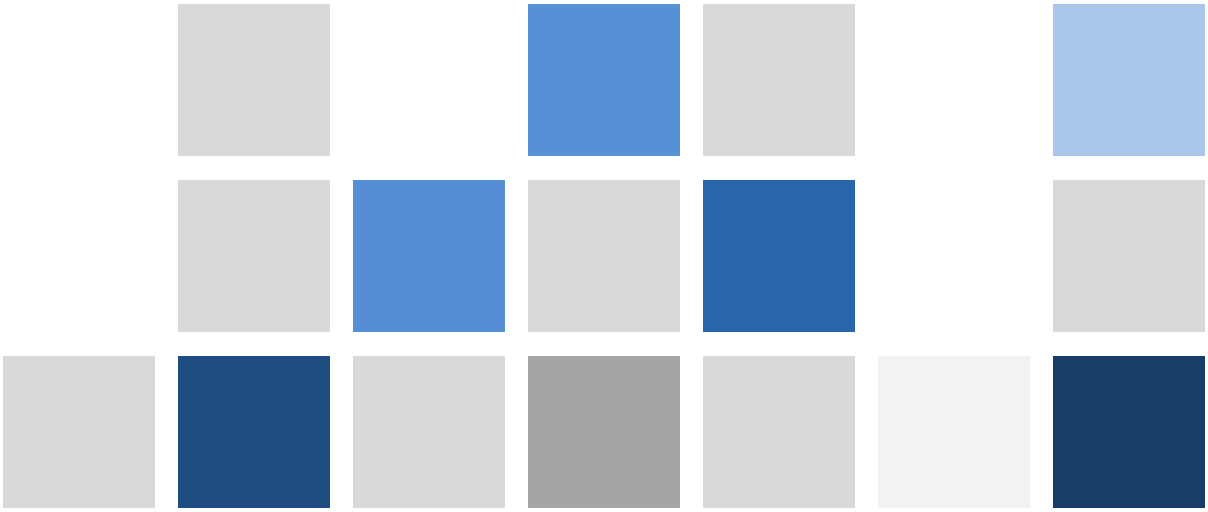
Dennis Gram | Pantelis Karapanagiotis | Jan Krzyzanowski |
Marius Liebald | Uwe Walz

An Extensible Model for Historical Financial Data with an Application to German Company and Stock Market Data

SAFE Working Paper No. 300

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de



Long-term data for Europe

EURHISFIRM

M5.1: Scientific Paper on data models and their extensibility

This project has received funding from
the European Union's Horizon 2020 research and innovation programme
under grant agreement N° 777489

An Extensible Model for Historical Financial Data with an Application to German Company and Stock Market Data

By

Dennis Gram[1], Pantelis Karapanagiotis[2], Jan Krzyzanowski[3], Marius Liebald[4], and Uwe Walz[5]

This version: December 2020

Abstract

Broad, long-term financial and economic datasets are a scarce resource, in particular in the European context. In this paper, we present an approach for an extensible, i.e. adaptable to future changes in technologies and sources, data model that may constitute a basis for digitized and structured long-term, historical datasets. The data model covers 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 company and stock market data, we discuss a relational implementation of this approach.

Classification codes: C81, C82

Acknowledgments: We are thankful to Wolfgang König and Johan Poukens who provided valuable comments on an earlier version of this paper. We are grateful for financial support to the Eurhisfirm consortium which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777489. The paper reflects only the author's view. The EU Commission is not responsible for any use that may be made of the information it contains.

[1] Leibniz Institute LIF-SAFE, Theodor-W.-Adorno Platz 3, 60323 Frankfurt, Germany; GRAM@SAFE-FRANKFURT.DE

[2] Goethe University Frankfurt, Theodor-W.-Adorno Platz 3, 60323 Frankfurt, Germany; PANTELIS.KARAPANAGIOTIS@HOF.UNI-FRANKFURT.DE

[3] Goethe University Frankfurt, Theodor-W.-Adorno Platz 4, 60323 Frankfurt, Germany; KRZYZANOWSKI@ECON.UNI-FRANKFURT.DE

[4] Goethe University Frankfurt, Theodor-W.-Adorno Platz 4, 60323 Frankfurt, Germany; LIEBALD@ECON.UNI-FRANKFURT.DE

[5] Goethe University Frankfurt and Leibniz Institute LIF-SAFE, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany; UWALZ@ECON.UNI-FRANKFURT.DE

1. Introduction

High-quality data is one if not the single most important input in empirical research in finance and economics. In the last decades, we have seen tremendous growth in structured data on firms, households, markets, at the micro- as well as the macro level. While data covering the US is easily accessible (see e.g. the CRSP and Compustat Databases starting 1925 and covering accounting as well as security prices), long-run structured and consistent databases for Europe are scarce (an exception covering broader European firm level data from 1980 onwards is the EUROFIDAI data. Base). This has two immediate consequences. First, the majority of empirical studies are based on US data. Transferring the conclusions of these studies to other legal systems or structures, such as those of continental European societies, often leads to potentially serious misconceptions given fundamentally different institutions and political systems (cf e.g. Acemoglu and Robinson (2013)). Second, not least in the aftermath of the financial crisis, it has been recognized that data with a short time horizon may focus on quite specific macroeconomic settings and is associated with overlooking long-term relations. For example, investigating only data for the "Great-Moderation" period overlooks structural changes that have occurred beforehand and afterward.

Hence, there is a need to build databases with financial data for Europe spanning over decades and centuries. Analyzing 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., Anderson et al. (2011)). It is essential for research to access datasets of this kind digitally.

However, the construction of long-term structured databases, comprised of historical datasets (in particular pre-WWII datasets), is a tremendous challenge. This paper highlights problematic aspects of such datasets' constructions and discusses potential solutions. Firstly, historical data coming from various sources that overlap with regard to the entities covered in a given time period can be ambiguous in the sense that either information can be incomplete in some sources or even information originating from different sources can be conflicting requiring a decision which source is more adequate. Secondly, 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, allowing us to overcome this problem. 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). Thirdly, studies in the fields of financial economics often rely on a mix of quantitative and qualitative data (see for this distinction, e.g. Costantino and Coletti (2008)). 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, analyzing them jointly is complicated.

The vast majority of historical data is stored in ooks, journals, loose-leaf editions, etc. Creating digital and structured datasets from these sources by hand requires huge resources. Additionally, this approach is not associated with a clear-cut cost-benefit trade-off. The evolution of machine learning over the last decade,

especially in the field of deep learning, together with quickly growing computing capabilities allow to digitize and structure historical sources in a fast and cost-efficient way. This is, in particular, the case if the underlying technologies exploit significant economies of scale. Having proper technologies in place is, however, only one requirement to digitize and structure historical data on a large scale. 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. Data models of databases covering more contemporary periods prove that this idea's importance is well understood. Databases including historic information, on the other hand, have found fewer applications of this idea in their data models. One reason is that most such databases were constructed in the scope of particular projects by individual researchers, who were not in coordination with each other and did not use common standards. A deeper reason comes from the nature of the sources themselves. 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. In most cases, one can deterministically assess whether and how a data transformation from one data format to another one is feasible. Such invariance points are missing from historical data because its sources were documented at times in which the standardization and identification theory was underdeveloped. Trying to approach historical data from the modern perspective is anachronistic. Although there is no invariance of standards in historical data, there is an invariance of sources. In contrast with contemporary data in which the sources are renewed, there are only a fixed finite number of input sources for historical data. This invariance can be used as a focal point when designing an extensible infrastructure.

Against this background, we outline a base for an extensible data model and a process to digitize and structure large historical datasets. We take into account that data collection often did not use pre-defined standards and that the ex-post harmonization, standardization, and verification of data cannot typically take place without having access to the original sources. This is not only true concerning national-level data but even more apparent regarding data from different countries. The proposed implementation aims to cope with the particular features of historical data allowing researchers to check the original data source. Building on this idea, we further elaborate on our project to digitize 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 organized as follows. In the next section, we review existing literature on the creation and build-up of historical financial datasets. We discuss research that is based on larger historical datasets, in particular on more aggregated data as well as on company data and stock market data. In the third section, the specific role of the original information leading to digitized historical databases is discussed and the principle of the preservation of the original historical source is proposed. The fourth section discusses the idea of a relational implementation of this principle while section five focuses on potential alternatives. The sixth and seventh sections discuss the collection of German company data that can populate this implementation. Specifically, section six delineates 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, the following section contains a review of reports on data collection projects as well as on literature on extensible data models. Furthermore, we provide insights into selected fields of research that are related to extensible data model structures as well as historical data in the area of economics and finance. We start with the former and then review existing data projects focusing on long term data in the area of financial economics. More precisely, we provide an introductory overview on four related fields, namely i) studies focusing on the collection process of historical datasets ii) research using long-term, historical and aggregated data in the context of economic and financial analyses, iii) papers employing long-term data micro-data on companies in economic and financial research as well all iv) studies which are based on long-term data in the area of stock market analyses. This sets the stage for our own project which is then outlined in subsequent sections.

2a Extensible data models

As extensibility represents a key feature of our data model, the following section aims at providing selected insights on the strand of literature that is related to the topic of extensible data model architectures. Xie, Lv, Qin, Du & Huang (2018) 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 occur in the context of the continuous integration of data from numerous sources. As Idreos, Maas & Kester (2017) point out, these challenges include the necessity to deal with the multitude of options that are based on the underlying system architectures. Furthermore, as the process of data system designing and tuning is complex, cumbersome, expensive, and there is no single data system architecture that fits the continuously increasing kinds of data-driven scenarios, these authors provide a vision for an evolutionary data system. They argue that this system represents a new way to think about architectures of data systems as all choices regarding the design may be made through an evolutionary process that does not require human intervention. With their prototype, they provide solutions that are able to interchange between a key-value store and a column-store architecture, thereby allowing them to adapt to changing workloads.

A comprehensive discussion of the features of various processes, formats, and systems, as well as their contribution to extensibility, is given in Ranft, Braswell, & König (2020). Instead, the present article narrows the discussion of extensibility in the context of system design over historical data sources and provides a concise exemplifying implementation using collected historical data from German sources. It introduces a design principle that is independent of the used technology and promotes the extensibility of historical, firm-level data models by focusing on the separation between sources and concepts instead of technologies and formats.

Further related literature points to other solutions in order to overcome the problem of static and data-independent decisions. These ideas on adaptive data system architectures aim at increasing performance through optimization of query plans that are based on the distinct properties of the included data. From this background, Nehme, Works, Lei, Rundensteiner & Bertino (2013) provide a modern data stream processing

approach which in its core aims at computing multiple data query routes that are individually designed for particular subsets of the data with distinct statistical properties. In another paper, Zoumpatianos, Idreos & Palpanas (2016) follow the approach to adaptively build auxiliary data structures that are required as big amounts of data series are continuously produced and need to be available for queries as soon as possible. As the authors argue, this is not possible with existing indexing methods for very big data series collections. Based on these considerations, the authors discuss the first adaptive indexing mechanism, which aims at solving the problem of indexing and querying in the context of very large and continuously evolving data series collections. The authors provide insights into the design and usage of query algorithms based on synthetic and real datasets. Finally, another strand of the literature argues that data should be kept in a flexible format structure that is embedded in a single system, allowing for taking multiple views on the data. From this background, Dittrich & Jindal (2011) describe their concept of storage views, which constitute secondary, alternative physical data representations that cover all or subsets of the primary log and which allow utilizing different storage views for different subsets of the data. Based on their considerations, the resulting database is capable to incorporate different types of systems that are based on row stores, column stores, or hybrid combinations, which would not be able to be achieved within traditional data bank management systems.

Abstraction and genericity are attributes that are commonly proposed in computer science literature as being central ingredients in the extensibility of systems. With this justification, they are employed in practice when designing programming languages, data models, and overarching systems. On many occasions, however, the necessity of introducing abstractions is overestimated, especially by novice designers, leading to more complicated designs and eventually hindering extensibility. Using an experimental design, Verelst (2005) provides statistical evidence of this industry wisdom, namely that abstractions are not always in favor of the extensibility of conceptual models. The author shows that there are two cases, which we find to be applicable in the context of historical data, in which the introduction of abstractions contributes to extensibility. Firstly, abstractions reduce the time needed to implement modifications in a conceptual model in cases of introducing complicated changes. In contrast, if modifications are simple, abstractions might increase the time needed to introduce changes. Taking into account the complexity of financial data, let alone of data with a historical depth, an abstraction separating sources from concepts can be proven to be helpful in the design of a firm-level data model. Secondly, the author also presents evidence that abstractions can be beneficial when it comes to the correctness of introducing changes to the conceptual model. This constitutes another feature of designing models with historical data, as the introduction of new historical sources might introduce needs for conceptual distinctions that were not previously conceived and, therefore, included in the data model.

The subjects of the experiment of Verelst (2005) are students, with the author suggesting that this isolates the statistical results from confounding effects regarding expertise. However, this might practically result in an overestimation of the effect that is of interest in our case, as the effect of the interaction between model complexity and abstraction on the time needed to perform extensions and mutations of a data model might be significantly smaller in a group of system designers, which is the target audience of our proposed system.

2b Processes of generating historical data

Generating historical datasets is a challenging underpinning that may be associated with severe potential deficiencies. Annaert, Buelens & Riva (2016) discuss some general potential flaws of 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, in particular with respect to 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 Jorion and Goetzmann (1999), Dimson, Marsh, and Staunton (2002, 2009) and Global Financial Data (2005). 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 the Antwerp University, most of the research is based on American financial micro-databases out of which the most widely used one is produced by the Centre for Research and Security Prices (CRSP) which is a platform managed by the University of Chicago. Recently, pursuant to Annaert, Buelens & Riva (2016), more research projects are aiming 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), as well as the collection of UK markets 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, da Costa, and Justino (2017) and Vaihekoski (2020)), . The DFIH initiative is targeting to build up a comprehensive long-run stock exchange database on the French markets from 1796 to 1976 (cf. Ducros & Grandi (2018)). The resulting database includes stock market information that was issued in Paris Bourse and covers all financial instruments, shares, and bonds which were issued by French and foreign governments and corporations and traded on the Paris Bourse. For the population of the DFIH database two main printed serial sources are utilized: the official lists of the Paris Stock Exchange publishing the information about the traded assets, and official and private stock exchange yearbooks publishing information about issuers. According to the online description of the framework of this historical database two data capture technologies were set up that both required the scanning of the printed sources. The first is characterized by manual data entry and the second by semi-automatic processing of the stock exchange yearbooks utilizing optical character recognition and artificial intelligence[6] .

Karapanagiotis (2019) points out that advanced database implementations 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 provides thorough discussions on functional requirements that relate to a proper identifier design, adequate documentation as well as quality assurance, and label validation. Furthermore, informational requirements in order to identify different classes of economic entities (natural persons, legal entities of official nature, and legal entities of non-official nature) as well as standards and governance are discussed. The author also points out the necessity for the

[6] See for more details: (HTTPS://WWW.PARISSCHOOLOFECONOMICS.EU/EN/ABOUT/HISTORY-OF-THE-PARIS-SCHOOL-OF-ECONOMICS/EQUIPEX-D-FIH-DONNEES-FINANCIERES-HISTORIQUES/SOURCES-AND-ORGANIZATION).

harmonization of data at a European level by arguing that the impact of research using data from each European country separately is lacking in comparison with that of using US data and that harmonization provides the opportunity to study relevant historical questions from a broader European perspective.

Finally, Rydquist & Guo (2020) utilize a newly constructed dataset of daily transaction prices and volume data from the Stockholm Stock Exchange for the period ranging from 1912 to 1978. In this context, the authors describe general methodological issues of missing values in historical portions of data (see also Vaihekoski (2020)). They point out that many papers on historical stock markets do not correct stock prices for capital operations effects as this information is not available. Furthermore, the authors state that their historical data contain complete information on capital operations from 1912 onwards due to the requirements of the Swedish Company law from 1910 according to which every capital operation must be filed with a government agency. The National Library of Sweden stores a hard copy of the Official Quotation List, which was scanned and coded afterward by the authors. Their newly generated database does not cover the entire history of the Stockholm Stock Exchange from 1901 onwards, due to the fact that before 1912 stock prices were reported in numerous newspapers on an ad hoc basis and the corresponding coverage depended on the editor of the newspaper.

2c Analysis of aggregated historical economic data

While long-term historical datasets that are available to the general public are scarce, in particular in the European context, there are numerous studies that rely on specific, often hand-collected data. We explore some of them, by focusing on studies with aggregated data first.

Starting with the seminal paper of Mehra and Prescott (1985) a number of researchers have investigated the behavior of returns in financial markets. While the initial research was heavily focused on the US (and partially the UK), the dataset constructed by Dimson, Marsh & Staunton (2002, 2009) which contains equity returns for 21 of the worlds' stock markets from 1900 through 2014 allowed an extension to further countries and a more wide-spread view. This broadening of the dataset allows for a more comprehensive view and the rejection of the hypothesis that the US returns were the consequence of a century-long phase of luck for the investors in US equity. This aggregate multi-country dataset has been used by many researchers in various directions. For instance, Goetzmann (2016) provides empirical evidence on historical price increases in global markets. He examines the frequency of sudden and large increases in market value for a broad panel of world equity market data from 1900 to 2014. Based on the contained information on indices of total real returns on equity, he is able to show that the overwhelming proportion of price increases in global markets were not followed by crashes and that the probability of a crash conditional on a boom is only slightly higher than the unconditional probability.

In a related paper, Danielsson, Valenzuela & Zer (2016) analyze the effects of volatility on financial crises by constructing a cross-country database based on Global Financial Data on historical volatilities and collecting monthly returns of stock market indices from which annual volatility estimates are obtained. The generated dataset covers 60 countries and up to 211 years from 1800 to 2010. However, only for the US such a long-term range (starting in 1800) is available, while for the other countries the time horizon is (significantly) shorter. Based on this, the authors find evidence that the level of volatility is historically not

a good indicator of a crisis but that relatively high or low volatility is. More precisely, low volatility increases the probability of a banking crisis, while high as well as low volatility have an impact on stock market crises and do not contribute to explaining currency crises. Along similar lines, Reinhart & Rogoff (2011) exploit a new long-term historical database in order to study debt and banking crises, inflation as well as currency crashes. Their data cover 70 countries stemming from Africa, Asia, Europe, Latin, and North America as well as Oceania and the time frame of their analyses covers over two centuries, ranging back to the date of independence or – for some countries – well into the colonial period. Based on this setup, the authors find that banking crises are preceded by rapidly rising private indebtedness. Furthermore, their analyses indicate that banking crises increase the likelihood of a sovereign default.

By building on the Macrohistory Database (see Jordà, Schularick & Taylor, 2017), Richter, Schularick & Wachtel (2018) 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. The authors point out that in modern economic history about one-quarter of credit booms are followed by systematic banking crises. The authors show that there are distinctive economic features of some credit booms which make them more likely to end in a crisis compared to other booms. More precisely, booms accompanied by house price booms and a rising loan-to-deposit-ratio are shown by the authors to be much more likely to end in a systemic banking crisis. Employing the same dataset, Jordà et al (2019) research the rate of return on everything including bonds, equities as well as real estate. Quite surprisingly they find that in the long-term the return on real estate equals on average the return on equity, with real estate returns having, however, a lower variance (see, however, Eichholtz, Korevaar, Lindenthal, & Tallec (2020) for substantially lower numbers for the same time period for historical Paris and Amsterdam micro data).

Another strand of the literature relies on regional data from the US. Calomiris, Mason, Weidenmier & Bobroff (2013) investigate the efforts of the US Reconstruction Finance Corporation (RFC) on bank failure rates in Michigan during the period between 1932-1934, which includes the Michigan banking crisis of early 1933 and its aftermath. For this purpose, the authors utilize a new dataset on Michigan banks by collecting detailed data on characteristics of Michigan banks, as well as characteristics of counties in which these banks operated. Bank financial data stem from the Federal Reserve member bank's Reports of Condition and Income. The final dataset includes county-specific data on numerous aspects of the economic environment. In another paper with a regional banking focus, Richardson (2011) analyzes how New York banks reacted to the financial crisis beginning in 1931 in Australia that subsequently struck Europe and finally migrated across the Atlantic. For this purpose, the authors made use of data from numerous sources which included memos of conversations between leading bankers as well as the governors of the New York Federal Reserve, articles written by commentators, and financial data.

2d Microeconomic, company-level historical data

In order to put our own project into perspective, we take a short tour d'horizon over historical studies using microeconomic data, in particular company (-related) data.

Moser, Voena & Waldinger (2014) perform a systematic analysis of the Jewish immigrants' effects on US innovation. Historical accounts suggest that these immigrants revolutionized US science and innovation,

while supporting empirical evidence has been scarce. For this purpose, the authors collected new datasets to measure aggregate changes in the US patenting across research fields and to investigate changes in research output at the level of individual US inventors between 1920 and 1970. Besides this, they utilized information on changes in patenting for individual US inventors across research fields with varying levels of exposure to the arrival of the German Jewish immigrants. The paper finds that patenting by US inventors increased substantially in émigré fields. Furthermore, inventor-level data indicate that immigrants encouraged innovation by attracting new researchers to their fields, rather than by increasing the productivity of incumbent inventors. Waldinger (2016) complements these analyses by examining the role of human and physical capital for the creation of scientific knowledge. For this purpose, he constructs a new panel dataset of physicists, chemists, and mathematicians at German and Austrian universities. The results indicate that the shock to human capital reduced output in the short run and had persistent effects in the long run. Regarding physical capital, the shock had a negative effect on output in the short run, while not being persistent in the long run. Lampe and Moser (2016) examine changes in patenting after the creation of a patent pool, by collecting a new dataset of 75,396 issued patents with information on application years which cover patents filed between 1921 and 1948 for 20 industries that were affected by the creation of a pool. A patent pool constitutes an arrangement by which two or more patent owners put their patents together while receiving a license to use them in return (Vaughan 1956). The empirical results across the 20 industries indicate a 14% decline in patenting for each additional patent that is included in a pool.

An example of a study on the financing side of companies is Barton and Waymire (2004) who investigate the relationship between financial reporting and investor protection preceding the market crash in 1929. They ask whether higher quality financial data reduce the losses of investors during the period of a stock market crash. For this purpose, the authors identify all firms listed in the monthly database of the CRSP that are traded on the New York Stock Exchange in October 1929, which results after some adjustments in a sample of 540 firms. Based on monthly stock returns data from CRSP, share trading volume data from the NYSE's Monthly and Yearly Record, and financial reporting data from Moody's Investment Manual, the authors generate a measure on reporting quality based on factor analysis on four determinants: the transparency of the income statement, the transparency of the balance sheet, the use of external auditors, and conservative accounting based on the values of nominal intangible assets. Based on this setup, the authors find that managers appear to have incentives to report higher quality financial information in the absence of mandatory regulatory requirements (see also Van Overfelt, Deloof, & Vanstraelen (2010)) and that such reporting is beneficial for investor protection.

2e Historical data and stock market analysis

Finally, this last subsection of the literature review discusses microeconomic analyses in the context of historical data that is related to the stock market. We start with some analyses of European markets which are more scarce, before referring to studies on the US markets.

Annaert, Buelens & Deloof (2015) analyze 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 which are taken from the SCOB database introduced in a previous section of the literature review. This database includes 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. Amongst others, the authors find that stock returns strongly depend on dividend income and that stocks were less risky in the nineteenth century compared to the twentieth century. In another analysis by Moortgat, Annaert & Deloof (2017), it was investigated whether investor protection and taxation regulation had an impact on dividend policy by utilizing the sample of listed Belgian firms between 1838 and 2012. Surprisingly, the authors find that the dividend policy was stable over time, thereby indicating that legislation on tax as well as investor protection changes tended to have rather little influence on firms' dividend policies. Braggion & Moore (2011) take a British perspective and analyze the effects of dividend policies of 475 British firms existing between 1895 and 1905 in an unregulated low tax regime. The authors find strong support that some dividend policies are used as signals but little support for agency models. In the same vein, the paper by Turner, Ye & Zhan (2013) utilizes a hand-collected dataset of firms on the London stock market between 1825 and 1870 and analyzes the underlying rationale for firms to pay dividends. The main finding of the paper is that dividends constituted a valuable tool for investors in order to communicate information about cash flows during a time where financial reporting and regulation were still in their infancy. Relying on data for the German stock market before World War I, Schlag and Wodrich (2000) seek to find empirical evidence on initial public offerings by investigating the pricing and long-run performance of IPOs. Their findings indicate that underpricing of IPOs has existed, but has significantly decreased over time in their sample. Using similar data Gehrig and Fohlin (2006) estimate effective spreads of securities traded at the Berlin Stock Exchange in 1880, 1890, 1900, and 1910. They find surprisingly tight effective spreads for the historical data, pointing to the rather pronounced efficiency of historical capital markets.

Turning to the US, Bernstein, Hughson, and Weidenmier (2019) utilize the establishment of a clearinghouse on the New York Stock Exchange in 1892 in order to analyze the impact of centralized clearing on counterparty risk. For this purpose, the authors collect 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. They find that the introduction of settlement through a clearinghouse substantially reduced volatility of NYSE returns which were caused by settlement risk and increased asset values. Thereby, these results indicate that a clearinghouse can improve market stability by reducing network contagion and counterparty risk. In another paper by Goetzmann, Ibbotson & Peng (2001), the authors estimate the power of past returns and dividend yields to forecast future long-horizon returns based on individual stock price data for New York Stock Exchange stocks from 1815 to 1925 and individual dividend data ranging from 1825 to 1870. The authors utilize information on more than 600 individual securities that results in a broad sample of NYSE stocks for the 19th century. Based on this dataset, the authors find some predictability of future returns in sub-periods but little predictability over the long term.

During recent years, substantial efforts have been made to reconstruct also indices from other countries. Regarding Chinese data, Fan (2010) describes the collecting process of individual stock data for the period from 1871 to 1940 from the North China-Herald, which was a local English newspaper that was published

in Shanghai and which contained a comprehensive collection of weekly share lists. Furthermore, Goetzmann and Huang (2015) utilize a dataset of hand-collected end-of-month stock prices of all companies listed on the St. Petersburg Stock Exchange from 1865 to 1914. For this purpose, information from five different sources, which contained the respective information in different time frames, are collected and made publicly available on the website of the International Center for Finance at the Yale School of Management.

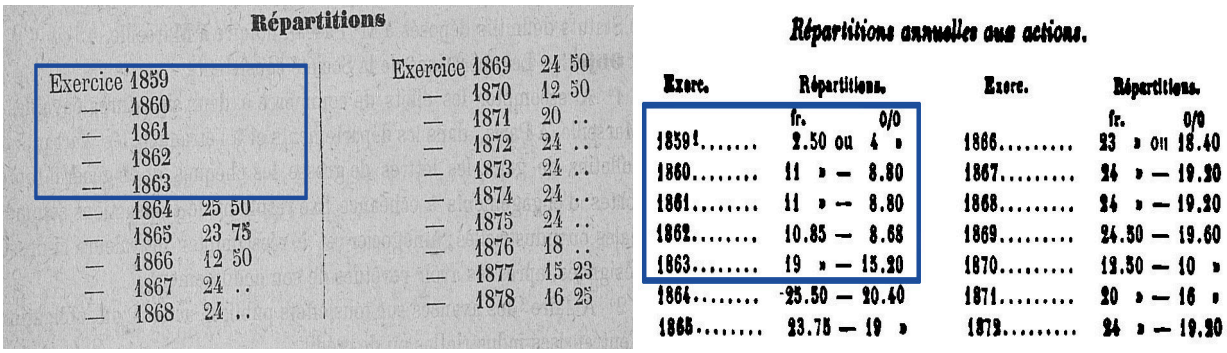
In summary - as Eichengreen (2016) points out - it can be said that research in financial history has enjoyed a renaissance in recent years, due to lower costs and greater efficiency regarding the extraction and digitalization of historical datasets. The existing datasets while developing quickly are, however, still fragmented, with little coherence and interlinkages. Establishing 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 (e.g. stock market and company data) but also refer to different jurisdictions (e.g. leading to data comparability but also the depiction of the global activities of economic agents) and different periods (even allowing to merge historical data with modern ones). In the next sections, we propose one step towards this direction, in general, but also in the way this could be implemented in a scalable manner.

3. The role of original information in historical databases

The set-up of a proper, high-quality, historical database that is flexible with regard to new additions of data sources requires a proper data model. This in turn has to be built with a proper understanding of the data input sources, their limitations, and the peculiarities that they come with.

One central issue 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 article, 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 $\mathcal{O}$ and obtaining the necessary information by investigating them, 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 researchers dealing 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 the ideal objects of $\mathcal{O}$ are conflicting.



Répartitions

Exercice	1859		Exercice	1869	24 50
—	1860		—	1870	12 50
—	1861		—	1871	20 ..
—	1862		—	1872	24 ..
—	1863		—	1873	24 ..
—	1864	25 50	—	1874	24 ..
—	1865	23 75	—	1875	24 ..
—	1866	12 50	—	1876	18 ..
—	1867	24 ..	—	1877	15 23
—	1868	24 ..	—	1878	16 25

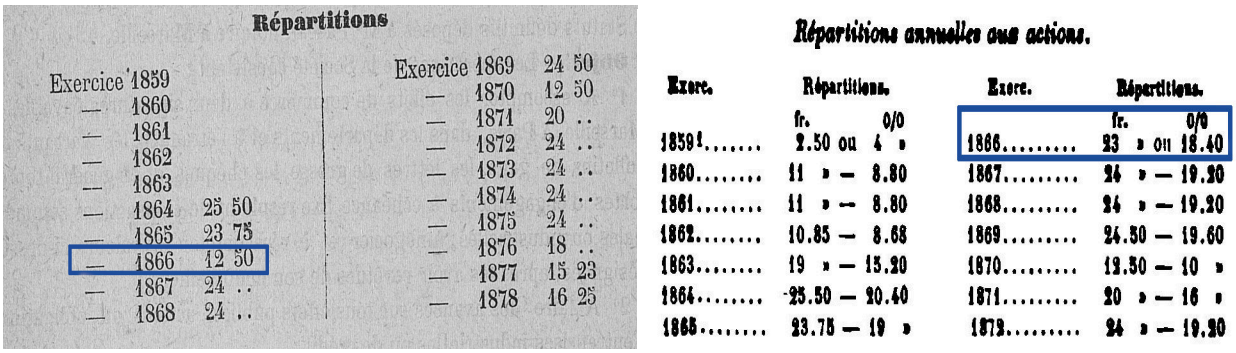
Répartitions annuelles aux actions.

Exerc.	Répartitions. fr.	0/0	Exerc.	Répartitions. fr.	0/0
1859!.....	2.50 ou	4 »	1866.....	23 » ou	18.40
1860.....	11 » —	8.80	1867.....	24 » —	19.20
1861.....	11 » —	8.80	1868.....	24 » —	19.20
1862.....	10.85 —	8.68	1869.....	24.50 —	19.60
1863.....	19 » —	15.20	1870.....	12.50 —	10 »
1864.....	25.50 —	20.40	1871.....	20 » —	16 »
1865.....	23.75 —	19 »	1872.....	24 » —	19.20

FIGURE 1. AN EXAMPLE OF MISSING INFORMATION. LEFT: COMPAGNIE DES AGENTS DE CHANGE YEARBOOK FROM 1880. RIGHT: COURTOIS YEARBOOK FROM 1874.

Figure 1 gives an example of missing information by presenting two snippets of different historical printed sources, both reporting 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 snippet from the "Courtois" yearbook from 1874. The records overlap for the years 1859 to 1872, they do not, however, contain the same information. In particular, dividend payments are not reported for the years 1859 to 1863 in the left snippet, whereas records exist in the right snippet.

From a data-model design perspective, the example of missing information in Figure 1 is relatively innocuous, because both sources dictate including 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 into the implementation. The situation becomes more complicated in the case of conflicting information.



Répartitions

Exercice	1859		Exercice	1869	24 50
—	1860		—	1870	12 50
—	1861		—	1871	20 ..
—	1862		—	1872	24 ..
—	1863		—	1873	24 ..
—	1864	25 50	—	1874	24 ..
—	1865	23 75	—	1875	24 ..
—	1866	12 50	—	1876	18 ..
—	1867	24 ..	—	1877	15 23
—	1868	24 ..	—	1878	16 25

Répartitions annuelles aux actions.

Exerc.	Répartitions. fr.	0/0	Exerc.	Répartitions. fr.	0/0
1859!.....	2.50 ou	4 »	1866.....	23 » ou	18.40
1860.....	11 » —	8.80	1867.....	24 » —	19.20
1861.....	11 » —	8.80	1868.....	24 » —	19.20
1862.....	10.85 —	8.68	1869.....	24.50 —	19.60
1863.....	19 » —	15.20	1870.....	12.50 —	10 »
1864.....	25.50 —	20.40	1871.....	20 » —	16 »
1865.....	23.75 —	19 »	1872.....	24 » —	19.20

FIGURE 2. AN EXAMPLE OF CONFLICTING INFORMATION. LEFT: COMPAGNIE DES AGENTS DE CHANGE YEARBOOK FROM 1880. RIGHT: COURTOIS YEARBOOK FROM 1874.

Figure 2 highlights a case of conflicting information using the same snippets that were used in Figure 1. In this case, the left snippet of Figure 2 suggests that the dividend paid in 1866 was 12.50 francs, while the right snippet of Figure 2 records that the dividend was 23 francs. From the scale of paid dividends of the surrounding years, one may be tempted to suggest that the "Courtois" yearbook records the correct dividend. With a more thorough examination, however, one observes that both sources agree that the paid

dividend was 12.50 francs in 1870, suggesting that a dividend of the same level is plausible to represent the real dividend value of 1866. It is, therefore, impossible to establish the actual historical value of 1866 without having any additional information regarding the dividends of the company for this year.

From a data-model perspective, this ambiguity poses a serious challenge to the typical modelling 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 is a hack that in some ways undermines the purpose of standardization. From an end-user perspective, a query that returns multiple values, although it can be well-defined in terms of a standard that allows more than one dividend values at a given time-point, can be potentially confusing and unexpected.

Another possibility to approach this problem from a standardization perspective is to assign weights to each dividend value. Besides increasing the cardinality to allow for more than one record, the recorded entries can be pairs of monetary values coupled with probabilities that signify their potential correctness. This approach will deliver expected, in the sense of compatibility with the standard, result sets that can also be reasonably comprehended by end-users. The difficulty concerning the implementation of this solution lies with determining the probabilities that accompany the dividend values, as it requires the innervation of experts that assign probabilities to each case. A uniform distribution of weights in all cases does not have any value-added 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 disregarding the archive space becomes impossible if one wants to provide users with accurate information content. Even in the case of assigning probabilities to values of various records, since these probabilities are manually established by historical expertise, there is a strong possibility that some end users would like to deviate and use weights that are based on their expertise. Any solution regarding the extensibility of the data model, therefore, 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 principle of preserving the historical sources

The design of this article is proposed, and is to be understood, in the context of a system, with multiple layers, that is powered by potentially multiple database technologies. Since the preservation principle that constitutes the basis of the design concerns the description of the historical sources and their association with data-items, such as company names, headquarter addresses, etc, the analysis here focuses on this part of the system only. Our working hypothesis is that the data digitization is performed by (semi-) automated optical character recognition tools. We adopt this assumption, although it is not necessary for the analysis and the proposed design can be easily adapted to systems with manual input, because our interest lies in the design of big-data systems that are unlikely to be viable if not based on automated input.

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, thereby, this layer is responsible for storing and associating the system's input data. The *input layer* does not handle concepts

such as companies, financial instruments, etc; this is the responsibility of other system layers that are built on top, using the *input layer* as a basis. Instead, the *input layer* represents a low-level abstraction that handles the sources and isolates them from the system's higher layers, which are susceptible to change through time as new technologies enable new representations or new sources are used.

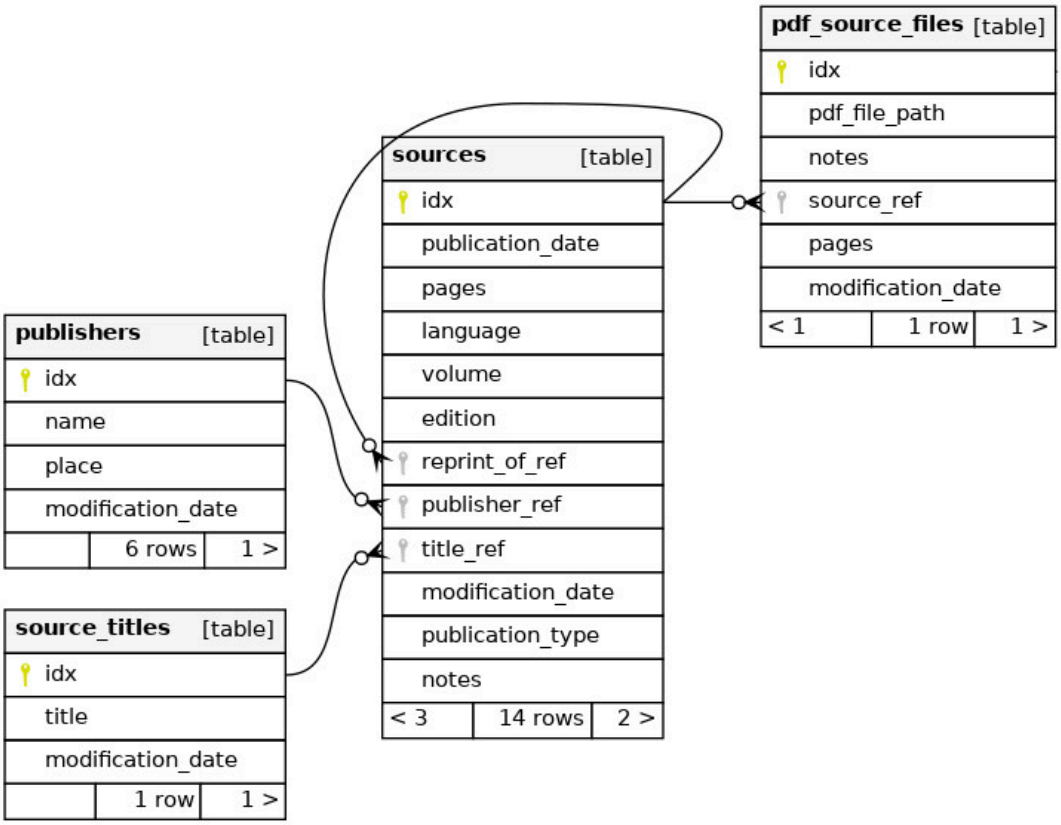


FIGURE 3. PUBLISHING INFORMATION RELATIONSHIPS

In terms of our concrete working hypothesis, the *input layer* is used to store information directly from the optical character recognition output. In our exemplifying relational implementation, the first part of which is presented in Figure 3, the *input layer* organizes information starting from the publisher. The same publishing house may have published multiple historical archives of interest, and in such a case, all of these sources are linked to the same publisher data-item. In order to cover cases in which a source is part of a publication series, source titles should be stored separately. Historical sources are also associated with digitized files, which are stored in their entirety. Each source can be associated with multiple digitized files, a point which becomes important when optical character recognition algorithms are used because differences in the quality of digitized files of the same source can lead to significantly different recognition results.

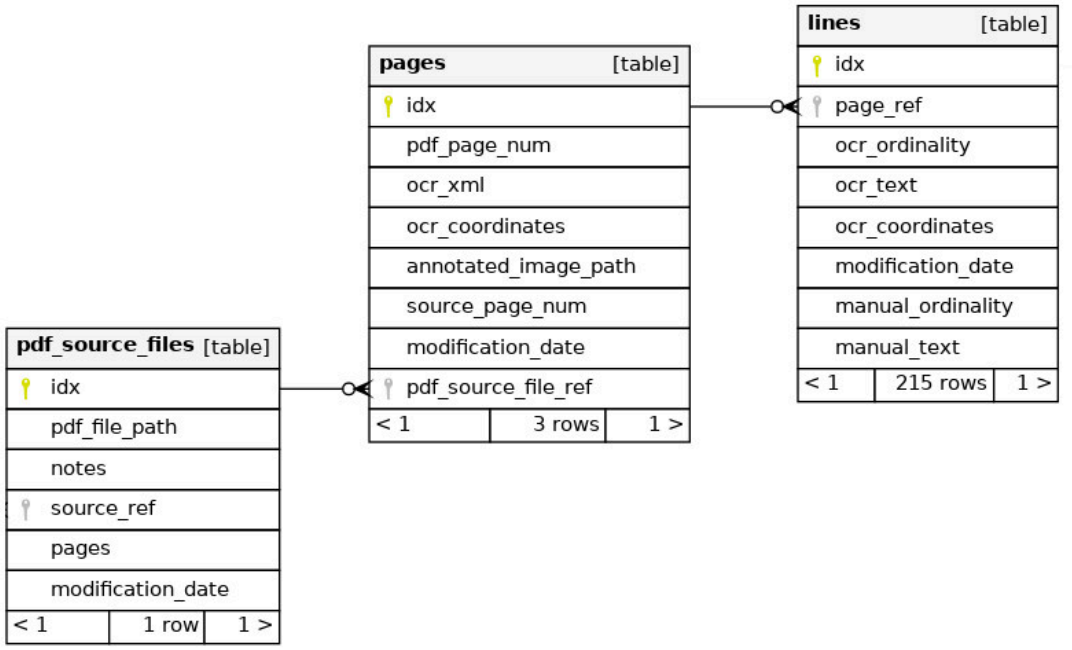


FIGURE 4. RAW DIGITIZED INFORMATION RELATIONSHIPS

The second part of the presentation of the implementation of our example is depicted in Figure 4 and concerns the organization of the raw digitized data. Every digitized file has multiple pages, which is reflected by a one-to-many relationship in the *input layer*. Every page data-item is linked to a digitized file. For every page data-item, the optical character recognition output in XML format, along with extraction metadata that are important in determining annotation boxes, are stored. In particular, our example stores the size, in terms of pixels, that the optical character recognition system attributes to the page. Moreover, since the page numbering in the electronic and in the physical files might be different, information about these numbers should also be 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 optical character recognition system, which implies that the stored line number does not necessarily reveal the actual line number of the source, but rather the number based on the ordering that the recognition algorithm attributed to this line. Besides the recognized text, the coordinates that the line is located in the page are stored. These coordinates are relative to the pixel coordinates of the associated page. The line data-item of our example allows also the possibility of storing manual corrections to the results of the automated recognition process. This can be seen in Figure 4, where the *lines* table allows to store both the *ocr_text* and the *manual_text*.

The last part of the *input layer* captures the essence of the principle of preserving the historical sources and, by associating concepts of interest in the data model with their origin, alleviates the ambiguity that characterizes the standardization of potentially conflicting archives. This is also the point that the implementation of our example departs from existing implementations that, as elaborated 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.

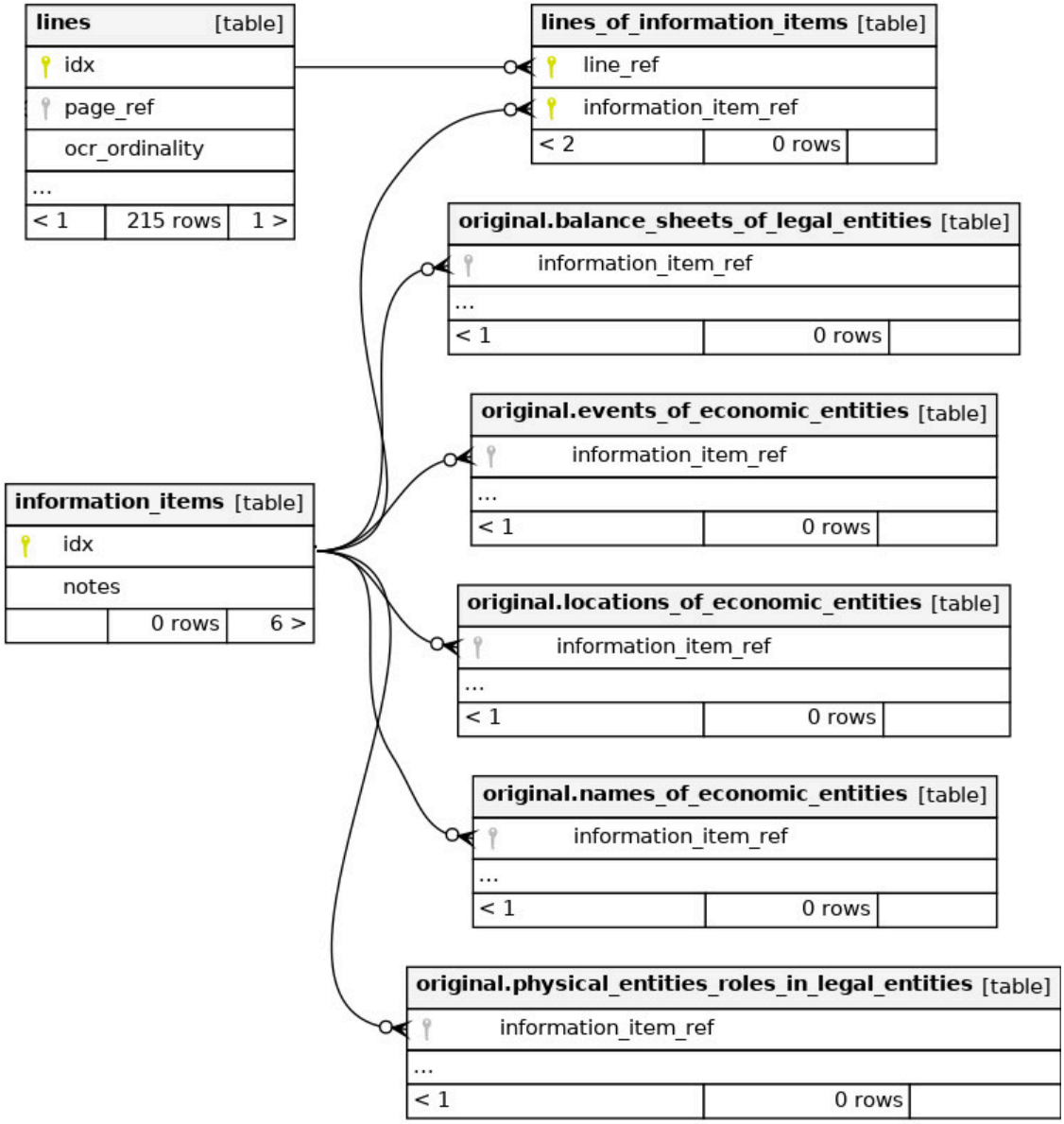


FIGURE 5. INFORMATION ITEMS RELATIONSHIPS

As illustrated by Figure 5. Information items relationships, 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 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 with 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, e.g. the name of a company, may be located to multiple lines that originate either from a single or multiple sources. As an example, in Figure 5. Information items relationships the information items are related with various, non-exhaustive concepts of the *original layer*, i.e. 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 upon the *input layer*, offering in this respect also 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 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 standardization is plausible. The former, however, are not completely explored, have complicated interconnections, and can have fuzzy and conflicting content; attributes that suggest that data-models that describe these higher concepts have to be frequently adjusted and updated. The benefit of the implementation of this article is that any updates of the layers that contain higher model concepts can be performed independently from the *input layer*. Moreover, it is possible to connect multiple higher-layer data models, with substantially different characteristics to a single *input layer*.

The higher layers can use different data 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 in cases of conflicting sources are. 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 concerning the plausibility of the data in the datasets that they construct and use.

5. A comparison of the preservation principle with existing approaches

There are two existing implementations of data models that aim to describe the information space of historical, firm-level data. The first one is developed in SCOB by the University of Antwerp and the second one in DFIH by the Paris School of Economics, with both of them being relying on relational technologies so far. While the data model of DFIH is a derivative of the SCOB's model, the implementations diverge in that the DFIH infrastructure is oriented towards semi-automated optical character recognition technologies for input, while the SCOB towards manual input[7]. The discussion in this section, although it is directly more applicable to DFIH's approach, is relevant to both of them, since as also noted in Section 4, the principle that this article proposes does not depend on the way that the data are collected.

[7] The development of the SCOB database started in 1998 whereas the application for the DFIH project was submitted in 2010.

In both of the aforementioned implementations, lines are associated with data model attributes. 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 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 implies 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 store only part of the line, which can potentially hinder the data provenance aspects of the model.

It is evident from the discussion of the two aforementioned problems that, although that both implemented systems move towards the direction of associating sources with model concepts, data provenance and model extensibility can be enhanced by applying the preservation principle proposed here. 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 that the principle of preserving the historical sources entails is needed to effectively achieve these effects. The amelioration of provenance in the example of this article comes from allowing associating data-items with more than one source in a standardized manner. The amelioration of extensibility comes from storing sources independently and allowing the association of each line with multiple concepts.

One might also consider metadata standards such as DDI as an alternative to the approach of this article. Albeit close, however, the purpose of such standards is different from the scope of the approach suggested by the preservation principle. DDI 2.5 standards aim to describe datasets and statistical data files at a study-level and not to associate historical sources with concepts. DDI 3.3 has an elaborate schema for associating study-level variables with various representations. However, the purpose of both versions is not to link historical printed sources with collected data-items. The scope of this article is narrower than that of DDI as it does not propose metadata standards that can describe and harmonize studies from various research projects of potentially different fields. Insteat, the article focuses at a more granular level as it aims to cope in particular with data-model evolvability and data-content verification issues that relate to information collected from historical sources. In particular, the association between the data of various model concepts and source lines that is discussed in this article is more specific, but also more granular than the scope of metadata standards and harmonization of datasets.

6. Collection of input data – The example of joint-stock German companies

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. We outline these first two steps in the following. Given our data model approach, further steps and historical data sources can be integrated in a rather straightforward and comprehensible manner.

The HdAG offers a detailed historical compilation of joint-stock companies in Germany. The series was published on an annual basis between 1896 and 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.

In the following paragraphs, we provide an overview of the steps in which text data from image-based files of the HdAG has been extracted. By doing this we outline also potential challenges of digitizing historical data beyond the challenge of integrating them in a flexible, extensible data model. The description of the steps sketches the procedure. The extraction of text data results in XML files containing the line-by-line extracted text data that can be used to populate the relational implementation that we propose.

The HdAG sources are scanned in a high resolution (600 dpi) format. Figure 6 shows a section of a scanned page. As can be seen in Figure 6, the pages are not scanned perfectly straight, so that the book fold provides a gradient effect of the text lines. In addition, the yellowed paper reduces the contrast between printed and unprinted areas. To improve the second deficiency, the scanned images are processed so that high contrast copies of the original images are produced.


*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.

FIGURE 6: SECTION OF A SCANNED HDAG PAGE

*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.

FIGURE 7: POST-PROCESSED SCAN

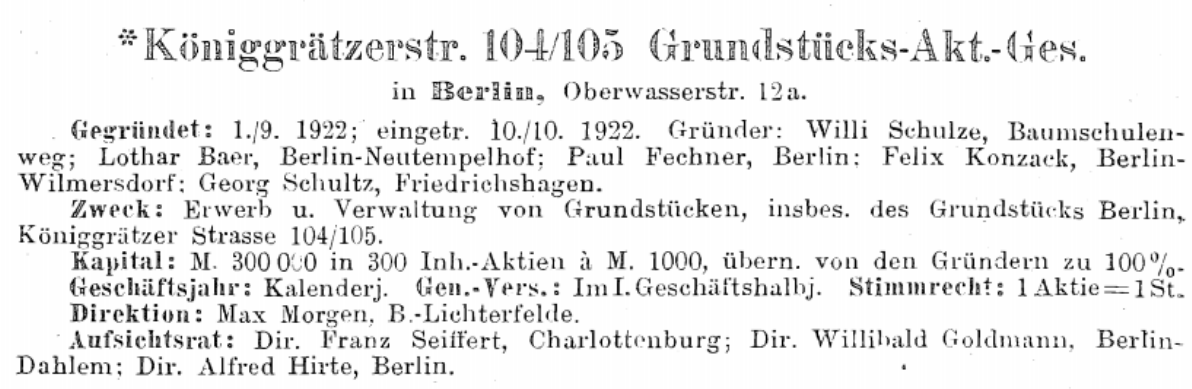
Figure 7 shows the same section of the image after pre-processing. The processed images are used in our optical character recognition (OCR) system. The ORC system used was designed especially for use extracting text data from parts of the HdAG series. By default, the OCR system that we use comes with a text recognition model for English characters. Based on training data, which is generated from manual transcriptions of text lines from the HdAG, the OCR system is trained to recognize the old German characters that are used in the printed books. The recognition is based on recurrent neural networks (LSTM) and is independent of any language model.

```
<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; Kapitänleutn. 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. Malzkaffee 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 />
```

FIGURE 8: EXTRACTED TEXT DATA

Figure 8 shows the output of the example snippet of Figure 7. In addition to the recognized text characters, additional information on the coordinates of the bounding box of the recognized text is stored in tags. The average error rate of the process is close to 3%. Furthermore, 18.7% of the errors concern over-recognition of spaces and they do not introduce difficulties in the subsequent processing of the extracted 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 is trained with a disproportionately large amount of training data that contains numbers. As a result, the error rate for numbers has been reduced to a mere 1%. However, the above-reported error rates are only valid for the scans that are of high-quality, i.e. there are not tilted, faded, or of low contrast. Figure 9 shows a low contrast example. Pages with such characteristics were corrected manually.



*Königgrätzerstr. 104/105 Grundstücks-Akt.-Ges.

in Berlin, Oberwasserstr. 12a.

Gegründet: 1./9. 1922; eingetr. 10./10. 1922. Gründer: Willi Schulze, Baumschulenweg; Lothar Baer, Berlin-Neutempelhof; Paul Fechner, Berlin; Felix Konzack, Berlin-Wilmersdorf; Georg Schultz, Friedrichshagen.

Zweck: Erwerb u. Verwaltung von Grundstücken, insbes. des Grundstücks Berlin, Königgrätzer Strasse 104/105.

Kapital: M. 300 000 in 300 Inh.-Aktien à M. 1000, übern. von den Gründern zu 100%.

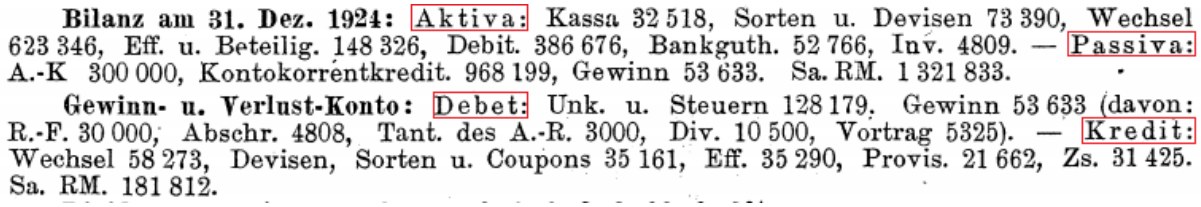
Geschäftsjahr: Kalenderj. Gen.-Vers.: Im I. Geschäftshalbj. Stimmrecht: 1 Aktie = 1 St.

Direktion: Max Morgen, B.-Lichterfelde.

Aufsichtsrat: Dir. Franz Seiffert, Charlottenburg; Dir. Willibald Goldmann, Berlin-Dahlem; Dir. Alfred Hirte, Berlin.

FIGURE 9: LOW CONTRAST EXAMPLE

There are some additional difficulties that arise even after correcting pages with low- quality scans. Balance sheet information and profit and loss statements were not always properly extracted. This was due to two factors. Firstly, 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. Secondly, given that the sources used to produce the reprint data was pretty old, parts of some pages were slightly folded, and, therefore, the text on the scan was compressed. To overcome this issue, we used the balance sheet and profit and loss statement layout.



Bilanz am 31. Dez. 1924: Aktiva: Kassa 32 518, Sorten u. Devisen 73 390, Wechsel 623 346, Eff. u. Beteilig. 148 326, Debit. 386 676, Bankguth. 52 766, Inv. 4809. — Passiva: A.-K 300 000, Kontokorrentkredit. 968 199, Gewinn 53 633. Sa. RM. 1 321 833.

Gewinn- u. Verlust-Konto: Debet: Unk. u. Steuern 128 179, Gewinn 53 633 (davon: R.-F. 30 000, Abschr. 4808, Tant. des A.-R. 3000, Div. 10 500, Vortrag 5325). — Kredit: Wechsel 58 273, Devisen, Sorten u. Coupons 35 161, Eff. 35 290, Provis. 21 662, Zs. 31 425. Sa. RM. 181 812.

FIGURE 10: BALANCE SHEET AND PROFIT AND LOSS STATEMENT

Figure 10 shows a typical example of a balance sheet and profit and loss statement of a medium-sized company. Common among all balance sheet information is that the statements contain the words “Aktiva:” und “Passiva:”. Similarly, for profit and loss statements the words “Debet:” and “Kredit:” are contained. Moreover, the order of appearance matters. “Passiva:” has to follow “Aktiva:” and “Debet:” has to follow “Kredit:”. Any balance sheet or profit and loss statements that did not fulfill this structure were manually checked.

7. Transformation of input data

The transformation process of the input is illustrated in Figure 11. In summary, the process leads to the creation of four different datasets. The following paragraphs sketch the process underlying dataset's *Dataset 1a* creation.

At the beginning of the process, multiple lines of the extracted text files that belong to one firm-year observation are identified (G_data_processing). Each line that marks the beginning of a potential company record is characterized based on six conditions. These are the

i. bounding box heights,

ii. vertical distances to text boxes of previous lines,

iii. the ratio of box widths to numbers of characters included in the box (i.e., horizontal space per character),

iv. the absolute number of characters of the box,

v. the fraction of non-alphabetical characters, and

vi. the location of the text boxes' center.

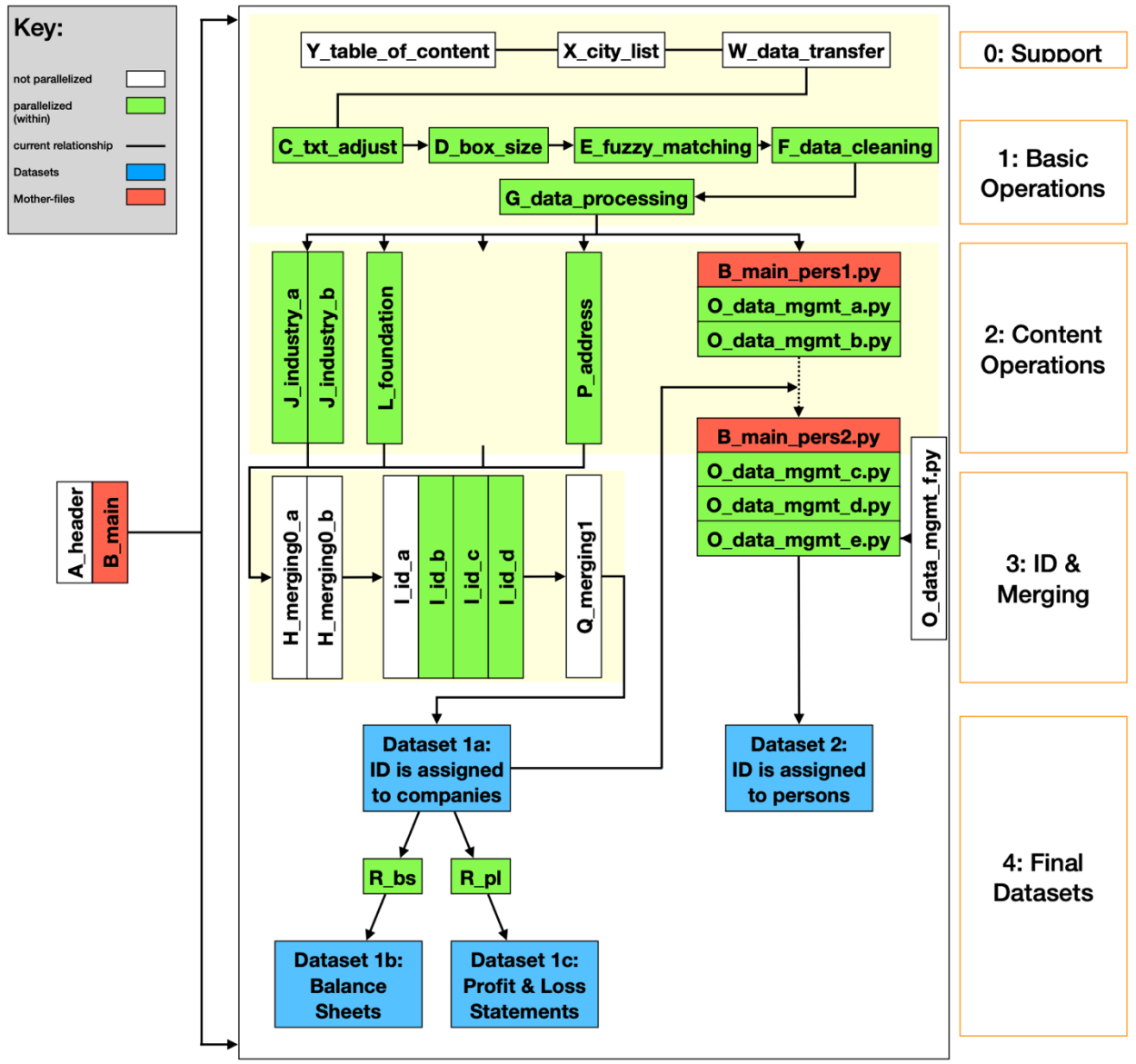


FIGURE 11. THE TRANSFORMATION PROCESS

Additionally, lines marking the beginning of other variables are located (E_ fuzzy_ matching), using a similar set of conditions as those stated above.[8] Because the initially described *company identification* is double-checked by the existence of variable duplications within one firm-year observation, the process of *variable identification* is implemented at first.[9]

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

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

The ID linking, represented by scripts *I_id_a - d*, constitutes the core part of the process. This part transforms repeated cross-sections of yearly data into a panel structure. Thus, each observation from t+1's cross-section is compared to all observations in time t.[10] Once the *linking score* exceeds a pre-defined threshold, two observations are considered to belong to the same legal entity and thus are assigned to the same ID. Various variables are used in each pair-wise comparison.[11] Depending on the variables' characteristics, either binary (e.g., dates) or continuous scores based on Levenshtein distances (e.g., company names) are calculated. The linking approach is insensitive to missing observations. In the end, individual linking scores for each variable are weighted to derive the final linking score.

There is a trade-off when setting the pre-defined linking threshold. On the one hand, a lower pre-defined threshold results in a higher number of false-positive matches. On the other hand, a high threshold leads to false-negative matches (i.e., that no match is found even though two observations belong to the same entity).

---

[8] Exemplary characteristics for the beginning of a company in the text flow is the line height, which is usually greater than 100 pixels (in Figure 3, the line height is 124 = 4551-4427 [difference in Y-coordinates]), a greater distance from the previous line, and the word "Gegründet:" in the following lines. The result of this processing step is a list of companies per book.

9 For example if two *foundation dates* are assigned to one firm-year entry, it is evident that one line corresponding to the start of a new company is not recognized as such.

10 In practice, observations are first compared to various subsets including more likely matches to reduce required computing resources. Subsets are, for instance, formed based on founding dates or industries.

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

The selected threshold aims at minimizing the number of false-positive links. Using a simulated test set, a threshold of 0.75 is set (see Figure 12).

FIGURE 12. DENSITIES OF LINKING SCORES

The approach partly results in fractional time-series meaning that one company is assigned to different IDs over time. A second linking round aims to mitigate this problem. In this round, observations previously assigned to one ID are only compared to IDs that do not include observations covering identical periods. This procedure allows for a lower matching threshold as the chance of false-positive matches is reduced by construction. After the automized linking, manual corrections are necessary to fill the remaining gaps in the time-series.

Descriptive statistics illustrate the resulting dataset's coverage. Considering only the raw data, Figure 13 shows that the number of observations extracted from the data source increases sharply until the middle of the 1920s. Afterward, the numbers exhibit a steady decline. One of the challenges of the dataset's construction is to deal with a change in reporting schemes. Until volume 25, 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 reflects one aspect of the above-mentioned lack of standardization of the data source. This change explains the drop of observations in volume 29, which includes many observations that are likewise covered by volume 30. Thus, duplicates are removed so that the panel structure's time dimension can be defined. Figure 14 illustrates the resulting observation count per year.

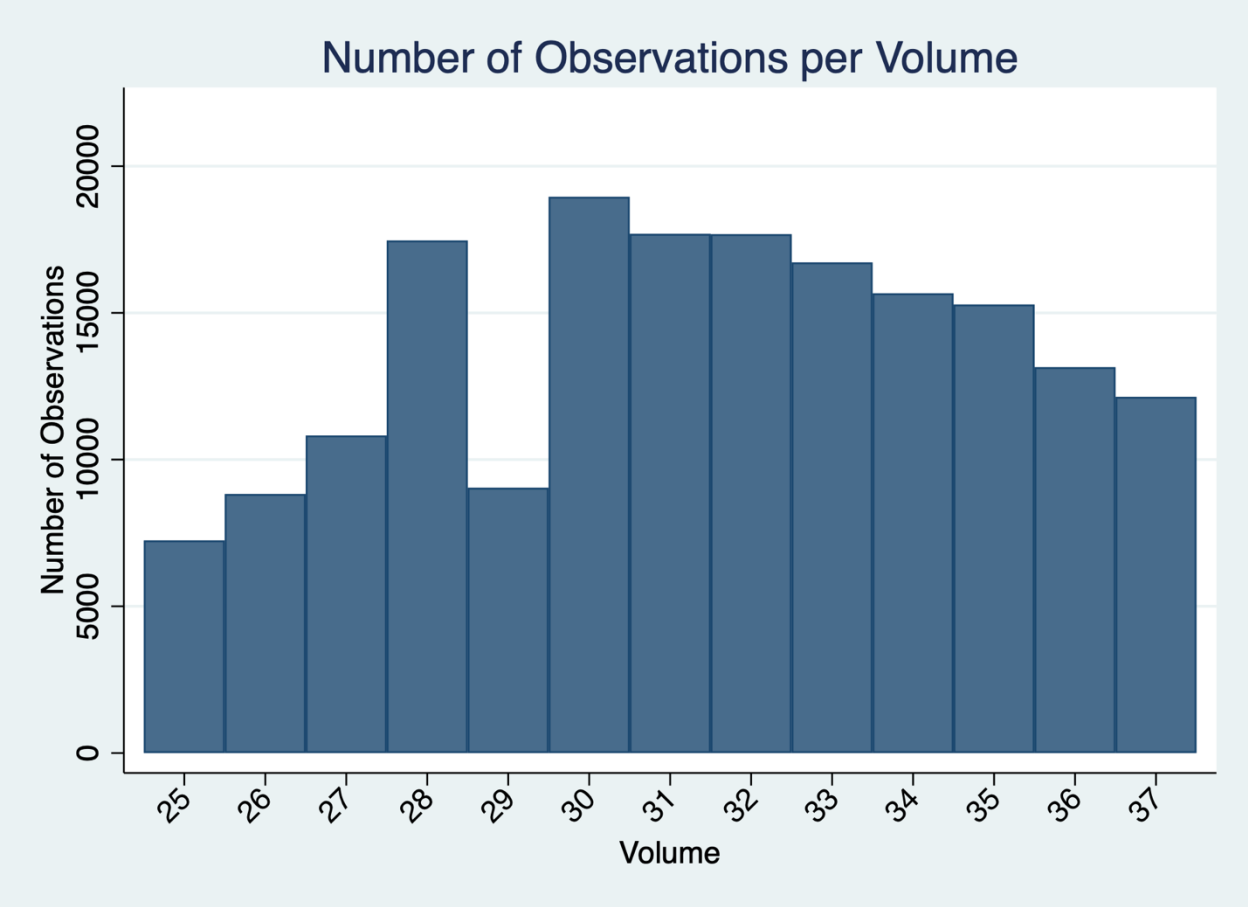


FIGURE 13. NUMBER OF OBSERVATIONS PER VOLUME

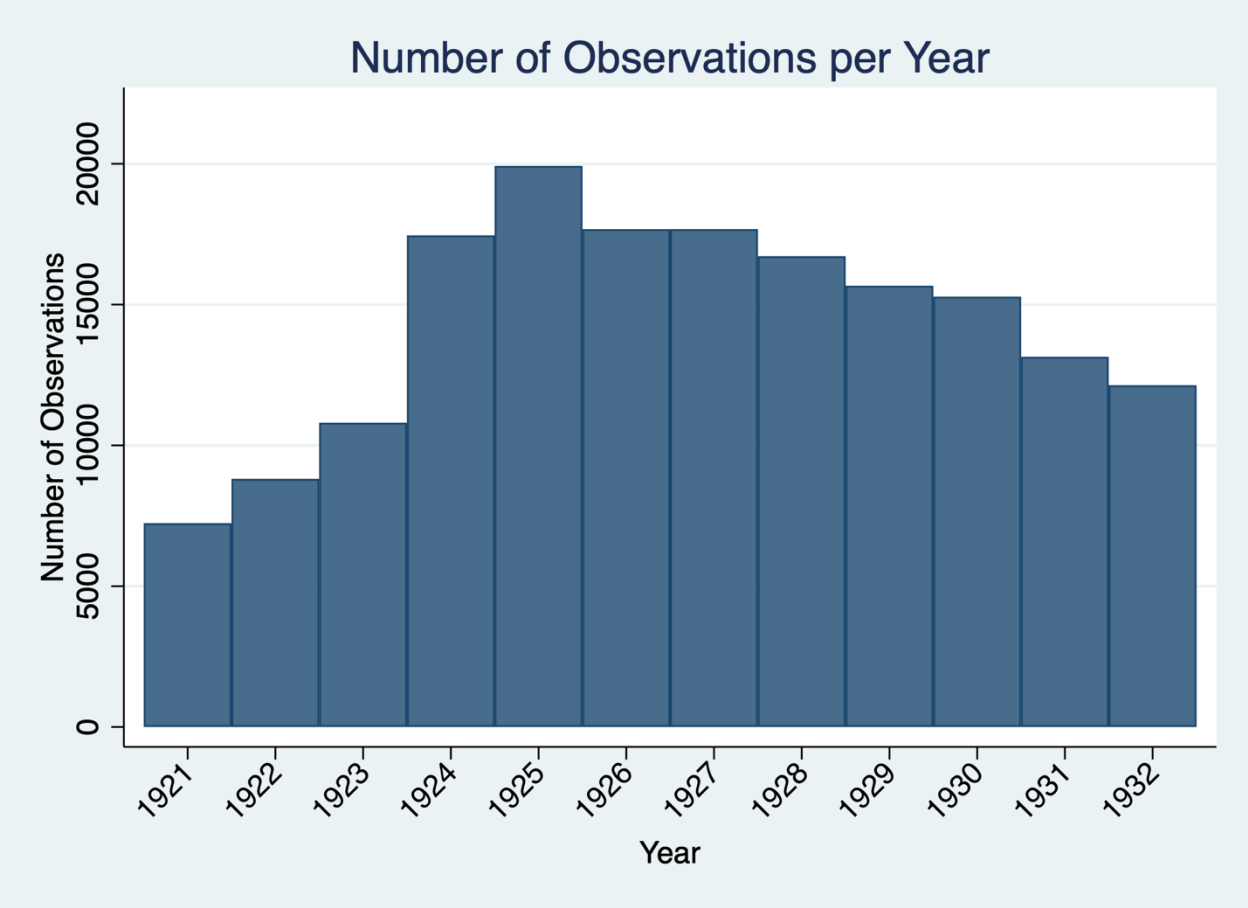
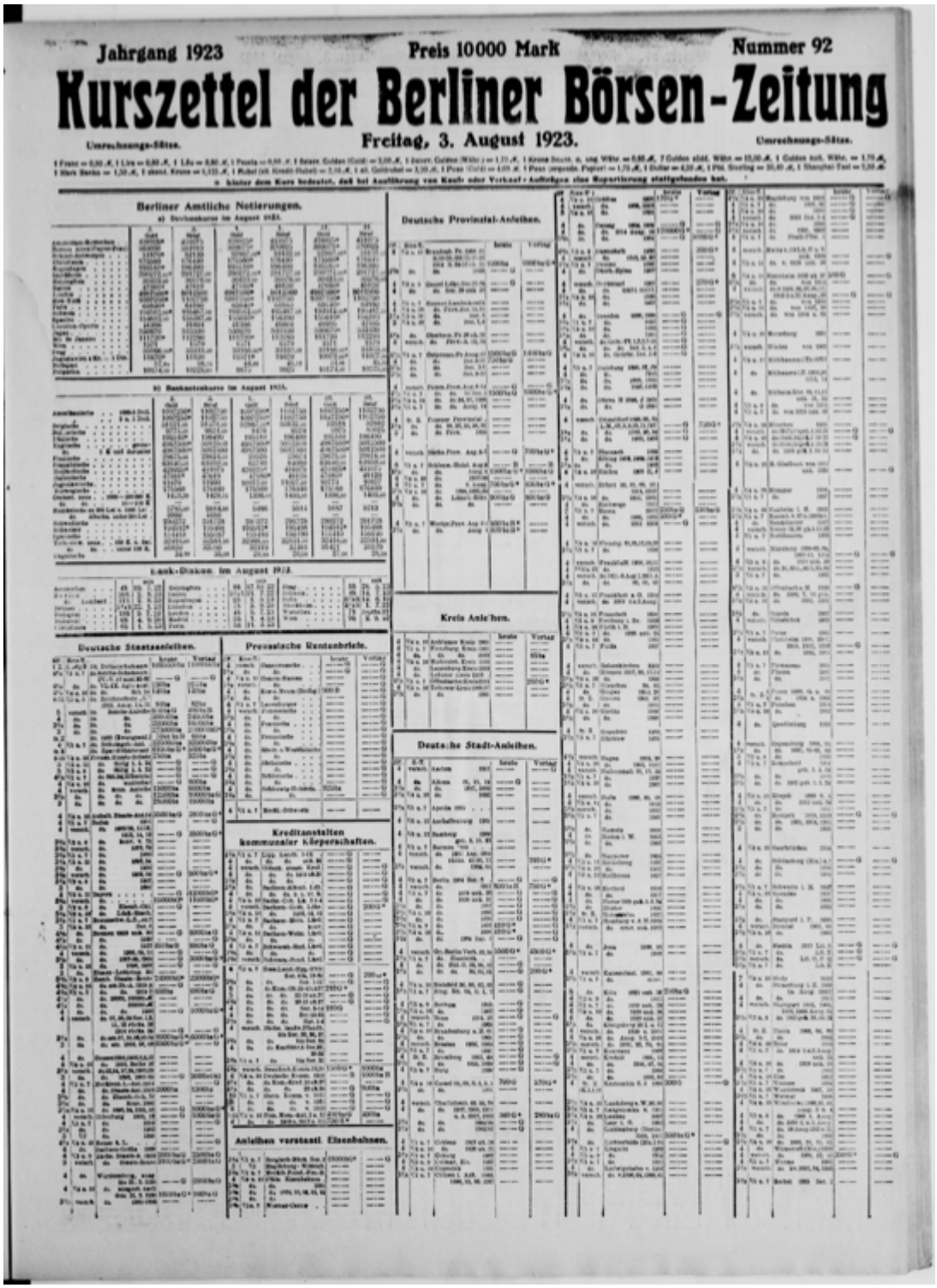


FIGURE 14. NUMBER OF OBSERVATIONS PER YEAR

8. Linking with security data

A subset of the joint-stock companies covered the HdAG was publicly listed at one or multiple German stock exchanges. Thus, matching daily stock market data extracted from the Berliner Börsen-Zeitung is potentially highly valuable for economic and financial research. Stock price information was printed in columns characterized by a high degree of variation in formats and content (see Figure 15). These inconsistencies together with challenges arising from inadequate horizontal segmentation of columns make an automated digitization and structurization process inapplicable. Instead, digitization by hand was applied. The resulting dataset was then matched to our company dataset.


Jahrgang 1923 Preis 10000 Mark Nummer 92

Kurszettel der Berliner Börsen-Zeitung

Freitag, 3. August 1923.

Berliner Amtliche Notierungen.

Deutsche Provinzial-Anleihen.

Kreis Anleihen.

Deutsche Stadt-Anleihen.

Deutsche Staatsanleihen.

Preussische Rentenbriefe.

Kreditanstalten kommunaler Körperschaften.

Anleihen verstaatl. Eisenbahnen.

FIGURE 15. NUMBER OF OBSERVATIONS PER YEAR

Matching data from different data sources is accompanied by other challenges and pitfalls compared to linking repeated cross-sections from one data source, as described in the previous section. For instance, probably due to space limitations, the Berlin Börsen-Zeitung's writing of company names used various kinds of abbreviations. Moreover, spellings of the two datasets relied on different sets of special characters.

The linking algorithm first homogenizes spelling styles before it performs the string comparison to mitigate issues potentially arising from these differences. As opposed to string transformations performed when cross-sections are linked, the transformation methods, in this case, focus more on abbreviations than potential OCR errors. The string comparison itself again relies on similarity measures referring to Levenshtein distances. It turned out that exact matching procedure did perform very poorly most likely due to different layouts of the underlying sources. Once corresponding IDs were linked, annual measures had to be derived from the daily stock prices included in the Berliner Börsen-Zeitung. For this reason, annual mean and median prices, as well as the prices' standard deviations, are calculated.

9. Conclusions

The paper discusses 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. While contemporary data models are built with standards in mind, applying this approach to historical data where the sources are highly non-standardized 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 article is based on the observation that the historical sources, which are finite in number and invariant in nature, constitute a solid basis to be used as 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 can 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.

The paper also discusses 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 variables that correspond to concepts of financial interest. Specifically, we describe the process of parsing and creating a company dataset. These variables can subsequently be used to create higher-level data model concepts that are built on top of the layer that is based on preserving the historical input.

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

References:

Acemoglu, D., & J. A. Robinson (2013). "Economics versus politics: Pitfalls of policy advice." Journal of Economic Perspectives 27.2, 173-92.

Anderson, H. M., Dungey, M., Osborn, D. R., & Vahid, F. (2011). Financial integration and the construction of historical financial data for the Euro Area. Economic Modelling, 28(4), 1498-1509.

Annaert, J., Buelens, F., & Deloof, M. (2015). Long-run stock returns: evidence from Belgium 1838–2010. Cliometrica, 9(1), 77-95.

Annaert, J., Buelens, F., & Riva, A. (2016). Financial History Databases: Old Data, Old Issues, New Insights?. Financial Market History, 44 - 65.

Annaert, J., König, W., Riva, A., & Yoo, L., EURHISFIRM - M1.2: Position paper (midterm), Horizon 2020 Research & Innovation Programme, No. 777489.

Barton, J., & Waymire, G. (2004). Investor protection under unregulated financial reporting. Journal of Accounting and Economics, 38, 65-116.

Bernstein, A., Hughson, E., & Weidenmier, M. (2019). Counterparty risk and the establishment of the New York Stock Exchange clearinghouse. Journal of Political Economy, 127(2), 689-729.

Braggion, F., & Moore, L. (2011). Dividend policies in an unregulated market: the London Stock Exchange, 1895–1905. The Review of Financial Studies, 24(9), 2935-2973.

Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. The Journal of Finance, 47(5), 1731-1764.

Calomiris, C. W., Mason, J. R., Weidenmier, M., & Bobroff, K. (2013). The effects of reconstruction finance corporation assistance on Michigan's banks' survival in the 1930s. Explorations in Economic History, 50(4), 526-547.

Costantino M., and Coletti, P. 2008. Information Extraction in Finance. WIT Press, Ashurst Lodge (Southampton).

Cordery, C. J. (2012). Funding social services: An historical analysis of responsibility for citizens' welfare in New Zealand. Accounting History, 17(3-4), 463-480.

Danielsson, J., Valenzuela, M., & Zer, I. (2018). Learning from history: Volatility and financial crises. The Review of Financial Studies, 31(7), 2774-2805.

Dimson, E., Marsh, P., & Staunton, M. (2002). Long-run global capital market returns and risk Premia. SSRN Working Paper No. 217849.

Dimson, E., Marsh, P., & Staunton, M. (2009). Triumph of the optimists: 101 years of global investment returns. Princeton University Press.

Dittrich, J., & Jindal, A. (2011). Towards a One Size Fits All Database Architecture. In *CIDR* (pp. 195-198).

Ducros, J., Grandi, E., (2018) Collecting and Storing Historical Financial Data: Project. In: Stuetzer, C. M., Welker, M., Egger, M.: Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications (350-372)

Eichengreen, B. (2016). Financial History in the Wake of the Global Financial Crisis. In E. Dimson & D. Chambers (Eds.), Financial Market History. CFA Institute Research Foundation.

Eichholtz, P., Korevaar, M., Lindenthal, T., & Tallec, R. (2020). The total return and risk to residential real estate. Available at SSRN 3549278.

EURHISFIRM (2019). D1.3: First yearly progress and strategy report to the General Assembly, Horizon 2020 Research & Innovation Programme, No. 777489.

Fan, W. (2004). Construction Methods for the Shanghai Stock Exchange Indexes: 1870-1940. The Shanghai Stock Exchange History Research Project of Yale SOM.

Gehrig, T., & Fohlin, C. (2006). Trading costs in early securities markets: the case of the Berlin Stock Exchange 1880–1910. Review of Finance, 10(4), 587-612.

Global Financial Data. (2005). GFD Encyclopedia of Global Financial Markets. 10th ed.

Goetzmann, W. N. (2016). Bubble Investing: Learning from History. In: **RESEARCH FOUNDATION BOOKS 2016 2016:3, (149-168**).

Goetzmann, W. N., & Huang, S. (2018). Momentum in imperial Russia. Journal of Financial Economics, 130(3), 579-591.

Goetzmann, W. N., Ibbotson, R. G., & Peng, L. (2001). A new historical database for the NYSE 1815 to 1925: Performance and predictability. Journal of financial markets, 4(1), 1-32.

Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production: Evidence from the collapse of international science. The Quarterly Journal of Economics, 133(2), 927-991.

Idreos, S., Maas, L. M., & Kester, M. S. (2017). Evolutionary Data Systems. CoRR abs/1706.0 (2017). arXiv preprint arXiv:1706.05714.

Jordà, Ò., Schularick, M., & Taylor, A. M. (2017). Macrofinancial history and the new business cycle facts. NBER macroeconomics annual, 31(1), 213-263.

Jordà, Ò., Knoll, K., Kuvshinov, D., Schularick, M., & Taylor, A. M. (2019). The rate of return on everything, 1870–2015. The Quarterly Journal of Economics, 134(3), 1225-1298.Jorion, P., & Goetzmann, W. N. (1999). Global stock markets in the twentieth century. The journal of finance, 54(3), 953-980.

Karapanagiotis, P., (2019). EURHISFIRM - D5.2: Technical Document on Preliminary Common Data Model, Horizon 2020 Research & Innovation Programme, No. 777489.

Lampe, R., & Moser, P. (2016). Patent pools, competition, and innovation—evidence from 20 US industries under the new deal. The Journal of Law, Economics, and Organization, 32(1), 1-36.

Mata, M.E., da Costa, J.R., & Justino, D. (2017), The Lisbon Stock Exchange in the Twentieth Century, Coimbra, Coimbra University Press.

Mehra, R., & Prescott, E. C. (1985). The equity premium: A puzzle. Journal of monetary Economics, 15(2), 145-161.

Moortgat, L., Annaert, J., & Deloof, M. (2017). Investor protection, taxation and dividend policy: long-run evidence, 1838–2012. Journal of Banking & Finance, 85, 113-131.

Moser, P., Voena, A., & Waldinger, F. (2014). German Jewish émigrés and US invention. American Economic Review, 104(10), 3222-55.

Nehme, R. V., Works, K., Lei, C., Rundensteiner, E. A., & Bertino, E. (2013). Multi-route query processing and optimization. Journal of Computer and System Sciences, 79(3), 312-329.

Van Overfelt, W., Deloof, M., & Vanstraelen, A. (2010). Determinants of corporate financial disclosure in an unregulated environment: evidence from the early 20th century. European Accounting Review, 19(1), 7-34.

Ranft, L. M., Braswell, J. & König W. (2020). EURHISFIRM - D5.5: Report on process for extendable data models, Horizon 2020 Research & Innovation Programme, No. 777489.

Reinhart, C. M., & Rogoff, K. S. (2011). From financial crash to debt crisis. American Economic Review, 101(5), 1676-1706.

Richardson, G. (2011). When the music stopped: Transatlantic contagion during the financial crisis of 1931. NBER Working Paper Series, 17437.

Richter, B., Schularick, M., & Wachtel, P. (2018). When to lean against the wind. WORKING PAPERS 18-10, New York University, Leonard N. Stern School of Business, Department of Economics.

Russ, R. W., Previts, G. J., & Coffman, E. N. (2006). The stockholder review committee of the Chesapeake and Ohio Canal Company, 1828–1857: Evidence of changes in financial reporting and corporate governance. Accounting Historians Journal, 33(1), 125-143.

Rydqvist, K., & Guo, R. Performance and development of a thin stock market: the Stockholm Stock Exchange 1912–2017. Financial History Review, 1-19.

Schlag, C., & Wodrich, A. (2000). Has there always been underpricing and long-run underperformance? IPOs in Germany before World War I (No. 2000/12). CFS Working Paper.

Schultz, S. M., & Hollister, J. (2014). The Delaware and Hudson Canal Company: Forming, Financing, and Reporting on an Early 19th Century Corporation. Accounting Historians Journal, 41(2), 111-151.

Turner, J. D., Ye, Q., & Zhan, W. (2013). Why do firms pay dividends?: Evidence from an early and unregulated capital market. Review of Finance, 17(5), 1787-1826.

Vaihekoski, M. (2020). Revisiting Index Methodology for Thinly Traded Stock Market. Case: Helsinki Stock Exchange; Case: Helsinki Stock Exchange.

Vaughan, F. L. 1956. The United States Patent System: Legal and Economic Conflicts in American Patent History. Norman, OK: University of Oklahoma Press.

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.

Waldinger, F. (2016). Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge. Review of Economics and Statistics, 98(5), 811-831.

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.

Zoumpatianos, K., Idreos, S., & Palpanas, T. (2016). ADS: the adaptive data series index. The VLDB Journal, 25(6), 843-866.

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 299	Ferdinand A. von Siemen	Motivated Beliefs and the Elderly's Compliance with COVID-19 Measures
No. 298	Calebe de Roure, Emanuel Moench, Loriana Pelizzon, Michael Schneider	OTC Discount
No. 297	Dimitrios Kostopoulos, Steffen Meyer, Charline Uhr	Ambiguity and Investor Behavior
No. 296	Reint Gropp, Thomas Mosk, Steven Ongena, Ines Simac, Carlo Wix	Supranational Rules, National Discretion: Increasing Versus Inflating Regulatory Bank Capital?
No. 295	Besart Avdiu, Alfons J. Weichenrieder	Financing Costs and the Efficiency of Public-Private Partnerships
No. 294	Christian Alemán, Christopher Busch, Alexander Ludwig, Raül Santaeulàlia-Llopis	Evaluating the Effectiveness of Policies Against a Pandemic
No. 293	Christoph Hambel, Holger Kraft, André Meyer-Wehmann	When Should Retirees Tap Their Home Equity?
No. 292	Andrea Modena	Recapitalization, Bailout, and Long-run Welfare in a Dynamic Model of Banking
No. 291	Loriana Pelizzon, Satchit Sagade, Katia Vozian	Resiliency: Cross-Venue Dynamics with Hawkes Processes
No. 290	Nicola Fuchs-Schündeln, Dirk Krueger, Alexander Ludwig, Irina Popova	The Long-Term Distributional and Welfare Effects of Covid-19 School Closures
No. 289	Christian Schlag, Michael Semenischev, Julian Thimme	Predictability and the Cross-Section of Expected Returns: A Challenge for Asset Pricing Models
No. 288	Michele Costola, Michael Nofer, Oliver Hinz, Loriana Pelizzon	Machine Learning Sentiment Analysis, COVID-19 News and Stock Market Reactions
No. 287	Kevin Bauer, Nicolas Pfeuffer, Benjamin M. Abdel-Karim, Oliver Hinz, Michael Kosfeld	The Terminator of Social Welfare? The Economic Consequences of Algorithmic Discrimination