The Joys and Sorrows of Data Normalization: Working with Sources from the Franciscan Custody of the Holy Land, XVII–XVIII Centuries

Antonio Iodice[1,*,†], Alessia De Benedictis[2,†]

[1]University of Roma Tre, Via Gabriello Chiabrera 199, 00145, Rome, Italy
[2]University of Roma Tre, Via Gabriello Chiabrera 199, 00145, Rome, Italy

Abstract

This paper presents the methodological workflow developed for HOLYLAB-DB, a digital database based on the account extracts sent by the commissariats of the Franciscan Custody of the Holy Land to Rome between 1654 and 1750. Building on previous work by Carnevali and Lastilla, the paper shifts the focus from the creation of the data-entry system implemented in a spreadsheet to the stages preceding and following it, examining the full pipeline from transcription to data cleaning, merging, and migration into MariaDB. Particular attention is devoted to the challenges posed by heterogeneous early modern accounting practices, the standardization of terminology, and the need to balance interoperability with the preservation of linguistic and contextual nuance. The paper also highlights the collaborative nature of the workflow, including the role of shared glossaries, controlled vocabularies, and weekly "Datathons." By reconstructing this workflow, the study offers a transparent model for the curation of complex historical datasets and for the creation of robust digital infrastructures for academic and non-academic users.

Keywords

Digital Humanities, Historical Databases, Data Cleaning, Early Modern Accounting, Franciscan Commissariats

1. Introduction

In 2024, Rebecca Carnevali and Lorenzo Lastilla published an article in *Umanistica Digitale* entitled *Technical Solutions for Data Systematization for Early Modern Accounting Sources. The HOLYLAB Database of Account Books of the Custody of the Holy Land* [1]. As members of the ERC project HolyLab, the authors described the processes of, and technical solutions adopted for, the conceptualization of their database (HOLYLAB-DB), arguing that this experience could serve as a "blueprint for approaching and systematizing historical data in early modern accounting sources in a vitally important flexible way" [1].

The use of spreadsheet-based systems for the digital representation of early modern accounting sources is not unprecedented. For instance, Carvalho has shown how eighteenth-century double-entry bookkeeping records can be modelled through relational spreadsheet databases conceived as an intermediate step towards SQL-based systems [2]. Historical databases have long been recognized as essential research tools in economic history and the humanities, requiring careful methodological choices in terms of data modelling, standardization, and long-term sustainability [3, 4]. They are increasingly cited as research outputs of large-scale projects such as ERC grants or, in the Italian context, funding schemes such as FIS and PRIN. At the same time, recent contributions in the digital humanities have drawn attention to the epistemological implications of data cleaning and standardization, questioning the assumption that these processes are purely technical or neutral [5, 6]. Within this growing literature, much less attention has been devoted explicitly to documenting the intermediate and often invisible

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.
†These authors contributed equally.
antonio.iodice@uniroma3.it (A. Iodice); ale.debenedictis@stud.uniroma3.it (A. De Benedictis)
0000-0002-5280-8664 (A. Iodice)

stages through which heterogeneous archival sources are transformed into interoperable data, particularly in terms of collaborative workflows, data cleaning practices, and quality control mechanisms. This is without taking into account further challenges concerning the long-term preservation of data and their interoperability across different digital infrastructures.

This paper positions itself at this intersection. Rather than proposing a new technological solution or data-entry interface, our approach resonates with recent calls in accounting history to foreground the concrete practices and workflows through which digital research is actually carried out, rather than focusing exclusively on tools or analytical outputs [7]. This paper offers a methodological and organizational case study of the end-to-end workflow adopted in the HOLYLAB-DB project, with particular attention to the stages preceding and following data entry. By making explicit the choices, constraints, and trade-offs involved in large-scale spreadsheet-based data curation, the paper aims to complement existing theoretical discussions with an empirically grounded account of how historical data are actually produced, standardized, and prepared for long-term reuse.

This paper takes up where the contribution of Carnevali and Lastilla ended. Their study focused on the conceptualization of the database, the building of the spreadsheets used for data entry and the critical analysis of the sources. We focus on the following stage of the creation of the HOLYLAB-DB, *i.e.*, the processes of data systematization, standardization, and merging. For the sake of clarity, we also discuss the stages that chronologically preceded the conceptualization phase outlined by Carnevali and Lastilla.

The choice of the term "stage" is not accidental. As HOLYLAB-DB is the product of a collaborative effort involving both the project's Principal Investigator, Felicita Tramontana, and several research fellows–some of whom joined the project after the initial phases and therefore required rapid training in data entry and data cleaning–the team soon adopted a clear and intuitive internal working structure through a well-defined sequence of stages, ranging from 00 to 4. For the sake of clarity, this same division is reproduced in the present paper.

Stage 00 refers to the creation of a shared glossary and other research aids essential to the workflow, which were regularly updated and discussed by the team. Stage 0 corresponds to the initial phase of transcribing the account books, carried out before the data-entry spreadsheet described by Carnevali and Lastilla was introduced. Stage 1 involves processing the data through the spreadsheets, while Stage 2 focuses on the data cleaning, through which every file produced in Stage 1 is reviewed. Stage 3 consists in the collection of the cleaned files, which are grouped and renamed according to the geographical origin of the information. Finally, Stage 4 marks the transition from the offline to the online model, that is, from Excel spreadsheets to the SQL language and the database developed through MariaDB. All files on which team members collaborated were uploaded to a shared folder on Google Drive and regularly updated. Two team members were in charge of editing most files and, in general, coordinating the group's efforts. Other project members could view the file and add comments that were later discussed during monthly team meetings. This protocol allowed to avoid issues related to the collaboration of different members on the same documents. The online publication of the database and its supplementary materials is scheduled in 2026. Some of the documents mentioned in the paper, such as the glossary, guidelines, and samples of controlled vocabulary lists and the journey section, can already be provided as supplementary material by writing to the project's PI.

This paper does not intend to dwell on the creation of the online database itself, but rather to address the challenges–as well as the opportunities for simplification in terms of data revision–offered by the serial processing of thousands of entries structured in a comparable form.

It is useful to briefly recall the nature of HOLYLAB-DB and the sources for which it was created, bearing in mind that reference should be made to the article by Carnevali and Lastilla for further details. The HolyLab project investigates the functioning of the Franciscan Custody of the Holy Land. The latter, headquartered in Jerusalem, has since the fourteenth century been entrusted by the papacy with the guardianship and maintenance of the Christian Holy Sites in Palestine. This responsibility required the gathering of alms across the Catholic world to sustain the friars and to finance the liturgical, charitable, and logistical needs of the Holy Places. To organize the collection and transfer of alms, starting in the sixteenth century the Custody established a network of branches known as commissariats. Each

Table 1
Total amount of transactions and years covered by the HOLYLAB-DB.

Commissariat	First Available Year	Last Available Year	N. of Years	N. of Transactions
Tuscany	1656	1750	91	5362
Genoa	1654	1735	57	1922
Rome	1655	1672	15	1090
Sicily	1657	1660	4	701
Naples	1651	1661	11	576
Turin	1717	1721	5	443
Milan	1651	1750	59	411
Holy Roman Empire	1654	1721	20	395
Malta	1655	1743	10	275
Ancona	1662	1664	3	21
Venice	1694	1723	9	20
Cyprus	1682	1683	2	4
Total	1651	1750	n/a	11220

commissariat was directed by a friar–the commissary–assisted by a lay accountant called apostolic syndic. Commissariats served as both an administrative and financial node within global religious and secular infrastructures. The syndics kept detailed account books recording incomes and expenditures, and the circulation of goods and people between America, Europe and the Levant. From 1654, after a specific decree by the Congregation of Propaganda Fide in Rome, commissariats were required to send to the Congregation extracts of their accounts on a regular basis for oversight and auditing purposes. These documents, preserved today in the Archivio Storico di Propaganda Fide, in the Vatican State, form a uniquely rich corpus for studying such an important early modern religious institution from a social and economic perspective.

The HOLYLAB-DB gathers the data extracted from the commissariats' accounting extracts dispatched to Rome between 1654 and 1750 [8]. As these are early modern historical sources, the data are uneven and not equally representative of all commissariats. Table 1 presents the number of transactions identified and recorded as of January 2026. The project is currently progressing towards finalizing the cleaning of additional account extracts from Tuscany, which have been added to those in the approximately 130 account extracts originating from various commissariats. Based on administrative correspondence and other archival documents, it is reasonable to assume that these records represent the entirety of the account extracts that were actually received and preserved in Rome during this period. Certain commissariats–such as those in Madrid, Paris, and, to some extent, Vienna and Naples–refused to comply with the directives issued from Rome and consequently either failed to send or ceased to send the requested documentation. Nevertheless, it should be noted that the database structure has been designed to accommodate future expansion: should new accounting records be discovered, further research initiatives could be developed to integrate this material, potentially through new funding opportunities, allowing the dataset to grow.

Extracts were drafted according to a simplified charge-and-discharge structure. This method of bookkeeping was usually used to categorize flows of money and/or other assets into ordered groups through three distinct stages. First, the agent (the apostolic syndic) was charged with what he had received or been paid; later on, he was discharged with what he had spent and paid. Finally, the difference, referred to as the balance, was computed. The balance denoted the value of what the agent owed his master (the commissariat) or vice versa [9]. We refer to a simplified structure given that the commissariats' accounting records have very few categories, being almost a simple list of revenues/charges and expenditures/discharges in chronological order.

2. Stage 00, Data Entry Supports

The initial stage was crucial in providing support to the researchers who, over time, participated in the data-entry process. Regular team meetings and discussions among group members allowed for the prompt revision and real-time updating of research support tools, in accordance with progress in data entry and with an evolving understanding–on a scientific level–of how the data were structured and how they changed over time.

The first file in Stage 00 consists of the instructions for the data-entry system implemented in the Excel spreadsheets. The instructions mirror the structure of the spreadsheets, providing a detailed, nine-page explanation of all forty-four fields, row by row and section by section. Over time, the instructions were expanded with examples and clarifications whenever questions arose. In particular, the procedure for recording the movements of goods and people in Section F of the template (Table 2) proved to be the most complex. This was especially the case for multiple transactions pertaining to a single journey, potentially involving different commissariats, where it was often challenging to identify the distinct steps of the journey relative to the commissariat where the transaction was registered. This ultimately led to the creation and upload of a sample file for consultation by the team.

The second file in the folder is made up of vocabularies related to the fields Transaction–Category (CTTC), Transaction–Object (CTTO), and Objects–Type of Object (DOT). As Carnevali and Lastilla also explained, such variables incorporate a "Data Validation" mechanism, which constrains the input entered by researchers to comply with certain conditions–specifically, belonging to a predefined set or vocabulary of options–and thereby helps to minimize typographical errors during data entry. As the data entry progressed, it became clear that capturing the variety of expressions and terminology used in the sources within a fixed range of options was challenging. This necessitated the regular updating and expansion of the vocabularies after discussion and approval during team meetings. For instance, in DOT, the option "paper" was added after it was observed that this type of commodity frequently appeared in records from certain commissariats, notably the Genoese one. It is always possible to verify the original terminology employed in the source, recorded in the Object Description field (DOD), as well as in the Transcription field (A), where the entire transaction is recorded. Feedback from work carried out in later stages, therefore, was integrated into Stage 00, leading to regular–collectively-approved–iterative changes.

The third file in this stage is the glossary. Even in this case, through regular team meetings and recourse to the literature, the best words or expressions were chosen as standard. It proved challenging, for instance, to distinguish the different ecclesiastical and courtesy titles employed by the sources, such as friar, father, layman, companion, general commissary, deputy commissary, and so on. Whenever in doubt, the original Italian terminology was preserved. Such glossaries will be published together with the database.

3. Stage 0, Sources' Transcription

Given the restrictions in place at the Historical Archive of Propaganda Fide–where, for instance, photographing documents is not permitted, nor is there an available Wi-Fi network–this stage proved essential in preparing for the first phase of data entry. Excel files were created corresponding to the various archival units in which the commissariats' account extracts are preserved, each containing the transcriptions of every transaction recorded.

The data were organized by location, with separate columns for dates, transaction amounts, and transcriptions. During this stage, preliminary attempts at categorizing charges and discharges were made, which would later serve as the basis for the data-entry spreadsheet described by Carnevali and Lastilla.

Table 2
Core structure of the data-entry template for a single entry/transaction in the Excel spreadsheet.

ID	Variable Code	Variable	ID	Variable Code	Variable
HL1	A	**Transcription**	HL1	DOQ	**Objects - Quantity**
HL1	BRRFS	**Record - Repository, fondo, and series**	HL1	DOU	**Objects - Unit**
HL1	BRUD	**Record - Unit and document**	HL1	DOT	**Objects - Type of Object**
HL1	BRD	**Record - Date - dd/mm/yyyy**	HL1	DOD	**Objects - Object Description**
HL1	BHA	**Holdings - Authority**	HL1	E	**People Circulation Data**
HL1	BHC	**Holdings - City**	HL1	EQ	**Quarantine - If**
HL1	BED	**Entry Date - dd/mm/yyyy**	HL1	EQD	**Quarantine - Duration**
HL1	CTTIE	**Transaction - Income/Expenditure**	HL1	EQW	**Quarantine - Where**
HL1	CTTC	**Transaction - Category**	HL1	FP	**Place A**
HL1	CTTO	**Transaction - Object**	HL1	FS	**Stage A**
HL1	CRVQ	**Recorded Value - Quantity**	HL1	FSP	**Specification of Place A**
HL1	CRVC	**Recorded Value - Currency**	HL1	FOD	**Object Description at Place A**
HL1	COC	**Original Currency**	HL1	FNTP	**Name of Travellers at Place A**
HL1	CTS	**Transaction - State**	HL1	FQTP	**Quantity of Travellers at Place A**
HL1	CTC	**Transaction - Commissariat**	HL1	FL	**Leg a**
HL1	CAGP	**Alms - Gathering Period**	HL1	FMT	**Means of Transport a**
HL1	CTEP	**Transaction - Expenditure Period**	HL1	FTT	**Type of Transport a**
HL1	CTHW	**Through Whom**	HL1	FSDL	**Starting Date for Leg a - dd/mm/yyyy**
HL1	CTW	**To whom**	HL1	FEDL	**End Date for Leg a - dd/mm/yyyy**
HL1	CFW	**For Whom**	HL1	FLQ	**Legs - Quantity**
HL1	CBW	**By Whom**	HL1	FLP	**Legs - Paid**
HL1	D	**Object Circulation Data**	HL1	NOTES	**Notes**

4. Stage 1, Data Entry

Stage 1 contains the Excel files compiled according to the official template, made of 44 variables nested in 5 layers (plus Notes): Archival Record (light grey), Financial Transaction (green), Object Circulation (blue), People Circulation (red), and Journey (yellow) (Table 2). The vertical structure allows for the addition of further fields, numbered progressively (*e.g.*, Object 2, Stage B, Leg 2), in order to record the presence of multiple objects, places, legs, and so forth.

Such detailed spreadsheet-based data-entry system allowed to break information into consistent, unique, and self-contained variables, in a process referred to as data atomization [1]. Each file corresponds to an extract from the account books of a specific commissariat, usually covering a two-year period. This level of granularity, however, comes at a significant cost in terms of time. At the outset of the project, we only had rough estimates of the exact number of documents to be processed. Based on calculations carried out after the completion of the first data-entry spreadsheets, a trained team member required, on average, approximately nine minutes to input a single transaction. This figure varies depending on the complexity of the entry and the experience of the data-entry operator. As a concrete example, the seventeenth-century spreadsheets for the Commissariat of Genoa–typically comprising around sixty entries per year–required approximately nine hours of work per year. Five team members were primarily dedicated to this task, over roughly one year and a half. Given the substantial volume of documentation, and in order to avoid duplication or the accidental omission of materials, an essential operation was the maintenance of a shared .docx file to keep track of both the work completed and the work assigned. Such task was entrusted to two team members in charge of coordinating the work on the database. In this file–listing all Excel documents–each extract was assigned to a specific member of the project, with a clear deadline to complete the assignment. Upon completion, each team member sent the spreadsheet to the DB supervisors, who marked the file with the team member's initials and the exact number of transactions contained in the completed Excel spreadsheet. The file name was then highlighted in yellow when forwarded to the next stage for data cleaning.

As noted above, most transactions follow a straightforward chronological order. Entries in the

syndics' records generally disregarded any connections between transactions: their priority was merely to demonstrate that the final balance of the extract was positive. One example concerns the manufacture of two baldachins for Jerusalem, one red and one white, produced in Genoa between 1685 and 1687. Information on the costs associated with these valuable liturgical objects was dispersed across hundreds of entries over a three-year period. Within the project, however, it was deemed essential–wherever possible–to establish links between supposedly-related expenses in order to maximize the usefulness of the data for future users. These links were recorded in the Notes field by specifying the univocal ID of each connected entry. When converting the data in SQL, such IDs were recognized and such connections were further checked.

Additional support in the standardization process, and in resolving uncertainties arising from the wide variety of terminology and information found in the sources, came from the regular organization of data-entry "Datathons". Project members met weekly in dedicated virtual sessions during which each participant worked for three to five hours on their assigned file, with the opportunity to consult colleagues in real time to address any question.

5. Stage 2, Data Cleaning

Following data entry, Stage 2 focused on cleaning the recorded data. While the practice of data cleaning has sometimes been met with suspicion in the humanities–where standardization is occasionally seen as risking the loss of complexity or historically meaningful variation [5, 6]–our approach aligns with recent calls for greater methodological transparency in data curation practices, particularly in projects dealing with heterogeneous historical sources. Minor transcription discrepancies can distort results in large datasets, and unchecked spelling or terminology variation can obstruct even basic quantitative operations. The task in Stage 2 was to find an appropriate balance between standardization and preservation of linguistic nuances. Standardization was necessary for interoperability and successful database searchability. Yet we also aimed to retain interpretive linguistic details paramount in historical research, particularly in fields such as DOD, FOD, and A. Far from suppressing diversity, this phase ensured that such heterogeneity remained legible and analytically usable within a digital environment.

A guiding principle throughout this workflow was the traceability of every intervention. To avoid any conflict caused by multiple members working on the same file concurrently, the team adopted a protocol whereby spreadsheets could be modified by only one member at a time. Each modification introduced was visually highlighted using a color code assigned to each reviewer (*e.g.*, yellow, green, orange, or blue), enabling immediate cross-verification and accountability.

The data cleaning operated at two complementary levels. The first involved what we could call a "rule-based" validation through built-in drop-down menus and formula constraints, which ensured that values conformed to the predefined closed vocabularies developed in Stage 00. The second level was fully manual, requiring a project member to review the Excel spreadsheets line by line. The decision to assign the preliminary cleaning work to a single reviewer for each commissariat was taken to ensure uniformity across the files containing data obtained from the same location. Based on the complexity of the transactions in each source, a second data-cleaning phase could be carried out by another reviewer, who would also verify the work of the first reviewer. Indeed, one challenge intrinsic to collaborative data entry in Stage 1 is that multiple researchers working independently may interpret ambiguous source material differently, leading to inconsistent categorization or transcription choices across files. By having a single person systematically review all entries for a given commissariat, it was possible to identify patterns of divergence. The reviewer could then consult with the two DB supervisors to establish a consistent approach, which was subsequently applied across the entire commissariat and, where appropriate, added to the glossary or guidelines to aid future data entry and cleaning. Thus, the iterative process of individual verification and cooperative standardization ensured better consistency within individual files and across the entire database.

One example of what had to be verified was the logical coherence between fields, for example, in group D, specifically between D (*i.e.*, Does the transaction concern an object?) and DOT (Type of

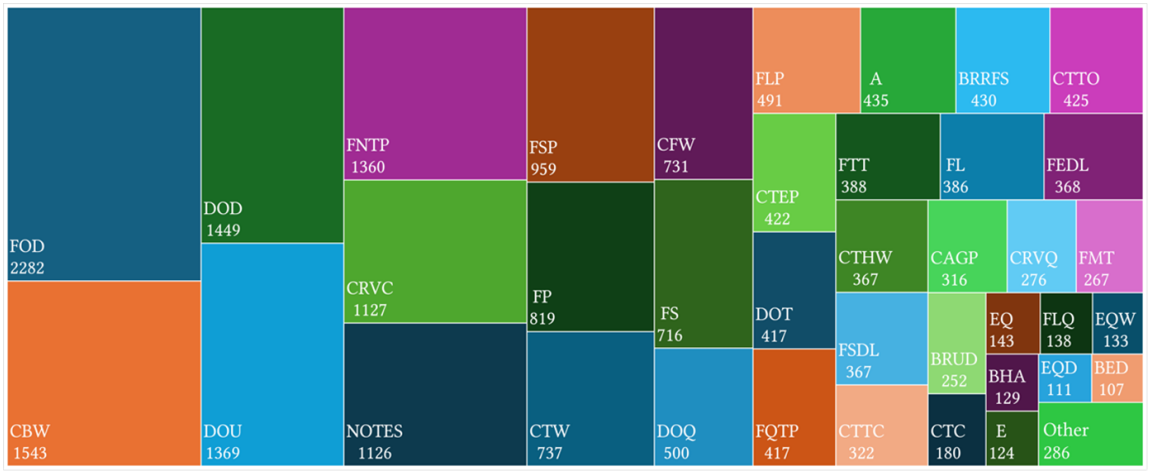

Figure 1: Hierarchical structure of revisions during Stage 2 data cleaning.

Object): whenever the value in D was "no," the corresponding DOT field had to be marked as n/a. When information was available, the DOT had to be coherent with DOD (Objects - Object Description) and DOU (Objects - Unit). Such checks often revealed deeper interpretive questions concerning the lexicon. A further level of cleaning focused on linguistic and terminological standardization, carried out in accordance with the shared glossary discussed in Stage 00. For instance, the reviewer would verify whether individual names or ecclesiastical and courtesy titles were transcribed in accordance with the pre-established template. Toponyms and currencies, instead, had to be translated into contemporary English for clarity and interoperability. When a place name corresponded to multiple possible locations, explicit disambiguation was introduced (*e.g.*, Tripoli [Libya] vs. Tripoli [Lebanon]). The original terminology is always shown in the Transcription field. Whenever uncertainty persisted, terms were recorded in a comparative table, where the original expression was paired with one or more proposed standardized forms or translations. Each proposal was then reviewed and approved by the DB supervisors, who provided guidance on interpretive questions, ensuring that collaborative decisions were discussed and approved before implementation.

A quantitative estimate of the time required for data cleaning alone proved difficult to isolate with precision. Cleaning efforts were influenced in part by the number of entries and by the density of possible ambiguities. Additionally, active revision of the single spreadsheet frequently intertwined with discussions among different team members to reach the best solutions. However, a quantitative assessment of the cleaning workflow provides a measure of its extent within the HOLYLAB-DB. The median proportion of modified cells from a representative sample of 34 files was approximately 16 per cent. However, the rate of intervention varied substantially–from about 1 to over 30 per cent–depending on the richness and heterogeneity of the data. More concise records, such as those from the Commissariat of Piedmont (Turin), required fewer revisions, whereas the more complex datasets from Genoa and Tuscany underwent more extensive corrections. Most modifications concerned descriptive variables (Figure 1), with the greatest concentration of changes occurring in the sections dedicated to Journey Data (F: 39.96 per cent), Money Circulation Data (C: 29.24 per cent), and Object Circulation Data (D: 17.02 per cent).

More broadly, the errors identified during Stage 2 fell into four principal categories: (1) terminological inconsistencies, such as synonymic variation or deviation from the standard lexicon; (2) formatting errors, particularly in numerical or monetary values; (3) minor graphical irregularities, including spacing, capitalization, or punctuation differences; and (4) occasional shifts of content between contiguous fields, particularly in those related to object and journey descriptions (*i.e.*, between DOD–DOT and FS–FSP). The resolution of such inconsistencies was essential for ensuring internal coherence and for enabling subsequent quantitative aggregation and automated querying within MariaDB.

Once these operations were completed, each file had reached a level of consistency and standardization

adequate for subsequent processing. The dataset was thus ready for the next phase of the workflow, Stage 3, in which cleaned files would be assembled and prepared for integration.

6. Stage 3, Data Merging

Stage 3 focused on the merging of the individual cleaned spreadsheets into a single Masterfile for each commissariat, the creation of which was entrusted to a single team member per commissariat. Rather than a preference for centralized work, assigning one operator per commissariat reflected the logic of the procedures employed, as these operations required to be performed on an entire commissariat at a time.

The first task of this stage involved copying the cleaned sheets into a new file, making sure to label them according to their temporal coverage. This would retain the logical segmentation but prepare the ground for their integration.

Secondly, the numerical progression of the entry identifiers was reconstructed. As each worksheet carried its own numerical order, identical identifiers recurred across files. To avoid conflicts and confusion once merged, Stage 3 involved decomposing the existing identifiers into their alphabetical prefix (HL) and numerical component, recalculating the latter according to the worksheet's position in the overall sequence, and generating a single uninterrupted progression for the entire commissariat. Subsequently, commissariat-specific codes of two or three letters were applied to the full set of identifiers (*e.g.*, from HL to HLGE for Genoa), ensuring that every record retained an immediate and unambiguous indication of its institutional origin. Notably, when the documentation of a commissariat extended across two centuries, the data were intentionally split into two separate files, one per century, to ensure their manageability. That said, to preserve chronological clarity, entry identifiers were treated as part of a single continuous sequence and the numerical progression established in the earlier file was carried forward into the subsequent one (*e.g.*, if the seventeenth-century Genoese corpus concluded with identifier HLGE1565, the eighteenth-century corpus began with HLGE1566).

After this step, a copy of the worksheet containing the shared validation parameters employed across the project (Rules) was added to ensure there would be no system or methodological errors.

Perhaps proving the common assumption that 80 percent of data analysis is devoted to cleaning and preparing the data, this commissariat-wide view also allowed us to resolve issues that had remained invisible at the scale of the individual file and to further uniform spellings, names, and internal references across entries [10, 11]. The Notes section was particularly affected by this process. The macro-viewpoint allowed for the detection of more links between journeys and transactions. Additionally, the cross-references that had already been signaled with the entry identifiers used in previous stages (*e.g.*, "Entry linked to HL34") now needed to be replaced with their commissariat-specific equivalents (*e.g.*, "Entry linked to HLGE245").

From a workload perspective, the time required for Stage 3–including the additional reviewing–averaged approximately seven to eight hours per merged spreadsheet, with overall working time varying according to the number of source files involved. This estimate is based on a sample of ten files, created by merging a total of thirty-two spreadsheets, and it refers to standard activities, excluding a limited number of outlier cases. By adding up the efforts of Stage 2 and Stage 3, it is possible to estimate that the cleaning and merging activities together accounted for a workload of roughly between one hundred forty and one hundred fifty hours per month for a year, underscoring the labor-intensive and iterative nature of this process.

By the conclusion of Stage 3, each commissariat's data had been processed into a unified and standardized corpus, ready for migration to SQL.

7. Stage 4, Data Migration

Once the Excel files are consolidated for each commissariat, they are ready to be transferred into SQL and MariaDB. The project's IT specialist, Inês de Sá, designed a relational database consisting of

twenty-seven tables, linked through various joins, with the transaction ID serving as both primary and foreign key. The choice of MariaDB was primarily driven by infrastructural considerations, as it is natively supported by the WordPress and Plesk environment provided by the University of Roma Tre. At the same time, MariaDB offers a significant advantage in terms of sustainability and openness, as it is fully open-source, whereas other relational database management systems, such as MySQL, combine GPL-licensed open-source versions with proprietary, paid functionalities.

Excel VBAs were used to convert the files produced by the team into CSVs, each corresponding to one of the database sections. While the Excel-based workflow adopted in HOLYLAB-DB proved effective for collaborative data entry and iterative revision, it also entails a number of limitations when transitioning to a relational database environment. Since Excel recognizes only perfectly identical values, one recurring issue concerns the management of variant spellings, particularly for personal names and place names that required a certain amount of manual intervention to ensure further standardization of the data. A further issue lies in the lack of synchronization between the spreadsheet files and the SQL database after data migration. Once the data have been converted and imported, any subsequent correction of minor errors in Excel does not automatically propagate to the database, requiring additional checks or manual interventions. This makes late-stage corrections more costly than they would be in a system based on direct data entry into a relational database. At the same time, the migration process itself functions as an additional layer of quality control. The additional review in Stage 4 was conducted by a revisor who prepared the files for migration, with the two DB supervisors intervening to address questions and resolve doubts as they arose. Once the final checks relating to individual names and other shared variables are completed for each merged file, the data will be imported into the online database hosted on Plesk. Users will then be able to explore the material through search queries applicable to any of the fields as defined in the original spreadsheets.

8. Conclusion

The experience gained through the development of HOLYLAB-DB highlights the methodological value of treating data entry, data cleaning, and data migration as tightly interconnected stages. Far from being a purely technical exercise, the workflow required continuous scholarly judgement, careful documentation, and collaborative problem-solving.

The challenges posed by heterogeneous early modern accounting sources–ranging from lexical variation to fragmented narrative and accounting structures–demonstrated the need for a workflow capable of balancing standardization with the preservation of historically meaningful detail. By establishing a transparent chain of operations, and foregrounding the organizational, temporal, and interpretative dimensions of data curation, we intended our paper to be useful not only for historians and digital humanists, but also for project designers and principal investigators involved in large-scale historical data infrastructures.

HOLYLAB-DB thus contributes to current discussions in the digital humanities on data curation, interoperability, and the long-term sustainability of historical datasets.

Acknowledgments

This research and work behind this paper have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme as part of the ERC COG-2020 project "HOLYLAB. A global economic organization in the early modern period: The Custody of the Holy Land through its account books (1600–1800)" (Grant Agreement 101001857). However, this text exclusively reflects the authors' views. The ERC is neither responsible for the content nor for its use.

The authors would like to thank Felicita Tramontana for her support from the early draft to the final version of the manuscript, as well as to the entire HolyLab team for their input and collaboration.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Online Resources

The HolyLab project can be consulted at https://holylab-erc.uniroma3.it/. The database will soon be available from the same webpage.

References

[1] R. Carnevali, L. Lastilla, Technical solutions for data systematization for early modern accounting sources. the HOLYLAB database of account books of the custody of the holy land, Umanistica Digitale 17 (2024) 25–45.

[2] J. M. de Matos Carvalho, From traditional accounting history to digital accounting history: An eighteenth century double-entry bookkeeping system represented in spreadsheet databases, in: XVI International Congress of Accounting and Auditing, 2017. URL: https://www.academia.edu/41125496/From_Traditional_Accounting_History_to_Digital_Accounting_History_An_Eighteenth_Century_Double_Entry_Bookkeeping_System_represented_in_Spreadsheet_Databases.

[3] M. Casson, N. Hashimzade, Introduction. research methods for large databases, in: M. Casson, N. Hashimzade (Eds.), Large Databases in Economic History. Research methods and case studies, Routledge, London, 2013, pp. 1–22.

[4] C. Harvey, J. Press, Databases in Historical Research. Theory, Methods and Applications, Palgrave Macmillan, Basingstoke, 1996.

[5] K. Rawson, T. Muñoz, Against cleaning, in: M. K. Gold, L. F. Klein (Eds.), Debates in the Digital Humanities 2019, 2019, pp. 279–292. doi:10.5749/j.ctvg251hk.26.

[6] A. Iodice, (poor) accounting for god: the tracking and monitoring of cash flows in the custody of the holy land's network (1650s–1680s), Accounting History Review 35 (2025) 117–147.

[7] M. Quinn, B. Murphy, Accounting history in a digital age: Empirical experiences, improving methods and unlocking new research avenues, Accounting History 30 (2025) 23–43.

[8] I. Llibrer Escrig, S. Villaluenga de Gracia, Learning from history. deconstructing the charge-and-discharge system within an accountability context, Accounting History Review 33 (2023) 1–27.

[9] L. Pandolfo, L. Pulina, Unlocking historical insights: Developing a dataset from historical archives, in: A. Dover, A. Formisano (Eds.), Proceedings of the 38th Italian Conference on Computational Logic, Udine, Italy, 2023. URL: https://ceur-ws.org/Vol-3428/paper17.pdf, june 21–23, 2023.

[10] H. Wickham, Tidy data, Journal of Statistical Software 59 (2014) 1–23. doi:10.18637/jss.v059.i10.

[11] T. Dasu, T. Johnson, Exploratory Data Mining and Data Cleaning, John Wiley & Sons, Hoboken, NJ, 2003.

[12] S. Marzagalli, W. Scheltjens, J. Land (Eds.), L'Histoire Maritime à l'épreuve des Humanités Numériques, volume XXXVIII of *Histoire et Mesure*, 2023.