Summary of the first annual EURHISFIRM Project Advisory Board and General Assembly meetings

15-16 March 2019, Wroclaw University of Economics

Participants: Project members and leaders, the Project Advisory Board (PAB), and external guests

NB: This is not an exact minute-by-minute record of the meetings; some parts of the discussions have been grouped into different themes for clarity.

Project Advisory Board (PAB) meeting

Presentation by the principal investigator Angelo Riva (Ecole d'Economie de Paris) on behalf of the executive committee (Jan Annaert [Universiteit Antwerpen] and Wolfgang König [Goethe-Universität Frankfurt am Main])

Presentation of main achievements

1. Human Resources: there were delays in the project progress in the beginning due to recruitment issues, but this is currently OK.
2. Economic history: types and origin of sources from the consortium countries have been analysed in Work Package (WP)4.
3. Transferring the report into a paper: we should be prepared for this.
4. Technical work: linking between Antwerp (**S**tudie**c**entrum voor **O**nderneming en **B**eurs [SCOB]) and Paris (**D**onnées **FI**nancières **H**istoriques [DFIH]) databases has started in WP6; standardisation and common data model has started in WP5.
5. Utility to users: survey was completed by WP8.
6. Communication: website created, and the communication plans have been delivered. Going forward, we need more contributions from the entire consortium for the communication tasks, such as the website contents. We are also trying to setup a web forum for discussion.

Main challenges

1. Coordination among inter-WPs: especially information technology (IT) tasks.
    - We need more frequent and complex interactions among WPs4-7, 9.
2. Optimal degree of the decentralisation.
    - We want to store data at the local sources but also a central source to manage them.
3. Community building: we need to increase the number of users.
    - Plan of community: culture of historical data needs to be increased with professional associations (perhaps to build together).
4. The first version of the project should be done 6 months earlier than the official deadline.

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

Discussion on the key questions

1. Prioritising the design study's dimensions and target results of interest for stakeholders.
2. The increasing importance of building a scientific community (also to prepare for the H2020 Integrating activities).
3. Overcoming challenges in obtaining funding for data collection purposes.

Feedback from Mike Staunton (London Business School), Member of the PAB

- Progress and results need to be more evident for WP7 (not enough advancements shown and the process includes too much manual interference). It seems that WP7 has started with the easier documents (the yearbooks). Task 7.2 for the moment focuses only on French sources.
- Report on sources is not comprehensive.
- There is currently a lack of coherence among some WPs (such as the disconnect between WP8's survey results and the sample selections for WP7).
    - Additionally, in the samples, UK data are not included.
- We should also be building our own indices instead of relying on external ones, starting from the index of the yearbooks to make lists of corporations.
- Proposals
    - Tasks should be more achievable. Perhaps the goals are too lofty and general.
    - The shift should change to being more user-centric and selective (in terms of data to analyse).
    - Focus more on scanning of sources.
    - Today the project is built around an OCR (optical character recognition) platform to harvest all market data to be inserted into a database. The focus should more user-oriented, more selective about the data to be analysed, more oriented toward the scanning of sources. The priorities should be equity prices to build stock indices documented with references to scanned sources.

Responses to Mike Staunton (Member of the PAB)

- Jeff Braswell, external consultant for WP5, says that the project is in the design phase, and this should not be confused with implementation; this is why we can only run tests currently. Moreover, even in the implementation phase, it will be impossible for EURHISFIRM to massively scan sources. It is not the purpose of an infrastructure. We set the stage to receive scanned sources and data scanned and digitalized by others. To do that, the project must anticipate the organization of data and sources.
- Francis Gross (European Central Bank), vice president of the PAB, says that the goal of the design phase is to test many ideas and concepts. To this end, the project must "play with real tools" as much as possible, but the project cannot skip work on the representations of concepts and languages. The European Infrastructures last for decades. We see the conflict between two approaches: a "pessimist" approach (the goal of the project should be to produce useful results like equity indices in the short run) and a more "visionary approach" (setting an infrastructure that will work for decades, then setting the stage for a long-run inclusive work). The OCR platform will work because technology advances faster; there is no doubt for that. It may be important to adjust the program slightly and produce some output in the short run to "convince pessimists".
- Angelo Riva (Ecole d'Economie de Paris), principal investigator, says that OCR tests are also "easy wins" and more achievable. This is why the starting point for the OCR treatments are yearbooks

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

and largely French sources. The selection of the samples to be treated in the design study has not been done on the basis of what the future users would prefer (see WP8's survey results), but on the availability of high quality scanned images and variety of languages/format. The goal is to develop a generic tool (rather than a custom tailored one for each source) so that it can easily adapt to many other kinds of sources. In WP6, the first step is being tackled, which is the database merge. Within this framework, the project can produce real outputs in the short term. Additionally, in line with the INFRADEV program, the funds are for the data infrastructure only, for the container; this cannot be used to finance data collection or scanning (the content): data and sources will be brought into the infrastructure by scholars and professionals. This is why we must set up a "general" infrastructure. Clearly the preferences of users must be taken into account in the prioritization of the works; however, this comes in the implementation phase. Regarding putting the sources online, WP10 is devoted to that. It is crucial. This is why within the French DFIH project for example, sources are already available on the web. Technology is a risk, but so is data entry by hand. To foster data collection on European firms over the long run we have to study technologies that make collection faster and cheaper. The aim of the inventory of the sources from WP4 is not to be comprehensive; this inventory wants to make a survey of the main serial sources used for data collection on listed companies used in the countries member of the consortium to set the stage for the meta-data and data format to be developed in further phases of the project. Such a comprehensive inventory would be impossible to do within the framework of a WP. The inventory produced within WP4 will be enriched thanks to the work of a community of economic and financial historians.

- Bertrand Coüasnon (INSA Rennes), WP7 leader, adds the following points:
    - The delays in the progress are due to the delays in the Human Resources hiring.
    - Regarding the OCR, this requires understanding the semantics of the text and cross validation. The challenge is not to improve the OCR technology in itself to recognize specific characters (e.g. the percentage sign (%), etc.). The way to increase the performance of the OCR is to embed knowledge about the sources from the data experts into the platform. The recognition of symbols, such as the percentage sign, is a minor issue.
    - There are plans to use the index to cross check the correctness and completeness of the information.
    - Regarding the progress so far, such as the percentage sign: this was the first results of the OCR/deep learning. It just needs more data to train the algorithms; it is indeed a bootstrapping process.
    - The "low" rate of results may have been an issue of presentation/communication: 40% is the success rate when taking into account the specific area mentioned alone. If we consider the document areas as a whole, these percentages are much higher.
    - For the document selection, mixed degrees of degradation and quality were chosen because printing qualities are different for different eras.
        - Mike Staunton comments that he has a photocopier for digitising and offers to let us use it for testing the UK data.
- Thierry Paquet (Université de Rouen Normandie), collaborator of WP7, also adds that there was no intention of designing an OCR technology through this project. The project's intention is to rather design an informational extraction system, and this extraction system has different

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

elements. Regarding the progress, the report currently only shows the primary results to show that it has been developed.

Synthesis of PAB discussion to the General Assembly (GA) by Leslie Hannah, President of the PAB

- Has significant progress been made?
- What are the goals to focus on going forward?
    - Is the OCR part too ambitious? There were both pessimists and optimists
    - Or should we actually be more ambitious?
- An important question concerns the funding for data collection. The INFRADEV program finances the infrastructure (container), but is it also important to take into account the need of funding for data collection? Should EURHISFIRM rely only on the data collected by the community, without concerning itself about the priorities and the timing of this data collection? Should the EURHISFIRM members undertake actions to raise funding for data collection in line with the EURHISFIRM goals and priorities?
    - Public institutions are not willing to fund data collections "per se" (i.e. to make data available for further research). Currently, international agencies will only fund data collection in order to answer specific research questions.

Synthesis of different discussions grouped by themes

The below is the synthesis on the series of discussion from both the PAB and GA combined by themes for coherence and ease of comprehension:

- **General approach/strategy to appeal to possible stakeholders**
    - Francis Gross (European Central Bank), vice president of the PAB, mentions the need to balance between pragmatism and being visionary. We should include practices such as experimentation: we should be failing and learning fast.
    - Gross also emphasizes that language (communication) is important: how do we represent/communicate the project to external audiences and future potential community participants?
    - Gross mentions that we should also review and adjust the road map of the WPs.
        - OCR is a new technology and it is progressing very quickly. We need to consider this aspect in the road map.
    - Leslie Hannah (London School of Economics), president of the PAB, mentions that the proposal and the annual report assumes that the US CRSP data are currently superior; but according to his opinion, DFIH's and SCOB's depth of coverage is better.
    - Ron Dekker (Consortium of European Social Science Data Archives [CESSDA]), external guest to the EURHISFIRM General Assembly, mentions that a big obstacle is reaching policy makers in the best way. We should engage their interest by emphasising the importance

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

of historical context in order to measure the current productivity. For example, the Venice Archives have made a visual representation of data [in order to appeal to the public].
    - Dekker also asks if we want collections of data in and/or outside of academics.
- **Resources and execution**
    - Project management methodology: debate between waterfall and agile methods (such as scrum). Oliver Watteler (GESIS—Leibniz Institute for the Social Sciences), project member, says that normally, for IT projects, agile methods are better suited, and this is proposed as a possible way to handle the forthcoming tasks. However, Wolfgang König (Goethe-Universität Frankfurt am Main), Executive Committee member and WP5 leader, points out that as this is an EU project with linear deadlines, it would not be easily possible to integrate agile methods for now.
    - The possibility of speeding up the project by 6 months is also brought up. The official project submission deadlines would not change, but internally, we would need to shift the deadlines if we agree to implement this change. Wolfgang König (Goethe-Universität Frankfurt am Main), Executive Committee member and WP5 leader, asks the question how we should implement this in reality/how to speed up?
    - Jesús Freire Costas (IBM Europe), PAB member, mentions that there are there are many resources in commercial and research sectors available for free (such as at IBM) including tools that can handle natural language processing.
    - Francis Gross (European Central Bank), vice president of the PAB, also suggests recruiting a PhD/group project to help with developing the technical infrastructure, if funding is available.
- **User-centric approach/importance of building a scientific community**
    - Leslie Hannah (London School of Economics), president of the PAB, also warns (contrary to the prior comments from Mike Staunton, PAB member) that being too user centric could be too narrow a vision. However, Francis Gross (European Central Bank), vice president of the PAB, then counters Hannah by mentioning that we need to think about what users will need in the future. We need to find users who will be keenly interested. We should find scientists, as well as linguists and philosophers.
    - Oliver Watteler (GESIS—Leibniz Institute for the Social Sciences), project member, also mentions that feasibility is as important as catering to users' needs, as well as long-term preservation of the data.
- **Overcoming challenges in obtaining funding for data collection purposes**
    - Francis Gross (European Central Bank), vice president of the PAB, mentions that policy makers are not always aware of research priorities. So we ourselves need to make them aware of the significance of data collection (so that we could increase the funding for that). We need to be more aggressive with the incentives. In general, a shift in paradigm regarding data use is necessary. Central institutions and financial institutions/academics do not recognise this. It may also be worth considering think tanks that would be relevant for these topics.
    - Joanna Sławatyniec (Rotterdam School of Management, Erasmus University), WP9 and WP11 collaborator, mentions that some private sectors (such as financial institutions) could be interested in the data; but rather than being interested in the utility, they could be interested in its value for marketing purpose such as patronage because "big data" and

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

related topics are trending issues in all sectors. This could be a way to grab the institutions' interest (and to encourage them to fund).

- Referring to WP10's report in D1.3: First yearly progress and strategy report to the General Assembly, Angelo Riva (Ecole d'Economie de Paris), principal investigator, suggests running a meta-analysis on the papers of economic history to see what kind of data are being used (to better assess potential users' needs and to better align towards funding criteria) and correct the recognized bias in results of the surveys, in addition to experts' interviews.
- Amélia Branco (Universidade de Lisboa), external guest to the EURHISFIRM General Assembly, mentions an organisation which provide trainings and tools on how to use the data to stakeholders and possibly interested future people.
    - Following up on Riva's meta-analysis idea above: after running an analysis of the contents, we could e.g. run workshops that explain how stakeholders can use the data to their benefit.

- **Dimensions of the design study to appeal to stakeholders (decentralisation)**
    - Federalist system (decentralisation): what is the optimal level for our project?
        - Centralised system optimises the unit cost, but it is not flexible
        - However, a decentralised system offers flexibility and increases local involvement and preserves the particularities of local institutions, but with higher costs than a centralised system
    - Resources for managing perennial archives
        - Sébastien Oliveau (PROGEDO), external guest to the EURHISFIRM General Assembly, says there are 2 versions:
            1. National archives (or also more "decentralized units at the level of Universities)
            2. EU archives: could also be an option for EURHISFIRM
        - Oliver Watteler (GESIS—Leibniz Institute for the Social Sciences), project member, mentions that CESSDA acts at policy level, but GESIS as a CESSDA Service Provider acts as a single point of longer term preservation" [GESIS is working on remote access for certain collections, but does not offer remote access points per se.]
    - Should we have one place to manage the data? is it possible that some systems are hosted by single institution (university)? If yes, to what extent and what degree of federalism should we imagine?
        - We would need one place for hardware, and this would need maintenance. For continued access, this would also need continuous funding
        - Francis Gross (European Central Bank), vice president of the PAB, suggests a system in which technology is maintained by one team with the network to access it. It should also be in multiple locations for safety. He also suggests that the "home" of the data needs budget. ECB can possibly be a place to be this home (or any other organisation).
    - Wolfgang König (Goethe-Universität Frankfurt am Main), Executive Committee member and WP5 leader, mentions that we should also collaborate with networks (such as CESSDA/DARIAH) to get new ideas and commit ourselves to the logical building.
    - Angelo Riva (Ecole d'Economie de Paris), principal investigator, says that we need the organisational support but also need local person(s) to manage. We need to start

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

discussions with national services providers (and CESSDA) and with policy-level organisations
    - Oliver Watteler (GESIS—Leibniz Institute for the Social Sciences), project member, confirms that CESSDA can help at higher levels

Presentations from external guests and PAB

Jochen Streb, Professor at Universität Mannheim

- Trends in economics research:
    - Going from macro-level (country) to micro-level (household or company) data
    - Topics such as innovation (patent data), social insurance, individual savings during the life cycle can benefit from micro-level data
    - Standardised micro-level information on publicly traded companies is also greatly needed

Erik Nowak, Professor at Università della Svizzera italiana

- Sharing of his experience on bankruptcy database building
    - Used OCR to read texts from the newspaper
- Switzerland should also be a part of the EURHISFIRM project because it has a huge equity market

Francis Gross (European Central Bank), vice president of the PAB, mentions that Switzerland could serve as an interesting benchmark for EURHISFIRM because the country was not interrupted by war.

Ron Dekker, Director of CESSDA (Consortium of European Social Science Data Archives)

- Dekker presents that there are several trends:
    - European open science cloud (EOSC)
        - Someone mentions that currently EOSC is currently top-down approach (i.e. centralised?) However, Brussels also want bottom-up initiatives
    - EC wants to combine the data infrastructures
        - They also want to get rid of silos, which should be more demand-oriented
        - The products should host small and viable communities that are improved
    - The current news indicates that the processes want to integrate into EOSC (ESFRI [European Strategy Forum on Research Infrastructures] → ERIC [European Research Infrastructure Consortium] → EOSC). This is a signal that the countries want to invest. E.g. Germany has invested 90 million € for the coming 10 years.
    - Question: where and how to organise the RIs? By nationalities or by categories?
- Platforms: all digital data will end up on platforms. The essential elements include:
    - Must have persistent identification freedoms in order to find data
    - Meta data
    - Historical repositories
    - Niches: secured platforms in social sciences, life science, other categories, etc.
    - Amalgamation across disciplines
- Presentation of CESSDA: a consortium of social science data
    - Mission is to distribute and sustain research data
    - Human elements are important such as teaching, training, etc.

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

- Benefits
  - Depositing of data
  - Visibility and credits
  - Compliance with founder requirements
  - Training
- Strategy
  - Training, technology, tools

Pierre-Cyrille Hautcoeur, L'École des hautes études en sciences sociales (EHESS) and President of the European Historical Economics Society

- Advice and keys to success based on his experience building the French financial database (DFIH)
  - Good partnership
  - Flexibility
    - In partnerships: such as adding and removing new partners
    - In technology
    - Adapting to shocks and unexpected events

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