginning of the research and would be able to undertake modifications, as knowledge of the object of study and the sources advances.

2 Why do current historians preferably use RDBMS?

When historians began to use databases in the last third of the 20th century, they did so in a technological context very determined by the development of SQL database systems. These were times when the DBMSs that were being developed incorporated the definition of relational structured data, since they were greatly influenced by the great success of the 1970 publication of Edgar Frank Codd's article 'A relational model of data for large shared data Banks' (Codd, 1970, 1971). Frank Codd established the so-called relational logic, in which data sets could be established in tables in a structured way of two dimensions: rows and columns. The rows represented a singular informative unit considered as a horizontally visible register or tuple, while the columns represented fields that can be interpreted as known aspects of reality. In the following decades, the RDBMS dominated the computer market for the use and massive management of data, and it spread among users who needed information storage and analysis tools. The potential of the new relational model (RM) was based on the fact that the tables were related to each other by sharing specific key fields. Using these relationships, SQL queries were implemented to extract information from the combination of a subset of tables. These queries could be complex, including the possibility of filtering and grouping data, carrying out mathematical operations, or ordering the information extracted. Besides, the RM included the referential integrity restriction to avoid data inconsistency during inserting, deleting, or updating actions. Before a relational database implementation, an initial data model needs to be designed, in which all the tables, their attributes, and their relationships are defined.

The database engines of the 1970s were based on Codd's definitions, and on his RM. Products like dBASE II, and its descendant dBASE III—as well as its Clipper compiler—by Wayne Ratliff, ORACLE by Larry Ellison, Ed Oates, and Bob Miner, or FoxPro or Harbor were based on the relational system and used SQL as the language of data access (Sarda, 1990). These early DBMSs were not user-friendly. The most obvious barrier was the absence of a graphical user interface that compensated, at least in part, for the difficulty of handling the software, since the instructions had to be entered through the command-line interface. This meant a lack of accessibility not only for the community of historians, but also for all possible users of the databases, including business organizations, which also did not have enough employees trained in the use of these tools. These difficulties led, in the mid-1980s, to the appearance of new more accessible and user-friendly database products, such as dBASE III PLUS, and its replacement dBASE IV, which included textual menus to allow a friendly use of the different database actions, like select, insert, update, and delete. All this would lead to the development of more advanced versions of DBMSs such as ORACLE, Microsoft's SQL SERVER, IBM's DB2, etc. Very soon some historians cognized the possibilities of the DBMSs distributed during the last quarter of the 20th century (Burton *et al.*, 1987). These digital pioneers of the Humanities made a significant effort to learn how to use these tools (Gutmann, 1987), attending technical courses or training programs at universities. A good example in the field of social history is the Fichoz relational database developed with Filemaker, a proprietary system designed by Jean Pierre Dedieu starting in the 1980s (Dedieu, 2000, 2004, 2012, 2013).

3 Relational database models, methodological dependence for historians?

The use of RDBMS was boosted with the publication of Codd's twelve relational rules (thirteen if number 0 is counted) (Codd, 1985), which established the mandatory requirements that a relational database model had to meet. Due to these requirements, most of the users of this technology, and among them historians, ended up attached to a certain theoretical, practical, and design servitude, with the limitations that this implied (Bradley and Pasin, 2013). The problem was underlined by some archaeologists: 'One of the major problems for archaeological data managers is keeping the computer subservient to the practice of archaeology and not vice versa' (Eve and Hunt, 2010, p. 1). To explain what we mean, let us think about the effects of the application of Codd's Rule 2. According to this standard, every single data may be accessed logically from a relational database using the combination of primary key value, table name, and column name. Suppose as an example that a researcher decides to use a date attribute, formatted 'dd/mm/yyyy' and linked to each individual (e.g. birth, death, etc.), as part of the primary key. If, after collecting an important series of

data, this researcher wants to enter data regarding an individual whose date is unknown, the options would be very complicated: she/he decides to dispense with this individual so as not to modify the database model, thus assuming the loss of scientific accuracy, or has to modify the database by rethinking the model, with the consequent loss of time and money (Harvey and Press, 1996). Some authors, such as Charles Harvey and Jon Press, have argued that historians can address this problem by building carefully conceived data models well before they begin to process the data and the relationships between them (Harvey and Press, 1996). As a matter of fact, this is only the initial phase in the life cycle of creating a database, which is made up of three stages: analysis, design, and development, or implementation. In the first phase, an analysis of the reality to be studied is carried out, generating the so-called Entity–relationship model (E/R). Such a model is characterized by being independent of technology, and by representing the data to be stored and how they relate to each other. This E/R model, developed by Peter Chen in 1976, was based on entities, attributes that define them, and relationships between these entities, as well as its own internal constraints and rules (Chen, 1976). In conclusion, the objective of the development of the E/R model was to obtain a diagram that would serve as a graphical summary of the interesting data for the user and the explicit relationships between said data, as well as the definition of all allowed objects (entities, attributes, domains, and relationships).

In the next stage, dedicated to database design, we generate an RM from the previous E/R model. This RM model included more details related to the technology in which the database will be implemented. In the end, processes, analysis, and design will result in a database made up of a set of tables and the relationships between them. According to Mark Merry, the historian would not have many problems when developing this model, since in historical methodology it is common to establish complex relationships (Merry, 2020). However, the same author points out that, although these complex relationships are gradually established as the historical sources are known and exploited, these relationships must be considered during the analysis phase of the database to reduce the risk of having to restructure it (Rossi *et al.*, 2014; Merry, 2020). It is probably these drawbacks that led David J. Bodenhamer to point out that, when explaining the multidimensional reality of the past, the historian prefers the word to the structured model of a database (Bodenhamer, 2008).

An example of the need to introduce modifications in the data models, as a consequence of the development of the research, can be found in the large-scale project The Trans-Atlantic and Intra-American slave trade databases (SlaveVoyages henceforth) (Project: The Trans-Atlantic and Intra-American Slave Trade Databases, 2020). Since 1960, Herbert S. Klein and other researchers began to store data from European ships or ports related to the slave trade in different software databases. But, beginning in the 1980s, project leaders observed that the data modeling did not exactly match the information obtained from the sources. The result was that starting in the 90s, a task of standardizing the data from books, articles, or spreadsheets had to be carried out, building a new data model in which all the data were integrated, standardized, and normalized, which meant a demanding job that multiplied the time spent and costs. On the other hand, in each modification of the data model, the historian is in danger of falling into redundancies and inconsistencies (Kantabutra, 2009; Kantabutra *et al.*, 2014). And the whole problem is aggravated when the database is not being managed directly by the historian and depends on a team of computer engineers, with whom, as usual, there may be communication problems (Molina Recio, 2002). Eve and Hunt have clearly referred to this (Eve and Hunt, 2010).

4 Recent solutions to RDBMS limits for historians

Despite the limitations of relational systems, there are many historians who use them optimally. An example is that of quantitative researchers in disciplines such as economic history, or those who carry out bibliographic or documentary works. In these cases, an RM satisfies their needs since their objects of study are perfectly adapted to this type of structure (Merry, 2020). The widespread use of RMs is also due to the fact that, in technological university careers, the teaching of these types of databases continues to be given priority, which in turn is due to their dominance in the market and to some inertia among many of the polytechnic instructors, who have also been trained in these systems (Viteri and Bayas, 2020). It is not surprising that the aforementioned case of Fichoz is not the only example available (Fichoz—A database for social history/*Base de données pour l'histoire sociale*, 2012). Another case is the Ancestry Library Edition, a database that gathers billions of genealogical records from multiple connected databases (Ancestry Library Edition, 2020). Its architecture mixes a relational database system with Apache Hadoop, a massive parallel data processing system (Morgan, 2014). Likewise, and also following this principle, the American National Biography Online Database contains American biographies that can be consulted by the identification fields of each individual (American National Biography, 2020). Another case is SlaveVoyages, based on a relational system, and which also has data visualization and

spatial representation tools (Smith *et al*., 2011; Project: The Trans-Atlantic and Intra-American Slave Trade Databases, 2020).

Also, in the field of Archaeology, the use of relational database systems has spread. That is the case of IDEA, developed by Madsen and Andresen in the mid-90s, or the closest in time, the ARK Toolkit framework. Both Andresen and Madsen (1996a,b) and Eve and Hunt (2010) are based on the Entity-Attribute-Value relationship for recording the information. Furthermore, ARK uses the MySQL relational database, which, according to Peter Jensen, discourages archaeologists from seeking new alternative technologies (Jensen, 2018). Another example is the IDEArq-C14 database, which preserves thousands of radiocarbon dates (Salas Tovar *et al*., 2016; Uriarte González *et al*., 2017) stored in a PostgreSQL relational database management system, with connections to PostGIS, a spatial database extender for PostgreSQL, that guarantee access to texts through hyperlinks to digitized documents (Uriarte González *et al*., 2017).

In the field of prosopography, the use of the first relational systems of the veteran Continental Origins of English Landholders, 1066-1166 database (COEL), implemented by Access and Paradox, has been improved. The same can be said for the Prosopography of Anglo-Saxon England (PASE), implemented by a MySQL relational database. Both projects are based on a factoid model,[1] composed of assertions that a source 'S' at location 'L' states something ('F') about person 'P' (Akoka *et al*., 2019, 2020). Also, in prosopography's studies, On-Line Analytical Processing or OLAP database engines have been used, which are systems used for the analysis of large amounts of data and decision-making in the business world (Tchounikine *et al*., 2018). And software suites such as Prosopange (Prosopography of Angevine officers, 13th–15th century http://base.angevine-europe.huma-num.fr/prosopange/), have even begun to be used to improve analysis capabilities on a relational database (Tchounikine *et al*., 2018, p. 8). The latest advances in this field of study have focused on the analysis of data with graphic visualization tools through the use of geographic information systems.

Other important innovations have been developed in large-scale historical databases thanks to technological advances related to the Internet, especially since the launch of the semantic web concept. Its principles were enunciated by Tim Berners Lee and James Handler, who developed the techniques that made it possible to handle and convert unstructured information, such as text, into structured data (Berners-Lee and Hendler, 2001). Among the techniques that currently take advantage of the concept of semantic web, and that are handled by engineers and historians, the Resource Description Framework (RDF) data models must be highlighted (Miller, 1998; Candan *et al*., 2001; Binding and Tudhope, 2011). In short, RDF models are based on the method of describing the terms of interest stored in the system, with a labeling, generally Extensible Markup Language (XML), constructing metadata that identifies them and, thus, being able to retrieve them from the query system. However, the approach of these models is like the classic E/R conceptual model, with the variation that the statements are structured in the triple: subject-properties-value. This scheme is known in RDF as triple semantics, and consequently databases are often referred to as Triplestores. This is the path that the SESHAT Databank Project (SESHAT: Global History Databank) has opened since 2015 (Turchin, 2014; Turchin *et al*., 2015; Turchin, 2017). Another recent case that follows this same line is Dacura, which offers datasets oriented to the semantic web (Feeney, 2016; Peregrine *et al*., 2018). As the results for 2017 of the Seshat Project indicate, the questions that can be addressed to this system are also subject to prior metadata markup, that is, data retrieved by queries must be properly codified. The Intentionally Linked Entities (ILE) is another good example of historical database innovation (Kantabutra, 2009). ILE is a general-purpose database management scheme that is particularly useful for representing complex social network systems in a dynamic geographical environment. It is a direct implementation of the E/R model, consisting of four major components: entities, entity sets, relationships, and relationship sets, and is based on a simple idea: replace all of the by-value linkages in the RM's tables with relationship objects that connect entities via pointers (Kantabutra *et al*., 2014). Technical advances in historical research methodology have accelerated in recent years. The newest bet has been, as Conny Kristel and Tobias Blanke point out, the use of NoSQL database systems (Kristel and Blanke, 2013). The explosive expansion of the Internet from the 1990s, and its impact on the traffic, storage, and consultation of millions of terabits of information, implied an exponential increase in the volume of the data, its variety and the velocity at which it is generated. For example, in 2003 it was estimated that the amount of digital data in the world was 5 exabytes (10^{18} bytes) but only 9 years later, in 2012, this amount had multiplied by 500, reaching 64.2 zettabytes (1021 bytes) in 2020. Because of this, database technology adapted to this new environment, improving the performance. The first advances would be consolidated with the proposal of non-relational data management systems, also known as NoSQL, because the way of interacting with the stored information was not necessarily based on the SQL language. Carlo Strozzi was the first referring to the term NoSQL to demonstrate the possibilities of his

database (NoSQL Relational Database Management System: Home Page, 2021) developed in 1998, which, although not really a non-relational system, showed relevant differences with the standardized models that followed the laws by Codd (NoSQL Relational Database Management System: Home Page, 2021). The term NoSQL was used again by Eric Evans during the meeting organized by Johan Oskarsson in San Francisco in June 2009, in which those data management architectures that did not necessarily use the SQL language were defined. Among these types of NoSQL databases, four main subclasses have been distinguished: Key-Value stores, Document-oriented databases or DoNoSQL, Columnar databases and Graph-Oriented databases or Property Graphs (Pivert, 2018). The Key-Value data stores are very simple but are quite efficient and powerful. The data consist of two parts, a string which represents the key, and the actual data which is to be referred as value, so we have sets of 'key-value' pairs. Columnar databases store data in columns. This may seem similar to traditional relational databases, but rather than grouping columns together into tables, each column is stored in a separate file or region in the system's storage. They are suitable for analytic applications. Graph databases store data in the form of a graph. The graph consists of nodes and edges, where nodes act as the objects and edges act as the relationship between the objects. The graph also consists of properties related to nodes. And, finally, DoNoSQL refers to databases that store their data in the form of documents. Documents are somewhat similar to records in relational databases, but they are much more flexible since they are schema-less. Document stores offer great performance and horizontal scalability options (Nayak *et al.*, 2013).

Among these subclasses, the first and most important in research projects was the type Key-Value, applied in the software CD/ISIS developed by UNESCO (CDS/ISIS Database Software, 1985). Its objective was to support libraries so that they could create and manage online catalogs. Second, there are the subclass graph-oriented databases, like TerminusDB or Neo4j, used in Beyond (2022) Virtual Record Treasury, Seshat: Global History Databank, China Historical Christian Database or STUDIUM Parisinum (China Historical Christian Database, 2020; Seshat: Global History Databank, 2020; Studium Parisiense Database, 2020; Beyond 2022 | Ireland's Virtual Record Treasury, 2020). These databases are based on an RDF model and develop a data query grounded on three fundamental elements: subject, predicate, and object (Angles and Gutierrez, 2008; Kristel and Blanke, 2013). This is the case of Studium Parisiense, which offers information on the members of Paris' schools and universities and whose system is based on XML text tagging (Genet *et al.*, 2016).

But, regarding NoSQL databases, it is very important to point out that those of the DoNoSQL subclass are not applied in historical projects. Its use is specialized in research related to literature or philology. More specifically, they are often implemented in critical editing of texts, as can be seen in Thebarum fabula, Search and Retrieval of Indic texts (SARIT), The Tibetan Buddhist Resource Center (TBRC), or for mass storage of ancient texts like Early English Books Online (Thebarum Fabula, 2020; SARIT: Search and Retrieval of Indic Texts (New), 2021; The Tibetan Buddhist Resource Center (TBRC), 2020; Early English Books Online, 2020). The most common engine is Exist-DB and they usually have a previous labeling of the texts according to the Text Encoding Initiative metalanguage.

To complete the analysis developed in this section, a study carried out on forty two historical research projects and their information stored in online databases is included.[2] To do so, the managers of these projects have been contacted by email to identify what type of database they use. From the forty two projects, only four of them do not use databases (9.5%). As it can be observed in Table 1, this study has revealed an important hegemony of relational systems. More than 70% of the projects continue to use SQL databases, while only 9% of them bet on the graph-oriented NoSQL database systems. In the rest of cases, 12%, different solutions not related to databases are used, like spreadsheets or datasets created by statistical software, for example SPSS. More details about the projects, along with the URL of their web sites, are included in Appendix 1.

We would like to make a proposal for the use of DoNoSQL, grounded on the experience of our own research on transnational aristocratic networks. This case study opens up wide possibilities in the field of analysis in social, economic, and cultural history that can be applied to many other examples, while meeting the most important needs of historians in this regard.

5 The problems of relational systems for *AtlantoCracies*

The *AtlantoCracies* Database project is produced by an interdisciplinary research group composed by specialists on the history of the European aristocracies in the early modern period as well as by digital engineers with a long expertise in relational and NoSQL databases. The main aim of this group is to go beyond current research on aristocratic networks, so far dominated by the analysis of individual case studies (Yun-Casalilla dir, 2008), or by the use of SQL databases

Table 1. Database technologies at historical projects

Database model	Type of database	Number of databases	Databases (%)	Projects (%)
SQL	Relational DBMS	29	87.87	78.37
NoSQL	Graph DBMS/RDF Stores	4	12.12	10.81
	Oriented-document Database	0	0	0

Source: AtlantoCracies Research Group.

(Aram *et al*., 2021a,b; Pérez-García and Díaz-Ordóñez, 2021) mainly grounded on social relationships (Yun-Casalilla, 2002, 2008, 2019, 2022; Yun-Casalilla and Redondo-Álamo, 2008). Therefore, researchers require a database that considers and integrates many different types of evidence, in order to study the relations between family strategies, on the one side, and of cultural transfers, political developments, and the reconversion of social, political, and economic capital, on the other side (Bourdieu and Passeron, 1977; Bourdieu, 1989). That is the reason why *AtlantoCracies* departs from the study of family networks but seeks to integrate many other aspects related to the way in which these networks shaped processes of social ascension, the development (or not) of a common aristocratic culture across the Atlantic, and other similar aspects. This Atlantic dimension, usually also neglected when studying trans-national aristocratic networks in Europe, is also essential. It is also intended to incorporate evidence as the compilation of data and the hypotheses that arise suggest new lines of analysis that have been unexpected up to now.[3]

In all research projects in which databases are involved two main stages can be distinguished: a first one, related with extracting data from different sources, its pre-processing and its storage, and a second one in which useful knowledge is extracted by using data queries or by applying more complex techniques like Data Mining. Besides, data can be classified according to its structure: structured data, semi-structured data, and unstructured data (Bărbulescu *et al*., 2013). Structured data refer to information with a high degree of organization and with a fixed structure, like the table organization supported by relational database models. Semi-structured data, also known as schemeless or self-describing structure, is a form of structured data that do not conform with the formal structure of relational database models but nonetheless contains labels to separate semantic elements and enforce hierarchies of records and fields within the data. In the case of unstructured data, there is no structure at all (emails, images, audio, video, etc.).

Based on what has been said above, the new approach proposed by our work is about the mentioned first stage and involves semi-structured data due to its flexibility and adaptability to the continuous changes that historical research implies. In other words, we refer to a database able to easily adapt to the different documentary sources (biographical documents, merit accounts of the historical actors, secondary bibliography, wills, post-mortem inventories, correspondence, etc.) and to the new hypotheses that arise during a research. Our aim is also to create a database powerful enough to model and to be interrogated to unveil the dynamic forms of social and spatial mobility and relationships of European elites (Owens and Kantabutra, 2020). These methodological requirements are even more important because of the differences among the diverse types of sources to be used, and more in particular between those related to the intra-European networks and the trans-Atlantic ones.

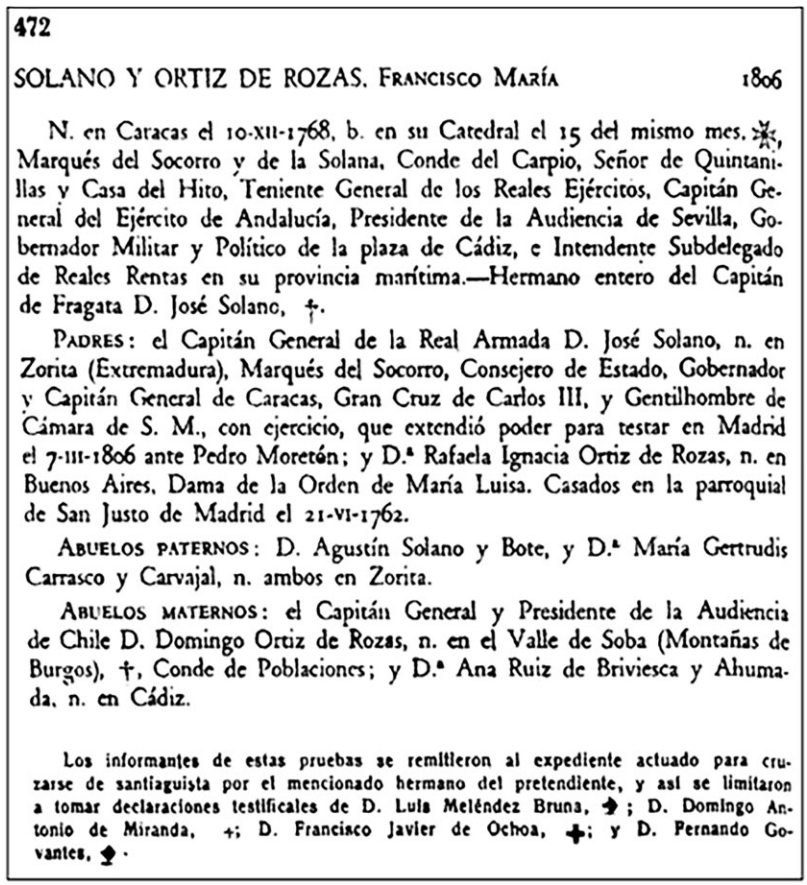
Very soon the immediate evidence found by the research group during the *AtlantoCracies* project was that the use of an RM limited the possibilities of handling very diverse sources as well as of understanding unforeseen dimensions of the history of this social group. This fact had already been verified by other authors on other topics within the field of history, who, however, did not seek solutions to these problems (Bradley and Pasin, 2013; Kristel and Blanke, 2013). Besides, we had previous experiences based on our involvement in two important research projects funded by European Research Council: ARTEMPIRE-648535 (*Conquest, commerce, crisis, culture, and the Panamanian Junction*. Consolidator grant of which the PI is Professor Bethany Aram (Aram *et al*., 2021a,b)) and GECEM-679371 (*Global Encounters between China and Europe: Trade Networks, Consumption, and Cultural Exchanges in Macau and Marseille, 1680–1840*. Starting grant of which the PI is Professor Manuel Pérez-García (Perez-García and Díaz-Ordóñez, 2021). In both projects, a relational database model was implemented to respond to the premises of the researchers: historians, computer engineers, biologists, archeologists, etc. However, during the development of ARTEMPIRE and GECEM, relational database models underwent several profound changes that involved a high cost in time and money. These changes responded to the need to add new data entities, attributes, and relationships after consulting new

documentary sources (Díaz-Ordóñez, 2020; Perez-García et al., 2022; Perez-García and Díaz-Ordóñez, 2022). And every time a change needed to be developed, it was necessary to hold long meetings with the computer engineers, stop the project to incorporate the changes, to increase the budget, etc.

From the beginning, it was decided in *AtlantoCracies*, which main entities, along with the attributes that define them, should be included in the E/R database model. The initial source of data selected was the book *Los Americanos en las órdenes nobiliarias* by Guillermo Lohmann Villena (Lohmann-Villena, 1993). The book contains a collection of historical characters, all of them members of the American and Peninsular nobilities who were awarded with a habit of the different military orders of the epoch. In Fig. 1, a piece of this document is shown regarding the noble *Don Francisco María Solano y Ortiz de Rozas*.

As it can be seen, the information, based on primary sources from the Section Military Orders (*Órdenes Militares*) of the *Archivo Histórico Nacional* (*AHN*) of Madrid is organized by the name of the awarded person and includes different data about his ancestors, normally two or three generations. Our initial objective was to build a database that serves to analyze the geographic, social, and economic mobility of the high nobility of the Hispanic monarchy in the modern age beginning with this group. Based on the aforementioned source, the most common data of each individual (aristocrat or non-aristocrat who would acquire a noble title) that appeared in the book were: names and surnames, place and date of birth, nobile titles, parents, etc. Figure 2 shows the E/R model, which is the result of the database analysis phase, and which responds to the scientific objectives pursued. Two fundamental elements stand out: the noble and his/her title. This RM generated from this E/R model was composed of three tables.

However, once the RM was obtained, it was observed that it did not answer many of our questions. How can one evaluate, beyond previous family relationships, the dynastic, religious, or political interests



472

SOLANO Y ORTIZ DE ROZAS. FRANCISCO MARÍA 1806

N. en Caracas el 10-XII-1768, b. en su Catedral el 15 del mismo mes. ✠, Marqués del Socorro y de la Solana, Conde del Carpio, Señor de Quintanillas y Casa del Hito, Teniente General de los Reales Ejércitos, Capitán General del Ejército de Andalucía, Presidente de la Audiencia de Sevilla, Gobernador Militar y Político de la plaza de Cádiz, e Intendente Subdelegado de Reales Rentas en su provincia marítima.—Hermano entero del Capitán de Fragata D. José Solano, †.

PADRES: el Capitán General de la Real Armada D. José Solano, n. en Zorita (Extremadura), Marqués del Socorro, Consejero de Estado, Gobernador y Capitán General de Caracas, Gran Cruz de Carlos III, y Gentilhombre de Cámara de S. M., con ejercicio, que extendió poder para testar en Madrid el 7-III-1806 ante Pedro Moretón; y D.ª Rafaela Ignacia Ortiz de Rozas, n. en Buenos Aires, Dama de la Orden de María Luisa. Casados en la parroquial de San Justo de Madrid el 21-VI-1762.

ABUELOS PATERNOS: D. Agustín Solano y Bote, y D.ª María Gertrudis Carrasco y Carvajal, n. ambos en Zorita.

ABUELOS MATERNOS: el Capitán General y Presidente de la Audiencia de Chile D. Domingo Ortiz de Rozas, n. en el Valle de Soba (Montañas de Burgos), †, Conde de Poblaciones; y D.ª Ana Ruiz de Briviesca y Ahumada, n. en Cádiz.

Los informantes de estas pruebas se remitieron al expediente actuado para cruzarse de santiaguista por el mencionado hermano del pretendiente, y así se limitaron a tomar declaraciones testificales de D. Luis Meléndez Bruna, ✠; D. Domingo Antonio de Miranda, ✠; D. Francisco Javier de Ochoa, ✠; y D. Fernando Govantes, ✠.

Figure 1. Data about Francisco María Solano y Ortiz de Rozas

Source: Guillermo Lohmann Villena, Los Americanos en las órdenes nobiliarias. Book I.

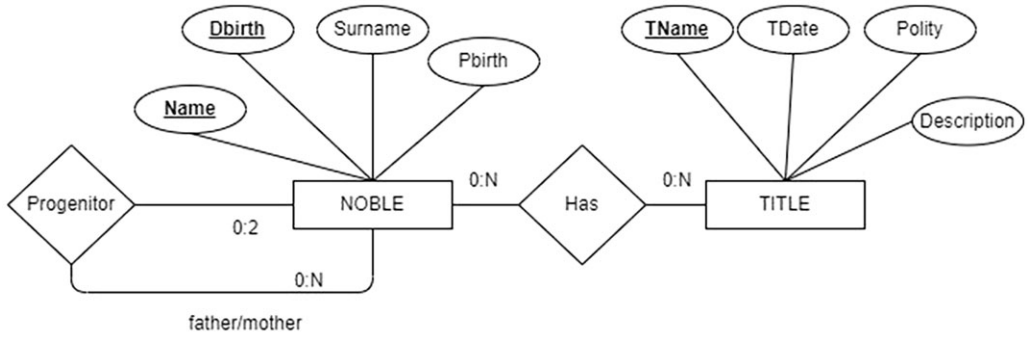


Figure 2. E/R model and RM developed for *AtlantoCracies*
Source: *AtlantoCracies* Research Group.

that were behind the elite marriage strategies? Ultimately, the aim was to go beyond existing studies, which have limited themselves to using databases that, by collecting only family relationships, leave aside the possibility that marital ties respond to other unknown external factors. It was therefore a matter of trying to capture this last block of factors among which we presume would be previous relationships of a spatial, cultural, religious, etc. type, and even other factors that, until now, may have gone unnoticed by historians (Carvalho and Campos, 2007). Further complicating analysis, we add the spatial component to consider how the marriage strategies were also related to geopolitics and the access or belonging of the nobles to certain European or American territories. So, it was necessary to add new sources of data and generate new hypotheses. Thus, the Spanish Historical Archive (AHN), where the original primary sources are, was consulted. Particularly, the part of the Military Orders Collection, known as the *Expedientillos*. In this document, new attributes and relationships were found, like marriages, important dates about the titles, the godparents of the candidates to obtain a noble title, the witnesses who participated in the granting of the title, etc. Figure 3 shows the data found in the *Expedientillos* about the same historical character presented in Fig. 1. As an example of the new information that the *Expedientillo* offers and that complements Lohmann's data, the witness Luis Menéndez Bruna was a knight of the Order of Alcántara, belonged to the Royal Council of Military Orders, and knew the suitor and his parents. For our project, the position and kinship and friendship relations of the historical characters were very important to analyze the connections between the Atlantic dimension and the local peninsular networks.

So, as in the GECEM and ARTEMPIRE projects, it became necessary to rebuild the E/R and relational database models. Thus, the database already implemented and in operation required a structural transformation that would imply paralyzing its use for a time, as well as the treatment of the new data to be entered and the modification of the software already implemented. Additionally, changes would impact the security and performance management of the DBMS. Thus, the adaptation to this new reality had an associated cost, as it can be observed in Fig. 4, in which the improved E/R model after adding the new data is shown.

This new E/R model implied ten new tables in the RM. As can be seen, every time the data source changes, the lack of flexibility of the RM implies a significant delay in the project. In the next section, a new methodology based on a Document-oriented database is presented as a solution to these drawbacks.

6 The new possibilities of document-oriented databases and their implementation at *AtlantoCracies*

As it has been explained with the examples presented in the previous section, a bottleneck was detected each time the relational database model needed to be changed when new sources were consulted and different attributes or even different research hypotheses appeared. We had to overcome the doubt that often arose about whether it was the researcher who had to adapt to the structure of the database or, on the contrary, it was necessary to modify the database model again to store the new attributes and relationships detected. In addition, neither of these two options appeared optimal, since either research opportunities were wasted or time and money were lost during the process of adapting the database. For all these reasons, it was considered that the most suitable solution was to move from structured data to semi-structured data, using a NoSQL database as a specialized tool for this type of data. The justification of why the Document-

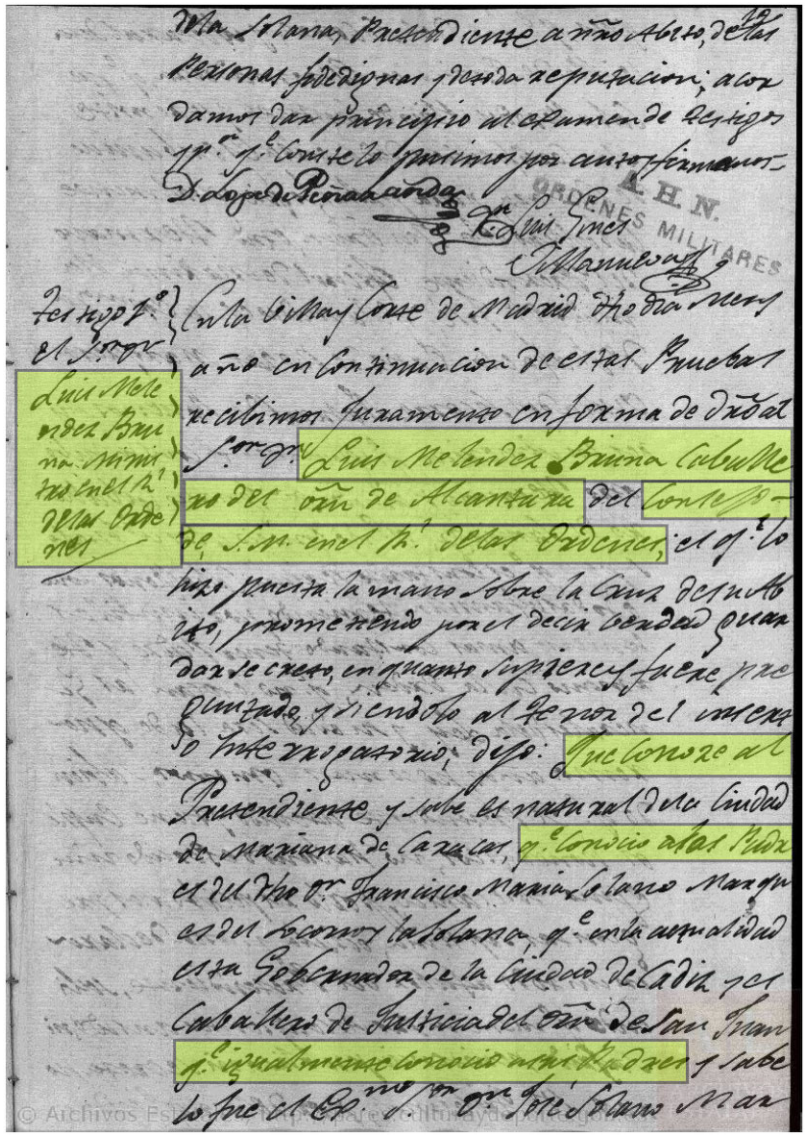


Figure 3. New data about Francisco María Solano y Ortiz de Rozas, including information about his witnesses highlighted.
Source: Archivo Histórico Nacional, Caballeros de Santiago. Mod. 74.

oriented databases have been selected from among the four types of NoSQL databases is presented below.

Basically, our project requires a database capable of working with the flexibility of semi-structured data but at the same time being as close as possible to the RM. This is because the *AtlantoCracies* database model is complex, since it is necessary to store different attributes from different entities that are also interrelated. Besides, the possibility of developing powerful data queries is required. Thus, on one hand, Key-Value databases and Columnar databases are the two types whose functionalities are furthest from our needs. Key-Value databases are actually pretty straightforward. Thanks to the simple data format that gives it its name, a key-Value store can be very fast for read and write operations, so it is suitable for cases in which huge amounts of simple data must be processed in real time. Columnar databases are used in data warehouses where businesses send massive amounts of data from multiple sources for business intelligence analysis. On the other hand, Graph databases and Document-oriented databases are the closest to our goals. Graph databases are specialized in the efficient management of heavily linked data. However, we are not only

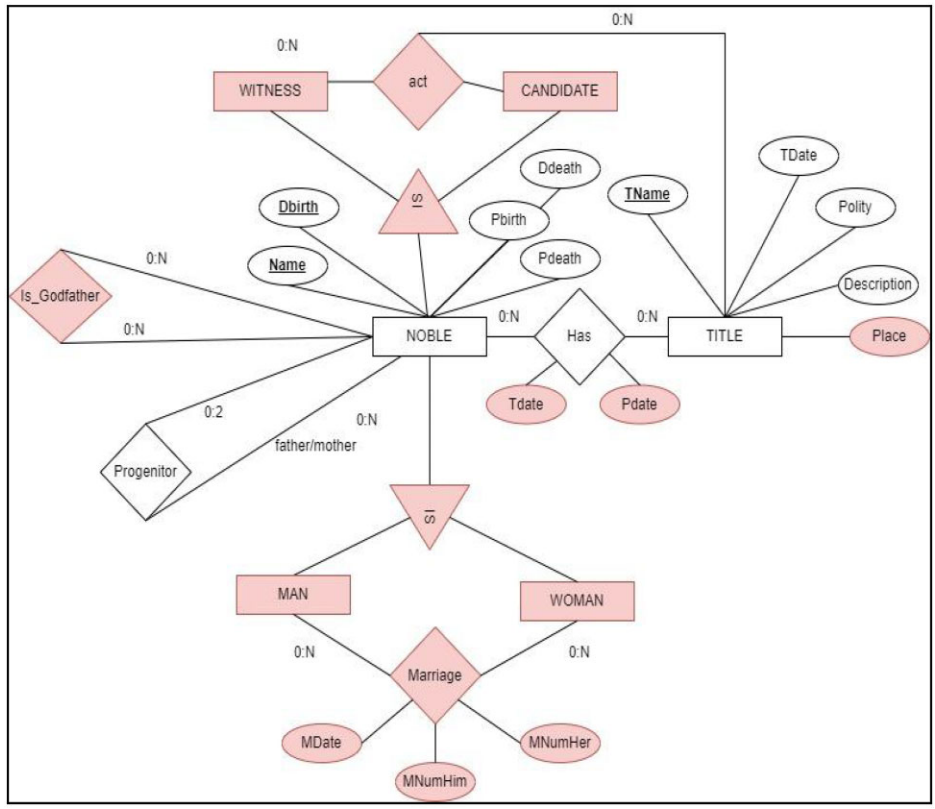


Figure 4. New E/R model, with the changes highlighted in red
Source: *AtlantoCracies* Research Group.

interested in the relationships between the different entities detected in our project, since they are just a part of the big picture. At the same time, graph databases sacrifice functionality in order to achieve speed and simplicity, and the way a graph data model is designed is very different from what we are used to, that is, the RM. In conclusion, the use of this type of NoSQL databases can be considered as a very promising option for future work in which the level of abstraction of our data will be increased, focusing only on relationships. Thus, the Document-oriented databases are the best option to achieve our current objectives.

Before explaining the reasons behind that decision, it is necessary to point out that in much of the scientific literature the term 'documentary database' is misused, since it tends to be confused with those digital repositories, websites, research projects, etc. that, regardless of their technology, store digitized historical documents (Giunta et al., 1996) This confusion probably comes from the fact that, for a long time, and given the hegemony of relational systems, when various authors refer to some software that stores documentation, they end up referring to it as a documentary database. For example, the Documentary Information Center, created in 1984, and which houses historical documents from Cuba, Italy, Chile, and France, is mentioned as a 'documentary database' in several works. Some handbooks also have used the term 'documentary databases' in their title, when, inside, the basic elements of RMs are explained (Abadal and Codina, 2005).

As it was said before, DoNoSQL databases work with semi-structured data but have functionalities and properties very close to the RM. First of all, DoNoSQL databases have an architecture based on sets of 'documents' called 'collections', equivalent to the tables of the RM. In turn, each 'document' (dDoD from now on) stored in the collection is the equivalent of the row stored in the RM tables, as can be seen in Fig. 5. A dDoD is organized into Key-Value pairs, like it is done also in markup languages (Friesen, 2019). That is, the dDoD becomes a container in which labeled metadata which make up its attributes is stored, and which is expressed indicating any possible value contained, either a text string, dates, binary data, or also arrays. Besides, the flexibility of this system allows that, although it is advisable to follow a basic scheme within a collection, each dDoD may have its own internal structure, being completely different

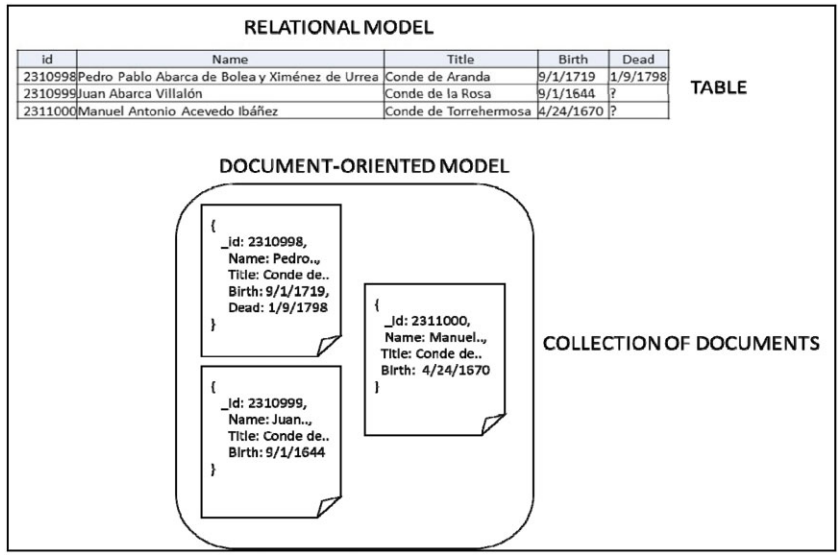


Figure 5. Relational versus Document-oriented database

Source: *AtlantoCracies* Research Group.

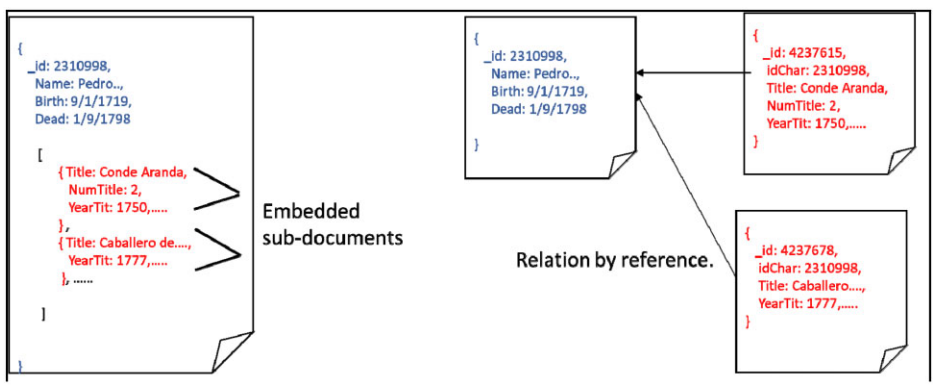


Figure 6. Relationships in a Document-oriented database. On the left side, the main document (in blue) is related to those sub-documents that are embedded (in red). On the right side, the same documents are related using the identification attribute of the main document as a reference.

Source: *AtlantoCracies* Research Group.

from any other dDoD stored in the same collection. In this way, the data model in a DoNoSQL database is as dynamic as the historian's own research is flexible and changing.

Second, in DoNoSQL databases, there are two ways to relate documents to each other: by reference or by embedded sub-documents (Fig. 6). This fact is very important when designing the database model because, although using references to other documents tries to emulate the way tables are related in the RM, the embedded sub-documents are the best choice to take advantage of the great performance of the DoNoSQL databases. So, it is necessary to get to a balance when using these two relationship methods.

Third, the DoNoSQL databases are very easy to use as well as very powerful. A good example is the database tool that is currently being used in the *AtlantoCracies* project: MongoDB (Edward and Sabharwal, 2015; Hows *et al.*, 2015). The data access language used by MongoDB is very rich and allows complex queries, even with several levels of embedded sub-documents: document filtering, projection of attributes, ordering, hierarchical queries, etc. (Jaraba Navas *et al.*, 2016) In addition, MongoDB includes optimization tools that are very similar to those we can find in RM databases: several types of indexes, an optimization tool that chooses the best execution plan for every query, multi-document transactions, etc.

All the reasons argued above were the ones that motivated the choice of a DoNoSQL database as the data processing tool required by the *AtlantoCracies* project. Once the decision was made, the database model was designed and an approach related to the pre-processing and storage of the research data was developed. Figure 7 shows the *AtlantoCracies* database model.

There is a main collection named Person. Its documents are composed of a set of attributes and four arrays of sub-documents. It is worth to say that, thanks to the flexibility of this type of databases, although the documents have a number of frequent and stable attributes, these attributes can vary in number, type, and meaning. Every array contains a list of sub-documents on the relations known about a person, the events in which a person was involved, the titles of nobility that a person has, and the different positions he/she held. As it can be observed, the different documents are related using the two methods explained in a previous paragraph (Fig. 6): by reference or using sub-documents. A person is related with his/her titles, positions, events, and relations using the embedded sub-documents. At the same time, a document Relation or a document Title is related to a document Person using the person id reference.

In Fig. 8, our new proposal is presented. It is a process designed to be as flexible as possible and thus facilitate the interaction between the researcher, the sources of data, and the database. First, the research data are extracted from the historical sources and stored in an intermediate data repository. Each extracted person can be different from the rest in terms of the number, type, and meaning of their attributes, although a basic scheme is followed. Next, a set of different .csv files are generated from the data repository, one for each historical person. This set of files is the input to a software developed in Java. This software automatically detects the data structure of each person, transforms it into a document, and stores this document in a MongoDB database located in the cloud. The main advantage of this methodology is that, thanks to the flexibility of the semi-structured data, it is not necessary to change the

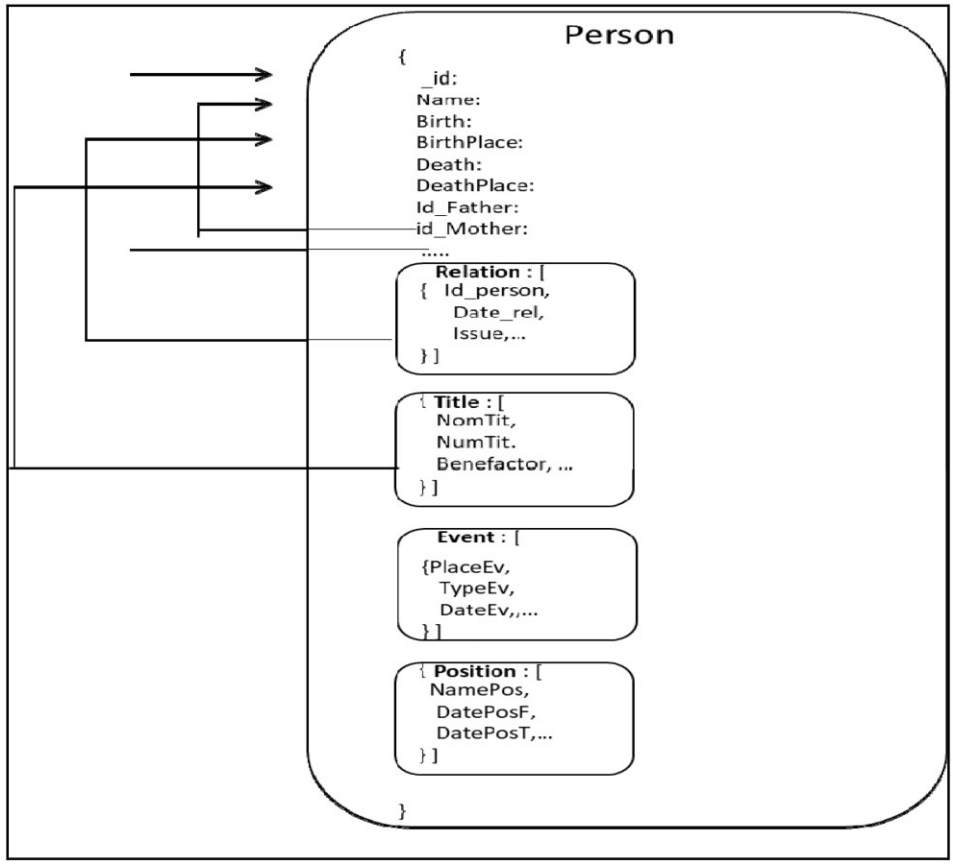


Figure 7. The *AtlantoCracies*'s database model is shown. A collection named Person is composed of a set of attributes and arrays of embedded documents.

Source: *AtlantoCracies* Research Group.

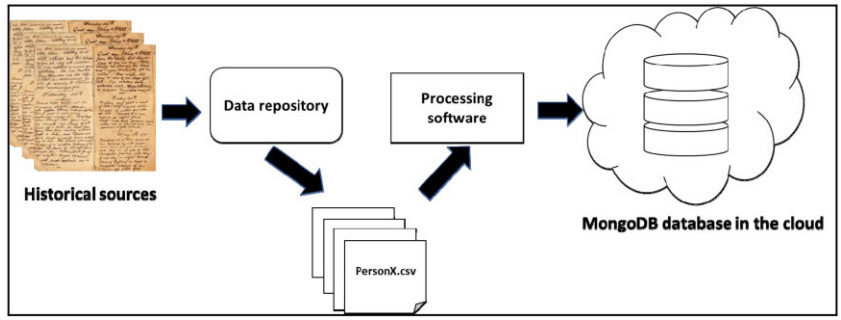


Figure 8. The proposed approach for the pre-processing and storage of the research data.
Source: *AtlantoCracies* Research Group.


```
_id: ObjectId('609cd66b9aab3a99799344bc')
datebirth: "1716-01-21"
placebirth: "Ciudad de México"
pbirthUTM: "19.419444/-99.145556"
countryDeath: "México"
datedeath: "1786-08-29"
gender: "H"
histDeath: "Virreinato de Nueva España"
name: "Francisco Javier de Aristoarena de Lanz"
pDeathUTM: "25.816667/-100.133333"
pageFr: "Null"
personId: 117
placeDeath: "Zacatecas"
source: "Ramón Maruri Villanueva Personal Database (American Spanish Nobility) …"
sourceType: "Database"
titles: Array
 0: Object
     continental: "América"
     countryTit: "España"
     dateTit: "1777"
     histTit: "Virreinato de Nueva España"
     nomTit: "Condado de Casa Fiel"
relations: Array
 0: Object
     idPerson: 710
     namePerson: "Josefa Tagle Bracho"
     placeRel: "Ciudad de México"
     placeRUTM: "19.419444/-99.145556"
     treatment: "Doña"
     typeRel: "Matrimonio"
 1: Object
     idPerson: 769
     dateRelF: "1770-12-03"
     namePerson: "María Guadalupe de la Campa Cos"
     placeRel: "Zacatecas"
     placeRUTM: "25.816667/-100.133333"
     treatment: "Doña"
     typeRel: "Matrimonio"
treatment: "Don"
countryBirth: "México"
hist30Birth: "Virreinato de Nueva España"
histBirth: "Virreinato de Nueva España"
continentBirth: "América"
```


Figure 9. An example of a document extracted from the *AtlantoCracies* MongoDB database.
Source: *AtlantoCracies* Research Group.

database model every time a new attribute, research line, or hypothesis is detected. As an example, in less than 2 months, about 3,000 historical people from different historical sources have been stored in our database without the need to modify it. In Fig. 9, an example of a document of our database is shown. This document is about *Don Francisco Javier de Aristoarena de Lanz*, with id 117, and it contains a list of specific attributes, one subdocument about his nobility title and two subdocuments about his relationships with other historical characters with id 710 and id 769.

7 Conclusion

Relational databases have been of great use to historians since their application to their research as early as the last third of the 20th century. However, the rules that determine their operation, in terms of its rigidity, structure, and treatment of the data, often become an obstacle. Historical research is, in most cases, dynamic and represents a multi-objective system; therefore, changes in the sources of data, research objectives, and hypotheses are added very frequently. As a consequence, a lack of flexibility in the database model implies a great cost in time and money and, in a project with a very limited shelf life, it could be an important issue. Such limitations have led historians to explore the technical possibilities of graphical models, proprietary software, and some typologies of NoSQL databases, such as graph-oriented databases. However, in the case of *AtlantoCracies*, a Document-oriented database, such as MongoDB, has been selected as a solution to the lack of flexibility of RMs. In our case, our new approach has allowed the insertion of different types of information in a totally heterogeneous way and according to the historical sources used. It is also evident that, given its advantages, this is not only the case for *AtlantoCracies* and that similar projects would benefit from it. It is worth mentioning that not all relational databases are capable of being adapted to a Document-based database model. That is, in those cases in which we have a large and complex relational database, made up of many tables, relations, and different entities, adapting to a Document-oriented model is not recommended, but it could be a good solution to transform only a specific part of it.

As it has been commented in this work, our proposal is about the extraction, pre-processing, and storage of data. However, once the data are ready, it is time to extract useful knowledge from the database. As a future work, using the data access language of MongoDB, several database queries will be implemented to ask questions related to the objectives of our project. Furthermore, the use of more advanced techniques, like Data Mining ones, to track patterns and hidden coincidences, opens up new and promising approaches to the subject. Finally, the possibilities of connecting MongoDB database with tools like Mahout (2000), SNA like GEPHI, or GIS like QGIS will be enormously beneficial.

Funding

This work has been financed by the Fondo Europeo de Desarrollo Regional (FEDER) and the Consejería de Economía, Conocimiento, Empresas y Universidades of the Junta de Andalucía (Regional Government of Andalusia) and it is part of the activities of the research group UPO-1264973 "In search for the Atlantic aristocracies. Latin America and the peninsular Spanish elites, 1492–1824," (PI, Bartolomé Yun-Casalilla), and also of the PAIDI research group HUM 1000 "The History of globalization: violence, negotiation and interculturality" (PI Igor Pérez Tostado).

Authors' contributions

Manuel Diaz-Ordoñez (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing), Domingo Savio Rodríguez Baena (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing), and Bartolomé Yun-Casalilla (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing).

Acknowledgement

The authors warmly thank María Jesús Milán Agudo for the collection and introduction of data into the data base, as well as the interface's development.

Notes

1. Factoid is a typical concept of Prosopography, introduced by Dion Smythe in the research project 'Prosopography of the Byzantine Empire'. It consists of the categorization of all the information referring to an individual that is found in the sources and that is stored. (Smythe, 2007, p. 30).
2. An internet search was done, using various repositories such as: Berkely Library (University of California), Smithsonian Libraries, Institute of Historical Research (University of London), Vanderbilt University or the European University Institute.
3. Some key publications on early modern European aristocracies in: (Asch, 2003; Dewald, 2007; Scott, 2007; Minvielle, 2009; Janssens and Yun Casalilla, 2005).

References

Abadal E. and Codina Ll. (2005). *Bases de datos documentales: características, funciones y método*. Madrid: Síntesis.

Akoka, J., Comyn-Wattiau, I., du Mouza, C., and Prat, N. (2019). Design science research for the humanities - the case

of prosopography. In Tulu, B., Djamasbi, S., and Leroy, G. (eds), *Extending the Boundaries of Design Science Theory and Practice*. Cham, Switzerland: Springer Nature Switzerland, pp. 239–53.

Akoka, J., Comyn-Wattiau, I., Lamassé, S., and du Mouza, C. (2020). Contribution of conceptual modeling to enhancing historians' intuition - application to prosopography. In Dobbie, G., Frank, U., Kappel, G., Liddle, S. W., and Mayr, H. C. (eds), *Conceptual Modeling. ER 2020. Lecture Notes in Computer Science*, vol. **12400**. Cham, Switzerland: Springer, pp. 164–173.

American National Biography. (2020). Research Ideas. https://www.anb.org/ (accessed 12 November 2020).

Ancestry Library Edition. (2020). Login page. https://ancestrylibrary.proquest.com/aleweb/ale/do/login (accessed 1 November 2020).

Andresen, J. and Madsen, T. (1996a). Dynamic classification and description in the IDEA. *Idea Archeologia e Calcolatori*, **7**: 591–602.

Andresen, J. and Madsen, T. (1996b). IDEA- the integrated database for excavation analysis. *Interfacing the Past: Computer Applications and Quantitative Methods in Archaeology CAA95: Analecta Praehistorica Leidensia 28.II, 1996*. Leiden: Leiden University Press, pp. 3–14.

Angles, R. and Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, **40**(1): 1–39.

Asch, R. G. (2003). *Nobilities in Transition, 1550-1700: Courtiers and Rebels in Britain and Europe*. London: Bloomsbury Academic

Aram, B., López Fernández, A., Muñiz Amian, D., *et al.* (2021a). ArtEmpire Database. https://artempire.cica.es/historic/documents/ (accessed 8 October 2021).

Aram, B., Lopez-Fernandez, A., and Muñiz-Amian, D. (2021b). The integration of heterogeneous information from diverse disciplines regarding persons and goods. *Digital Scholarship in the Humanities*, **36**(2): 255–67.

Bărbulescu, M., Grigoriu, R. O., Halcu, I., *et al.* (2013). Integrating of structured, semi-structured and unstructured data in natural and build environmental engineering. In *11th RoEduNet International Conference*, Sinaia, Romania, 17–19 January 2013.

Berners-Lee, T. and Hendler, J. (2001). Publishing on the semantic web. The coming internet revolution will profoundly affect scientific information. *Nature*, **410**(6832): 1023–24.

Beyond. (2022). *Ireland's Virtual Record Treasury*. https://beyond2022.ie/ (accessed 5 October 2020).

Binding, K. C. and Tudhope, D. (2011). A star is born: some emerging semantic technologies for archaeological resources. In Jerem, E., Redő, F., and Szevérenyi, V. (eds), *On the Road to Reconstructing the Past: Computer Applications and Quantitative Methods in Archaeology (CAA): Proceedings of the 36th International Conference*, Budapest, 2–6 April 2008, pp. 111–6.

Bodenhamer, D. J. (2008). History and GIS: implications for the discipline. In Knowles, A. K. (ed) *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands CA: ESRI Press, pp. 219–34.

Bourdieu, P. and Passeron, J. C. (1977). *Reproduction in Education, Society and Culture*. Beverly Hills, CA: Sage.

Bourdieu, P. (1989). *La Noblesse d'État: Grandes Ecoles et Esprit de Corps*. Paris: Éditions de Minuit.

Bradley, J. and Pasin, M. (2013). Structuring that which cannot be structured: a role for formal models in representing aspects of medieval Scotland. In Hammond, M. (ed), *New Perspectives on Medieval Scotland, 1093–1286*. Woodbridge, Suffolk: Boydell and Brewer, pp. 203–14.

Burton, V., Blomeyer, R., Fukada, A., and White, S. J. (1987). Historical research techniques: teaching with database exercises on the microcomputer. *Social Science History*, **11**(4): 433–48.

Candan, K. S., Liu, H., and Suvarna, R. (2001). Resource description framework metadata and its applications. *ACM SIGKDD Explorations Newsletter*, **3**(1): 6–19.

Carvalho, J. and Campos, R. (2007). Interpersonal networks and the archaeology of social structures; using social positioning events to understand social strategies and individual behaviour. *Revista de História da Sociedade e da Cultura*, **7**: 175–93.

CDS/ISIS Database Software. (1985). UNESCO and Information processing tools. https://wayback.archive-it.org/10611/20160102101854/http://portal.unesco.org/ci/en/ev.php-URL_ID=2071&URL_DO=DO_TOPIC&URL_SECTION=201.html (accessed 1 November 2020).

Chen, P. P. S. (1976). *The Entity-Relationship Model: Toward a Unified View of Data*. Cambridge, MA: M.I.T. Center for Information Systems Research.

China Historical Christian Database. (2020). Mapping China's Christian Past. https://chcdatabase.com/ (accessed 7 December 2020).

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, **13**(6): 377–87.

Codd, E. F. (1971). Further normalization of the data base relational model. *Research Report/RJ/IBM*, **909**: 1–33.

Codd, E. F. (1985). Is your DBMS really relational? *Computerworld*, **19**(41): 1–2.

Dedieu, J. P. (2000). Un instrumento para la historia social: la base de datos Ozanam. *Cuadernos de Historia Moderna*, **XXIV 24**: 185–205.

Dedieu, J. P. (2004). Les grandes bases de données: una nouvelle approche de l'histoire sociale: le système Fichoz. *Revista da Faculdade de Letras Historia*, **5**(1): 101–14.

Dedieu, J. P. (2012). *Fichoz – A Database for Social History/Base de Données Pour l'Histoire Sociale*. https://fichoz.hypotheses.org/ (accessed 21 November 2020).

Dedieu, J. P. (2013). Fichoz 2011. Balance de una base de datos sobre la España moderna. In Jiménez Estrella, A., Lozano Navarro, J. J., Sánchez Montes, F., and Birriel Salcedo, M. M. (eds), *Construyendo Historia. Estudios en Torno a Juan Luis Castellano*. Granada: Editorial de la Universidad de Granada, pp. 185–200.

Dewald, J. (2007). *The European Nobility, 1400-1800*. Charlesbourg, Québec: Braille Jymico.

Díaz-Ordóñez, M. (2020). GECEM database, digital humanities and scientific interdisciplinary: understanding global history from the historical document to binary computer language. *GECEM Newsletter*, **4**: 5–6.

Early English Books Online. (2020). Text creation partnership. https://quod.lib.umich.edu/e/eebogroup/ (accessed 9 December 2020).

Edward, S. G. and Sabharwal, N. (2015). *Practical MongoDB: Architecting, Developing, and Administering MongoDB*. New York, NY: APress.

Eve, S. and Hunt, G. (2010). ARK: a developmental framework for archaeological recording. In ed. Posluschny, A., Lambers, K., and Herzong, I. (eds), *Layers of Perception Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, Berlin, Germany, 2–6 April 2007.

Feeney, K. (2016). The dacura data curation system. *IFIP Advances in Information and Communication Technology*, **482**: 15–20.

Friesen J. (2019). *Java XML and JSON: Document Processing for Java SE*. Berkeley: Apress.

Genet, J. P., Idabal, H., Kouamé, T., Lamassé, S., Priol, C., and Tournieroux, A. (2016). General introduction to the 'studium' project. *Medieval Prosopography*, **31**: 156–72.

Giunta, M. A., Hartgrove, J., and Dowd, M-J. M. (1996). *The Emerging Nation: A Documentary History of the Foreign Relations of the United States Under the Articles of Confederation, 1780-1789*. Washington D.C.: National Historical Publications and Records Commission.

Gutmann, M. P. (1987). *Managing a Large Historical Database with Relational Database Software: A Report on the Fredericksburg Project*. Austin: Texas Population Research Center and University of Texas at Austin.

Harvey, C. and Press, J. (1996). *Databases in Historical Research: Theory, Methods and Applications*. New York, NY: St. Martin's Press.

Hows, D., Membrey, P., Plugge, E., and Hawkins, T. (2015). *The Definitive Guide to MongoDB: A Complete Guide to Dealing with Big Data Using MongoDB*. New York, NY: APress.

Janssens, P. and Yun-Casalilla, B. (2005). *European Aristocracy and Colonial Elites: Patrimonial Management Strategies and Economic Development, 15th-18th Centuries*. Aldershot and Burlington: Ashgate.

Jaraba Navas, P. C.Guacaneme Parra, Y. C., and Rodríguez Molano, J. I. (2016). Big data tools: haddop, mongoDB and weka. In Tan, Y. and Shi, Y. (eds), *Data Mining and Big Data First International Conference, DMBD 2016*, Bali, Indonesia, 25–30 June 2016, Proceedings, pp. 449–456.

Jensen, P. (2018). *Approaching Reality: Integrating Image-Base 3D Modelling and Complex Spatial Data in Archaeological Field Recording*. PhD dissertation, Aarhus University.

Kantabutra, V. (2009). *Intentionally-Linked Entities: A General-Purpose Database System*. FreePatentsOnline 20090319564, filed June 23, 2009, and issued December 24, 2009.

Kantabutra, V., Owens, J. B., and Crespo Solana, A. (2014). Intentionally-Linked Entities: a better database system for representing dynamic social networks, narrative geographic information, and general abstractions of reality. In Crespo Solana, A. (ed), *Spatio-temporal Narratives: HGIS and the Study of Trading Networks (1500–1800)*. Cambridge: Cambridge Scholars Press, pp. 56–78.

Kristel, C. M. and Blanke, T. (2013). Integrating holocaust research. *International Journal of Humanities and Arts Computing*, **7**(1–2): 41–57.

Lohmann-Villena, G. (1993). *Los Americanos en las Ordenes Nobiliarias*. Madrid: CSIC.

Mahout, A. (2020). Apache Mahout (TM). https://mahout.apache.org/ (accessed 3 December 2020).

Morgan, T. P. (2014). 'How Ancestry.com Branched out and Embraced Scale.' *EnterpriseAI*. January 9, 2014.

Merry, M. (2020). 'Designing Databases for Historical Research.' *Postgraduate Online Research Training*. https://port.sas.ac.uk/mod/book/view.php?id=75 (accessed 11 January 2021).

Miller, E. J. (1998). An introduction to the resource description framework. *Bulletin of the American Society for Information Science*, **25**(1): 15–9.

Minvielle, S. (2009). *Dans l'intimité Des Familles Bordelaises: Les Élites et Leurs Comportements Au XVIIIe Siècle*. Bordeaux, France: Éd. Sud ouest.

Molina Recio, R. (2002). De la utilidad y los inconvenientes de la informática para la historia. *Ámbitos Revista de Estudios de Ciencias Sociales y Humanidades*, **8**: 107–16.

Nayak, A., Poriya, A., and Poojary, D. (2013). Type of NOSQL databases and its comparison with relational databases. *International Journal of Applied Information Systems*, **5**: 16–9.

NoSQL Relational Database Management System: Home Page. (2021). NoSQL: a non-SQL RDBMS. http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/NoSQL/Home%20Page (accessed 12 January 2021).

Owens, J. B. and Kantabutra, V. (2020). A research scheme for a world history of the world. *Entremons: UPF Journal of World History [Universitat Pompeu Fabra, Barcelona]*, **11**: 69–98.

Peregrine, P. N., Brennan, R., Currie, T., *et al.* (2018). Dacura: a new solution to data harvesting and knowledge extraction for the historical sciences. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **51**(3): 165–74.

Perez-García, M. and Díaz-Ordóñez, M. (2021). 'GECEM Project Database Version 2021.' *GECEM Project Database*. https://gecemdatabase.eu (accessed 22 November 2021).

Perez-García, M. and Díaz-Ordóñez, M. (2022). GECEM project database: a digital humanities solution to analyse complex historical realities in early modern China and Europe. *Digital Scholarship in the Humanities*, **38**: 296–312.

Perez-García, M., Wang, L., Svriz Wucherer, P. M. O., Fernández de Pinedo, N., and Díaz-Ordóñez, M. (2022). The GECEM database and big data applications to "new" global history: circulation of global goods and trade networks in China and Europe, Seventeenth-Eighteenth Centuries. *Itinerario. Journal of Imperial and Global Interactions*, **46**(1): 14–39.

Pivert, O. (ed) (2018). *NoSQL Data Models: Trends and Challenges*. Vol. **1**. London: ISTE.

Project: The Trans-Atlantic and Intra-American Slave Trade Databases. (2020). *Database*. https://www.slavevoyages.org/voyage/database (accessed 16 January 2021).

Rossi, F., Villa-Vialaneix, N., and Hautefeuille, F. (2014). Exploration of a large database of French notarial acts with social network methods. *Digital Medievalist*, **9**: 1–16.

Salas Tovar, E., Fernández Freire, C., Uriarte González, A., Fraguas Bravo, A., del Bosque González, I., and Vicent García, J. M. (2016). 'IDEARQ.' *Infraestructura de Datos Espaciales de Investigación Arqueológica*. https://digital.csic.es/handle/10261/139851 (accessed 22 November 2020).

Sarda, N. L. (1990). Extensions to SQL for historical databases. *IEEE Transactions on Knowledge and Data Engineering*, **2**(2): 220–30.

SARIT. (2021). Search and Retrieval of Indic Texts (New). https://sarit.indology.info/exist/apps/sarit/works/ (accessed 14 January 2021).

Scott, H. M. (2007). *The European Nobilities in the Seventeenth and Eighteenth Centuries*. **2 2**. Basingstoke: Palgrave Macmillan.

Seshat. (2020). Global History Databank. http://seshatdatabank.info/ (accessed 22 November 2020).

Smith, J. A., Owen, J. F., and Gray, J. R. (2011). CloudCAP: a case study in capacity planning using the cloud. *Lecture Notes in Computer Science*, **6966**: 110–17.

Smythe, D. C. (2007). A whiter shade of pale: issues and possibilities in prosopography. In Keats-Rohan Katharine S. B. (ed), *Prosopography Approaches and Applications: A Handbook*. Oxford: University of Oxford, pp. 127–37.

Studium Parisiense Database. (2020). Projet Studium Parisiense. http://studium.univ-paris1.fr/?action=index (accessed 16 October 2020).

Tchounikine, A., Miquel, M., Pécout, T., and Bonnaud, J. L. (2018). Prosopographical data analysis. Application to the angevin officers (XIII–XV centuries). *Journal of Data Mining and Digital Humanities*, **2018**: 1–12.

The Tibetan Buddhist Resource Center (TBRC). (2020). The Buddhist Digital Archives. https://library.bdrc.io/ (accessed 4 January 2021).

Thebarum Fabula. (2020). Biblioteca digital del mito tebano. http://thebarumfabula.usc.es/ (accessed 24 November 2020).

Turchin, P. (2014). The SESHAT databank project: the 2014 report. *Cliodynamics*, **5**(1): 58–64.

Turchin, P. (2017). Seshat: global history databank publishes first set of historical data. *Cliodynamics*, **8**(1): 75–9.

Turchin, P., Brennan, R., Currie, T., *et al.* (2015). Seshat: the global history databank. *Cliodynamics*, **6**(1): 77–107.

Uriarte González, A., Fernández Freire, C., Fraguas Bravo, A., *et al.* (2017). IDEArq-C14: una infraestructura de datos espaciales para la cronología radiocarbónica de la prehistoria reciente ibérica. *CEUR Workshop Proceedings*, **2024**: 209–25.

Viteri, H. S. and Bayas M. A. (2020) La transición del manejo de bases de datos entre el modelo SQL al NOSQL en la enseanza de carreras tecnológicas. *Journal of Science and Research Science and Research Magazine*, **5**: 29–48.

Yun-Casalilla, B. (2002). *La Gestión del Poder. Corona y Aristocracia en Castilla, Siglos XVI-XVIII*. Madrid: Akal.

Yun-Casalilla, B. (2019). *Iberian World Empires and the Globalization of Europe, 1415–1668*, London: Plagrave-Macmillan.

Yun-Casalilla, B. (2022). Early modern Iberian empires, global history and the history of early globalization. *Journal of Global History*, **17**: 539–61.

Yun-Casalilla, B. (dir). (2008). *Las redes del Imperio. Elites Sociales en la Articulación del Imperio Español, 1492–1714*. Madrid: Marcial Pons.

Yun-Casalilla, B. and Redondo-Álamo, A. (2008). Aristocracias, identidades y espacios políticos en la monarquía compuesta de los Austrias. La Casa de Borja (ss. XVI y XVII). *Homenaje a Don Antonio Domínguez Ortíz*. Granada: Universidad de Granada, pp. 759–71.

Appendix 1

The data gathered from all the projects that have been asked for information about the technology used in their databases is presented in the following table. The first column represents the name of the project, the type of database technology is shown in the second column and, finally, the last column contains the URL of the online resources of the projects.

Project	Database model	URL
RIBApix. RIBA Architecture Image Library Database	Relational	https://www.architecture.com/image-library/
Continental Origins of English Landholders, 1066-1166 database	Relational	http://www.coelweb.co.uk/coeldatabase.html
Catalunya durant el franquisme	NA	http://basedadesfranquisme.uab.cat/
China Rural Reconstruction Dataset	NA	https://www.shss.ust.hk/lee-campbell-group/projects/csscd-project/
Pompeii Archaeological Research Project: Porta Stabia (PARP: PS)	Relational	https://classics.uc.edu/pompeii/index.php/home.html
Prosopographie der mittelbyzantinischen Zeit	Relational	http://www.pmbz.de/splashscreen.ger.php
Proyecto de Investigación Arqueológico Regional Ancash mobile database (PIARA)	Relational	http://www.piaraperu.org/digital-archaeology.php
Excavation FastiOnline	Relational	http://www.fastionline.org/excavation/index.php
GECEM database (China and Europe: Trade Networks, Consumption and Cultural Exchanges in Macau and Marseille (1680-1840)	Relational	https://www.gecemdatabase.eu/
Unknown No Longer: Virginia Slave Names Database	NA	https://www.virginiahistory.org/collections/unknown-no-longer-database-virginia-slave-names
The Cambridge Group for the History of Population and Social Structure	Relational	https://www.campop.geog.cam.ac.uk/datasets/
Sonoma Historic Artifact Research Database (SHARD)	Relational	http://web.sonoma.edu/asc/shard/
Artefacts©, Online Encyclopaedia of Archaeological Small Finds.	Relational	https://artefacts.mom.fr/
ArtEmpire Database ERC CoG 648535	Relational	https://artempire.cica.es/
BRASILHIS Database. Personal Networks and Circulation in Brazil during the Hispanic Monarchy (1580-1640)	Relational	http://brasilhis.usal.es/en
Cartells de la Guerra civil espanyola	Relational	https://www.bib.uab.cat/comunica/cedoc/cartellsgc/
DICOBJ. Dictionnaire des objets protohistoriques de Gaule méditerranéenne	Relational	http://syslat.fr/SLC/DICOBJ/d.index.html
Old Bailey Proceedings Online	Relational	https://www.oldbaileyonline.org/static/Data.jsp
Origins of English Landholders, 1066-1166 database (COEL)	Relational	http://pase.ac.uk/jsp/index.jsp
Sound Toll Registers online	Relational	http://www.soundtoll.nl/index.php/en/over-het-project/str-online
The Trans-Atlantic and Intra-American slave trade databases	Relational	https://www.slavevoyages.org/voyage/database
Indianola Immigrant Database	Relational	https://vrhc.uhv.edu/manuscripts/indianola/
China Historical Christian Database	Graph DBMS/RDF Stores	https://chcdatabase.com/
Making the History of 1989	Relational	https://chnm.gmu.edu/1989/
Canmore—Royal Commission Ancient and Historial Monuments of Scotland	Relational	https://canmore.org.uk/
PARES. Portal Archivos Españoles en Red	Relational	http://pares.mcu.es/ParesBusquedas20/catalogo/search
IDEARQ. Infraestructura de Datos Espaciales en Arqueología	Relational	http://www.idearqueologia.org/visualizador_idearq/
Beyond 2022 Virtual Record Treasury	Graph DBMS/RDF Stores	https://beyond2022.ie/

(continued)

(continued)

Project	Database model	URL
Seshat: Global History Databank	Graph DBMS/RDF Stores	http://seshatdatabank.info/
STUDIUM Parisinum	Graph DBMS/RDF Stores	http://studium.univ-paris1.fr/
Histories of the National Mall	Relational	http://mallhistory.org/
Southwest Harbor Public Library Digital Archive	Relational	http://swhplibrary.net/home/
Marple Local History Society Archives	Relational	https://www.marplelocalhistorysociety.org.uk/archives/
DIVA-HisDB	NA	https://diuf.unifr.ch/main/hisdoc/diva-hisdb
Caffè Lena History Project's Searchable Database	Relational	http://caffelenahistory.org/index.php?25
Online Medieval Sources Bibliography	Relational	http://medievalsourcesbibliography.org/
Medieval Londoners database (MLD)	Relational	https://medievallondoners.ace.fordham.edu/search/

NA = Not Applicable.