Long-term data for Europe
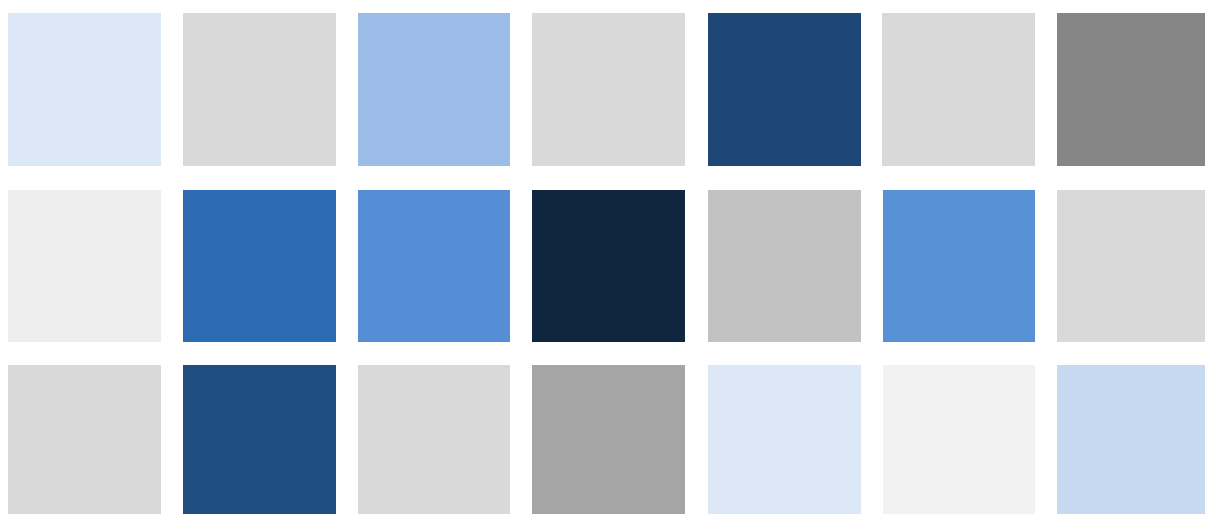
EURHISFIRM

AUTHOR(S):

Robin ADAMS *(Queen's University Belfast)*

Gareth CAMPBELL *(Queen's University Belfast)*

Christopher COYLE *(Queen's University Belfast)*

John TURNER *(Queen's University Belfast)*

APPROVED IN 2019 BY:

Jan ANNAERT *(Universiteit Antwerpen)*

Wolfgang KÖNIG *(Goethe-Universität Frankfurt)*

Angelo RIVA *(École d'Économie de Paris)*

Table of Contents

Summary of Recommendations:

Based on an online questionnaire surveying the considered opinions of 121 potential users of the EURHISFIRM platform (D8.2, carried out between October 2018 and January 2019) and qualitative interviews with 12 potential users and stakeholders (D8.3, carried out between June and August 2019), we make the following recommendations for the development of the EURHISFIRM project:

1. We recommend that priority should be given to data relating to the twentieth century, as this was the most popular time period among our respondents.
2. Regarding forms of company data, we recommend that priority should be given to ordinary equity market data because it was by far the most popular with respondents.
3. There was also a desire expressed for accounting data, specifically, total assets, total debt, revenues, and profits.
4. Respondents were less keen on data related to government and corporate bonds.
5. In terms of frequency of share prices, we recommend daily and monthly data take priority over weekly or yearly data, where possible.
6. As regards geography, the United Kingdom was the most popular country, followed by Germany and France.
7. Turning to usability, we recommend that the EURHISFIRM platform should as far as possible enable users to manipulate the data themselves. Specifically, users should be able to download the data in bulk, in MS Excel format, and with minimum restrictions on downloads per day. We recommend that the EURHIFIRM platform use Wharton Research Data Services as an example of best practice in this area.
8. However, for less popular data, we recommend that EURHISFIRM provide simple, non-tabulated PDF scans of the original source document, where possible.
9. For reassurance as to the accuracy of data, users should be able to 'click through' to a scan of the original document and EURHISFIRM should provide a full citation of any source material.
10. We also recommend that EURHISFIRM provide an explanation of the methodology and rationale for any interpretation or manipulation of data carried out by EURHISFIRM researchers.
11. Finally, regarding feedback, an email address would be sufficient.

Background, objectives, and method

Online questionnaire (D8.2)

The D8.2 survey sought to explore the preferences of potential users and stakeholders, enabling effective prioritisation and design of EURHISFIRM's data and services.

The survey was carried out from 9 October 2018 to 7 January 2019 via online questionnaire. There were 125 respondents to the survey, 109 (89%) of whom were academics, the remainder government employees (4), corporate employees (2), students (4) and other (3). A further three declined to answer. Of the academics, 67 were in finance, 22 were economic historians, nine were economists, four were in accountancy and two were historians. The majority, 72 of respondents specialised in finance, while 30 were historians or economic historians. A further 11 were economists, the remainder focussed on accountancy (4) or business (1).

Of the 119 respondents who gave their nationality, 38 (32%) were from the United Kingdom, 9 from France, 8 from Poland, 5 from Ireland, 5 from the United States, and the rest from the Netherlands (4), Belgium (3), Germany (3), Switzerland (3), Australia (2), Italy (2), Portugal (2), Spain (2), Sweden (2), Denmark (1), Hong Kong (1), Mexico (1), and Norway (1).

Interviews (D8.3)

Building on survey results analysed in D8.2, we sought to explore further the preferences of potential users and stakeholders.

We approached 35 potential respondents, identified by referral from members of the EURHISFIRM advisory committee and thereafter by snowball sampling. In response to feedback from the questionnaire conducted in D8.2, particular care was taken to include interviewees from outside the UK and beyond academia. Mindful of this, we spoke to 12 potential users of the EURHISFIRM data platform, including academics (5), practitioners (7), and database experts (3), some of whose expertise transcended these categories. Geographically, interviewees offered expertise and experience of Austria (2), France (3), Germany (3), Poland (2), Switzerland (1), and the UK (3); again, with some overlapping categories.

The interviews were conducted in a semi-structured form, over telephone and Skype, in June, July and August 2019. Interviews were, in the main, 15-20 minutes in length, and were conducted on the understanding that results would be anonymised and personal data kept confidential, in line with the Queen's University Belfast research ethics policy.

Findings

The findings from D8.2 and D8.3 can be divided into two categories: The first explores respondents' preferences regarding content to be provided, and the second their preferences regarding usability of the platform.

http://www.eurhisfirm.eu

Content

Periods, geographies and companies of interest

Respondents were asked which periods would be of most use to them in their research. The most popular periods were 1900-49 and 1950-99, which 95% and 97% of respondents reported would be quite useful or very useful. However, when asked which period they would find most useful, the largest share, 43% of respondents answered 2000 onwards.

Respondents were given a choice of countries and asked for which of them corporate or financial data would be useful. The UK was the first choice of 50.42% of respondents and the second choice of a further 13.56%, although 32% of survey respondents were from the UK. The next most popular countries were Germany and France. In addition to the countries mentioned in the questionnaire, respondents suggested Austria (incl. Austria-Hungary), Greece, Ireland, Italy, Portugal, Russia, Scandinavia, and Switzerland. Nine respondents suggested all EU countries, EU15 countries, or Eurozone countries.

When asked which kinds of companies' price data would be most useful, most respondents (57.8%) restricted their preference to price data for all companies 'in a particular country'. Only 23.85% expressed a preference for the broadest option provided, which was price data for all companies in Europe.

Data of interest

Ordinary equity market data was considered the most useful form of company data, gaining the first preference of 36.75% of respondents and the second preference of a further 20.51%. Within ordinary equity market data, prices were by far the most popular data, with 57.94% respondents giving it their first preference. The next most popular first preference was adjusted prices, at 10.28%. Dividends gained the most second preferences, at 16.98%, followed by Trading Volume (11.32%), Mergers and Acquisitions (10.38%), and Bid-Ask Spreads (also 10.38%).

As regards frequency, almost 45.45% of respondents expressed a preference for daily data and 35.45% for monthly data, while only 13.73% and 6.36% preferred annual and weekly data respectively.

Considering securities that are traded on multiple exchanges, 49.55% of respondents stated that they would be most likely to use prices of all the markets where a given security was traded. Another 45.95% of respondents indicated that they would be most likely to use prices of the main market only, while only 3.6% gave the average price of the all the markets where a security is traded as the data they most likely would use.

Accounting data was the second most popular choice, gaining 24.79% of first preferences and 18.8% of second preferences. When asked which kinds of accounting data would be most useful, total assets, total debt, revenues, and profits were all rated at a similar level, but total assets attracted by far the most first preferences, at 38.79%. Total debt attracted the most second preferences from respondents, at 30.09%. Beyond the data listed in the questionnaire, some respondents suggested the division of revenue, assets and debt into their short-term and long-term components, while others simply called for access to all accounting data. When asked whether they would prefer accounting data to be adjusted or unadjusted,

the results were split: 38.94% of respondents indicated a preference for original balance sheet and P&L item names, while 37.17% declared a preference for uniformly adjusted balance sheets and P&L items.

Government or corporate bonds were identified as the first preference of 12.82% of respondents and the second preference of 29.91%. Respondents were asked which aspects of bonds they would find most useful to have data for. The majority (55.77%) indicated price as their metric of first preference, while yield to maturity was the first preference of 19.23% of respondents. Yield to maturity also made up the largest share (18.63%) of second preferences. Coupon rate gained 14.71% of the share of second preferences, and bid-ask spreads 12.75%.

Turning to macroeconomic data, this was the first preference of 11.11% of respondents and the second preference of 8.55%. GDP, inflation, exchange rates, and interest rates all received a similar amount of support from respondents, but GDP attracted the largest share of first preferences (34.31%). Interest rates were the first preference of 21.57% of respondents and the second preference of 30.69%. While inflation was the first preference of only 9.8% of respondents, it gained 25.74% of second preferences.

Governance data was the first preference of 7.69% of respondents and the second choice of 14.53%. Within governance data, the names of directors gained the most first preferences (27.36%), followed by board size (23.58%). While directors' pay received the least amount of first preferences (13.21%, joint last with Address of Headquarters), the largest share of respondents named it as their second preference (24.04%). In addition to those options provided by the questionnaire, respondents indicated that demographic information on directors would be useful, as would directors' shareholdings and information on the composition of board committees.

The respondents were invited to make other suggestions as to desirable content. The inclusion of measures of operating profit, such as EBIT and EBITDA, was suggested, as was the amount of tax paid by a company. One respondent suggested breaking assets down into short- and long-term assets. Other suggestions included the provision of information on mergers, demergers and name changes; historical ownership and bankruptcy information; and information on companies' patent ownership.

Usability

Datastream was the most widely used data service provider among the respondents to D8.2, with 69.57% of them claiming to have used it. CRSP/Compustat (53.91%), Bloomberg (52.17%), and Wharton Research Data Services (49.57%), were also commonly used. Wharton Research Data Services was the provider of preference for the largest share (21.9%) of respondents, followed by Datastream (19.05%) and CRSP/Compustat (18.1%). Other data service providers used by respondents included Thomson Reuters, S&P Capital IQ, Bureau van Dijk (Amadeus), and Morningstar.

The respondents were asked to rate their experience of the data service that they use most often. Bloomberg was ranked the best for ease of use, followed by Wharton Research Data Services and CRSP/Compustat. CRSP/Compustat scored best for data availability, followed by Bloomberg and Global Financial Data. For bulk downloads, CRSP/Compustat was considered the best data service, followed by Wharton Research Data Services and Bloomberg. Clarity was important to respondents, as was accuracy, and ability to manipulate raw data. Several respondents cited Wharton Research Data Services as best practice in usability and data availability. In the words of one respondent, 'WRDS gets it right - easy web interface, easy bulk downloads,' while another asserted, 'WRDS is a great example of how things should be', and another described it as 'a good service to emulate'.

The ability to manipulate data, check the accuracy of data, and give feedback were recurring themes in responses to both D8.2 and D8.3.

Manipulation of data

When asked about the consumption of their time in using data services, the respondents to the D8.2 online questionnaire cited the rearranging of data into a format that can be analysed as the most time-consuming task. Given a selection of popular data management programmes, the respondents were asked how useful it would be if a data service was integrated into each of them. Excel was rated to be the most useful, followed by STATA and R.

The respondents to D8.2 were also invited to share some examples of good practice that, in their experience, make data services easier to use. Of the 34 respondents who answered this question, 12 emphasised the importance of bulk downloads, particularly in MS Excel format. When the respondents were invited to share examples of things that make some data services difficult to use, bulk downloading again featured prominently in their answers. Of the 32 respondents who answered this question, 10 cited restrictions on downloading data as impediments to the use of data services; in particular the speed and quantity of data downloads. One respondent cited restrictions on downloading data via machines.

There were also a number of respondents who identified problems that are pertinent to historical research, such as gaps in the data or survivorship bias. Some respondents also mentioned specific problems to do with clarity, including a lack of clear definitions and labels, poor layout, confusing interface, too many options, and having to negotiate many screens. In this regard, one respondent identified Datastream as 'the best example of unnecessary complication'. Obstacles to matching and joining data from various databases were also cited as an impediment to the effective use of data services. In this

respect, two respondents noted the need for standardised identifiers. Only one respondent cited financial cost as an impediment to use.

Users' preference for the ability to manipulate the raw data themselves was also a recurring theme in the D8.3 interviews. An example given by one of our interviewees was the functionality to build charts; 'I never use the charts the platforms produce – I always just download the data and make my own'. This reflected a general preference among our interviewees for control. There was a consensus that the ability to manipulate data oneself, ideally by bulk downloads to excel, was very important. In the words of one interviewee, 'The closer you are to raw data the better'.

However, cognizant of limited resources, we also asked interviewees what they considered to be the minimum acceptable service provision from a platform such as EURHISFIRM. This revealed low expectations, reflecting the gap in the market EURHISFIRM seeks to fill. 'Most people are so grateful for long-run data,' argued one interviewee, 'they will take what they can get'. The interviewees from an academic background agreed that even un-tabulated PDF scans of original documents, if accurately labelled, would constitute a meaningful contribution. As one interviewee put it, this would be 'a huge step forward'.

Accuracy of data

Checking the accuracy of data was reported as the second most time-consuming activity in D8.2. Information on the data collecting methodology was considered most useful in resolving this issue, and providing a link to the original image was also considered to be of use. This concern over the accuracy of historical data was highlighted elsewhere in the questionnaire: where respondents were invited to give original suggestions, three of them noted the importance of access to original documents. In the words of one respondent, 'for historical data errors are inevitable. Therefore, users should be able to trace outliers back to the originals sources. This is more than useful; it should be a requirement.'

The reliability of data was also a recurring issue raised by the interviewees in D8.3. In the words of one respondent, 'the main thing is that people trust the integrity of your data – *that*'s the value-added'. This was considered to be particularly important in a project like EURHISFIRM, which faces challenges in standardising data that transcends not only time but also borders. The use of OCR technology was also thought to heighten users' concerns over accuracy of data.

Loyalty to the original source material was considered to be essential. In the opinion of one interviewee with considerable experience of historical data, 'If you are constructing databases out of any historical material, the aim must be to be as close to the original as possible – [this is] extraordinarily important'. Echoing the suggestions of some respondents to D8.2, interviewees with an academic background said that they would highly value the ability to 'click through' to a PDF scan of the original document. This would not only give assurance as to the quality of the transcription, but also show the data in its original context and empower the user to use his or her own judgement. In the words of one interviewee, 'Any good economic historian would be wanting to go in and clean the data themselves'. As noted by another interviewee, displaying the data in its original format could also lead the users to other, related data, on

the same page but not in the database. As another interviewee explained, 'it is not just a matter of trust; you might be interested in the original data for reasons relating to your research question'.

Interviewees with a business background were less interested in viewing the original documents, but agreed that a citation of the source document(s) would offer reassurance as to the data's provenance. In cases where some judgement has been deployed by EURHISFIRM, for example in order for data to be uniform across time or borders, interviewees suggested that a 'blurb' describing the methodology by which data was manipulated would be useful. As argued by one interviewee, 'your users will have some understanding of the sources that you are using – and should be able to transparently be able to understand decisions that you have made.' Other details to be included in such a blurb, according to one interviewee, would be an acknowledgement of any weaknesses in the source material or methodology. In the opinion of two interviewees, these descriptions could in and of themselves be useful to researchers, and if openly searchable could drive traffic to the main EURHISFIRM site.

Feedback

Related to the issue of reliability, we canvassed D8.3 interviewees' opinions on means of user feedback. There was a consensus that EURHISFIRM should at least offer an email address whereby users could give suggestions or make complaints. Interviewees were doubtful that such an email address would attract much traffic, but they agreed that the option of emailing feedback would reassure them of quality. The possibility of a users' forum was also discussed, but interviewees doubted that EURHISFIRM would have a large enough community of users for this to be viable. Nevertheless, one of the interviewees suggested that the core subgroup of highly active users could be harnessed as a source of feedback. As a variation on the theme, one interviewee suggested a 'comments section' attached to each particular source material, allowing open discussion as to quality of data. Alternatively, another interviewee suggested a Twitter account for feedback.