Design and Implementation of a Historical German Firm-level Financial Database

DENNIS GRAM, Leibniz Institute for Financial Research SAFE, Germany
PANTELIS KARAPANAGIOTIS, EBS Business School, Germany
MARIUS LIEBALD and UWE WALZ, Goethe University Frankfurt, Germany

Broad, long-term financial, and economic datasets are scarce resources, particularly in the European context. In this article, 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 digitized and structured long-term historical datasets for different jurisdictions and periods. 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, we outline a relational implementation of this approach based on historical German firm and stock market data from 1920 to 1932.

CCS Concepts: • **Information systems** → **Database design and models**; **Deduplication**; **Digital libraries and archives**; • **Applied computing** → **Document management and text processing**; **Economics;**

Additional Key Words and Phrases: Databases, economic history, cliometrics, financial data, Germany

ACM Reference format:
Dennis Gram, Pantelis Karapanagiotis, Marius Liebald, and Uwe Walz. 2022. Design and Implementation of a Historical German Firm-level Financial Database. *J. Data and Information Quality* 14, 3, Article 19 (June 2022), 22 pages.
https://doi.org/10.1145/3531533

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

Uwe Walz also with Leibniz Institute for Financial Research SAFE.
Authors' addresses: D. Gram, Leibniz Institute for Financial Research SAFE, 3 Theodor-W.-Adorno-Platz, 60323, Frankfurt am Main, Hesse, Germany; email: gram@safe-frankfurt.de; P. Karapanagiotis, EBS Business School, 1 Rheingaustraße, 65375, Oestrich-Winkel, Hesse, Germany; email: karapanagiotis@ebs.edu; M. Liebald and U. Walz, Goethe University Frankfurt, 4 Theodor-W.-Adorno-Platz, 60323, Frankfurt am Main, Hesse, Germany; emails: liebald@econ.uni-frankfurt.de, uwalz@econ.uni-frankfurt.de.

1936-1955/2022/06-ART19 $15.00
https://doi.org/10.1145/3531533

US is easily accessible, long-run structured and consistent databases for Europe are scarce.[1] This lack has two immediate consequences. First, most empirical studies in economics and finance use US data. Transferring the conclusions of these studies to other legal systems or structures, such as those of continental Europe, potentially leads to serious misconceptions because of fundamentally different institutions and political systems (see e.g., Acemoglu and Robinson [2013]). Second, not least in the aftermath of the 2008 financial crisis, the research community has recognized that data with a short time 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. Hence, there is a need to build databases with financial data (in a particular company and financial market data) for Europe that span decades and eventually centuries.[2] Our approach aims at narrowing this gap.

However, the construction of long-term structured databases that comprise historical datasets is a tremendous challenge. First, historical data from various sources with overlapping entities for a given time period can be ambiguous in that different sources can either have incomplete information or even conflicting information with other sources. Second, linking and merging different data sources without the existence of common identifiers is difficult (see also Coletti and Murgia [2015]). Thus, proper metadata standards must be set up and agreed on. Third, studies in the fields of financial economics often rely on a mix of quantitative and qualitative data (see e.g., Costantino and Coletti [2008] for this distinction). For instance, company reports, which are frequently used in economic research, entail numerical data (e.g., balance sheet, profit and loss statement) as well as textual data (e.g., executive boards, supervisory boards). While accessing combined datasets is essential for research, merging is complicated.

Fourth, the vast majority of historical data is stored in items such as books, journals, or loose-leaf editions. Creating digital and structured datasets by hand from these sources requires immense resources. The development of machine learning, and especially deep learning, over the last decade, however, has made the digitization and structuring of historical sources possible in a faster and more cost-efficient way. Fifth, another crucial feature is to have a data model that is flexible in nature while providing sufficient space for proper standards and identification of sources. Datasets that include historical information often lack this feature since individual researchers constructed most of these datasets in the scope of their individual projects and did not adhere to common standards. A deeper reason for the tremendous challenge of constructing databases that comprise historical datasets are due to the nature of the sources. The standards and persistence of identification constitute invariance points in data-model designs. Once one sets these basic properties, the extensibility of the data model follows. Such invariance points are often missing from historical data because the researchers documented their sources at times in which the standardization and identification theory was underdeveloped. Although there is no invariance in the standards for historical data, there is an invariance in sources. This invariance can be used as a focal point when designing an extensible infrastructure.

Our article aims at addressing all these issues and therefore has the following main objectives. First, we aim at outlining a base for an extensible data model and a process to digitize and structure large historical datasets. Second, we offer a proof of concept by constructing a database for German historical firm-level and stock market data. Third, we describe how this database can become part

[1]For example, concerning the US, the **Centre for Research and Security Prices** (**CRSP**) and Compustat Databases provide accounting data and security prices starting from 1925. For Europe, the exception is the EUROFIDAI database that covers broader firm-level data from 1980 to today.

[2]One could consider Refinitiv's Datastream as a potential candidate. This historical financial database claims to cover 70 years of data for 175 countries. A more granular view, however, reveals that data coverage for numerous countries is deficient and that routinely old data is dropped.

of a pan-European historical database for firm-level and stock market data which is accessible and open to the wider research community. In this spirit, we implement the German historical database on the basis of the FAIR principles (Findable, Accessible, Interoperable, and Re-usable data) setting the pace for a flexible infrastructure.

In our approach, we take into account that data collection often does not use predefined standards and that the ex-post harmonization, standardization, and verification of data cannot typically take place without having access to the original sources. The proposed implementation aims at coping with the particularities of historical data that allows researchers to validate the digitized data against the original data source. Building on this idea, we further elaborate in more detail on our project to digitize and structure historical company data from Germany based on our proposed extensible data model. This provides a use case for our extensive data model.

Finally, we would like to note that analyzing long-term historical data provides not only a better understanding of the underlying mechanisms of European societies from a historical perspective. It might also be helpful for the analysis of contemporary (economic) challenges, not least because it allows to cover significant exogenous shocks and hence identify causal relationships (see e.g., Doerr et al. [2021]; Ferguson and Voth [2008]; Huber et al. [2021]), which is important for future economic progress (e.g., Anderson et al. [2011]).[3] Therefore, researchers need to be equipped with a digital access to datasets of this kind.

The article is organized as follows. In the next section, we review the literature on the creation and development of historical financial datasets. In the third section, we discuss the specific role of the original information that leads to digitized historical databases, and then we propose the principle of the preservation of the original historical source. The fourth section has a discussion on our relational implementation of this principle, while section five compares it with potential alternatives. In section six, we delineate the collection of input data by referring in detail to our use case—the German historical firm-level database. The seventh section focuses on the key steps of the *input layer*'s transformation into the *original layer*. We provide summary statistics of our German historical firm-level database. Furthermore, we discuss the linking of this database with the corresponding stock market data. 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 long-term data, we provide an overview on four related fields: (i) studies that focus on the collection process of historical datasets, (ii) research that uses long-term historical and aggregated data in the context of economic and financial analyses, (iii) articles that use long-term microdata on companies, and (iv) studies that are focusing on long-term data in the area of stock market analyses. This overview sets the stage for our own study.

2.1 Extensible Data Models

Xie et al. [2018] point to the challenges occurring in the context of the highly desirable continuous integration of data from numerous sources. Against the background of these challenges, including the necessity to deal with the multitude of options that are based on the underlying architectures of the systems, Idreos et al. [2017] provide a vision for an evolutionary data system according to which the evolution does not require human involvement.

[3]Historical data might also shed new light on contemporary discussion such as the importance to analyze the determinants of capital structure decisions with long-term data (see e.g., DeAngelo and Roll [2015]) and the impact of previous shocks on decision-makers on their future risk-taking (see e.g., Bernile et al. [2017]), a matter which can be potentially well addressed with data from interwar Germany with its many severe shocks.

Ranft et al. [2021] provide a comprehensive discussion of the features of various processes, formats, and systems as well as their contribution to extensibility. Our article narrows this general discussion of extensibility in the context of a 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. These ideas on adaptive data architectures aim at increasing performance through optimization of query plans that are based on the distinct properties of the included data. Nehme et al. [2013] build on this and provide a modern approach to processing a data stream. Their approach aims at computing multiple routes of data queries that are individually designed for particular subsets of the data. Zoumpatianos et al. [2016] follow the approach to adaptively build auxiliary data structures that are required as large numbers of data series are continuously produced. Further, another strand of the literature argues that data should be kept in a flexible format structure, embedded in a single system that provides multiple data views (see e.g., Dittrich and Jindal [2011]).

Abstraction and genericity are attributes that the computer science literature commonly identifies as central ingredients in the extensibility of systems. However, on many occasions, the designers have overestimated the necessity of adding abstractions that hinder extensibility. Using an experimental design, Verelst [2004] provides statistical evidence of this industry wisdom, namely that abstractions are not always in favor of the extensibility of conceptual models. The author shows two cases, which we find to be applicable in the context of historical data. First, abstractions reduce the time needed to implement modifications in a conceptual model in cases of adding complicated changes. In contrast, if modifications are simple, abstractions might increase the time needed to make changes. Considering the complexity of financial data, an abstraction separating sources from concepts can be helpful in the design of a firm-level data model. Second, the author also presents evidence that abstractions can be beneficial to the correctness of adding changes to the conceptual model. This benefit constitutes another feature of designing models with historical data, as the addition of new historical sources might introduce needs for conceptual distinctions that designers did not include in the data model.

2.2 Processes of Generating Historical Data

Generating historical datasets in the area of finance is a challenge, which is worthful to address but may be associated with severe potential deficiencies. Annaert et al. [2016] argue that the weak empirical foundations of economic and financial analytical models are in parts attributable to the scarce availability of long-run consistent financial microdata. Country indices that reflect the performance of bond and equity markets are on the other hand more easily available (see e.g., Dimson et al. [2002, 2009]; Jorion and Goetzmann [1999] or the *Global Financial Data* database). The scarcity of long-run financial microdata 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 such as CRSP, managed by the University of Chicago. Recently, and pursuant to Annaert et al. [2016], 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., Mata et al. [2017]; Vaihekoski [2021]). The DFIH initiative has developed a comprehensive long-run stock exchange database on the French markets from 1796 to 1976 [Ducros et al. 2018]. The DFIH database uses two main printed serial sources for its population: the official lists of the Paris Stock Exchange and official and private stock exchange yearbooks. Two technologies to capture data

were set up. The first was characterized by manual data entry and the second by semi-automatic processing of the stock exchange yearbooks that utilized **optical character recognition** (**OCR**) and artificial intelligence.[4]

Karapanagiotis [2020] discusses whether the *EurHisFirm* initiative, which aims at covering data from numerous European countries, requires an overarching identification system for European, historical firm-level data. For this purpose, he provides thorough discussions on 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 to identify different classes of economic entities as well as standards, governance, and the need to harmonize across countries.

Finally, Rydqvist and Guo [2021] utilize a newly constructed dataset of daily transaction prices and volume data from the Stockholm Stock Exchange for the period from 1912 to 1978.

2.3 Analysis of Aggregated Historical Economic Data

There are numerous studies that rely on specific, often hand-collected long-term **aggregate** datasets. Starting with the seminal article of Mehra and Prescott [1985], several researchers have investigated the behavior of returns in financial markets. While the initial research was heavily focused on the US, the dataset constructed by Dimson et al. [2002, 2009] extents the data to equity returns for 21 of the world's stock markets from 1900 through 2014. Many researchers have used this aggregate multi-country dataset to explore various directions (see e.g., Goetzmann [2015]).

In a related article, Danielsson et al. [2018] analyze the effects of volatility on financial crises by constructing a cross-country dataset. This dataset covered 60 countries with varying observation periods. For the US, the country with by far the longest observation period, data is available for 210 years from 1800 to 2010. Along similar lines, Reinhart and Rogoff [2011] exploit a new long-term historical dataset to study debt, banking crises, and inflation as well as currency crashes. Their data covered 70 countries and the time frame of their analyses covered over two centuries.

By building on the Macrohistory Database, Jordà et al. [2017] and Richter et al. [2020] investigate the relation between credit cycles and banking crises. The Macrohistory Database contains economic and financial data for 17 advanced economies from 1870 to 2013. Using the same database, Jordà et al. [2019] research the rate of return on everything including bonds, equities, and real estate.

2.4 Microeconomic, Firm-level Historical Data

To put our own project into perspective, we take a short tour of the historical studies that use microeconomic data, in particular company (-related) data.

By collecting new datasets to measure aggregate changes in the US patenting across research fields, Moser et al. [2014] investigate changes in research output at the level of individual US inventors between 1920 and 1970. Waldinger [2016] complements these analyses by examining the role of human and physical capital in the creation of scientific knowledge. For this purpose, he constructed a new panel dataset of physicists, chemists, and mathematicians at German and Austrian universities. Lampe and Moser [2016] examine changes in patenting after the creation of a patent pool by collecting a new dataset, which covered the patents filed between 1921 and 1948 for 20 industries that were affected by the creation of a pool.

An example of a study on the financing side of companies is Barton and Waymire [2004] who use firms listed in the monthly database of the CRSP that were traded on the New York Stock

[4]See for more details: https://www.uantwerpen.be/en/research-groups/scob/about and https://dfih.fr/about.

Exchange in October 1929. They investigate the relationship between financial reporting and investor protection that preceded the market crash in 1929.

2.5 Historical Data and Stock Market Analysis

In this subsection, we discuss microeconomic analyses of historical stock market data. We start with some analyses of European markets which are scarcer, before referring to studies on the US markets.

Annaert et al. [2015] analyze the monthly returns of Belgian stocks listed on the Brussels stock exchange in the period from 1838 to 2010. They utilize data on stock returns which are based on official quotation lists of the stock exchange that are taken from the SCOB database. This database comprises end-of-the-month stock prices, dividends, interests, ex-dividend day, corporate actions, and the number of stocks for all stocks ever quoted on the Brussels stock exchange. In another analysis, Moortgat et al. [2017] use a sample of listed Belgian firms between 1838 and 2012 to investigate whether investor protection and taxation regulation had an impact on dividend policy. Braggion and Moore [2011] analyze the effects of dividend policies on 475 British firms from 1895 to 1905 in an unregulated low tax regime. In the same vein, the article by Turner et al. [2013] utilizes a hand-collected dataset of firms on the London stock market between 1825 and 1870 to analyze firms' underlying rationale behind paying dividends. Relying on data for the German stock market before World War I, Schlag and Wodrich [2000] seek empirical evidence on initial public offerings by investigating the pricing and long-run performance of IPOs. Using similar data, Gehrig and Fohlin [2006] estimate effective spreads of securities traded on the Berlin Stock Exchange in 1880, 1890, 1900, and 1910.

Turning to the US, Bernstein et al. [2019] utilize the establishment of a clearinghouse on the New York Stock Exchange in 1892 to analyze the effect of centralized clearing on counterparty risk. For this purpose, the authors collected end-of-month data on transaction prices as well as bid and ask closing prices, and trading volumes for all stocks included in the Dow Jones Industrial Average from September 1886 to December 1925. In another article, Goetzmann et al. [2001] estimate the power of past returns and dividend yields to forecast future long-horizon returns based on data for individual prices of New York Stock Exchange stocks from 1815 to 1925 and individual dividend data ranging from 1825 to 1870.

In recent years, studies have made substantial efforts to reconstruct indices from other countries. Regarding Chinese data, Fan [2004] describes the collecting process of individual stock data for the period from 1871 to 1940 from the North China-Herald, a local English newspaper. Furthermore, a study by Goetzmann and Huang [2018] relies on a dataset of hand-collected end-of-month stock prices of all companies listed on the St. Petersburg Stock Exchange from 1865 to 1914.

In summary—as Eichengreen [2016] points out—the research in financial history has enjoyed a renaissance in recent years due to lower costs and greater efficiency regarding the extraction and digitalization of historical datasets. However, while developing quickly, the existing datasets are still fragmented, with little coherence and interlinkages. Designing an extensible common data model with a set of metadata may be a promising route towards more coherent and interrelated data. This interrelation may not only reflect different types of data but also refer to different jurisdictions and different periods.

3 THE ROLE OF ORIGINAL INFORMATION IN HISTORICAL DATABASES

One central issue of setting up historical databases is that of deduplication. The problem of deduplication is neither new nor specific to databases with historical information (see Madnick et al. [2009] for a literature overview). Deduplication commonly refers to processes that establish whether two or more records in a collection of data represent the same object of interest. We

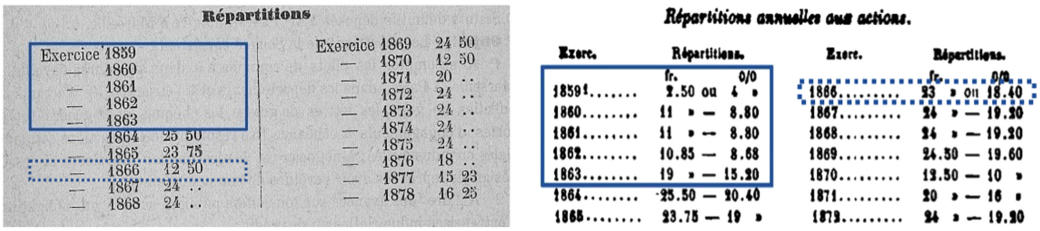


Fig. 1. Examples of missing (solid line) and conflicting (dashed line) information.

complement this literature by specifying a relational data model with some metadata characteristics that are particularly relevant when dealing with historical non-standardized and non-verifiable information sources. Our solution is based on capturing the provenance characteristics of historical information.

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 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 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 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 rely on different semantics and formats or, even worse, their descriptions of objects of O are conflicting.

Figure 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 1880s 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 1859–1872; however, they do not contain the same information (see solid line rectangles). Specifically, dividend payments are not reported for the years 1859–1863 in the left snippet, while records exist in the right snippet.

From the perspective of a data-model design, the example of missing information in Figure 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.

Figure 1 also highlights a case of conflicting information (see dashed line rectangles). In this case, the left snippet of Figure 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 reveals that both sources agree 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 modeling approach that standardizes 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 standardization. From an end users 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.

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 to emerge.

Our relational implementation is based on data whose digitization is performed by (semi-) automated OCR tools. We adopt this approach because our goal is to provide a solution that is scalable both in terms of including additional historical archives and allowing a greater number of researchers parallelly working on the data. Our digitization approach is not necessary for the analysis, and the proposed design can be easily adapted to systems with manual input. However, big-data systems are unlikely to be viable if not based on automated input. In our implementation, the *input layer* is used to store information directly from the output of the OCR. We present the *input layer* schema in three parts.

The first part, depicted in Figure 2, concerns publishing information. The same publishing house may have published multiple historical archives of interest, and in such a case, all of these sources are linked to a single publisher's record. In order to cover cases in which a source is part of a publication series, source titles are stored separately. Historical sources are also associated with digitized files, which are stored in their entirety. Each source can be linked to multiple digitized files. This one-to-many relation is essential for our OCR approach because differences in the quality of digitized files of the same source can lead to significantly different recognition results.

In terms of notation, each rectangle in Figure 2 represents a table in our relational implementation. The fields preceded by a yellow key define unique identifiers for each table. The fields preceded by grey keys represent foreign keys. The last row of each rectangle summarizes, going from left to right, the number of table fields referencing other tables, the number of records in the table, and the number of other tables referring to this table.

The second part of the *input layer* is depicted in Figure 3 and concerns the organization of the raw digitized data. Every digitized file has multiple pages, an association that is reflected by a one-to-many relation between the *pdf_source_files* and the *pages* tables of the *input layer*. Every page record corresponds to a digitized page file. For every page record, the output of the OCR is stored in XML format, along with extraction metadata that is important in determining annotation boxes.

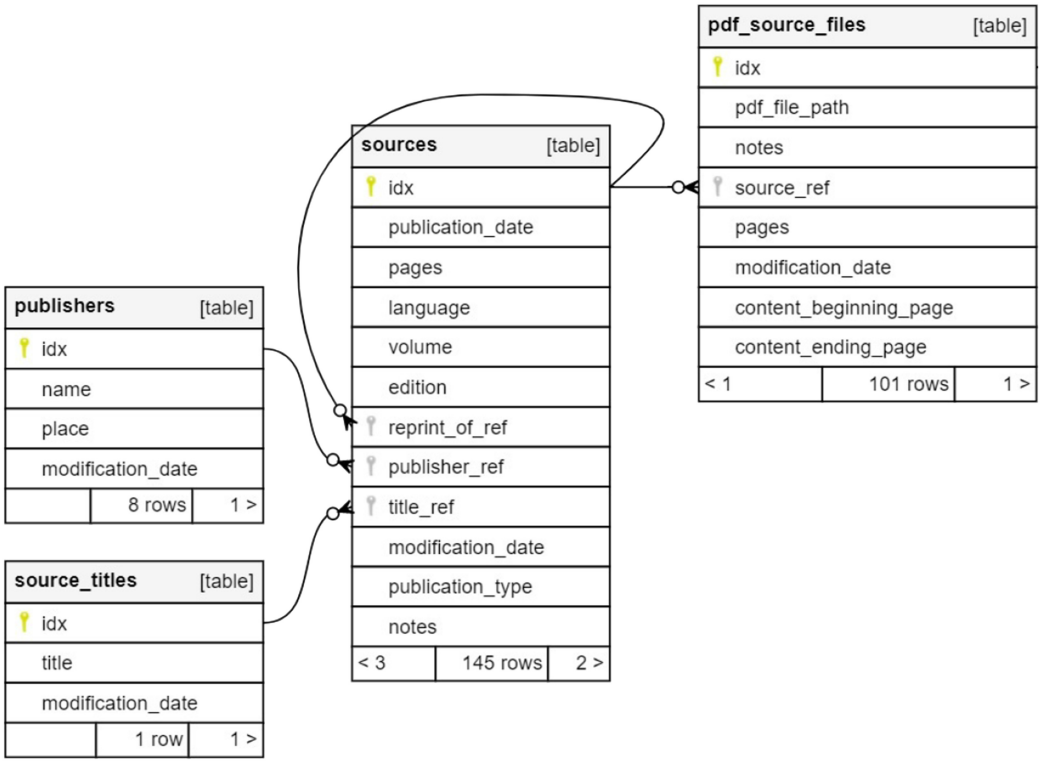


Fig. 2. Publishing information relations.

In particular, our implementation stores the size, measured in pixels (*ocr_coordinates* column), that the OCR system assigns to the page. Moreover, since the page numbering in the electronic and in the physical files might be different, information about these numbers is stored separately.

The extracted lines of the files are stored in an analogous manner. Every page contains multiple lines. The digitized lines are stored as they are recognized by the OCR system, which implies that the stored line number does not necessarily reflect the actual line number of the source but rather the number based on the ordering that the recognition algorithm assigns to this line (*ocr_ordinality* column). Besides the recognized text, the coordinates that identify the line in the page are stored. These coordinates are relative to the pixel coordinates of the associated page. Our example also considers the possibility of storing manual corrections to the results of the automated recognition process. This can be seen in Figure 3 in which the *lines* table allows storing both automated and manual input in the *ocr_text* and *manual_text* columns.

The last part of the *input layer* captures the essence of the principle of preserving the historical sources by associating concepts of interest in the data model with their origin, which alleviates the ambiguity characterizing the standardization of potentially conflicting archives. This is also the point at which our implementation 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 buffer that enhances the extensibility capacity of the system.

As illustrated in Figure 4, the design introduces the concept of information items. An information item is an abstraction that comes in-between the sources and the data records that are

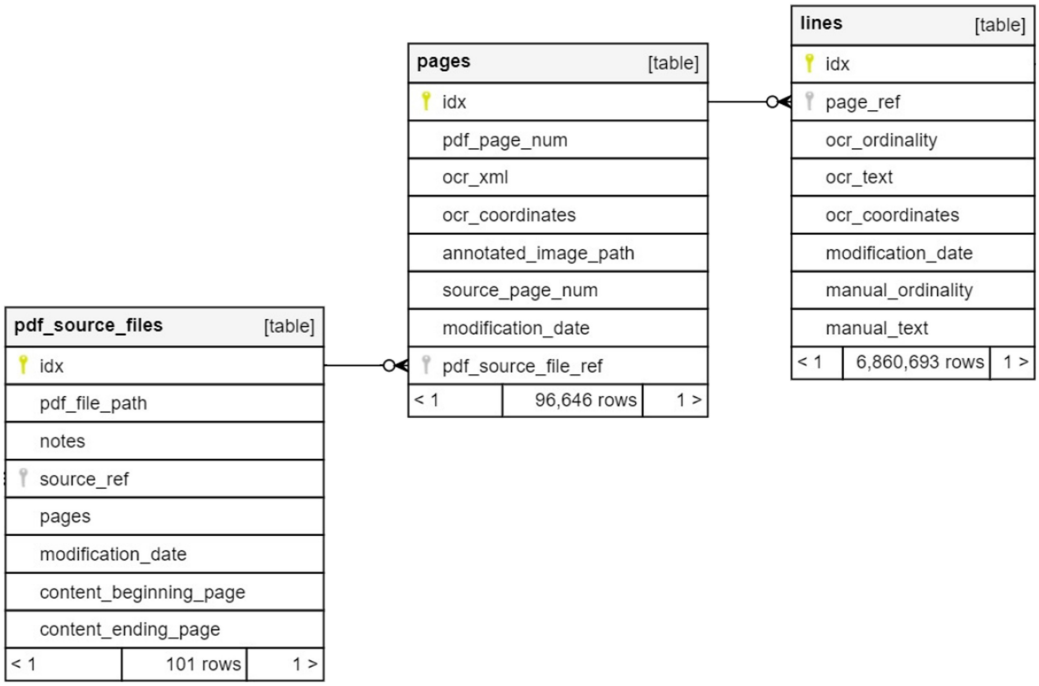


Fig. 3. Raw digitized information relations.

conceptualized 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. Each information item may contain data relating to one or more concepts of the data model and produce one or more data records. For example, multiple board member names are located in a single line, or both the address and the managing director of a company 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 Figure 4 the information items are related to 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 relating to economic entities of interest. Specifically, our implementation uses the information item abstraction to associate lines with company names. Subsequently, the collected company names are also linked with the corresponding industries found in the historical sources.

The design of the *input layer* brings together three fundamental components from a top-down, infrastructure point of view. Each component corresponds to one of the previously presented parts. The first component concerns the organization of historical archives and their potentially multiple digitized copies. The second component essentially associates the OCR output with each of the digitized sources. The last component is the abstraction buffer that underlies the preservation principle and enables some provenance and verification novelties in the design of systems with historical context.

In essence, this approach fully captures the provenance of the data records that are found in system layers that are built on top of the *input layer* and also offers the ability to develop verification processes within the system. Technically, the abstraction uses a simple many-to-many relationship to describe the association between sources and higher model concepts. The information item abstraction also creates a modeling buffer between these higher order modeling concepts and the

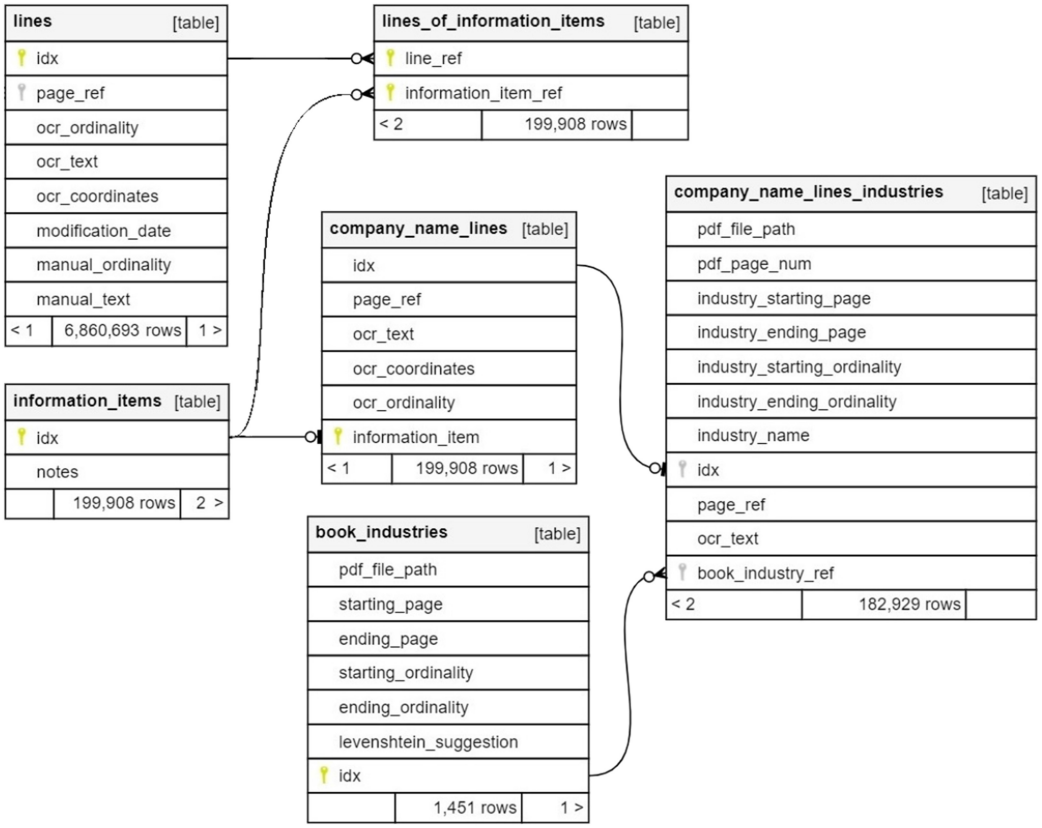


Fig. 4. Information items relations.

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 standardization is plausible. The former, however, are not entirely explored, have complicated interconnections, and can have fuzzy and conflicting content; attributes which suggest that data models describing these higher order 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 order 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*.[5]

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]We have already put this modularity feature into practice in our implementation. Our database's in-development and explorative extensions are performed in *devel layers* built on top of the *input layer* while operating in parallel with the main *original layer* presented in this article.

5 A COMPARISON OF THE PRINCIPLE WITH EXISTING APPROACHES

There are two existing implementations of data models that aim at describing 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 input.[6] The discussion in this section, although it is directly 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 the 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 3 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 records 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 records, 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.

Content-wise, our implementation of the German database is complementary to the information contained in the DFIH and SCOB databases. The DFIH database covers French, SCOB documents Belgian, and ours focuses on German companies. All three databases provide data on financial instruments, financial statements, and corporate governance of their firms. In particular, DFIH and SCOB are two of the flagship implementations of the *EurHisFirm*, an interdisciplinary research consortium aiming at promoting research based on previously unexplored, European micro level, historical data.[7]

6 COLLECTION OF INPUT DATA—THE EXAMPLE OF GERMAN JOINT-STOCK COMPANIES

In the following section, we outline our approach which builds on a two-step automation process to populate the *input layer* of our implementation of the extensible, historical German database, 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.

[6]The development of the SCOB database started in 1998 while the application for the DFIH project was submitted in 2010.
[7]See also https://eurhisfirm.eu/.

Table 1. Data Distribution Among Volumes

HdAG Volume	Digitized Pages	OCR Pages	OCR Lines	Pages corrected manually
25	5,784	5,783	433,601	74
26	6,354	6,351	480,708	256
27	6,748	6,741	508,267	380
28	8,256	8,246	614,277	240
29	4,508	4,505	339,634	172
30	9,061	9,051	620,636	28
31	8,476	8,475	586,072	23
32	8,494	8,488	560,925	35
33	8,180	8,177	553,483	115
34	7,894	7,893	530,650	92
35	7,850	7,849	533,929	72
36	7,512	7,512	551,066	9
37	7,574	7,574	547,445	8
Total	96,691	96,645	6,860,693	1,504

*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. 5. Section of a scanned HdAG page.

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.

We have 13 years of firm level data available during the interwar period. We extract the variables from digital reprint data for 1920–1932 (volume 25–37) of the HdAG. The 13 years are spread over 101 editions (each volume contains between 5 and 10 editions) containing 96,645 pages in total. Table 1 shows the data distribution among volumes. Differences in Digitized Pages and OCR Pages are primarily due to white pages that are ignored by the OCR engine.

Although the HdAG sources were scanned in a high resolution (300 dpi) format, the original purpose of the scans was to generate data to reprint the old books. Thus, the scan settings were not chosen to be optimal for OCR. As a result, we lose some OCR accuracy due to lower contrast scans.

Figure 5 shows a section of a scanned page. The pages were not scanned perfectly straight, and the book fold attributes a gradient effect of the text lines. In addition, the yellowed article reduces

***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. 6. Post-processed scan example.

the contrast between printed and unprinted areas. To improve the second deficiency, the scanned images were processed so that higher contrast copies of the original images could be produced.

Figure 6 shows the same section of the image after pre-processing. The processed images are used in our OCR system. The OCR system was designed especially for extracting text data from parts of the HdAG series. By default, the OCR system that we used as a basis came with a text recognition model for English characters.[8] Based on training data, which was generated from manual transcriptions of 80 randomly sampled pages (~0.000828% of available pages), the OCR system was trained on 4,047 lines (60 pages) of training data to recognize the old German characters that were used in the printed books. We kept 1,340 lines (20 pages) as test data to evaluate the performance of the OCR model.

The evaluation of the error rate on the test data showed that after 200,000 iterations on the training data, the error rate increased. We identified this as the threshold to overfitting, stopped the training process, and stored the parameters. The average error rate of the process was 3.021%. 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 3368 4639'>Gegründet: 21./8. 1923; eingetr. 31./10. 1923. Gründer: Fabrikbes. Max Gauser, Fabrikbes.</span><br />
<span class='ocr_line' title='bbox 204 4639 3371 4712'>Friedrich Krauth, Kulmbach; KapitÄĆÄdnleutn. a. D. Karl Siegfried Ritter von Georg, Ham–</span><br />
<span class='ocr_line' title='bbox 204 4711 2698 4786'>burg; Bücherrevisor Karl Ehemann, Bamberg; Gustav Popp, Kulmbach.</span><br />
<span class='ocr_line' title='bbox 349 4800 3368 4872'>Zweck: Herstell. und der Vertrieb von Malz u. Malzkaftee sowie der Handel mit Ge–</span><br />
<span class='ocr_line' title='bbox 202 4873 3365 4948'>treide, Futtermitteln u. dgl., die Eingehung von Interessengemeinschaften mit anderen</span><br />
<span class='ocr_line' title='bbox 202 4944 3366 5018'>Unternehm. gleicher oder verwandter Art, die mittelbare oder unmittelbare Beteilig. an der–</span><br />
<span class='ocr_line' title='bbox 203 5016 1968 5085'>artigen Unternehm. sowie der Erwerb von solchen.</span><br />
<span class='ocr_line' title='bbox 346 5101 3368 5179'>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 203 5173 831 5249'>Gründern zu pari.</span><br />
```

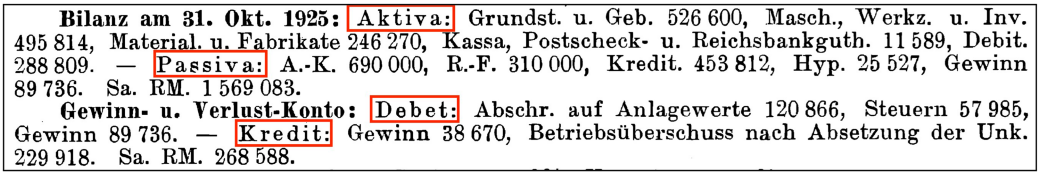
Listing 1. Extracted text data.

Since errors in numbers (e.g., instead of a "1", a "7" is recognized) 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 1.094%. In addition to the recognized text characters, metadata on the coordinates of the bounding box of the recognized text was stored in tags.[9] This information will be used later in the process to identify company names and other relevant information items.

During the process of identifying firm names, further quality issues with some pages have emerged. 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. We identified such pages by searching for repeating string artifacts that were produced by the OCR system on low-quality pages. Overall, 1,504

[8]See https://github.com/ocropus/ocropy.

[9]See Listing 1 for an OCR output example. The output corresponds to the content shown in Figures 5 and 6.


Bilanz am 31. Okt. 1925: Aktiva: Grundst. u. Geb. 526 600, Masch., Werkz. u. Inv. 495 814, Material. u. Fabrikate 246 270, Kassa, Postscheck- u. Reichsbankguth. 11 589, Debit. 288 809. — Passiva: A.-K. 690 000, R.-F. 310 000, Kredit. 453 812, Hyp. 25 527, Gewinn 89 736. Sa. RM. 1 569 083.

Gewinn- u. Verlust-Konto: Debet: Abschr. auf Anlagewerte 120 866, Steuern 57 985, Gewinn 89 736. — Kredit: Gewinn 38 670, Betriebsüberschuss nach Absetzung der Unk. 229 918. Sa. RM. 268 588.

Fig. 7. Balance sheet and profit and loss statement example.

pages (column five in Table 1 shows the distribution among the volumes) with such characteristics were identified and corrected manually.

There were some additional difficulties that arose even after correcting pages with low-quality scans. Balancesheet information and profit and loss statements were not always properly extracted. This issue was due to two factors. First, on some occasions, the OCR system did not manage to properly identify the lines in the balance sheet and profit and loss areas. Thus, the needed information was not extracted. Second, given that the sources used to produce the reprint data are old, parts of some pages were slightly folded, and therefore the text on the scan was compressed. To overcome this issue, we used the page format of the balance sheet and profit and loss statement.

Figure 7 shows a typical example of a balance sheet and profit and loss statement of a medium-sized company. Common among all balancesheet information was that the statements contained the words *Aktiva* und *Passiva*. Similarly, for profit and loss statements contained the words *Debet* and *Kredit*. Moreover, the order of appearance mattered as *Passiva* had to follow *Aktiva*, and *Debet* had to follow *Kredit*. Any balance sheet or profit and loss statement that did not fulfill this structure were manually checked.

7 TRANSFORMATION OF INPUT DATA

The goal of the database is to provide firm and person-level data in a panel structure, enabling econometric analyses. The following section describes a selection of the most fundamental key steps of the *input layer*'s transformation into the *original layer*, provides summary statistics, and highlights difficulties.

The core of the transformation is assigning the lines stored in the *input layer* to individual companies. Given that companies are listed in the HdAG as continuous blocks (see Figure 5), we needed to identify company blocks' starting and ending lines to parse the OCR output into standardized concepts. More precisely, as a new beginning of a company block (or the end of the book) represents the end of the previous block, the challenge is reduced to identifying starting lines correctly. This logic applies to most periodical economic handbooks. However, the implementation depends on individual layouts and structures that vary within and across handbooks. In the case of the HdAG for our observation period, the starting line of a company block is the company's name horizontally centered and printed in a bold and large font relative to the remaining text body. Moreover, the vertical distance between a company name and the first subsequent non-company name line is larger than between two ordinary text body lines. Relying on page coordinate information provided by the recognized bounding boxes' geometries and contents, we generate numeric variables capturing the aforementioned conditions. For instance, for a line to be considered a company name, its bounding box's height must lie between 92 and 190 pixels, and the pixel-to-character ratio must exceed 46.[10]

Aside from company segmentation, the HdAG's content is divided into industries. Each volume is divided into sections summarizing companies corresponding to the (unfortunately

[10] Aside from layout-specific thresholds, the database design allows for an easy and effortless expansion to other sources.

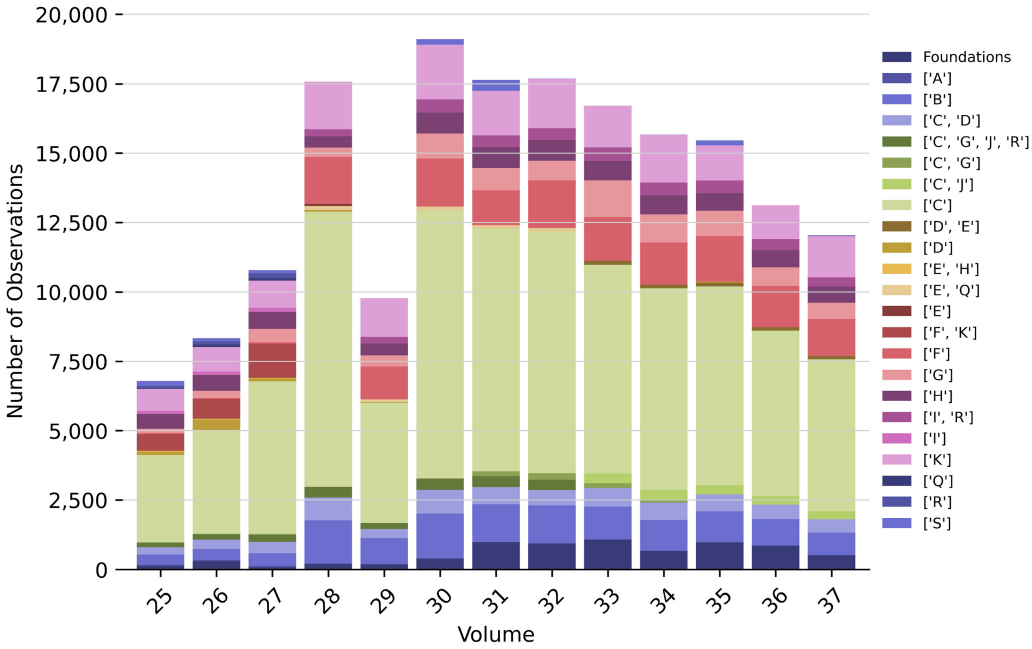


Fig. 8. Distribution of firms per industry and volume.

non-standardized) used industry classification. Thus, it is reasonable to take advantage of this characteristic and assign an *input layer*'s line to not only a company but also an industry early on. A reader can determine the industry of a random page based on two criteria. First, wherever a new industry starts, the name of the industry spans in bold, large letters over the content part of a page. Second, a relatively small textbox at the top of a page indicates the industry to which companies on that page belong. Our industry assignment approach follows the logic applied for company entries: We determine industry ranges by identifying their starting lines. The duality of industry information outlets serves as a validity check. We preselect boxes that potentially represent industries based on manually defined sets of geometric conditions. Then, we apply a string similarity metric to calculate distances between our preselection and elements from a manually compiled list of all industries included in our observation period. This step is crucial in alleviating the inconsistencies stemming not only from OCR errors but also from the non-standardized industry names used in the sources. Finally, we assign the pages to the industries with the lowest string distance.

Company and industry assignment is sensitive to noise resulting from various sources. Most notably, advertisement pages lead to frequent false positive identifications of company and industry starting lines. These pages commonly have a unique layout that is substantially different from other HdAG's content pages. Another source of bias results from company names spanning over multiple lines. Our algorithm accounts for both of these issues and avoids an artificial increase in the number of identified firms.

Implementing these steps leads to an impression on the firm-industry distribution per volume. This distribution is illustrated in Figure 8.[11] Because industry labels vary significantly over time,

[11]Note that newly founded companies are not assigned to industries. Instead, the geographic location that commonly indicates a firm's industry specifies that the company was founded during the previous year, explaining the inclusion of *Foundations* in Figure 8.

Table 2. HdAG Volume—Time Coverage

Volume	Edition	Period				Year
		Start	End	Middle	Duration (Days)	
27	1	09.04.22	30.09.22	05.07.22	174	1922
27	2	01.10.22	15.04.23	07.01.23	196	1923
28	1a	16.04.23	10.12.23	13.08.23	238	1923
28	1b	11.12.23	01.02.24	06.01.24	52	1924
28	2a	02.02.24	20.05.24	27.03.24	108	1924
28	2b	02.02.24	20.05.24	27.03.24	108	1924
29	1a	21.05.24	15.11.24	18.08.24	178	1924
29	1b	21.05.24	15.11.24	18.08.24	178	1924
30	1	16.11.24	05.04.25	25.01.25	140	1925
30	2	06.04.25	10.07.25	23.05.25	95	1925
30	3	11.07.25	15.10.25	28.08.25	96	1925
30	4	16.10.25	25.01.26	05.12.25	101	1925
31	1	26.01.26	15.04.26	06.03.26	79	1926
31	2	16.04.26	10.07.26	28.05.26	85	1926
31	3	11.07.26	15.10.26	28.08.26	96	1926
31	4	16.10.26	17.01.27	01.12.26	93	1926

we harmonize the collected data by mapping the industry names included in the HdAG to sections from the NACE Rev. 2 classification scheme. Sections are the least granular level offered by the NACE classification scheme. It is apparent that Manufacturing (NACE Section C) is the largest industry in terms of company count throughout the observation period. The figure additionally highlights two factors impeding more granularity in industry-based analyses:

(1) Non-standardized labeling of industries included in the HdAG prohibits a full industry classification using current-day classification schemes. For instance, the industry label *Bau-Banken, Bau-, Terrain- und Immobilien- Gesellschaften etc.* encompasses banks that provide loans for the construction of real estate as well as real estate companies. Accordingly, this industry falls within the NACE Sections K (Financial and Insurance Activities) and F (Construction).
(2) Labels and scopes of industries included in the HdAG are time-inconsistent. In the course of our observation period, the previously stated industry changes to *Bau-, Terrain- und Immobilien-Gesellschaften etc*, now excluding financial institutions. In other cases, new industries emerge or dissolve, explaining the dynamics in Figure 8.

Panel data entail a time dimension. The HdAG is organized in annually published volumes divided into various editions. The volume, however, cannot be used as an accurate time variable because it contains entries spanning over multiple calendar years. Aside from this, linking to other databases would be ruled out. The reporting scheme that the HdAG follows changes in our observation period. Until 1925, a volume approximately covers the second half of a given calendar year and the first half of the subsequent calendar year. Afterward, the coverage of a volume roughly corresponds to the calendar year (see Table 2 for volumes 27–31).

We introduce a time variable by assigning data to calendar years depending on the edition's stated reporting window. To be precise, a data point is allocated depending on the reporting window's midpoint year. Figure 9 shows the number of identified firms per volume (left panel) and year (right panel). It is important to note that the change in reporting schemes and the

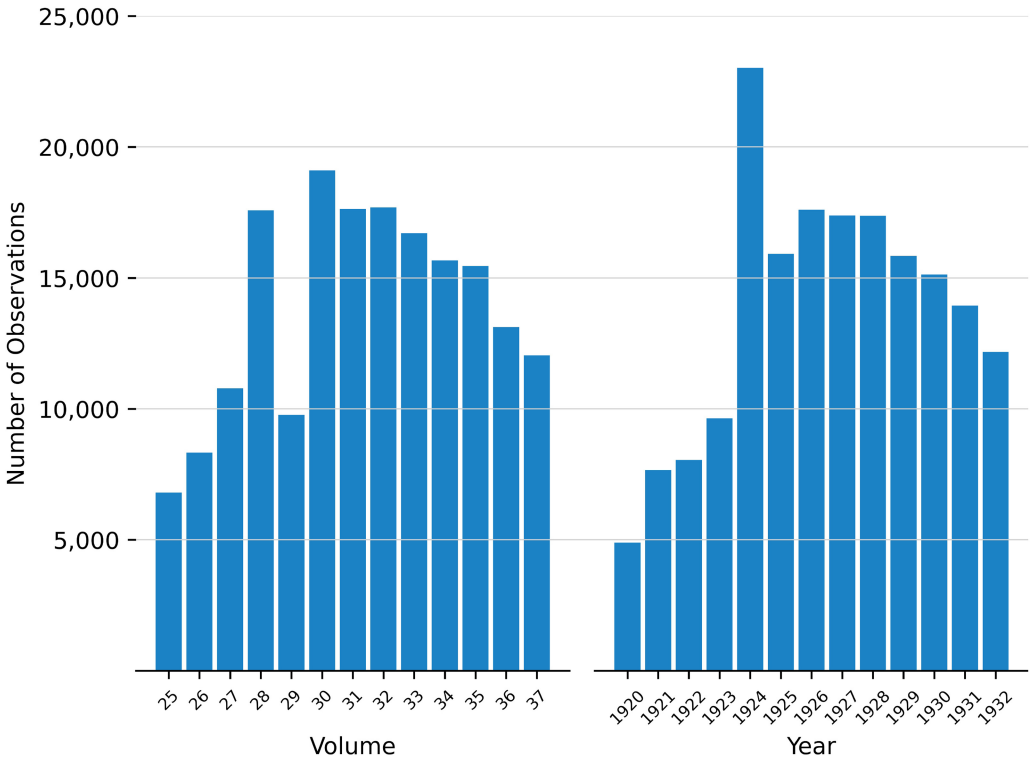


Fig. 9. Identified firms per volume and year.

associated coverage of volume 29 of only half a year causes subsets of firms to either report twice or not at all, explaining—together with the fact that there was massive company formation in the hyperinflation period (1923)—the jump in the right panel.

The following paragraphs sketches additional in-development steps that have not been fully implemented in the database yet. For identified company entries, we isolate a set of pre-defined variables. Among others, the pre-defined variables comprise foundation date(s), the address, balance sheet, profit and loss statement, and members of various corporate boards. Regarding the average firm in our sample, we identify approximately 7.5 variables per year. We additionally parsed, structured, and mapped a substantial fraction of these variables, allowing for econometric analyses. For instance, we observe that the mean Return on Assets (RoA) peaks at 8.5 percent in 1922 and that the mean age of a company in 1931 equals 18.33 years.

Another critical step is the creation of a panel ID needed to link annual cross-sections. By referring to company names and other tendentially time-consistent variables (e.g., foundation date and harmonized industry class), we generate a unique identifier on a firm-level. Depending on the nature of each variable, we either rely on string similarity metrics (e.g., company names) or an exact matching approach (e.g., foundation dates) in this process.

Moreover, as company entries include personal information on board members (i.e., name(s), place of living, occupation, academic title), we create a second unique identifier at a person-level. Based on this, we can assess that the most engaged person in our sample was affiliated with 143 firms in 1925. We further see that, for example, the fraction of board members having a Ph.D. varies from 12.6 percent in 1921 to 19.8 percent in 1932 (see Table 3).

Table 3. Person-level Data Overview

Year	Obs.	Title (%)			Position (%)			Distinct firms			
		Ph.D.	Professor	Nobility	Supervisory board	Management board	Procurator	Max	Mean	Median	Std. Dev.
1921	65156	0.126	0.007	0.022	0.582	0.254	0.163	56	3.445	1	5.706
1922	74853	0.136	0.007	0.022	0.631	0.262	0.107	59	3.730	1	6.271
1923	82634	0.152	0.007	0.021	0.714	0.285	0.002	62	4.276	1	7.103
1924	123354	0.162	0.007	0.020	0.722	0.276	0.002	79	4.525	2	7.785
1925	192575	0.173	0.007	0.021	0.731	0.268	0.001	143	5.234	2	9.994
1926	108326	0.177	0.007	0.020	0.733	0.250	0.017	73	4.300	1	7.508
1927	107551	0.176	0.007	0.019	0.720	0.244	0.036	71	4.066	1	7.036
1928	108372	0.177	0.007	0.018	0.704	0.238	0.058	68	3.928	1	6.911
1929	102656	0.186	0.007	0.018	0.696	0.236	0.067	71	3.950	1	6.988
1930	100915	0.194	0.008	0.017	0.691	0.236	0.073	77	4.013	1	7.267
1931	98065	0.195	0.008	0.017	0.676	0.226	0.098	77	3.750	1	6.797
1932	89422	0.198	0.007	0.017	0.673	0.229	0.099	42	3.110	1	4.658

Lastly, we linked the HdAG data to a second data source that offers daily (January 1921–December 1924) and monthly (January 1925–December 1927) stock market prices of joint-stock companies listed on the Berlin stock exchange. We manually extracted the prices from a newspaper named *Berliner Börsen-Zeitung* which contains the prices of the stocks of all companies listed on the stock exchange of Berlin, at that time the largest of its kind in Germany. The linking again relies on string similarity metrics. A challenge arose from different sets of special characters and varying the usages of abbreviations between the sources. Thus, we first harmonized the spellings before we calculated string similarities. If the linking score exceeds a threshold, we accept matches between the HdAG and *Berliner Börsen-Zeitung*. Additionally, the relatively low number of firms listed on the stock exchange allows for manual verification of recommended matches, which prohibits false positive links.

In total, we linked 1,106 companies from the HdAG to 1,625 stocks series, implying that if a company has shares outstanding for trading, it has 1.47 series listed on the Berlin Stock Exchange on average. It becomes evident that it was common for firms to have preferred aside common stock traded in Berlin in the 1920s. Some firms even had multiple traded series of one kind. In such cases, we incorporated additional information included in the series' names and tracked the stock series' history to improve the linking quality.

8 CONCLUSION

In this article, we discuss the principle of preserving the historical sources as a potential solution to the standardization difficulties that arise from the potential ambiguity that accompanies the data collections from historical sources and hinders the design of extensible data models.

We sketch and realize a relational implementation of the principle, and we examine how this approach could incrementally enhance existing historical databases such as SCOB and DFIH, which have complementary content. 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 detail a proof of our concept for the use case of establishing long-term historical German firm-level data. Our data is extracted from granular historical sources documenting the German companies of the interwar period and are used to populate the relational implementation that we have developed.

Furthermore, we describe the process of parsing the text data and creating variables that correspond to concepts of financial interest. Specifically, we describe the process of parsing and creating a company database, classifying the collected companies into harmonized industries, and linking them with stock market data.

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

Finally, we would like to point out that our database generated is thought to be made available to the research community according to the so-called FAIR principle. Specifically, a two-step procedure is aimed at for accessing. In a first layer, the data will be made (and already is made) accessible in a betaversion to scientists on the application. They can apply for data use in pilot projects and gain access to the data in a simple procedure. This early data access is linked to the obligation to cooperate in improving the data quality. After the finalization of this step, the database will be open access to the general research community.

Lastly, while we have stressed the importance of long-term historical datasets in particular for Europe, we also would like to point out the limitation of such datasets for the analysis of long-term development, not least due to the fact that the number of companies varies a lot over time, typically increasing substantially, making interpolating exercises (e.g., comparing a small number of companies in historical times with many more today) very problematic. Nevertheless, we think that such databases lend themselves—if used with caution and care—to very valuable analyses, which also increases our knowledge of mechanisms that are relevant today.

REFERENCES

Daron Acemoglu and James A. Robinson. 2013. Economics versus politics: Pitfalls of policy advice. *Journal of Economic Perspectives* 27, 2 (2013), 173–92. DOI: https://doi.org/10.1257/jep.27.2.173

Heather M. Anderson, Mardi Dungey, Denise R. Osborn, and Farshid Vahid. 2011. Financial integration and the construction of historical financial data for the euro area. *Economic Modelling* 28, 4 (2011), 1498–1509. DOI: https://doi.org/10.1016/j.econmod.2011.02.027

Jan Annaert, Frans Buelens, and Marc Deloof. 2015. Long-run stock returns: Evidence from belgium 1838–2010. *Cliometrica* 9, 1 (2015), 77–95. DOI: https://doi.org/10.1007/s11698-014-0109-7

Jan Annaert, Frans Buelens, and Angelo Riva. 2016. Financial history databases: Old data, old issues, new insights? In *Proceedings of the Financial Market History*. David Chambers and Elroy Dimson (Eds.), CFA Institute Research Foundation, Charlottesville, VA.

Jan Barton and Gregory Waymire. 2004. Investor protection under unregulated financial reporting. *Journal of Accounting and Economics* 38 (2004), 65–116. DOI: https://doi.org/10.1016/j.jacceco.2004.06.001

Gennaro Bernile, Vineet Bhagwat, and P. Raghavendra Rau. 2017. What doesn't kill you will only make you more risk-loving: Early-life disasters and CEO behavior. *The Journal of Finance* 72, 1 (2017), 167–206. DOI: https://doi.org/10.1111/jofi.12432

Asaf Bernstein, Eric Hughson, and Marc Weidenmier. 2019. Counterparty risk and the establishment of the new york stock exchange clearinghouse. *Journal of Political Economy* 127, 2 (2019), 689–729. DOI: https://doi.org/10.1086/701033

Fabio Braggion and Lyndon Moore. 2011. Dividend policies in an unregulated market: The london stock exchange, 1895–1905. *The Review of Financial Studies* 24, 9 (2011), 2935–2973. DOI: https://doi.org/10.1093/rfs/hhr026

Paolo Coletti and Maurizio Murgia. 2015. Design and construction of a historical financial database of the italian stock market 1973–2011. *Journal of Data and Information Quality* 6, 4 (2015), 1–23. DOI: https://doi.org/10.1145/2822898

Marco Costantino and Paolo Coletti. 2008. *Information Extraction in Finance*. WIT Press, Southampton, UK.

Jon Danielsson, Marcela Valenzuela, and Ilknur Zer. 2018. Learning from history: Volatility and financial crises. *The Review of Financial Studies* 31, 7 (2018), 2774–2805. DOI: https://doi.org/10.1093/rfs/hhy049

Harry DeAngelo and Richard Roll. 2015. How stable are corporate capital structures? *The Journal of Finance* 70, 1 (2015), 373–418. DOI: https://doi.org/10.1111/jofi.12163

Elroy Dimson, Paul Marsh, and Mike Staunton. 2002. Long-run global capital market returns and risk premia. (2002). Retrieved 28 Dec., 2021 from http://papers.ssrn.com/abstract=217849.

Elroy Dimson, Paul Marsh, and Mike Staunton. 2009. *Triumph of the Optimists*. Princeton University Press, Princeton, NJ.

Jens Dittrich and Alekh Jindal. 2011. Towards a one size fits all database architecture. In *Proceedings of the CIDR 2011, 5th Biennial Conference on Innovative Data Systems Research*. 195–198. Retrieved from https://bigdata.uni-saarland.de/publications/DJ11.pdf.

Sebastian Doerr, Stefan Gissler, Jose-Luis Peydro, and Hans-Joachim Voth. 2021. Financial Crises and Political Radicalization: How Failing Banks Paved Hitler's Path to Power. (2021). Retrieved 28 Dec., 2021 from https://www.bis.org/publ/work978.pdf.

Jérémie Ducros, Elisa Grandi, Raphaël Hékimian, Emmanuel Prunaux, Angelo Riva, and Stefano Ungaro. 2018. Collecting and storing historical financial data: The DFIH project. In *Proceedings of the Computational Social Science in the Age of Big Data*. Cathleen Stuetzer, Martin Welker, and Marc Egger (Eds.), Herbert von Halem Verlag, 355–377.

Barry Eichengreen. 2016. Financial history in the wake of the global financial crisis. In *Proceedings of the Financial Market History*. David Chambers and Elroy Dimson (Eds.), CFA Institute Research Foundation, Charlottesville, VA.

Wenzhong Fan. 2004. Construction Methods for the Shanghai Stock Exchange Indexes: 1870–1940. (2004). Retrieved 28 Dec., 2021 from https://som.yale.edu/sites/default/files/2021-12/SSE-CC.pdf.

Thomas Ferguson and Hans-Joachim Voth. 2008. Betting on hitler—the value of political connections in nazi germany. *The Quarterly Journal of Economics* 123, 1 (2008), 101–137.

Thomas Gehrig and Caroline Fohlin. 2006. Trading costs in early securities markets: The case of the berlin stock exchange 1880–1910. *Review of Finance* 10, 4 (2006), 587–612. DOI: https://doi.org/10.1007/s10679-006-9010-y

William N. Goetzmann. 2015. *Bubble Investing: Learning from History*. Technical Report. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w21693.

William N. Goetzmann and Simon Huang. 2018. Momentum in imperial russia. *Journal of Financial Economics* 130, 3 (2018), 579–591. DOI: https://doi.org/10.1016/j.jfineco.2018.07.008

William N. Goetzmann, Roger G. Ibbotson, and Liang Peng. 2001. A new historical database for the NYSE 1815 to 1925: Performance and predictability. *Journal of Financial Markets* 4, 1 (2001), 1–32. DOI: https://doi.org/10.1016/S1386-4181(00)00013-6

Kilian Huber, Volker Lindenthal, and Fabian Waldinger. 2021. Discrimination, managers, and firm performance: Evidence from "aryanizations" in nazi germany. *Journal of Political Economy* 129, 9 (2021), 2455–2503. DOI: https://doi.org/10.1086/714994

Stratos Idreos, Lukas M. Maas, and Mike S. Kester. 2017. Evolutionary Data Systems. arXiv e-prints (2017); Retrieved from https://arxiv.org/abs/1706.05714.

Òscar Jordà, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M. Taylor. 2019. The rate of return on everything, 1870–2015. *The Quarterly Journal of Economics* 134, 3 (2019), 1225–1298.

Òscar Jordà, Moritz Schularick, and Alan M. Taylor. 2017. Macrofinancial history and the new business cycle facts. *NBER Macroeconomics Annual* 31 (2017), 213–263.

Philippe Jorion and William N. Goetzmann. 1999. Global stock markets in the twentieth century. *The Journal of Finance* 54, 3 (1999), 953–980. DOI: https://doi.org/10.1111/0022-1082.00133

Pantelis Karapanagiotis. 2020. *Technical Document on Preliminary Common Data Model*. Technical Report. DOI: https://doi.org/10.5281/zenodo.3686930

Ryan Lampe and Petra Moser. 2016. Patent pools, competition, and innovation-evidence from 20 US industries under the new deal. *The Journal of Law, Economics, and Organization* 32, 1 (2016), 1–36. DOI: https://doi.org/doi.org/10.1093/jleo/ewv014

Stuart E. Madnick, Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. 2009. Overview and framework for data and information quality research. *Journal of Data and Information Quality* 1, 1 (2009), 1–22. DOI: https://doi.org/10.1145/1515693.1516680

Maria Eugénia Mata, José Rodrigues da Costa, and David Justino. 2017. *The Lisbon Stock Exchange in the Twentieth Century*. Imprensa da Universidade de Coimbra/Coimbra University Press. DOI: https://doi.org/10.14195/978-989-26-1303-1

Rajnish Mehra and Edward C. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15, 2 (1985), 145–161. DOI: https://doi.org/10.1016/0304-3932(85)90061-3

Leentje Moortgat, Jan Annaert, and Marc Deloof. 2017. Investor protection, taxation and dividend policy: Long-run evidence, 1838–2012. *Journal of Banking & Finance* 85 (2017), 113–131. DOI: https://doi.org/10.1016/j.jbankfin.2017.08.013

Petra Moser, Alessandra Voena, and Fabian Waldinger. 2014. German jewish émigrés and US invention. *American Economic Review* 104, 10 (2014), 3222–55. DOI: https://doi.org/10.1257/aer.104.10.3222

Rimma V. Nehme, Karen Works, Chuan Lei, Elke A. Rundensteiner, and Elisa Bertino. 2013. Multi-route query processing and optimization. *Journal of Computer and System Sciences* 79, 3 (2013), 312–329. DOI: https://doi.org/10.1016/j.jcss.2012.09.010

Lukas Manuel Ranft, Jefferson Braswell, and Wolfgang König. 2021. *EURHISFIRM D5.5: Report on Process for Extendable Data Models*. Technical Report. DOI: https://doi.org/10.5281/zenodo.4616475

Carmen M. Reinhart and Kenneth S. Rogoff. 2011. From financial crash to debt crisis. *American Economic Review* 101, 5 (2011), 1676–1706. DOI: https://doi.org/10.1257/aer.101.5.1676

Björn Richter, Moritz Schularick, and Paul Wachtel. 2020. When to lean against the wind. *Journal of Money, Credit and Banking* 53, 1 (2020), 5–39. DOI: https://doi.org/10.1111/jmcb.12701

Kristian Rydqvist and Rong Guo. 2021. Performance and development of a thin stock market: The stockholm stock exchange 1912–2017. *Financial History Review* 28, 1 (2021), 26–44. DOI: https://doi.org/10.1017/S0968565020000104

Christian Schlag and Anja Wodrich. 2000. Has There Always Been Underpricing and Long-Run Underperformance? - IPOs in Germany Before World War I. (2000). Retrieved 28 Dec., 2021 from https://www.ifk-cfs.de/fileadmin/downloads/publications/wp/00_12.pdf.

John D. Turner, Qing Ye, and Wenwen Zhan. 2013. Why do firms pay dividends? Evidence from an early and unregulated capital market. *Review of Finance* 17, 5 (2013), 1787–1826. DOI: https://doi.org/10.1093/rof/rfs048

Mika Vaihekoski. 2021. Revisiting Index Methodology for Thinly Traded Stock Market. Case: Helsinki Stock Exchange. (2021). Retrieved 28 Dec., 2021 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3716682.

Jan Verelst. 2004. The influence of the level of abstraction on the evolvability of conceptual models of information systems. In *Proceedings of the 2004 International Symposium on Empirical Software Engineering*. (Redondo Beach, CA). IEEE, 17–26. DOI: https://doi.org/10.1109/ISESE.2004.1334890

Fabian Waldinger. 2016. Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge. *The Review of Economics and Statistics* 98, 5 (2016), 811–831. DOI: https://doi.org/10.1162/REST_a_00565

Zhipu Xie, Weifeng Lv, Linfang Qin, Bowen Du, and Runhe Huang. 2018. An evolvable and transparent data as a service framework for multisource data integration and fusion. *Peer-To-Peer Networking and Applications* 11, 4 (2018), 697–710. DOI: https://doi.org/10.1007/s12083-017-0555-7

Kostas Zoumpatianos, Stratos Idreos, and Themis Palpanas. 2016. ADS: The adaptive data series index. *The VLDB Journal* 25, 6 (2016), 843–866. DOI: https://doi.org/10.1007/s00778-016-0442-5

Received March 2021; accepted February 2022