An extensible model for historical financial data with an application to German company and stock market data

Dennis Gram¹, Pantelis Karapanagiotis², Jan Krzyzanowski², Marius Liebald², and Uwe Walz^{2,*}

¹Data Centre, Leibniz Institute LIF-SAFE, Frankfurt, Germany

²Management and Microeconomics, Goethe University, Frankfurt, Germany

Abstract. Broad, long-term financial, and economic datasets are a scarce resource, particularly in the European context. In this paper, we present an approach for an extensible data model that is adaptable to future changes in technologies and sources. This model may constitute a basis for digitised and structured long-term historical datasets. The data model covers the specific peculiarities of historical financial and economic data and is flexible enough to reach out for data of different types (quantitative as well as qualitative) from different historical sources, hence, achieving extensibility. Furthermore, based on historical German firm and stock market data, we discuss a relational implementation of this approach.

1 Introduction

High-quality data are one of the most important inputs in the empirical research in finance and economics. Over the last few decades, we have seen tremendous growth in structured data on firms, households, and markets at the micro- as well as the macro-level. While data covering the US is easily accessible, long-run structured and consistent databases for Europe are scarce¹. This lack has two immediate consequences. First, most empirical studies use US data. Transferring the conclusions of these studies to other legal systems or structures, such as those of continental Europe, often leads to potentially serious misconceptions because of fundamentally different institutions and political systems (cf., e.g., [1]). Second, not least in the aftermath of the 2008 financial crisis, the research has recognised that data with a short horizon may focus on quite specific macroeconomic settings and might overlook long-term relations. For example, only investigating the data on the “Great Moderation” period overlooks structural changes that have occurred before and after that period.

* Corresponding author: uwalz@econ.uni-frankfurt.de

¹ For example, in the US there is the Centre for Research and Security Prices (CRSP) and Compustat Databases starting 1925 that cover accounting as well as security prices. For Europe, the exception is the EUROFIDAI database that covers broader firm-level level data from 1980 to today.

Hence, there is a need to build databases with financial data for Europe that span decades and centuries. Analysing long-term historical data provides not only a better understanding of the underlying mechanisms of European societies but is also crucial for the contemporary development of economic policies, which is important for future economic progress (cf., e.g., [2]). Therefore, research needs to access datasets of this kind digitally.

However, the construction of long-term structured databases that comprise historical datasets (in particular pre-WWII datasets) is a tremendous challenge. This paper highlights problematic aspects of such datasets' constructions and discusses potential solutions. First, historical data from various sources with overlapping entities in a given time period can be ambiguous in the sense that either the information can be incomplete in some sources, or sources can be conflicting, which requires a decision on which source is more adequate. Second, linking and merging different data sources without the existence of common identifiers is difficult. Thus, proper metadata standards have to be set up and agreed on that makes overcoming this problem possible. The linking of different data sources offers huge potential for economic research. For instance, linking company information with stock market data enables sophisticated firm-level analyses (e.g., the Compustat/CRSP merged database in the US). Moreover, linking data across jurisdictions is a necessary condition to study economic phenomena while accounting for differences in institutional set-ups (which is necessary for Europe). Third, studies in the fields of financial economics often rely on a mix of quantitative and qualitative data (see e.g., [3] for this distinction). For instance, company reports, which are frequently used in economic research, entail numerical data (e.g., balance sheet, profit and loss statements) as well as textual data (e.g., executive boards, supervisory boards). While accessing combined datasets is essential for research, merging them and, hence, analysing them jointly is complicated.

Against this background, we outline a base for an extensible data model and a process to digitise and structure large historical datasets. We take into account that data collection often does not use pre-defined standards and that the ex-post harmonisation, standardisation, and verification of data cannot typically take place without having access to the original sources. This is not only true for national-level data but is even more apparent for data from different countries. The proposed implementation aims to cope with the particular features of historical data that allows researchers to check the original data source. Building on this idea, we further elaborate on our project to digitise and structure historical company data from Germany and merge them with stock market data on the basis of our proposed extensible data model.

The paper is organised as follows: In the next section, we review the literature on the creation and build up of historical financial datasets. We discuss research that is based on larger historical datasets. In the third section, we discuss the specific role of the original information that leads to digitised historical databases, and then we propose the principle of the preservation of the original historical source. The fourth section has a discussion on the idea of a relational implementation of this principle, while section five focuses on potential alternatives. In section six, we delineate the population of the input layer with data obtained from historical sources. The seventh section focuses on the construction of panel datasets that are built on top of the historical sources layer. In the eighth section, we provide a short discussion of the resulting stock market data and the development of its digital structure as well as the potential merge with the company dataset. The last section concludes.

2 Literature review

In order to provide insights into existing models for extensible databases as well as data projects aiming at generating and employing long-term data, we provide an overview on four

related fields: i) studies on extensible data models and ii) studies that focus on the collection process of historical datasets. This overview sets the stage for our own study.

[4] point out that the combination of data from multiple sources aims at helping to unlock the potential value of the underlying data. However, according to the authors, challenges occurred in the context of the continuous integration of data from numerous sources. As [5] point out that these challenges include the necessity to deal with the multitude of options that are based on the underlying architectures of the systems. With their prototype, these authors provide solutions that are able to interchange between a key-value store and a column-store architecture that permits adaptation to changing workloads.

[6] provide a comprehensive discussion of the features of various processes, formats, and systems as well as their contribution to extensibility. By contrast, our paper narrows the discussion of extensibility in the context of system design to provide a concise exemplifying implementation that uses the collected historical data from German sources.

Further related literature points to other solutions to overcome the problem of static and data-independent decisions. From this background, [7] provide a modern approach to processing a data stream which in its core aims at computing multiple routes of data queries that are individually designed for particular subsets of the data with distinct statistical properties. In another paper, [8] follow the approach to adaptively build auxiliary data structures that are required as large numbers of continuously produced data series need to be available for queries as soon as possible. Further, another strand of the literature argues that data should be kept in a flexible format structure that is embedded in a single system that provides multiple views of the data. In this context, [9] describe their concept of storage views that constitutes secondary, alternative physical data representations covering all or subsets of the primary log by utilizing of different storage technologies for different subsets.

Abstraction and genericity are attributes that the computer science literature commonly identifies as being central ingredients in the extensibility of systems. Using an experimental design, [10] provides statistical evidence of this industry wisdom. The author shows that there are two cases, which we find to be applicable in the context of historical data in which the addition of abstractions contributes to extensibility. First, abstractions reduce the time needed to implement modifications of a conceptual model in cases of adding complicated changes. Second, the author also presents evidence that abstractions can be beneficial to the correctness of adding changes to the conceptual model.

Generating historical datasets is a challenge that may be associated with severe potential deficiencies. [11] discuss some potential general flaws in historical financial datasets. They argue that the weak empirical foundations of economic and financial analytical models are in parts attributable to the scarce availability of long-run financial micro data that particularly exists for databases that contain financial-instrument-specific information over broader periods. Country indices that reflect the performance of bond and equity markets are on the other hand more easily available, which is particularly due to the contributions of [12], [13], [14] and [15]. The scarcity of long-run financial micro data is particularly prevalent at the European level. Apart from exceptions such as the Studiecentrum voor Onderneming en Beurs (SCOB) database of Antwerp University, most of the research is based on American financial micro-databases of which CRSP, managed by the University of Chicago, is the most widely used. Recently, pursuant to [11], more research projects have aimed at providing more historical data in a European context; for instance, the project at the Paris School of Economics on “Data for Financial History” (DFIH), the collection of UK market data at the Centre for Economic History of Queen’s University Management School Belfast as well as the initiative on the Helsinki, Lisbon and Stockholm Stock Exchanges (see e.g., [16] and [17]). The DFIH initiative has developed a comprehensive long-run stock exchange database on the French markets from 1796 to 1976 (cf., [18]). Two technologies to capture data were set up that both required the scanning of the printed sources. The first was characterised by

manual data entry and the second by semi-automatic processing of the stock exchange yearbooks that utilised artificial intelligence methods.

[19] points out that implementing advanced databases such as the above described SCOB and DFIH have similar characteristics. He discusses whether the EURHISFIRM initiative, which aims to cover data from numerous European countries, requires an overarching identification system for European, historical firm-level data. For this purpose, he examines functional requirements that relate to a proper identifier design and adequate documentation as well as quality assurance and label validation. Furthermore, he also discusses informational requirements in order to identify different classes of economic entities.

Finally, [20] utilise a newly constructed dataset of daily transaction prices and volume data from the Stockholm Stock Exchange for the period from 1912 to 1978. In this context, the authors describe general methodological issues of missing values in historical portions of data (see also [17]). They point out that many papers on historical stock markets did not correct stock prices for effects of capital operations as this information was not available. Furthermore, the authors state that their historical data contained complete information on capital operations from 1912 onwards due to the requirements of the 1910 Swedish Company law that required that every capital operation must be filed with a government agency.

3 The role of original information in historical databases

One central issue of setting up historical databases is that of deduplication. The problematic of deduplication is neither new nor specific to databases with historical information. Deduplication commonly refers to processes that establish whether two or more records in a collection of data represent the same object in the context of interest. To have a precise formulation of the concept for the cases that we are considering in this study, suppose that we are interested in designing a data model for a set of ideal objects denoted by $\mathcal{O}$. For instance, let this set contain all the companies that operated in Europe in the last three centuries. Albeit convenient when designing, the ideal set $\mathcal{O}$ contains elements that are typically not perfectly identifiable since these elements, being historical, might not exist anymore and records of their existence might be unavailable or erroneous.

Instead of having direct access to the objects of $\mathcal{O}$, only historical archives that describe them are available in most cases. In contrast with contemporary best practices, historical sources are neither written with standards in mind nor can be validated by examining the original object. As a result, the researchers who deal with them often encounter situations in which the descriptions in different sources, due to the lack of standards, use different semantics and formats or, even worse, their descriptions of objects of $\mathcal{O}$ are conflicting.

Fig. 1 gives an example of missing information by presenting two snippets of different historical printed sources that both report dividends paid by “Société générale de crédit industriel et commercial”. The left snippet is taken from the 1880’s yearbook published by the governing body of the exchange, while the right one comes from the “Courtois” yearbook from 1874. The records overlap for the years 1859 to 1872; however, they do not contain the same information (see solid line rectangles). In particular, dividend payments are not reported for the years 1859 to 1863 in the left snippet, while records exist in the right snippet.

Fig. 1. Examples of missing (solid line) and conflicting (dashed line) information.

From the perspective of a data-model design, the example of missing information in Fig. 1 is relatively innocuous, because both sources dictate the inclusion of a dividend concept in the data model. If any missing information is located in an alternative source in the future, there is no need to update the data model, but instead only to add the new data to the implementation. The situation becomes more complicated if there is conflicting information.

Fig. 1 also highlights a case of conflicting information (see dashed line rectangles). In this case, the left snippet of Fig. 1 shows that the dividend paid in 1866 was 12.50 francs, while the right snippet shows that the dividend was 23 francs. The scale of paid dividends in the surrounding years suggests that the “Courtois” yearbook records the correct dividend. However, a more thorough examination indicates that both sources agreed that the paid dividend was 12.50 francs in 1870, suggesting that a dividend of the same level plausibly represents the real dividend value of 1866. Therefore, it is impossible to establish the actual value of the dividend in 1866 without having any additional information for that year.

From a data-model perspective, this ambiguity poses a serious challenge to the typical modelling approach that standardises the accepted values of data fields such as dividends. There is the possibility to increase the cardinality of the number of records that the dividend field accepts; however, this approach in some ways undermines the purpose of standardisation. From an end user’s perspective, a query that returns multiple values can be potentially confusing and unexpected, although it can be well-defined in terms of a standard that allows more than one dividend value at a given time-point.

Another possible way to approach this problem from a standardisation perspective is to assign weights to each dividend value. Besides increasing the cardinality to permit more than one record, the recorded entries can be pairs of monetary values coupled with probabilities that signify their potential correctness. This approach, in the sense of compatibility with the standard, can deliver expected result sets that can also be reasonably comprehended by end users. The difficulty of this solution lies with determining the probabilities that accompany the dividend values, as this determination requires the innervation of experts that assign probabilities to each case. A uniform distribution of weights in all cases does not add any value since it essentially corresponds to simply increasing the cardinality of the field.

This discussion should have convinced the reader that the actual historical information and the historical archives are entangled in a way that attempting to separately model the information space while disregarding the archive space becomes impossible if one wants to provide users with accurate information. Even in the case of assigning probabilities to values of various records, there is a strong possibility that some end users would like to deviate and use weights that are based on their expertise. Therefore, any data model extensibility solution should take this constraint into consideration. This consideration is the starting point of the principle of preserving the historical sources.

4 A relational implementation of the preservation principle

We refer to the layer of the system that we are focusing on as the *input layer*, because the historical sources constitute input data from the overarching research infrastructure perspective and because this layer is responsible for storing and associating the system's input data. The *input layer* does not handle concepts such as companies and financial instruments; those are the responsibility of other system layers that are built on top of the *input layer*. Instead, the *input layer* represents a low-level abstraction that handles the sources and isolates them from the system's higher layers that are susceptible to change through time as new technologies enable new representations or new sources emerge.

In terms of our concrete working example, the *input layer* is used to store information directly from the output of the OCR (see Section 6). The implementation of the input layer begins by organizing information from the publisher. Each published title is linked to its corresponding scanned pages. In turn scanned pages are linked to extracted OCR output. These parts of the model are omitted from the presentation since they are not required for describing the preservation principle.

The part of the *input layer* that captures the essence of the principle of preserving the historical sources associates concepts of interest in the data model with their origin, which alleviates the ambiguity that characterises the standardisation of potentially conflicting archives. This is also the point at which the implementation of our example departs from existing implementations that, as discussed in Section 5, do not fully capture the nature of the association between sources and concepts in their data models. Besides the exact association of sources with concepts, the implementation that we propose here acts as an abstraction layer that enhances the extensibility capacity of the system.

As illustrated by **Fig. 2**, the design introduces the concept of information items. An information item is an abstraction that comes in-between the sources and the data items that are conceptualised in the higher layers of the data model. Every line is associated with one or more information items. In turn, these information items are linked to one or more concepts. The information items may concern the same concept of the data model; for example, multiple board member names are located in a single line, or different concepts, for example, the address of the company and the managing director, are found in the same line. Conversely, the same concept, for example, the name of a company, may be located in multiple lines that originate either from a single or multiple sources.

As an example, in **Fig. 2** the information items are related with various non-exhaustive concepts of the *original layer*; that is, a separate schema from that of the *input layer* that contains fields to describe concepts such as the various roles of physical entities, names and locations of economic entities, events related to economic entities, and balance sheets of legal entities. In essence, this approach fully captures the provenance of the data items that are found in system layers that are built on the *input layer*, and also offers the ability to develop verification and validation processes within the system. Technically, the abstraction uses a many-to-many relationship to describe the association between sources and higher model concepts. The abstraction also creates a modelling buffer between these higher concepts and the concepts relating to the sources. The latter are invariant, relatively simple to describe, and their description is non-conflicting; attributes that suggest with high confidence that their thorough standardisation is plausible. The former, however, are not completely explored, have complicated interconnections, and can have fuzzy and conflicting content; attributes which suggest that data models that describe these higher concepts have to be frequently adjusted and updated. The benefit of the implementation of this study is that any updates of the layers that contain higher model concepts can be performed independently from the *input layer*. Moreover, the possibility exists to connect multiple higher-layer data models with substantially different characteristics to a single *input layer*.

Fig. 2. Information items relationships

The higher layers can use different definitions to describe the linked data, and definitions from different models can be updated independently from each other and the input layer. Moreover, higher layers can use probabilities and confidence intervals to signify how probable the values that they contain are in cases of conflicting sources. The benefit is that these probabilities can be assigned in a distributive manner at a research level. Researchers with different beliefs about the probabilities can be accommodated as they can retrieve the original data and decide on their own the plausibility of the data in the datasets that they construct and use.

5 A comparison of the principle with existing approaches

There are two existing implementations of data models that aim to describe the information space of European, historical firm-level data. The first one was developed in SCOB by the University of Antwerp and the second one in DFIH by the Paris School of Economics. Both of them relying on relational technologies so far. While the data model of DFIH is a derivative of the SCOB model, the implementations diverge in that the DFIH infrastructure is oriented towards semi-automated technologies of OCR for input, while the SCOB is oriented towards manual input². The discussion in this section, although it is directly

² The development of the SCOB database started in 1998 while the application for the DFIH project was submitted in 2010.

more applicable to DFIH's approach, is relevant to both of them since the principle that this study proposes does not depend on the way that the data are collected.

In both of these implementations, lines are associated with data model's concepts. For instance, a company name record is accompanied by the text of the line in which it was located. This design allows a record of information to be linked with a single historical source. However, this approach cannot innately handle all cases that were discussed in Section 2 and are frequently found in historical data. The first problem with this approach is that the same informational content can be located in multiple sources. For example, a company name can be in multiple handbooks of listed firms. Since, in most cases, official historical company registries, based on which companies can be unambiguously identified, are not available, any choice of a handbook as the authoritative source is arbitrary. Thereby, this is a potential source of inconsistencies in the content of the system and a burden to its maintenance. The second problem with this approach is that a single line of a source may contain multiple, distinct data items from a content perspective. For example, a single line may contain multiple board member names or multiple financial statement items. The one-to-one design means that in such cases either the same line should be stored for many data items, which leads to data duplication and raises difficulties in keeping data consistency when updating such records, or only part of the line should be stored, which can potentially hinder the data provenance aspects of the model.

It is evident from the discussion of these two problems that, although both implemented systems move towards the direction of associating sources with model concepts, data provenance and model extensibility can be enhanced by applying the proposed preservation principle. Provenance is improved by the ability to trace back all the originating sources of a data item, and extensibility is promoted by separating the semantics for the sources from those of the information space of interest. The many-to-many relationship between lines of sources and data items is needed to effectively achieve these effects. The amelioration of provenance comes from associating data items with more than one source in a standardised manner. The amelioration of extensibility comes from storing sources independently and associating each line with multiple concepts.

6 Collection of joint-stock German company data

In the following section, we outline our approach which builds on a two-step automation process to populate the proposed implementation of the extensible data model with historical German data, including company and stock market observations. The historical data come from printed sources typewritten in old German fonts: the series of "Handbuch der deutschen Aktiengesellschaften" (HdAG) as the main source for the company data and the "Berliner Börsen-Zeitung" for the stock market data.

The HdAG offers a detailed historical compilation of joint-stock companies in Germany. The series was published on an annual basis from 1896 to 2001. Each book contains extensive information on all German joint-stock companies (listed and non-listed), such as date of foundation, purpose, corporate structure, management board, supervisory board, balance sheets, and profit and loss statements.

*** Bayerische Malzfabrik Akt.-Ges., Kulmbach.**
Gegründet: 21./8. 1923; **eingetr.** 31./10. 1923. **Gründer:** Fabrikbes. Max Gauser, Fabrikbes. Friedrich Krauth, Kulmbach; Kapitänleutn. a. D. Karl Siegfried Ritter von Georg, Hamburg; Bücherrevisor Karl Ehemann, Bamberg; Gustav Popp, Kulmbach.
Zweck: Herstell. und der Vertrieb von Malz u. Malzkaffee sowie der Handel mit Getreide, Futtermitteln u. dgl., die Eingehung von Interessengemeinschaften mit anderen Unternehm. gleicher oder verwandter Art, die mittelbare oder unmittelbare Beteilig. an derartigen Unternehm. sowie der Erwerb von solchen.
Kapital: M. 10 Mill. in 900 St.-Akt. u. 100 Vorz.-Akt. zu M. 10 000, übern. von den Gründern zu pari.

Fig. 3. Post-processed scan

The HdAG sources were scanned in a high resolution (600 dpi) format. To improve page-folding and low-contrast deficiencies, the scanned images were processed so that high contrast copies of the original images could be produced. **Fig. 3** shows a section of a resulting image after pre-processing. The processed images are used in our OCR system, which was designed especially for extracting text data from parts of the HdAG series. By default, the OCR system that we used came with a text recognition model for English characters. Based on training data, which was generated from manual transcriptions of text lines from the HdAG, the OCR system was trained to recognise the old German characters that were used in the printed books. The recognition was based on recurrent neural networks (LSTM) and was independent of any language model.

Fig. 4 shows the OCR output of the example snippet in **Fig. 3**. In addition to the recognised text characters, additional information on the coordinates of the bounding box of the recognised text was stored in tags. The average error rate of the process was close to 3%. Furthermore, 18.7% of the errors were the over-recognition of spaces, and they did not add difficulties to the subsequent processing of the extracted data.

```
<span class="ocr_line" title="bbox 541 4427 3035 4551">*Bayerische Malzfabrik Akt.-Ges., Kulmbach.</span><br />
<span class="ocr_line" title="bbox 353 4565 856 4699">Gegründet: 21./8. 1923; eingetr. 31./10. 1923. Gründer: Fabrikbes. Max Gauser, Fabrikbes.</span><br />
<span class="ocr_line" title="bbox 204 4713 837 4847">Friedrich Krauth, Kulmbach; Kapitänleutn. a. D. Karl Siegfried Ritter von Georg, Ham-</span><br />
<span class="ocr_line" title="bbox 284 4861 748 4995">burg; Bücherrevisor Karl Ehemann, Bamberg; Gustav Popp, Kulmbach.</span><br />
<span class="ocr_line" title="bbox 349 5009 837 5143">Zweck: Herstell. und der Vertrieb von Malz u. Malzkaffee sowie der Handel mit Ge-</span><br />
<span class="ocr_line" title="bbox 282 5157 837 5291">treide, Futtermitteln u. dgl., die Eingehung von Interessengemeinschaften mit anderen</span><br />
<span class="ocr_line" title="bbox 282 5305 866 5439">Unternehm. gleicher oder verwandter Art, die mittelbare oder unmittelbare Beteilig. an der-</span><br />
<span class="ocr_line" title="bbox 283 5453 666 5587">artigen Unternehm. sowie der Erwerb von solchen.</span><br />
<span class="ocr_line" title="bbox 346 5601 827 5735">Kapital: M. 10 Mill. in 900 St.-Akt. u. 100 Vorz.-Akt. zu M. 10 000, übern. von den</span><br />
<span class="ocr_line" title="bbox 283 5749 519 5883">Gründern zu pari.</span><br />
```

Fig. 4. Extracted text data

Since errors in numbers (e.g., instead of a "1", a "7" is recognised) have distorting effects on the data quality, the OCR model was trained with a disproportionately large amount of data that contained numbers. As a result, we reduced the error rate for numbers to 1%. However, the above-reported error rates were only valid for the scans that were of high-quality, that is, they were not tilted, faded, or of low contrast.

7 Harmonisation and linking of input data

The transformation process of the input data is illustrated in **Fig. 5**. In summary, the process leads to the creation of four different datasets. At the beginning of the process, multiple lines of the extracted text files that belonged to one firm-year observation were identified (G_data_processing). Each line that marked the beginning of a potential company record was characterised based on six conditions. These were: (i) the bounding box heights, (ii) the vertical distances to boxes of previous lines, (iii) the ratio of box widths to numbers of characters included in the box, (iv) the absolute number of characters of the box, (v) the fraction of non-alphabetical characters, and (vi) the location of the boxes' centre.

Additionally, lines that marked the beginning of other variables were located (*E_fuzzy_matching*) by using a similar set of conditions as those stated above. Every potential company entry was validated by checking for variable duplications within each firm-year observation³.

Common to all subprocesses, summarised as *1: Basic Operations*, was the necessity to account for a wide range of potential OCR extraction errors. Therefore, similarity measures that were based on Levenshtein distance scores of the extracted text data were used and the observed misspellings were corrected. Moreover, the possibility to manually correct observed misspellings of critical terms was provided.

Scripts assigned to *2: Content Operations* aimed at extracting interpretable values of variables from their unstructured string representations. At this stage, the only variables necessary for the subsequent ID creation and linking (*3: ID & Merging*) were constructed.

Fig. 5. The transformation process

The ID linking, represented by scripts *I_id_a - d*, constituted the core part of the process. This part transformed repeated cross-sections of yearly data into a panel structure. Thus, each observation from $t+1$'s cross-section was compared to all observations in time t . Once the *linking score* exceeded a pre-defined threshold, two observations were considered to belong to the same legal entity and thus were assigned to the same ID. Various variables were used

³ For example, if two foundation dates are assigned to one firm-year entry, then that means that one line corresponds to the start of a new company but is not recognized as such.

in each pair-wise comparison⁴. Depending on the variables' characteristics, either binary (e.g., dates) or continuous scores based on Levenshtein distances (e.g., company names) were calculated. The linking approach was insensitive to missing observations. In the end, individual linking scores for each variable were weighted to derive the final linking score.

The approach partly resulted in fractional time-series which meant that one company was assigned to different IDs over time. A second linking round aimed to mitigate this problem. In this round, observations previously assigned to one ID were only compared to IDs of observations that covered different periods. This procedure used a lower matching threshold as the chance of false-positive matches was reduced by construction. After the automatised linking, manual corrections were necessary to fill the remaining gaps in the time-series.

Descriptive statistics illustrate the resulting dataset's coverage. Considering only the raw data, the left part of **Fig. 6** shows that the number of observations extracted from the data source increase sharply until the middle of the 1920s. Afterward, the numbers show a steady decline. One of the challenges of the dataset's construction was to deal with a change in reporting schemes. Until Volume 29, the HdAG covered reports from the beginning of July to the end of June. From Volume 30 onwards, however, the HdAG reporting scheme became linked to calendar years. This change reflects one aspect of the earlier mentioned lack of standardisation of the sources. The change explains the drop of observations in Volume 29 that includes many observations that are likewise covered by Volume 30. Thus, duplicates were removed so that the panel structure's time dimension could be defined. The right part of **Fig. 6** illustrates the resulting observation count per year.

Fig. 6. Number of observations per volume (left) and year (right)

8 Linking with security data

A subset of the joint-stock companies covered by the HdAG was publicly listed on one or multiple German stock exchanges. Thus, matching daily data on the stock market extracted from the Berliner Börsen-Zeitung is potentially highly valuable for economic and financial research. The information on stock prices was printed in columns characterised by a high degree of variation in formats and content. These inconsistencies together with the challenges that arose from inadequate horizontal segmentation of columns made an automated digitisation and structuration process inapplicable. Instead, digitisation by hand was applied. The resulting dataset was then matched to our company dataset.

Probably due to space limitations, the Berliner Börsen-Zeitung's used various kinds of abbreviations for company names. Moreover, the spellings of the two datasets relied on different special characters. The linking algorithm first homogenised spelling styles before

⁴ While founding dates do not vary over time, other variables (e.g., company names) might occasionally do so. Due to the wide range of variables covered in the linking process, such changes in variables should not lead to false-negative links.

performing string comparisons to mitigate issues potentially arising from these differences. As opposed to string transformations performed when cross-sections were linked, the methods, in this case, focussed more on abbreviations than potential OCR errors. The string comparison itself again relied on similarity measures such as Levenshtein distances.

9 Conclusions

In this paper, we discuss the principle of preserving the historical sources as a potential solution to the standardisation difficulties that arise from the potential ambiguity that accompanies the data collections from historical sources and hinders the design of extensible data models. While contemporary data models were built with standards in mind, applying this approach to historical data where the sources are highly non-standardised and often conflicting is rather anachronistic. Using such contemporary data models hinders the extensibility of research infrastructures that are based on such approaches. Instead, the principle of this paper was based on the observation that the historical sources, which are finite in number and invariant in nature, constituted a solid basis to be used as a linchpin for the design of extensible database-driven systems.

We sketch and develop a relational implementation of the principle, and we examine how this approach could incrementally enhance existing historical databases such as SCOB and DFIH. We also highlight that the scope of the principle, which is to accurately associate sources and higher-level concepts, is different from that of applying metadata standards to datasets, which is to holistically describe the content of a dataset.

We also discuss the process of extracting data from images of historical sources for German companies. These data can be used to populate the relational implementation that we have developed. Furthermore, we describe the process of parsing the text data and creating harmonised variables that correspond to concepts of financial interest. Specifically, we describe the process of creating a company dataset and linking it with a security dataset.

Lastly, our analysis paves the way for dealing with historical European sources of which most are non-harmonised, unstructured, and highly heterogeneous. Thus, we contribute with new insights and tested approaches to the book-to-database paradigm.

References

1. Acemoglu, D., & Robinson, J. (2013). Economics versus politics: Pitfalls of policy advice. *Voprosy Ekonomiki*, 2013(12), 4-28.
2. Anderson, H., Dungey, M., Osborn, D., & Vahid, F. (2011). Financial integration and the construction of historical financial data for the Euro Area. *Economic Modelling*, 28(4), 1498-1509.
3. Costantino, M., & Coletti, P. (2008). *Information Extraction in Finance*.
4. Xie, Z., Lv, W., Qin, L., Du, B., & Huang, R. (2018). An evolvable and transparent data as a service framework for multisource data integration and fusion. *Peer-to-Peer Networking and Applications*, 11(4), 697-710.
5. Idreos, S., Maas, L., & Kester, M. (2017). Evolutionary data systems. *arXiv*.
6. Ranft, L., Braswell, J., & König, W. (2021, 3). EURHISFIRM D5.5: Report on process for extendable data models. Retrieved from <https://doi.org/10.5281/zenodo.4616475>
7. Nehme, R., Works, K., Lei, C., Rundensteiner, E., & Bertino, E. (2013). Multi-route query processing and optimization. *Journal of Computer and System Sciences*, 79(3), 312-329.
8. Zoumpatianos, K., Idreos, S., & Palpanas, T. (2016). ADS: the adaptive data series index. *VLDB Journal*, 25(6), 843-866.

9. Dittrich, Jens, A. (2011). Towards a One Size Fits All Database Architecture (OctopusDB). *Cidr*, 195-198.
10. Verelst, J. (2005). The influence of the level of abstraction on the evolvability of conceptual models of information systems. *Empirical Software Engineering*, 10(4), 467-494.
11. Annaert, J., Buelens, F., & Riva, A. (2016). Financial History Databases: Old Data, Old Issues, New Insights? *Financial Market History*, 44, – 65.
12. Jorion, P., & Goetzmann, W. (1999). *Global Stock Markets in the Twentieth Century* (Vol. 54). Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00133>
13. Dimson, E., Marsh, P., & Staunton, M. (2005). Long-Run Global Capital Market Returns and Risk Premia. *SSRN Electronic Journal*(217849).
14. Dimson, E., Marsh, P., & Staunton, M. (2009). Triumph of the optimists: 101 years of global investment returns. Princeton University Press.
15. Global Financial Data. (2005). *GFD Encyclopedia of Global Financial Markets* (10th ed.).
16. Mata, M., Costa, J., & Justino, D. (2017). The Lisbon stock exchange in the twentieth century. *Coimbra*: Coimbra University Press.
17. Vaihekoski, M. (2020). *Revisiting Index Methodology for Thinly Traded Stock Market. Case: Helsinki Stock Exchange*. Case: Helsinki Stock Exchange.
18. Ducros, J., Grandi, E., Hékimian, R., Pruniaux, E., Riva, A., & Ungaro, S. (2018). *Collecting and storing historical financial data: the DFIH project*. Retrieved from <https://econpapers.repec.org/RePEc:hal:journl:halshs-01884372>
19. Karapanagiotis, P. (2020). *Technical Document on Preliminary Common Data Model*. Retrieved from <https://doi.org/10.5281/zenodo.3686930>
20. Rydqvist, K., & Guo, R. (2020). *Performance and development of a thin stock market: The Stockholm Stock Exchange 1912-2017*.