Models of Data Extraction and Architecture in Relational Databases of Early Modern Private Political Archives
edited by Dorit Raines

Extraction, Architecture and Recovery of Family Correspondence Data
The Platform "EpiCAT. Family Letters from Catalonia (Sixteenth-Nineteenth Centuries)"

Javier Antón Pelayo
Universitat Autònoma de Barcelona, España

Abstract The optimal approach for documenting information within a uniform documentary series involves the establishment of a database. To ensure the functionality of datafication, it is recommended to implement a design that encompasses accessibility, interoperability, and efficient data retrieval. An exemplar instance of this is represented by the EpiCAT platform, designed for the curation of family correspondences originating from Catalonia during the span of the sixteenth to nineteenth centuries. While this platform facilitates the association of metadata, its utilization in the context of text mining remains unexplored. This emerging paradigm necessitates novel viewpoints in the formulation of historical discourse.

Keywords Datafication. Data retrieval. EpiCAT platform. Catalonia. Family correspondences. Interoperability.

Summary 1 Introduction. –2 The Databases. – 3 The EpiCAT Platform. – 4 The Horizon of Text Mining. – 5 Conclusions.

Edizioni Ca'Foscari

Studi di archivistica, bibliografia, paleografia 8
e-ISSN 2610-9093 | ISSN 2610-9875
ISBN [ebook] 978-88-6969-919-1 | ISBN [print] 978-88-6969-920-7

Peer review | Open access
Submitted 2023-10-03 | Accepted 2024-01-22 | Published 2025-05-22

DOI 10.30687/978-88-6969-919-1//009

1 Introduction

The most efficient approach to document and record information in a homogeneous documentary series is to establish a comprehensive database. For effective datafication, it is crucial to implement a design that considers accessibility, interoperability, and agile data retrieval.

EpiCAT is a specialized platform designed to manage a vast collection of family letters dating from 1500 to 1850, housed in public archives in Catalonia.[1] Catalonia serves as an extensive yet manageable framework, acting as a laboratory that offers substantial evidence to verify specific historical processes. The primary objective of this project is to demonstrate the utility of this resource when approached systematically and on a large scale. Through this approach, the statements found in official documents can be supplemented, qualified, and at times, amended. Moreover, it allows for an adequate examination of individuals' private and intimate spheres (Borkosky 2002), as well as an in-depth exploration of seemingly mundane cultural practices and historical banalities. Additionally, when exploring these extensive documentary materials, it becomes common to encounter unprecedented epistolary sociability, revealing political, economic, social, cultural, and various other strategic pursuits.

For instance, in the examination of a marriage, official documents such as sacramental marriage books from parishes or notarial marriage chapters certify the final decision with clarity, devoid of any hesitations or consideration of alternative choices. Conversely, letters exchanged between relatives and friends before the wedding unveil the array of potential candidates the family was contemplating, including their strengths, weaknesses, economic status, and the level of compatibility between potential spouses. This broader scope of marriage possibilities allows for an understanding of the relational aspirations of a household, their evaluation of different options, and the underlying family strategy intended for implementation.

To collect and systematize this scattered and diverse information from various sources, a multifunctional database is essential. This is where the expertise of historians intersects with that of information resource professionals. While digital tools can be learned and employed, highly sophisticated computer procedures may necessitate collaboration with specialized technical personnel.

Recently, the Swedish historian Mats Fridlund classified historians into three categories: Historian 1.0 utilizes search engines and personal databases; Historian 2.0

1 https://epicat.uab.cat/.

> systematically use[s] various digital applications and quantitative methodologies for big-data text and data mining, calculations and visualisations, such as topic modelling, network analysis and text and data scraping. Most of these methods necessitate investments in acquiring expertise in or collaborators skilled in coding and database methodologies. (Fridlund 2020, 77)

On the other hand, Historian 1.5 adopts a hybrid methodology that combines quantitative and qualitative, automatic and manual methods, without the explicit use of programming and coding. The question one must ask is, at which level do I find myself? Which level of expertise is required for my research?

2 The Databases

Databases play a fundamental role in historiographical research. In essence, a database system enables the storage, recording, and retrieval of information from documentary sources. Typically, the data recorded in a database is sourced from specific primary sources, organized, and classified according to logical criteria and research objectives. The abundance of quantitative, serial, or numerical data offers numerous advantages, such as providing precise and verifiable evidence to support analyses and arguments, enhancing the consistency of formulated hypotheses, enabling more rigorous simulation models, and ultimately reinforcing the research's conclusions.

Nonetheless, the methodological appropriateness of databases and digital humanities, in general, has been the subject of profound reflections, which have also brought attention to some weaknesses of the digital paradigm. These include concerns about the infallibility of algorithms, the idea of a universal digital archive as a utopian ideal, and the potential lack of contextualisation in the retrieved data (Milligan 2022).

Databases, as excellent tools for historians, can be categorized into two types: relational and non-relational databases. Non-relational databases are structured following a hierarchical approach. On the other hand, relational databases are systems that organize information in tables, with the ability to connect these tables to one another. When dealing with structured and homogeneous data, the use of relational databases yields high-quality results (Gil 2021). A prime example of this would be the *EpiCAT. Family Letters of Catalonia (Sixteenth-Nineteenth Centuries)* database (Antón Pelayo et al. 2023).

3 The EpiCAT Platform

3.1 The Architecture

The development of the EpiCAT digital project was spearheaded by Alicia Calvo Burés. From 2017 to 2021, she served as the technical manager, overseeing the database structure, design, and web application development. Subsequently, she continued to be responsible for updating and maintaining the portal **[fig. 1]**.

Figure 1 The EpiCAT home page

The technical implementation of this project involved the creation of an internal application, comprising several interconnected relational databases, designed for handling all aspects related to the letters (back-end). Additionally, a web application was developed to facilitate the exploration and access of the collected information (front-end). The internal application allows researchers to work with letters stored in different archives in a decentralized and simultaneous manner. On the other hand, the web application primarily caters to users searching for documentary records and navigating between these records and their associated entities. While the tools integrated into the portal are directly aligned with the research objectives, considerations were also given to enabling direct access for citizens and supporting the didactic use of the epistolary materials that have been incorporated. To cater to a broader audience, the consultation is available in three languages: Catalan, Spanish, and English **[fig. 2]**.

The chosen data model for this project is of the Entity-Relationship type, a widely established model for conceptual database design since its introduction by Peter Chen in 1976. Its implementation was carried out using MySQL, which remains one of the most widely used database management systems and is supported by the web hosting service of the Universitat Autònoma de Barcelona.

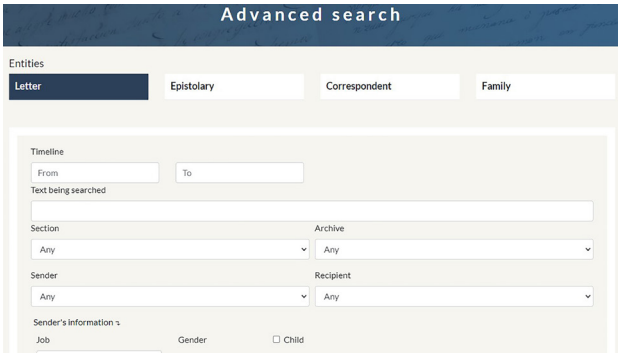

Figure 2
A glimpse of a part of the Advanced research functionality

The primary entities included in the data model are 'Epistolary', 'Letter', 'Correspondent', 'Family', and 'Fund'. The creation of the 'Epistolary' entity was driven by the necessity to consolidate letters from the same family, even if they were located in different archives. For instance, the correspondence of the Burguès family includes letters found in the National Library of Catalonia, as well as the Municipal Archive of Girona and the Archive of the Cathedral of Girona (Antón Pelayo 2005; 2013; 2019a) **[fig. 3]**.

Figure 3 Epistolary of the Burguès family

Epistolary pacts

Significant epistolary pacts.

Show 10 entries Search:

#	Start	End	Overview	Total letters	Correspondent 1	Correspondent 2
8	1811-9-26	1812-5-2	Tres cartes amoroses el 1811, quan eren nuvis, i set cartes el 1812, ja casats, separats per malaltia de Narcís.	10	Burguès i de Guàrdia, Narcís de	Caramany i de Camps, Maria dels Dolors
9	1818-7-14	1818-7-26	Cartes a la seva dona, des de Sant Martí Sescorts a Girona, on li explica les seves activitats i pregunta pels fills i la família.	6	Burguès i de Guàrdia, Narcís de	Caramany i de Camps, Maria dels Dolors
11	1826-8-30	1826-9-13	Narcís explica a la seva dona la gestió del patrimoni que fa a Coromines i pregunta per la família i amics.	4	Burguès i de Guàrdia, Narcís de	Caramany i de Camps, Maria dels Dolors
10	1823-9-27	1824-4-13	Comunicació entre Jaume, presoner a França, i el seu germà Narcís (hereu), a Girona. Descripció de la seva situació i dels seus sorollosos compromisos matrimonials.	6	Burguès i de Guàrdia, Jaume	Burguès i de Guàrdia, Narcís de

Figure 4 A part of the epistolary pacts of the Burguès family

Related letters

Letters included in the epistolary of the family Burguès

Show 10 entries Search:

#	Date	Sender	Origin	Recipient	Destination
Letter [2003]	18 de febrer de 1840	Pombo, José Ignacio de	Barcelona	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [229]	28 de gener de 1840	Burguès i de Caramany, Maria Teresa de	Reus	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [2002]	26 de gener de 1840	Burguès i de Caramany, Maria Ignàsia	Girona	Pombo, José Ignacio de	Barcelona
Letter [2001]	17 de gener de 1840	Claveria y de Haro, Rafaela de	Madrid	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [228]	16 de juliol de 1839	Burguès i de Caramany, Maria Teresa de	Reus	Burguès i de Caramany, Maria Ignàsia	Girona
Letter [227]	28 de maig de 1839	Burguès i de Caramany, Maria Teresa de	Reus	Burguès i de Caramany, Maria Ignàsia	Girona

Figure 5 Part of the related letters of the Burguès family

3.2 The Metadata

Family letters, dating back to antiquity, are typically composed on a simple support, often a sheet of paper, and adhere to a set of elements and formalities that have remained relatively unchanged over time. Despite each letter being a spontaneous and autonomous creation of an individual, this consistency allows for the extraction of abundant and standardized factual information (Antón Pelayo 2019b).

The "factual metadata" (Méndez Rodríguez 2000) that can be derived from a letter includes the names of the sender and recipient, creation date, place of origin, often the intended destination, document measurements, number of sheets in the document, a description of the document (overwritten sections, mail marks, paper

characteristics, etc.), and the language of the text. Additionally, it is beneficial to gather other factual metadata that may not appear in every letter but can help in characterizing the correspondents, such as gender, age, profession, and kinship.

On the other hand, the 'descriptive metadata' of the letters has posed one of the most challenging aspects of the EpiCAT project due to the heterogeneity, uniqueness, and variable nature of the content found in family letters. Initially, controlled descriptors or pre-established vocabularies, designed by project members, were applied. However, the dynamic content of the letters necessitated the inclusion of free descriptors to capture the nuanced aspects of epistolary communication. To ensure a minimal level of control and terminological consistency, the Tesaurus d'Història de Catalunya[2] was adopted as the reference vocabulary (Cuadrado et al. 1994).

EpiCAT incorporates two levels of descriptive metadata: 'topics' and 'subjects'. The 'topic' encompasses terms or expressions that encapsulate the overall subject of the letter, with 45 topics currently available, allowing the application of only one per letter. Labelling the global content of a letter is not always straightforward; however, certain letters may have a specific purpose, making categorization easier. For instance, love letters, travel correspondence, letters of consolation, recommendation, gratitude, and marriage are examples where labelling is more straightforward. Remarkable research has been carried out by Montserrat Jiménez Sureda (2020) on love letters **[fig. 6]**.

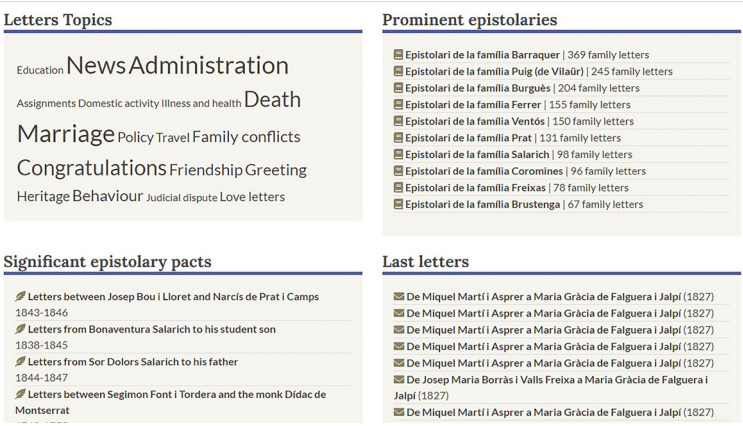

Figure 6 Lists of letters topics, epistolaries and epistolary pacts

2 https://webs.uab.cat/sibhilla/tesaurus-dhistoria-de-catalunya-0/.

The 'subjects' comprise a broader set of labels (currently 183 terms) intended to describe most, or all of the content found in each letter. There is no limit to the number of descriptors that can be applied in this field, which represents the most qualitative and refined content indexing process throughout the registration process. Thanks to this tool, searching and retrieving information on the minutest details of epistolary documentation attains historical significance. Some examples of minimal themes frequently reported in several letters include affection, domestic animals, dances, silkworms, board games, lightning, donkey's milk, lottery, pig slaughter, fashion, private oratories, locust plague, laughter, and sadness. Additionally, there are more prevalent and extensive 'subjects' communicated in the letters, encompassing children's activities, love, Carlism[3], marriage, weather, mail, death, women, education, political events, family news, and health, among others.

3.3 Data Retrieval

The design of the information retrieval system for the EpiCAT platform incorporates three distinct approaches: a simple search engine, an advanced search engine, and faceted navigation. The advanced search engine offers various filters in dropdown format, enabling users to obtain answers to specific questions. For instance, researchers can inquire about letters sent from Barcelona between 1820 and 1835, the number of letters written by women during that period, digitized letters centred around marriage themes, correspondents identified in EpiCAT with a legal profession, and the quantity of letters written by girls in the French language.

The search engine results table allows for flexible modification of the listed elements order criteria, enabling chronological sorting in ascending or descending order, as well as alphabetical sorting based on any of the listed columns. Additionally, a text box is provided for users to apply free-text filters. Moreover, data can be exported in CSV, Excel, or PDF format, if desired.

Faceted navigation, available on the exploration page, facilitates dynamic filtering of records associated with charts, allowing users to narrow down the dataset to highlight the most relevant information. This dynamic filtering functionality is implemented using Vue.js, a versatile and widely used JavaScript library **[fig. 5]**.

3 Spanish political and ideological movement that supported the candidacy of the Infante Carlos against his niece Isabel, in 1833, unleashing a civil war in which the dynastic question was a pretext to try to keep intact the structures of the Old Regime and fight against the expansion of liberalism.

Regarding the files for each entity, careful attention has been given to establishing links between them, enabling independent access to the files of individual letters, families, correspondents, or epistolary relationships.

The epistolary-type records contain various data points, such as transcriptions, subjects, themes, and family relationships between senders and recipients when available. In addition, the database identifies epistolary relationships that link two individuals in specific time and space, revolving around particular themes. These stable epistolary connections have been termed 'relevant epistolary relationships', akin to authentic 'epistolary pacts' **[fig. 4]**.

Most of the letters have undergone complete transcription. The applied transcription criteria aim to ensure the collected material's utility to both historians and philologists. To this end, the criteria seek to facilitate reading while respecting the original text's idiosyncrasies. Thus, essential interventions have been made to serve both purposes. Guiding principles include applying the contemporary punctuation system to the original texts, adopting current conventions for upper- and lower-case usage in the document transcripts, accenting transcribed words following modern regulations, and preserving the phonetic features of the originals as much as possible. Furthermore, square brackets […] have been introduced to indicate added letters that aid readability, while angle brackets <…> denote words, syllables, or letters appearing in the originals that may distort the reading.

The platform also includes some representative scans of charts, constituting approximately 10% of the registered letters. These scans, stored in PDF format, are archived in the UAB (Universitat Autònoma de Barcelona) Digital Document Repository – an open access tool – with policies for document preservation and version control using Git system,[4] ensuring their integrity and long-term accessibility.[5]

4 `https://lore.kernel.org/git/xmqqleh3a3wm.fsf@gitster.g/T/#u`.

5 `https://www.uab.cat/web/our-collections/uab-digital-repository-of-documents-345777080660.html`.

4 The Horizon of Text Mining

A more intensive and detailed approach to annotating the elements of a letter involves using the XML-TEI system, which has been employed in projects like the *Digital Archive of Letters in Flanders*[6] and *P.S. Post Scriptum*.[7] However, it should be noted that the TEI system primarily serves linguistic purposes, and its implementation demands a significant time investment.

While there are automatic marking procedures and experimental methods for the automatic transcription of historical documents, manual review remains essential. In the case of EpiCAT, the TEI procedure was recently considered by our team, and it was concluded that the time involved does not justify the benefits, particularly for historians. The primary objective of analysing letters in this context is to provide insight into historical processes or subjects that are challenging to document (Mora Mellado 2022). These may include parent-child relationships, expressions of emotions, manifestations of friendship, female sociability, educational strategies, family solidarity, and more.

Family letters incorporated into EpiCAT are transformed into structured data upon registration, and the attached textual transcriptions are intended to facilitate reading while preserving the unique writing characteristics of each author from a time when languages lacked standardisation. As such, specific transcription criteria are applied to capture the lexical peculiarities of individual letter writers. Conversely, text mining necessitates text data preprocessing, involving cleansing and transforming the text into a usable format. This process risks discarding valuable information pertinent to historiographic work in favour of identifying patterns and extracting information from vast datasets. It is important to recognize that analysing the texts of a social network, such as Twitter-X, differs significantly from studying private epistolary correspondence from the Modern Age.

6 https://ctb.kantl.be/project/dalf/.

7 http://teitok.clul.ul.pt/postscriptum/index.php?action=home.

5 Conclusions

Digital tools offer immense potential in automating certain processes, granting historians more time to focus on other aspects of their research. These digital tools introduce novel paradigms and innovative approaches to investigating the past. Therefore, engaging in methodological discussions among specialists and disseminating these techniques and procedures to students in History and Humanities is crucial.

When utilizing a database to extract and manage data, careful design, structure, and articulation of research objectives are paramount. Clearly stating the criteria for including information in a documentary set and the assumptions guiding the decision to incorporate or discard specific materials is essential. The results of a search provide isolated data in the form of numbers and percentages, requiring contextualisation to facilitate comprehension and analysis.

Whenever possible, projects should embrace collaboration and interdisciplinarity to address more complex and unforeseen research questions. Such an approach enhances the capacity of collected materials to shed light on previously unimagined inquiries.

Bibliography

Antón Pelayo, J. (2005). *La sociabilitat epistolar de la família Burguès (1799-1803)*. Girona: Quaderns del Cercle.

Antón Pelayo, J. (2013). *La correspondència epistolar de la família Burguès (1750-1850)*. Bellaterra: Universitat Autònoma de Barcelona.

Antón Pelayo, J. (2019a). *La comunicació epistolar de la família Burguès durant l'estada a Coromines (1727-1774)*. Bellaterra: Universitat Autònoma de Barcelona.

Antón Pelayo, J. (2019b). "La teoría de la carta familiar (siglos XVI-XIX)". *Revista de Historia Moderna. Anales de la Universidad de Alicante*, 37, 95-125. https://doi.org/10.14198/RHM2019.37.04

Antón Pelayo, J. (2023). "Ordenando cartas: EpiCAT, el portal para la gestión de cartas familiares en Cataluña". Acosta, F.; Duarte, A.; Lázaro, E.; Ramos Roví, M.J. (eds), *La historia habitada. Sujetos, procesos y retos de la Historia Contemporánea del siglo XXI. Actas del XV Congreso de la Asociación de Historia Contemporánea* (Córdoba, 9-11 septiembre 2021). Córdoba: Universidad de Córdoba, 1575-81.

Antón Pelayo, J. et al. (2023). "EpiCAT. Una plataforma para la gestión de cartas familiares". *Vínculos de Historia*, 12, 358-69. https://doi.org/10.18239/vdh_2023.12.19

Borkosky, M.M. (2002). "Epistolarios: la intimidad expuesta". *Cahiers du GRIAS*, 10, 27-45.

Chen, P.P. (1976). "The Entity-Relationship Model. Toward a Unified View of Data". *ACM Transaction on Database Systems*, 1(1), 9-36. https://doi.org/10.1145/320434.320440

Cuadrado, M. et al. (1994). "Thesaurus d'Història de Catalunya: creació i anàlisi d'un llenguatge documental aplicat a la historia". *Item*, 15, 134-59.

Fridlund, M. (2020). "Digital History 1.5: A Middle Way Between Normal and Paradigmatic Digital Historical Research". Fridlund, M.; Oiva, M.; Paju, P. (eds), *Digital*

Histories. Emergent Approaches Within the New Digital History. Helsinki: Helsinki University Press, 69-87.
https://doi.org/10.33134/HUP-5-4

Gil, T.L. (2021). *How to Make a Database in Historical Studies*. Cham: Springer.

Jiménez Sureda, M. (2020). *Amb el cor al paper. Història i teoria de les cartes d'amor*. Bellaterra: Universitat Autònoma de Barcelona.

Méndez Rodríguez, E.M. (2000). "Metadatos y tesauros: aplicación de XML/RDF a los sistemas de organización del conocimiento en intranets". *La gestión del conocimiento: retos y soluciones de los profesionales de la información*. Bilbao: Universidad del País Vasco, 211-20.

Milligan, I. (2022). *The Transformation of Historical Research in the Digital Age*. Cambridge: Cambridge University Press.
https://doi.org/10.1017/9781009026055

Mora Mellado, S. (2022). *Uso de bases de datos y herramientas de marcación y minería de texto para la preservación y estudio de documentos y hechos históricos* [MA thesis]. Barcelona: Universitat Autònoma de Barcelona.