Long-term data for Europe

EURHISFIRM

D1.5: Project midterm report

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

http://www.eurhisfirm.eu

AUTHORS:

Sébastien ADAM (UNIVERSITÉ DE ROUEN NORMANDIE)
Robin ADAMS (THE QUEEN'S UNIVERSITY OF BELFAST)
Jan ANNAERT (UNIVERSITEIT ANTWERPEN)
Miguel ARTOLA BLANCO (UNIVERSIDAD CARLOS III DE MADRID)
Stefano BATTILOSSI (UNIVERSIDAD CARLOS III DE MADRID)
Simon BOUVIER (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Frans BUELENS (UNIVERSITEIT ANTWERPEN)
Gareth CAMPBELL (THE QUEEN'S UNIVERSITY OF BELFAST)
Bertrand COÜASNON (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Christopher COYLE (THE QUEEN'S UNIVERSITY OF BELFAST)
Marc DELOOF (UNIVERSITEIT ANTWERPEN)
Jérémy DUCROS (ÉCOLE D'ÉCONOMIE DE PARIS)
Coen FIERST VAN WIJNANDSBERGEN (ERASMUS UNIVERSITEIT ROTTERDAM)
Nathalie GIRARD (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Renata GWOŹDZIEWICZ-PĘCHERZEWSKA (UNIWERSYTET EKONOMICZNY WE WROCŁAWIU)
Stefan HOUPT (UNIVERSIDAD CARLOS III DE MADRID)
Krzysztof JAJUGA (UNIWERSYTET EKONOMICZNY WE WROCŁAWIU)
Abe de JONG (ERASMUS UNIVERSITEIT ROTTERDAM)
Joost JONKER (KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN - KNAW)
Pantelis KARAPANAGIOTIS (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Wolfgang KÖNIG (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Katarzyna KUZIAK (UNIWERSYTET EKONOMICZNY WE WROCŁAWIU)
Aurélie LEMAITRE (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Anita MAKOWSKA (UNIWERSYTET EKONOMICZNY WE WROCŁAWIU)
Thierry PAQUET (UNIVERSITÉ DE ROUEN NORMANDIE)
Alexander PEUKERT (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Johan POUKENS (UNIVERSITEIT ANTWERPEN)
Lukas Manuel RANFT (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Emmanuel RAVIART (ÉCOLE D'ÉCONOMIE DE PARIS)
Yann RICQUEBOURG (INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES)
Angelo RIVA (ÉCOLE D'ÉCONOMIE DE PARIS)
Andres ROJAS CAMACHO (UNIVERSITÉ DE ROUEN NORMANDIE)
Gabriel SCHNEIDER (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Joanna Kinga SŁAWATYNIEC (ERASMUS UNIVERSITEIT ROTTERDAM)
John TURNER (THE QUEEN'S UNIVERSITY OF BELFAST)
Lana YOO (ÉCOLE D'ÉCONOMIE DE PARIS)

APPROVED IN 2019 BY:

Jan ANNAERT (UNIVERSITEIT ANTWERPEN)
Wolfgang KÖNIG (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Angelo RIVA (ÉCOLE D'ÉCONOMIE DE PARIS)

List of terms and acronyms

CDM	Common data model
CESSDA	Consortium of European Social Science Data Archives
CRSP	The Center for Research in Security Prices
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DDI	Data Documentation Initiative
DFIH	Données Financières Historiques
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
EUROFIDAI	Institut Européen des données financières
FAIR	Findability, accessibility, interoperability, and reusability (data principles)
FIBO	Financial Industry Business Ontology
GLEIF	Global LEI Foundation
IPR	Intellectual property rights
LEI	Legal Entity Identifier
RDF	Resource description framework
RI	Research infrastructure
SCOB	Studiecentrum voor Onderneming en Beurs
SSHOC	Social Sciences and Humanities Open Cloud
TOGAF	The Open Group Architecture Framework
WGIS	Work Group on Identification and Standardisation

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

Table of contents

I. Context

What is the problem/issue being addressed?

With economic growth still slow in some parts of Europe, the key societal challenges facing the European Union are investment, growth, and job creation. Unstable capital markets had undermined corporate investments and had led to increased unemployment and social inequality, harming citizens' well-being and sowing mistrust of public decision-makers and academic experts. In order to promote strong, sustainable growth and to meet these urgent social and economic challenges, the European Union needs sound scientific evidence.

Big data are promising tools in science today. However, in spite of the crucial advantages offered by "born-digital" big data, they still lack the historical depth that "born-on-paper" long-term data can provide. Scientific research, government policy, and society as a whole must explore the historical data necessary to understand the dynamics of the past and how these structure the present and the future. However, because we lack these empirical foundations, this crucial historical understanding of our society remains unfulfilled.

Why is it important for society?

The current lack of high quality long-term empirical European data prevents the usage and testing of models for analysing structural and cyclical changes, which are crucial for understanding the interactions between financial, economic, and social evolutions. To further illustrate, the 2008 Financial Crisis began in the mortgage-backed security (MBS) sector, where risk models went astray because they were calibrated on only five to ten years of historical data, taken from a very benign period. Moreover, macro-financial models failed to take into account several frictions: in the period over which these models were estimated/calibrated, such frictions had not manifested themselves. In other words, we now retrospectively understand that the models were based on a relatively small time-period and thus gave biased and narrow snapshots of the overall picture. If longer-term historical data had been available at that time to build empirical models based on larger timespans, we can assume that we could have had a more robust understanding of the underlying currents and risks, which could have potentially prepared us better for (or perhaps could have forewarned us better on) the crisis.

As such, creating sound future policy and preparation for future financial events requires the understanding of both past and current dynamics. Creating the data to develop this knowledge requires sharp interdisciplinary skills, some of which are specific to a country, or even to a region, because of the heterogeneity of historical business rules and practises. These peculiarities call for an ad hoc Research Infrastructure that can also connect to other existing systems.

IT research must therefore develop innovative models and technologies that push forward the technological frontier and spark a big data revolution in historical social sciences: the scaling up of the variety, quantity, and quality of available long-term data. Digitalized historical sources as part of the European cultural heritage represent a shared wealth in terms of citizenship, cultural growth, and economic potential.

II. Project objectives and impact

EURHISFIRM aims to design the most comprehensive long-run economic and financial database to-date. It envisions creating a world-class research infrastructure (RI) to ***connect, collect, collate, align, and share detailed, reliable, and standardized long-term company-level data for Europe*** to enable researchers, policymakers and other stakeholders to ***analyse, develop, and evaluate effective strategies to promote investment and economic growth*****.**

It will handle data on European companies such as accounting, funding and investment, stock exchange data, governance rules, directors, patents, and headquarter locations. The creation of a vibrant European community will support the project's development based on innovative technologies, which will provide reliable and standardized long-term company-level data for European stakeholders: policy makers, scholars, and private companies. To achieve this goal, EURHISFIRM is developing innovative tools to spark a "big data revolution" in the historical social sciences and to open access to cultural heritage in close cooperation with existing RIs. This design enables ESFRI, member states and other funding bodies to decide on the further preparation and implementation of the RI.

In terms of impact, the EURHISFIRM project meets the need for establishing state-of-the-art benchmark research infrastructures in Europe, particularly in the social sciences and humanities, in which Big Data have yet to be fully exploited.

The long-term impact of the project is expected in four areas:

1) *The scientific domain:* The cross-cooperative research infrastructure design would foster more fruitful Pan-European research in the economic, financial and historical studies, helping to unleash the untapped research potential in these areas, particularly in terms of long-term comprehensive data.
2) *Policy bodies and public organisations:* As a result of the greater exploitation of European financial and economic data and the resulting research advancements, policy makers would be in a position to gain deeper insights to the European economy and society, which would further aid in their policy-making decisions.
3) *The private sector:* EURHISFIRM would enable financial institutions to develop better strategies to cope with economic and financial risks. It would spur the development of more efficient and more robust new financial products, tailored to the needs of European investors. Moreover, the technological innovation EURHISFIRM will bring to the field of data extraction and enrichment will further the European technological frontier.
4) *European citizens:* Through policies designed to better resist economic shocks based on the insights gained from robust historical data, European citizens will benefit from a society that is more prepared to bounce back more quickly from social, economic, and other issues that result from financial crises. The promotion of the European cultural heritage will also bring to fruition the collective treasury formed by the histories of European companies and strengthen the European identity, citizenship, diversity and cultural growth. Moreover, EURHISFIRM data are invaluable to enhance financial literacy among European savers. By providing educational-based services, EURHISFIRM will cooperate with prudential authorities, financial advisors, wealth

managers, and investor associations to build simple and interactive graphical user interfaces that allow citizens to browse the data and learn about the consequences of investment decisions.

III. Details of the work carried out

III.1 Overview of the progress from the beginning of the project to the end of the midterm period covered by the report and main results achieved so far

The EURHISFIRM project is on-schedule according to the project proposal. The accomplishments for each Work Package are summarised below in section 1.2 for this reporting period (1 April 2018 – 30 September 2019). All required deliverables and milestones to-date have been submitted (17/54 [31%] and 4/21 [19%], respectively).

The project's Work Packages can be grouped into several core categories:

- ***Economic history foundations***: To ensure that the developed technology properly accommodates the data that it will handle, Work Package 4, along with the project's other economic historians, worked on cataloguing EURHISFIRM's data sources and established a common standard of documentation for current and future data; the selected standard is the DDI-Lifecycle standard. Work Package 4 has completed 86% of its tasks (6/7 deliverables and milestones) as of the writing of this report and will complete its entire Work Package by the end of September 2019. In Work Package 7, the economic historians also work with the technical developers to ensure that the tools developed can process the data elements correctly and can take into account historical nuances. The economic history work, such as verifying the results of the tools developed and producing reference documents with regards to the sources being treated, also form a crucial part of the data extraction work in Work Package 7. Six sources in total (price lists from three markets (Belgium, France, Germany) and yearbooks from three markets (France, Germany, Spain)) have been chosen as the representative samples for this work based on the existing or potential capacities for digitalisation in the sources' current states.

- ***Information technology work (data extraction technologies, common data modelling, infrastructure architecture)***: The technical Work Packages of the project are Work Packages 5, 6, 7, and 9, working on common data modelling, data connecting and matching, data extraction and enrichment system, and infrastructure policy and architecture, respectively. The official deliverable deadlines for most of these Work Packages begin on or after the midterm, but the work for these tasks have been in progress, as described in the summaries for Work Package 5, 6, 7, and 9. Work Package 5 has studied the models of existing database models and describes key design elements to be considered for the EURHISFIRM common data model; it also established a methodology for common data model evaluation. It coordinates the Working Group on Identification and Standardization (WGIS) to increase the communication and collaboration between the different work packages to facilitate the implementation of technical standards in the project. Work Package 6 has also recently begun its work in data connection and linkage among the independent databases sources, in line with the project timeline. This work will logically use the consortium's most advanced databases (the Brussels and Paris financial databases) as a first test case by linking them and testing their abilities to retrieve the queried data in a synchronised and accurate way. (For example, if a user searches for a company existing in

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

both the Brussels and Paris databases, the system should return the information associated to this company in both databases by correctly identifying this company in both sources and recognising that this is indeed the same company despite possible variations in the name.) Work Package 6 has already started discussions with other relevant institutions regarding further test cases. Work Package 7 has begun developing artificial intelligence-based systems to recognise the tabular structures and the texts of the data sources through collaboration with the consortium's economic historians. The Work Package has also begun to conduct tests on the tools developed in order to validate and improve the results, and web linking tools (to connect the company entities in the databases to the data existing in the wider world wide web) have begun to be evaluated under task 7.4. Work Package 9—while it is officially scheduled to start in the third quarter of 2019—has begun to investigate the architectural framework of the project by investigating materials from existing European research infrastructures and participating in related technical discussions with other Work Packages.

- ***Practical aspects of the infrastructure operations (business plan [including user target research], legal plan, and cultural heritage of the data produced)***: Work Package 3 has begun its work concerning ownership and property rights, especially in the context of open science, which will address the overall legal recommendations, issues, constraints, and solutions concerning the EURHISFIRM infrastructure design in all of its aspects (including but not limited to data rights, user and access rights, privacy laws). Work Package 8, in charge of understanding the infrastructure's potential interaction with users and identifying the target users and their needs, has conducted quantitative and qualitative studies with relevant subject pools. Work Package 10's core work in business model and governance will begin in the third quarter of 2019, but it has nonetheless already made progress in defining alternative business model concepts, as well as conducting a preliminary assessment of business and governance model alternatives. Work Package 11's work in evaluating alternative strategies to use digitized material (from historical sources) for the promotion of European cultural heritage will also begin in the third quarter of 2019, and preparatory work has been in progress to ensure a solid start to these tasks.

- ***Project administration (communication, community building, logistics and strategy/vision [including compliance to open science frameworks and FAIR data])***: Work Package 1 runs the project's logistics and administration, coordinates with other Work Packages and partner institutions in the overall project management of EURHISFIRM, and works with the Executive Committee and the Steering Committee to drive the project strategy and direction (including compliance to open science frameworks and FAIR data protocols through a continuously updated Data Management Plan). Work Package 2 leads the project's communication and outreach tasks, including the maintenance of the project communication tools (website, social media), the project identity and logo, and organisation of the project's annual General Assembly meetings. Finally, the community building task, which concerns both Work Packages 1 and 2, is indeed a project-wide effort from all consortium members and a key part of EURHISFIRM's overall ambitions.

III.2 Explanation of the work carried out per Work Package

Work Package 1: Project management

Work Package 1 is responsible for all coordination tasks for the project in: (1) overall strategy and administration, (2) scientific aims, and (3) technical milestones. The priority for the first year was to establish the foundations for these parts. Thereafter, Work Package 1 has been focusing on building a community of collaborators, stakeholders, and interested persons from the research, private and public sectors; citizens; and European research infrastructures and service providers. These efforts come from the fact that EURHISFIRM recognises the importance of strong community and networks to strengthen the European research infrastructure ecosystem. In this regard, EURHISFIRM has agreed to join the SSHOC (Social Sciences and Humanities Open Cloud), part of the EOSC (European Open Science Cloud) initiative. CESSDA (Consortium of European Social Science Data Archives), which is a coordinator for SSHOC, has also agreed to join the EURHISFIRM consortium. EURHISFIRM further aims to concretise and expand upon these community-building efforts by applying for the Horizon 2020 INFRAIA-02-2020 call (Integrating Activities for Starting communities).

In terms of the overall strategy and administration, the project continues its execution, keeping in mind FAIR (findability, accessibility, interoperability, and reusability) data principles and European-level collaborations. All required deliverables and milestones to-date have been submitted (17/54 [31%] and 4/21 [19%], respectively). With the Executive Committee, the Work Package also drives the project strategy and overall vision and the organisation of tasks to support those objectives (e.g. defining the short- and long-term goals of the project, identification of future events and conferences to publicise the project, communication strategies, and other tasks to advance the project visibility).

Regarding scientific coordination, Work Package 1 is responsible for the Data Management Plan, continuously updated by aligning with other Work Packages on these points. Four versions will be submitted as project deliverables, of which two have been submitted so far. The latest version of the data management plan was the second version (D1.7: Second Data Management Plan) completed at the 12-month mark of the project. This public deliverable is available for viewing on the EURHISFIRM website at https://eurhisfirm.eu/wp-content/uploads/2019/04/EURHISFIRM-D1.7-DataManagementPlan_v2.3.pdf (DOI: https://doi.org/10.5281/zenodo.3246339). The third version will be delivered at the end of March 2020 and the last version will be delivered at the end of March 2021.

To coordinate technical milestones, Work Package 1 tracks the technical Work Packages' progress and facilitates strategic and coordination discussions as necessary. These tasks are done in collaboration with all other Work Packages. To facilitate the work in the consortium, we have implemented tools such as file sharing technologies, video-conferencing tools, and instant messaging systems (hosted on Huma-Num [https://www.huma-num.fr], the French national services provider of DARIAH).

Work Package 1 also coordinates the collective reports of the project (such as the annual report, the midterm report such as this deliverable) and also maintains continuous communication with the administrative staff of the consortium partners.

Work Package 2: Dissemination and communication

Work Package 2 centres around the communication about the project and dissemination of its results, both internally, but more importantly, also to the broader stakeholder community. The Work Package started by creating an identity for the project, by designing a logo, templates, and a website (www.eurhisfirm.eu), which is one of the centrepieces for the project's communication. In addition, the first steps on social networks were established. Further, a dissemination and communication plan was developed to define the target audiences and the channels by which to communicate with them, as well as what to disseminate and to whom. To this end, with input from the project partners, an inventory was made of the potential stakeholders, which include policy makers, academia, business and society.

The following tasks have been accomplished so far:

- Developing the project's identity and brand: This task was aimed to create a unique identity for the project, including the project website and identity material: creation of recognizable project Identity materials (logo, indication of guidelines for a coherent brand image (fonts, colours, backgrounds), development of ready-to-use templates for project documents and presentations); development of a unique brand for the project, taking into account its characteristic features; adaptation to both substantive requirements and project guidelines; setting guidelines and consistency for the project.

- Creation of website for external communication of the project: creation of website layout; creation of subpages dedicated to relevant project requirements and subjects; regular implementation of information connected with the project. The website hosting is provided by Huma-Num [https://www.huma-num.fr], the French national services provider of DARIAH.

- Creation of social media/social networking profiles: profile creation, implementing information connected with the project.

- Development of a dissemination and communication plan: this plan was designed to outline the dissemination and communication activities and to raise awareness amongst the different target audiences. This plan was prepared in the first months of the Project. The following elements were specified: all the key messages as well as the target audiences; appropriate tools, platforms and channels; participations in conferences and workshops to meet the information needs of the target audiences and the project's communication objectives; ensuring maximum outreach of all news and project results.

- Inventory of European and national distribution networks: the task aims to complete, with the support of the members and of identified stakeholders, an overview of the stakeholders in the project. The task was completed by the following activities: overview of the communities with a stake or interest in the Project, stakeholder groups from different domains: policy makers, academia, business, society; contacts to key/strategic from Project point of view institutions in Poland.

The following activities are in progress:

- Building the project's community; project promotion and dissemination: the most important activity within this task has been the organization of the first General Assembly Meeting (March 15-16 2019 in Wroclaw).

- Expand on social media tools and communication channels e.g. Research Gate. These are conducted on a continuing basis.

Work Package 3: Legal and Ethical Issues

Open data and open access are key requirements of the European Commission for developing a research infrastructure that will become part of its infrastructure roadmap. Nevertheless, the collection and sharing of data and images can be linked to ownership and property rights issues, which need to be identified and solved. Ownership rights on images and data may place restrictive conditions on access and prevent data linkage. The aim is to design rules and policies for handling ownership rights with individual researchers, research institutes, data vendors and owners of historic paper sources. This task will also propose a policy for the data access in close collaboration with Work Packages 9, 10 and 11.

The preliminary study of German copyright and unfair competition law showed that much of the raw data is indeed in the public domain and can thus be used for any lawful content. There are, however, exceptions to this rule, in particular, if longer text passages are used. It is also evident that the relevant laws of copyright and – even more so – unfair competition differ among the EURHISFIRM countries. It will thus be necessary to study the legal situation in each of these seven jurisdictions. As regards to future topics of consideration, the different possible alternative uses of EURHISFIRM database can be used as scenarios to discuss legal matters within task 3.1.

The writing of the final Work Package 3 report has started. The report will identify and examine in-depth potential issues with ownership or property rights and lay out policies to deal with them. The subject matter of the rights, the actual rights persisting and the limitations and exceptions to copyright protection will be examined respectively. To the extent possible, the report will refer to the source materials presented in Deliverable 4.2 (Report on the Inventory of Data and Sources) and examine whether, and if so which, rights might persist in the materials. Further, it will examine whether the possible utilizations of the materials in the context of the database project might infringe those rights or whether they fall within the scope of a limitation or exception of the exclusive rights.

Work Package 4: Data and sources inventory and documentation

Work Package 4 has built up expertise regarding existing datasets and historical printed serial sources on publicly traded companies from 1815 onwards, as well as on data documentation (metadata) standards. The tasks set in Work Package 4 are key inputs for the future work of several other Work Packages that will draw on its output for the elaboration of a common metadata model (Work Package 5), data connecting and matching technologies (Work Package 6) and a system for automated data extraction (Work Package 7). In short, these tasks included drawing up an inventory of the sources that are available for building the databases our research infrastructure aims to do, describing their contents and semantics, assessing their quality, and ultimately defining the project's documentation standard.

All of Work Package 4's deliverables and milestones have currently been completed within the specified timeframe. This would not have been feasible without the indispensable input from all of the participating teams.

- Having appropriate data documentation or metadata standards is crucial for the tasks of Work Package 4. Metadata is defined by the International Organisation for Standardisation (ISO) as "data that defines and describes other data" or, simply, "data about data". After a review of several standards, it was concluded that the Data Documentation Initiative (DDI) provides the most appropriate one, namely the DDI-Lifecycle standard, which can be produced and edited by the Colectica Designer software.

- The data and sources inventory (D4.2, December 2018) identifies the principal official publications of company information, stock exchange price lists and yearbooks with summary governance and financial information on publicly traded companies in each of the participating countries, as well as existing datasets for financial history. The inventory also provides a summary description of their contents.

- The report on data and sources semantics (D4.3, March 2019) contextualises the categories of information commonly found in stock exchange yearbooks and price lists. It does so by providing clear definitions and a detailed, historical overview of legislation, regulation and customs in the fields of company identification, corporate governance, securities trading and financial reporting.

- A selection of printed stock exchange price lists and yearbooks, one of each category for every country in the consortium, as well as several existing databases and datasets on financial history (including SCOB, D-FIH, Eurofidai and LSPD) were documented according to the DDI Lifecycle standard with the Colectica software (D4.4, May 2019). The report furthermore includes a ranking of source types and a methodology for assessing the quality of datasets by means of a questionnaire.

- Building on the results of task 4.1 and the experiences of task 4.3, the DDI family of standards was chosen as EURHISFIRM's standard of documentation (D4.5, September 2019). DDI is widely used in the social sciences research community (including CESSDA). In the report, we propose different standards for different tasks: DDI Codebook (DDI 2.5) for documenting data during upload and DDI Lifecycle (DDI 3.2) for the subsequent harmonisation of data. We also evaluated two software packages for documenting datasets in DDI: Dataverse and Colectica Designer.

- The protocol of data documentation (M4.1, September 2019) is a manual for uploading data and metadata with the Dataverse platform. It is intended for researchers who want to contribute their data to EURHISFIRM and facilitates meeting the FAIR guiding principles for research datasets.

- For the final scientific paper on data and sources (M4.2, September 2019), we built on the information collected during tasks 4.2 and 4.3 to give a complete overview of provisions on the mandatory disclosure and publication of information on the governance and ownership of joint-stock companies in the company laws of the consortium countries.

Work Package 5: Common data model

Work Package 5 focuses on the development of concepts, architecture, and design of an overarching European level, Common Data Model (henceforth CDM). It progressively sets standards, identifies best practices and facilitates harmonization processes for the integration of European, long-term, firm-level data from heterogeneous, historical, national sources. Towards this goal, we firstly document models available from within as well as outside the institutions of the consortium and evaluate their strong points and their weaknesses. Secondly, based on the observations of the first step, Work Package 5 designs an initial model of historical, European firm-level data with information that is spanned in three dimensions: financial information, accounting information and management information.

We have started Work Package 5 earlier than originally planned. The approach so far indicates that previous work on model designs done in institutions of the consortium such as SCOB and DFIH provides solid building blocks for the design of the CDM. There are various challenges when adapting these national models into a European level and at this point, the characteristics that lead to the success of established systems, such as CRSP and EUROFIDAI, can be incorporated into the CDM. The overarching identification system and the semantic structure are future focal points of the CDM design.

A central characteristic of our initial design is the distributive nature of the CDM. Our goal is to propose a design that is minimally intrusive to existing implementations, respects national idiosyncrasies and gives the ability to national research centres to grow collaboratively but also independently. Towards this orientation, the CDM is initially designed as a hyper-national-structure that coordinates, consists of, and is led by the national research centres.

To facilitate the coordination and standardisation among the technical Work Packages of the project, Work Package 5 also coordinates the **Working Group on Identification and Standardization (WGIS).** Standards are indispensable fundamentals of any ICT-based (information and communications technology-based) system integration and system interaction design. In this environment, the Working Group on Identification and Standardization (WGIS) of EURHISFIRM aims to increase the communication and collaboration between the different work packages to facilitate the implementation of standards in the project by exchanging problem descriptions and offering solutions regarding their individual Work Packages as well as for the overarching EURHISFIRM goals and objectives.

The Architecture Team, as a subgroup of the WGIS, currently discusses certain technical topics in more depth. These topics include the Common Data Access Service of Work Package 5, the use of identification regimes and data matching methodologies, and the consideration of potential technology platforms and architectures for network data management.

Work Package 6: Data connecting and matching

The key idea behind the research infrastructure (RI) is that users should be able to query it without needing to know in which databases the required information is to be found. Behind the screens, the infrastructure therefore needs to be able to locate and connect the relevant databases and to retrieve information from them in a consistent way. In Work Package 6, the technologies that allow this will be developed and tested. As envisaged, this work has just begun. For the first test case, the two most advanced databases hosted by member-institutions, the Paris and Antwerp databases (DFIH (Données Financières Historiques) and

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

SCOB (Studiecentrum voor Onderneming en Beurs), respectively) covering the Paris and Brussels exchanges, will be linked and integrated. To this end, the two teams involved have started to look into the technical details to achieve this. At the same time, preparatory investigations have started to single out other databases, both internal and external to the project, for testing purposes. Members of the Antwerp team have reached out to Mike Staunton of the London Business School (EURHISFIRM Project Advisory Board member) who has agreed to include the LSPD (London Share Price Database) in Work Package 6 to test the matching and connecting of databases. Members of the Paris team have likewise been in contact with the CNRS about Eurofidai. Different scenarios for matching and connecting data from Eurofidai to consortium members' databases (i.e. SCOB and DFIH), from a basic exchange of identifiers to a complete merger, are being discussed. In preparation of the tasks within Work Package 6, the LSDP and Eurofidai have also been documented in D4.4 (see above) and existing literature on matching (historical) data on companies and persons from different datasets has been reviewed by economic historian Johan Poukens. This preparatory work will ensure that work on the actual programming of the matching and connecting will take off immediately when a database expert joins the Antwerp team on October 1, 2019.

Work Package 7: Data extraction and enrichment system

The aim of Work Package 7 is to develop an intelligent and collaborative system for the extraction of structured information from images of historical documents related to companies' financial and economic activities. To keep the Work Package manageable, the system only takes into account historical printed sources related to listed companies such as yearbooks and exchange lists.

We first defined and selected the document sample dataset for testing and validating the recognition system we develop. Six corpora have been selected within the consortium to conduct the experiments for information extraction on yearbooks and price lists. After a final decision from the Steering Committee, they are the German Yearbook 1913-1914 (Handbuch der deutschen Aktiengesellschaften), the Madrid daily bulletin 1929-1930, the French Desfossés Yearbook 1962; the official price lists Brussels 1912, Madrid 1934 and Paris 1961-1962. All in all, this selection covers three different periods of time (pre-World War I, the inter-war period, and post-World War II), in three different languages (French, German, Spanish).

This work is being realised via close collaboration with the economic historians in the consortium to ensure proper knowledge transfer of the sources being treated, especially those from the source countries' institutions (Universiteit Antwerpen, Universidad Carlos III de Madrid, École d'Économie de Paris, Johann Wolfgang Goethe-Universität Frankfurt am Main). The economic historical work in this Work Package include creating "specifications" or documents that explain the structures and contents of each type of source (to be used to "teach" the artificial intelligence how to read the data) and providing other input to improve the systems (such as described in the section "Yearbook experiments" below) as well as verifying that the tools' test results show correct extractions of the data.

The French Desfossés Yearbook 1962, the Belgium and French official price lists have been digitized. For the Handbuch Yearbook 1913-14 the consortium has decided (at the Steering Committee meeting of 20 April 2019) to use a re-edition of it which is available in digitized version. No legal problems should cover the use of this version during the experiment phase of Work Package 7. At the time of writing, digitization of the Madrid bulletin and price lists is running.

We started to work on some of these sample tests: French Yearbooks (Desfossés, 1962), and on French Official Price Lists, to specify and design elements of the software libraries: a library of document component detectors for structure recognition and a general-purpose text recognizer. These constitute deliverable D7.1 and will be used to build different prototypes of document recognition and understanding systems adapted to different types of documents. We have also begun some preliminary work on the German Yearbook (such as the extraction of one of the rubrics/entities [the Capital entity]).

Deliverable D7.1: Document Components Detectors for Structure Recognition – Task 7.2 (IRISA)

We developed a library containing document component detectors and tools used for recognizing structures on different kinds of corpuses, yearbooks and stock price lists, including table rulings, table separators without rulings, text lines, contextual segmentation of text lines, connection with a commercial OCR (optical character recognition). This library is shared with a French national project (HBDEX (Exploitation of Historical Big Data for the Digital Social Sciences: application to financial data, funded by the *Agence nationale de la recherche*) focusing on the lists of the Paris over-the-counter (OTC) market (the *Coulisse*); thus, part of this work is done in this context. It is planned that this joint library will keep evolving throughout both of the projects. This library has been developed using the DMOS-PI method, a generic multi-resolution method for document collection recognition, with perceptive vision mechanisms and interactive process.

Deliverable D7.1: General-purpose text recognizer (OCR) – Task 7.3 (LITIS)

We design our own Deep Learning based-OCR platform. It is built on convolutional neural networks (CNN) combined with bilateral recurrent Neural Networks layers. Training is performed using the CTC (Connectionist Temporal Classification) loss function. Outputs can be parsed (Viterbi Beam Search) using different language models depending on the context of use within the document. Language models can encode lists of possible stock names, or the syntactic rules used to write specific information such as prices, dates, etc.

Yearbook experiments

IRISA developed a structure detection system for the Desfossés yearbook, using the software library for document components, by defining a grammatical description of company sheets (Volume 2) to detect company name, the different rubrics organized in paragraphs for each company. Another description has been done for the table of all the administrators (Volume 1). We are currently working on the extension of the library by retraining the deep learning line detection to improve the detection of text lines in balance sheets, to then improve the detection of left and right alignments. Indeed, they are important to understand and to correctly extract the balance sheet contents. An application to a new yearbook (like the Handbuch) will be done only by defining the physical specificities while keeping the logical description, which will be common to different kinds of yearbooks. These structure detection systems will be combined with LITIS's Named Entities Extraction process presented thereafter.

LITIS has implemented a generic pipeline of processes that can run similarly on the various Yearbooks that are considered within the consortium. The inputs of the pipeline are images of documents and the outputs are information attached to each company the Yearbook is reporting. The information extracted is structured in rubrics composed of lists of named entities (i.e. list of person names, as is the case of the “governing board” rubric), or list of linked named entities (i.e. [date, amount, currency] as is the case of the “capital” rubric).

This pipeline (see Figure 1) is composed of OCR followed by text analysis for Named Entities extraction and linking. This pipeline only considers running text and ignores table structures. Table extraction is conducted by IRISA within Work Package 7. For the experiments conducted so far, a general-purpose industrial OCR was used and proved to give sufficiently good results so that LITIS mostly concentrated on the extraction process of named entities in yearbooks for implementing the whole pipeline.

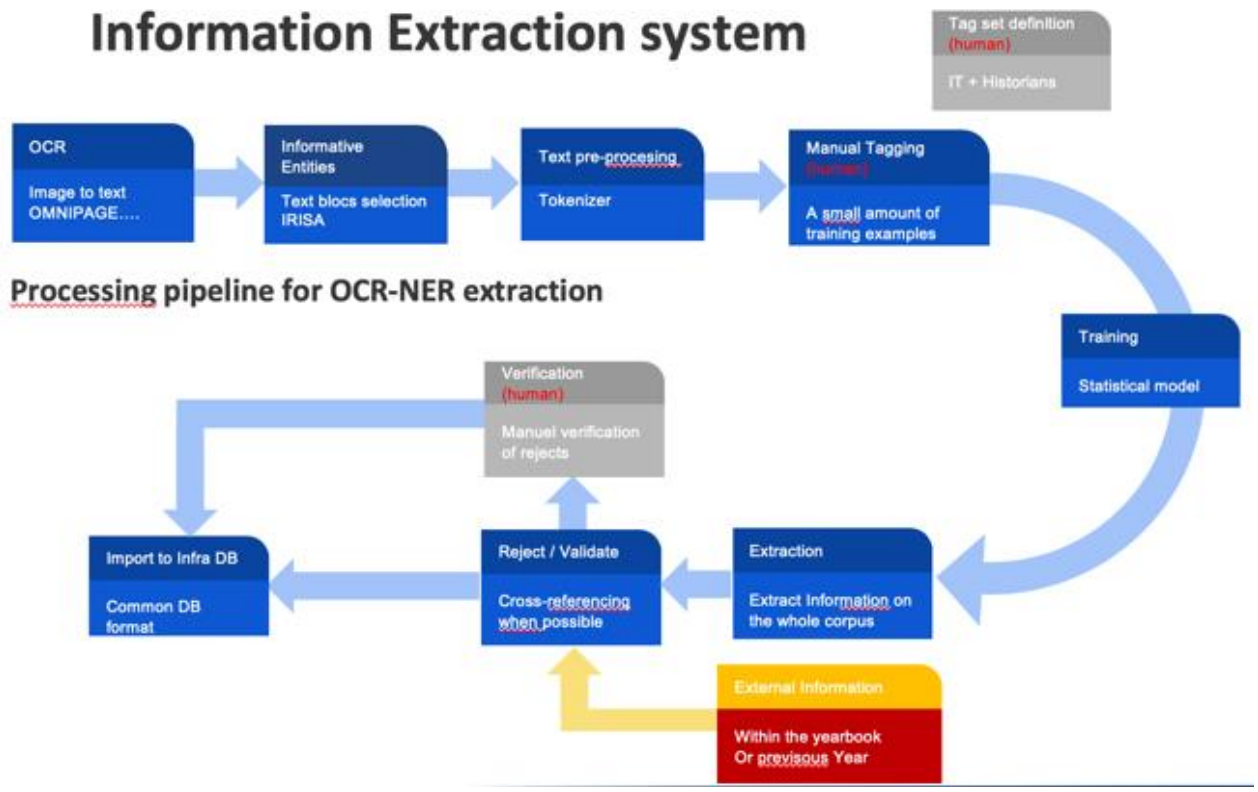
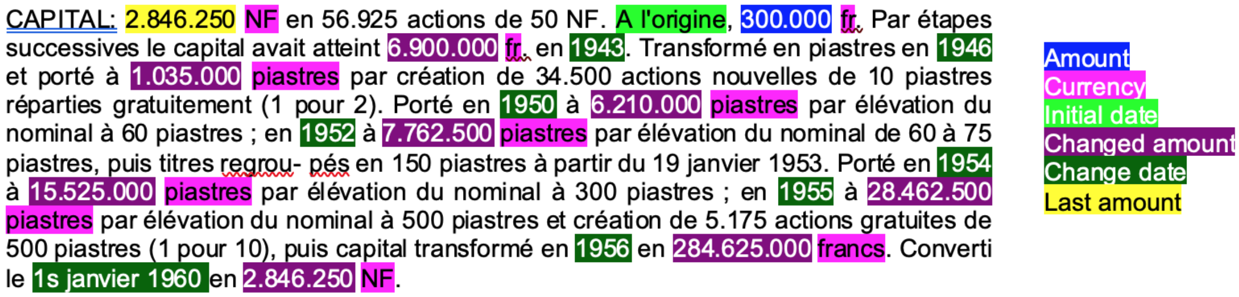


Figure 1: processing pipeline for OCR-NER.

For the sake of genericity, the Named Entities Extraction process bares on machine learning technologies. This provides a general scheme for the design of AI software. First, the economic historians (experts) are asked to tag some text samples by highlighting the different categories (classes) and sub-categories of information to be extracted. Naming conventions for tags are specified first with the IT team. Tags are chosen sufficiently generically to apply to all the corpora as much as possible (see Figure 2 for one example of tagging and naming convention, in the case of “Capital” rubric on the Desfossés corpus).



CAPITAL: 2.846.250 NF en 56.925 actions de 50 NF. A l'origine, 300.000 fr. Par étapes successives le capital avait atteint 6.900.000 fr. en 1943. Transformé en piastres en 1946 et porté à 1.035.000 piastres par création de 34.500 actions nouvelles de 10 piastres réparties gratuitement (1 pour 2). Porté en 1950 à 6.210.000 piastres par élévation du nominal à 60 piastres ; en 1952 à 7.762.500 piastres par élévation du nominal de 60 à 75 piastres, puis titres regrou- pés en 150 piastres à partir du 19 janvier 1953. Porté en 1954 à 15.525.000 piastres par élévation du nominal à 300 piastres ; en 1955 à 28.462.500 piastres par élévation du nominal à 500 piastres et création de 5.175 actions gratuites de 500 piastres (1 pour 10), puis capital transformé en 1956 en 284.625.000 francs. Converti le 1s janvier 1960 en 2.846.250 NF.

Amount
Currency
Initial date
Changed amount
Change date
Last amount

Figure 2: Example of tagging a "Capital" rubric on the Desfossés corpus, and the associated naming convention.

Experiments have been conducted on the Desfossés corpus to extract Company names and the information content in six of the most important rubrics reported in the yearbook using this general framework, while only the Company names and Capital rubric have been processed on the Handbuch Yearbook. We obtain very promising results on both corpora which demonstrate the genericity of the methodology adopted among corpus and languages. We also show that the level of performance of the extraction process depends on the quantity of data initially tagged manually by historians (order of magnitude of 200 tagged examples per rubric, for the most difficult rubrics, to obtain good results).

Price List Experiments

Using the software library which is built in common with HBDEX, we defined a structure detection which has been applied on the French Official List. The structure recognition is able to detect typed columns, the price sections and the price lines (Figure 3). It is also combined with the general-purpose text recognizer. We are currently working on a generic strategy for integrating the day sequences to detect stabilities in the column structure and later on the names (column titles, section title, prices names and values).

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

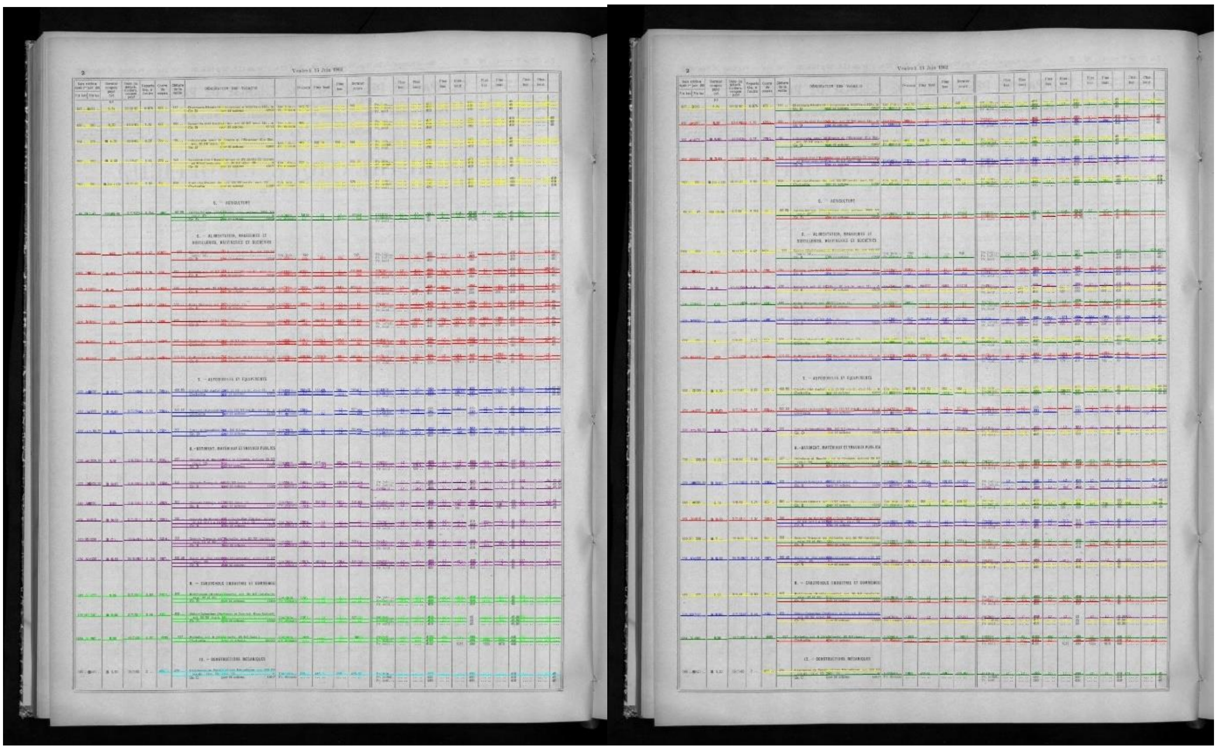
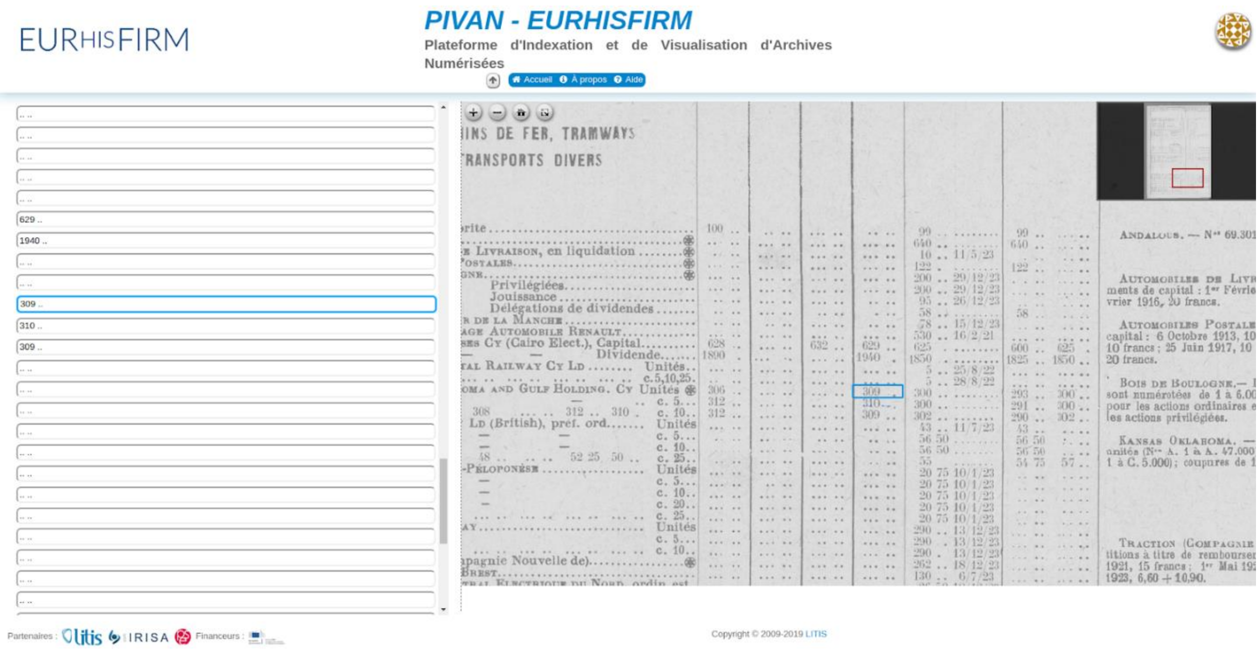


Figure 3: Example of structure detection on French Official Price Lists 1961-1962 with columns, price sections (left), price lines (right)

Man-to-Machine Interface to Browse Correct and Validate extraction results – Task 7.3 (LITIS)

We designed a first prototype of a web visualizer to browse the OCR results and the information extracted from a collection of documents. The visualizer allows editing of the textual content where human correction of the transcriptions is needed. When alerts are generated by the system (e.g. "309" field on Figure 4), the GUI shows the values of this field for the day before and the day after (see Figure 5 below) in order to help the operator (human interference) to validate the OCR output. In this example, the operator can read the "preceding day" value of the stock "KANSAS OKLAHOMA...." on D+1 image to correct the closing value of day D.

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


EURHISFIRM
PIVAN - EURHISFIRM
Plateforme d'Indexation et de Visualisation d'Archives Numérisées
Accueil
À propos
Aide
Partenaires : litis IRISA Financeurs :
Copyright © 2009-2019 LITIS

629 ..

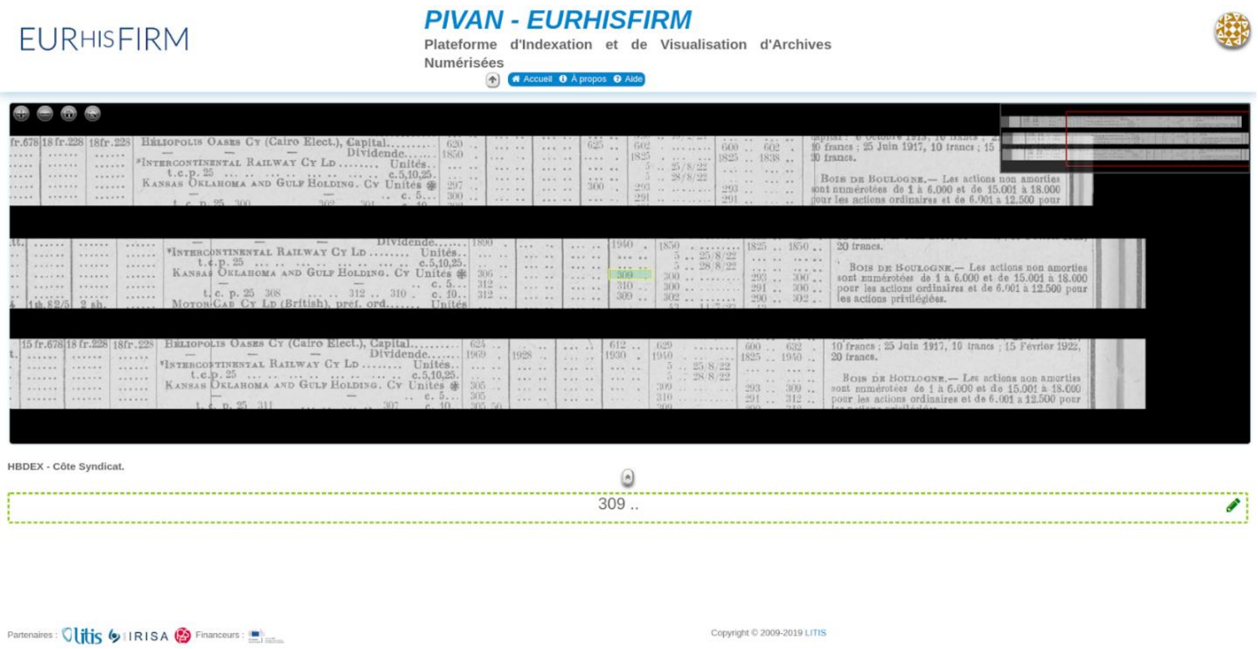
1940 ..

309 ..

310 ..

309 ..

Figure 4: GUI highlighting on possible miss-recognized field in the table.


EURHISFIRM
PIVAN - EURHISFIRM
Plateforme d'Indexation et de Visualisation d'Archives Numérisées
Accueil
À propos
Aide
Partenaires : litis IRISA Financeurs :
Copyright © 2009-2019 LITIS

HBDEX - Côte Syndicat.

309 ..

Figure 5: GUI showing the values of the field the day before and the day after to assist the operator in validating the question highlighted.

Automated Linking of Named Entities – Task 7.4 (HeBIS)

As part of the evaluation of linking tools, task 7.4 has already worked on and with existing datasets from EURHISFIRM project partners, external data sources and authority files. Linking of company data between data sets from Germany (SAFE - Sustainable Architecture for Finance in Europe), Belgium (SCOB) and France (DFIH) and with authority files GND and DBPedia showed that manual work is still needed to gain a high quality of links between data sets.

Work Package 8: Interaction with users

The objective of Work Package 8 is to determine the optimal design of the data and services that EURHISFIRM RI should provide by gathering and analysing the preferences of potential end-users and key stakeholders (academics, practitioners, regulators, etc.). A large-scale survey, via an online questionnaire, was developed in order to identify the preferences of potential end-users and key stakeholders of the EURHISFIRM project (D8.1, submitted August 2018). The survey was conducted and its results were analysed (D8.2, submitted January 2019). Building on the survey, the Work Package identified qualified people and conducted semi-structured interviews with them regarding their perspectives and preferences for the design of EURHISFIRM. The findings of these interviews were then shared with the rest of the EURHISFIRM group (D8.3, submitted August 2019).

Work Package 9: Infrastructure policy and architecture

Work Package 9 designs the architecture and the operation of the RI, with regards to access, security, support and maintenance, in cooperation with ESFRI (European Strategy Forum on Research Infrastructures) Landmark CESSDA (Consortium of European Social Science Data Archives). Users' preferences on data and service design guide the platform's architecture and operation. Accordingly, the security system, the maintenance and the desk management of the platform are designed and estimated. The platform's architecture and operation are compatible with the National Focus Points' and site's policies.

The Work Package also assesses the optimal level of integration of EURHISFIRM with existing RI such as CESSDA and DARIAH, following the user requirements' specification and RI policies. As per EURHISFIRM agenda, the Work Package is to officially commence late 2019, so it is in a very preliminary stage at this point. Most efforts have been put into designing policies and requirements for usage of the future system. Aspects such as user types; access rights; GDPR consequences; classifications for Confidentiality, Integrity and Accessibility have been proposed. The basis for this work comes from materials from CESSDA, standards such as ISO 25010 (software quality), and reference materials for Research Infrastructures such as EOSC and FAIR.

Work Package 9's output will contain a description of required functionalities with as much detail as possible based on the input from other Work Packages. In accordance with CESSDA, an estimate will be made for the required technical infrastructure.

Work Package 10: Business model and governance

The aim of Work Package 10 is to develop a business and governance model that contributes to the articulation of EURHISFIRM's platform design. Significant progress has been made concerning two tasks: 1) the definition of alternative business model concepts, and 2) a preliminary assessment of business and governance model alternatives. While business and governance model can vary across countries, the main source of revenues for data repository is large structural public funding.

In March 2019, a call for applications to a position of Business & Strategic Planning Manager was published and the selection process (including personal interviews) was successfully completed. The hired professional will be required to engage in the following activities: survey existing concepts of business and governance models for Research Infrastructures in general, as well as data repositories more specific to the social sciences and humanities; establish contacts and manage exchanges with experts running Research Infrastructures and policy-making bodies, such as ESFRI (European Strategy Forum on Research Infrastructures); establish contacts and manage exchanges with EURHISFIRM's prospective stakeholders (academic institutions, scholars, research funding agencies, public institutions, private companies); negotiate prospective funding commitments with public and private research infrastructure funding institutions; cooperate with other Work Packages within the EURHISFIRM project on all issues related to the formulation of the business and governance model; design EURHISFIRM's long-run business plan based on estimated costs and streams of revenues that ensure its financial sustainability; draft preliminary and final reports on EURHISFIRM's business and governance model; attend EURHISFIRM's meetings.

Work Package 11: Cultural heritage

Work Package 11 explores concepts and tools to stimulate the lasting conservation of the digitized material and provides guidelines for making those materials publicly accessible. It also explores innovative ways to use digitized images as documentation for the data extracted from them and evaluates alternative strategies to use digitized material. More specifically, Work Package 11 has three main objectives:

- The use of digital images to document data and inspire further research and Identify sources of interest for cultural heritage;
- The promotion of Europe's cultural heritage by facilitating digital preservation and online accessibility of sources with a unique historical value;
- The mobilization of digitized images of historical sources as an exceptional additional documentation for the data (including the exploration of ways to make materials accessible and connected to EURHISFIRM data).

Officially, i.e. as per the EURHISFIRM planning found in the "Deadlines and Meetings" document, Work Package 11 is to commence with Task 11.1 (Evaluating strategies and practices to value cultural heritage) in November 2019. Steps taken so far are all in preparation for the official start of the project sections. Logistical and operational progress have thus been carried out to support and develop future goals and milestones. Over the coming weeks, work will be carried out to prepare the first milestones and deliverables (i.e. D11.1 Strategies and practices to value cultural heritage).

IV. Conclusion

This document reported on the results achieved so far from the beginning of the EURHISFIRM project until its midterm (1 April 2018 – 30 September 2019). As stated in sections I and II, the EURHISFIRM project addresses the crucial need for *reliable and standardized long-term company-level data in Europe* and will benefit European citizens/society, the scientific domain, public policy and public organisations, and the private sector. The EURHISFIRM project designs a research infrastructure compatible with open science and FAIR data principles that will provide a comprehensive platform to access heterogeneous European long-term company level data in a standardised, reliable, scientifically sound and technologically advanced way.

To accomplish this goal, EURHISFIRM is comprised of 11 Work Packages that can be roughly grouped into the following categories: economic history foundations, Information technology work (data extraction technologies, common data modelling, infrastructure architecture), practical aspects of the infrastructure operations (business plan [including user target research], legal plan, and cultural heritage of the data produced), project administration (communication, community building, logistics and strategy/vision [including compliance to open science frameworks and FAIR data]). These Work Packages continue to progress in accordance with the timeline set out in the project proposal.

For the remaining second half of the project term, EURHISFIRM will continue executing the remaining tasks according to the project proposal and the deliverable/milestone schedule. At the same time, EURHISFIRM will prioritise a further sharpening of the project vision, in relation to our own ambitions for the project itself as well as in relation to the project's place in the overall European research infrastructure landscape overall in order to help advance and promote this community.