Pan-EU historical financial database (EURHISFIRM) project

Dr. Owen Sims

20th August 2018

bit.ly/EURHISFIRM

Existing databases (1)

- The United States has been investing resources to build and link databases suited for research over the long-run.
  - **Collaborative for Historical Information and Analysis** (CHIA) is a collaboration of academic and research institutions for the purpose of constructing and populating a world-historical data resource.

    "Purpose of CHIA is to create a single, comprehensive archive linking social and natural variables, development statistics, governance and social structure." (Zadorozhny et al., 2013: 4).
  - **Wharton Research Data Services** (WRDS) provides access to over 250TB of data across multiple disciplines including accounting, banking, economics, insurance and healthcare.
  - **Center for Research in Security Prices** (CRSP), the most widely used database in finance, contains prices and dividends for shares listed on the New York Stock Exchange since 1926.

Existing databases (2)

- Due to the USA's hegemony in data production, American companies are frequently and implicitly deemed *"representative"*.
- Further, the use of US data precludes any understanding of the features of the European economy.
- Few disjointed, centralised databases have been built so far in Europe by the academic community:
  - London Business School: **London Share Prices Database**;
  - University of Antwerp: **SCOB database**;
  - Paris School of Economics: **Equipex Data for Financial History**.

  And by private companies:
  - Thompson Reuters: **US Datastream**.

Establishing a feature set (1)

What type of information should the database contain?

- The resulting feature set will depend on the potential stakeholders and end-users. These may include:
  - Researchers & academics;
  - Practitioners;
  - Policy makers;
  - Regulatory agencies; and
  - Interested members of the public.
- Focus is placed on **financial and economic data** that will inform the main feature set of the EURHISFIRM database.
- However, given the identified end-users and the discrepancies between European economies, extending the dataset to include more **cultural, institutional, political and social data** could be beneficial.

Establishing a feature set (2)

- Optimally, a **panel dataset**—a mixture of time-series and cross-sectional data—would be required that contains **identification** and **time-sensitive data** regarding the firms **balance sheet** and **financial statements**.

  - **Firm identifying information.** Database UID, company ID, company name,date (year) established, ticker symbol, address, country of headquarters, country code, CIGS industry, CIGS sector, CIGS sub-industry, industry classification code.
  - **Company description.** Number of employees, market names, directorate.
  - **Company status.** Acquired, merged.
  - **Balance sheet items.** Earnings, total assets, total liabilities, operating income, operating expenses, debt, equity, land, buildings, equipment.
  - **Share characteristics.** Identity of markets, share par value, dividend payer, dividend yield, liquidity, preference shares, uncalled shares, number of shares, maximum value, share prices, dividend payments, shares traded and market value.
  - **Shareholder characteristics.** Acheson et al. (2017) found that different characteristics of shareholders could be exploited; such as institutional investors, rentiers, etc.
  - **Metadata.** Sources of data.

Establishing a feature set (3)

- The type of data advocated mimics the **Compustat** accounting and financial data model[1].
- Evaluating any extension to the data store would depend on what is needed to be calculated.
- For example, if the user is interested in comparing companies and industries over time one may want to calculate **accounting ratios** such as *liquidity*, *profitability* and *market ratios*.
- Further, economic and financial market data would be important.

[1] For more information on the Compustat model, see: http://web.utk.edu/~prdaves/Computerhelp/COMPUSTAT/Compustat_manuals/user_05r.pdf

Foreseeable issues

- Technical issues, such as inconsistencies between firms and industries over time, may arise when constructing a common data format.
- For example, **differences in financial and accounting standards**, which can change over time and space. Likewise, the frequency in which financial statements are recorded may also vary over time and economy.
- Therefore, data structures may be incomplete when pulled from the database.
- To compensate for this, it may be good to complement the financial data with **metadata on standards** for the time periods and economies that have been collected.

Extending the database (1)

What do end-users want from the database in terms of information?

- Need to define **user requirements documentation**, which requires negotiation with stakeholders to determine what is technically and economically feasible.
- The user requirements documentation should include:
  - Data requirements: *Feasible data that should be included.*
  - Technical requirements: *Technologies required to get (additional) data.*
  - Interface requirements: *Steps required to access the data.*
  - Migration of any electronic data: *Use of technologies (Tesseract).*
  - Operational requirements: *Performance measures and the actions taken in effecting the results that are desired to address requirements.*
- Prioritising requirements (mandatory, beneficial, nice to have)

Extending the database (2)

- **Ask.**
  One-on-one and group interviews with relevant stakeholders (scholars, practitioners, policy makers, regulatory agencies), questionnaires / surveys, prototyping and use cases.
- **Read.**
  Literature review of how equivalent US databases have been used will highlight case studies.
- Any desired requirements should be **reviewed** and **ratified** by the stakeholders and the subject matter experts. This can be done though:
  - **Presenting** the intent and current state of the database development at conferences and workshops.
  - **Open source** some aspects of the development. For example, create a Github account or forum that allows potential users to contribute data sources or suggestions (track issues).

Functionality (1)

What do end-users want from the database in terms of functionality?

- Further questions to ask when considering functionality:
  - *What is the user community used to with respect to other databases?*
  - *How do we expect analysis be performed on large datasets?*
  - *Will the user be able to port-forward the database directly?*
  - *What is the most scalable and extendable approach if more data becomes available?*
- Within industry, the **best practice** to interact with a database is through the use of an **Application Programming Interface (API)**.
- The API is **RESTful** in that the queries and path parameters within a URL request are used to access components of data in a predictable way. Documentation can be automatically generated.

Functionality (2)

- For example, querying for the full dataset of all firms through the API would be done with the following request:

GET .../api/firm-data/all

- For the full dataset of all firms that existed between 1930 and 1940 we would extend the request to the following:

GET .../api/firm-data/all?years=[1930,1940]

- Also, if each firm is given a UID, this identifier can be used to query resources attached to the specific firm(s). For example:

GET .../api/firm-data/uid/abc12345?features=[earnings,noEmployees,noShares]

- In each case the API would respond with a **JSON** containing the appropriate data.

  This approach would be suitable for statistical programming languages—such as *R*, *Julia* and *Python*—and software such as *Excel* and *Stata*.

Any questions?