F.A.Q. BEDOFIH

Revision Version 2.0: 02/02/2016

eurofidai
CNRS UPS 3390

BEDOFIH
Base Européenne de Données Financières à Haute Fréquence

1. BEDOFIH General Questions

How to get access to BEDOFIH data?

Please contact us via the *BEDOFIH access form*. Note that a call for projects to use BEDOFIH data was organized in June 2014. As of today, EUROFIDAI is not planning to launch another call for projects. The only way to get access to BEDOFIH data is to contact us and submit your research project with a complete CV.

Does BEDOFIH data include limit order book?

Deutsche Boërse Data (Eurex and Xetra) includes a limit order book with the 20 best limits provided by Deutsche Boërse. For the other markets (Euronext Paris, Bats Chi-X, LSE), EUROFIDAI computes the limit order book on-demand for an extra fee. *Please contact us* to discuss your needs in detail (frequency of the order book, number of limits, etc.).

How does EUROFIDAI obtain BEDOFIH data?

EUROFIDAI collects data directly from stock exchanges to build the BEDOFIH database. EUROFIDAI receives a dump of the flow of market messages with exhaustive information. In some cases, we obtain data directly from Regulatory Authorities with enhanced data.

In which format is BEDOFIH data delivered?

BEDOFIH data is provided on-demand on CSV files. It can also be delivered via a Postgres database with access from a local virtual machine.

How much does BEDOFIH data cost?

EUROFIDAI charges extraction fees for extracting the data, computing the order book or providing virtual machines for example. Additionally, for some markets EUROFIDAI pays the providers to obtain the data.

What is the difference between EUROFIDAI and BEDOFIH?

EUROFIDAI (European Financial Data Institute) is the name of the institution. BEDOFIH is the name of the High Frequency Financial Database developed by EUROFIDAI.

What are the added values of BEDOFIH?

BEDOFIH provides many added values from getting the data from stock exchanges or regulators until delivering it to researchers. Data is decrypted from different formats to human readable files. The cost of the data is negotiated and mutualized for academic use. BEDOFIH provides additional data such as order books computed at high frequency. A scientific support is also offered to researchers during their use of the data.

How big is the high-frequency data by market?

On raw data files it is about:
Xetra : 2.5 To per year
Eurex : 25 To per year
Bats : 1 To per year
Chi-X : 2 To per Year
LSE : 1.5 To per year
Euronext : 1To per year

Note that these numbers are approximate and may be higher if supplementary data is requested such as order books.

2. BEDOFIH – Questions about AMF (Euronext Paris)

What is the difference between the two IDs?

The two IDs (fundamental and characteristic) identify any order when forming a couple. The characteristic ID begins at 1 and is incremented at every modification of the order. Thus the modified order is seen as a new order in our dataset which is shown by a new row.

What is the difference between all the dates?

There are several dates concerning the Euronext Paris data, but basically we may focus on four which are in a chronological order:

- o_dt_be stands for order/date time/book entrance and corresponds to the date and time of the first validity (here validity has the sense of active), that means for the characteristic ID equal to 1.
- o_dt_va stands for order/date time/validity and corresponds to the date time at which the order become valid (or active). For orders having the characteristic ID equal to one, this date time is the same as o_dt_be.
- o_dt_mo stands for order/date time/modification and correspond to the date time at which the order is modified (if it is). Normally this date time is equal to the date time o_dt_va for the "next" order (i.e.: having the characteristic ID +1) but sometimes there is a thin time lag between these two.
- o_dt_br stands for order/date time/book release and correspond to the date time of the "end" of the order. This end may appear because of various action (all quantities have been traded, the trader cancelled his order, etc...)

One order is valid/active during the gap [o_dt_va ; o_dt_mo] if o_dt_mo is fulfilled and [o_dt_va ; o_dt_br] if not.

How are the partial executions treated in the Euronext Paris dataset?

The information about partial executions must be taken carefully. Indeed, the variable o_state has a modality meaning "partially executed". However, in our database no order has this modality. That is because there is always an 'action' after the partial execution. Since we receive the data quite a time after the real period, either the order has been fully negotiated or the trader cancelled the order or...

So the variable o_state will have the modality corresponding to the last 'action', i.e. the 'action' who 'kills' the order. It is theoretically possible that an order has the modality 'partially executed' but for now we don't have any.

Can we have more information about investor categories?

There are two investor categories:

- o_member
- o_account

The first one, which classifies the member in three categories (HFT, MIXED and NON HFT), is made by the AMF and is based on the frequency of orders submitted. The second classifies members relatively to whom they trade with. This information comes from an auto-reference and thus must be taken carefully.

3. BEDOFIH - Questions about BATS and CHI-X

What are the differences between the “multicast” data and the “Europe” data?

BATS and Chi-X save their data in two ways, the first is to distinguish by country on a total of 12 separate streams. This method saves the data with more detail and includes a more important number of messages.
The second recording method is to unify all the data in a same stream. In this system, there are fewer types of messages, but there is of course the same general information.

What are the frequencies of BATS and Chi-X observations?

The frequency of the data does not depend on the database but on the data format. In the format “multicast”, data is given in nanoseconds, whereas in the “Europe” format, the precision level is the millisecond.
However, recording is often restricted to the millisecond, and the nanosecond data is relatively marginal.

4. BEDOFIH - Questions about London Stock Exchange (LSE)

What are the differences between "T_OrderDetails.CSV" and "T_OrderHistory.CSV" files?

There are many differences between the two files, all can be found in the specification file to identify the different variables of each message component of these files. From a more practical point of view, we can see these files as complementary files to build the order book but also to follow the evolution of the market over the day.
We can find at the beginning of the "T_OrderHistory.CSV" file some messages that recap the activity and that we have to consider if we want all the messages that impact one specific day.

What is the unit used in the order books for the variable "Time" and why?

The unit of the variable "Time" is the millisecond, this unit was retained for the time display as it allows manipulating entire data in a decimal system, thus more intuitive to operate various operations of calculation. Moreover, we can find a millisecond precision in all BEDOFIH project databases. This facilitates the work with data from different databases.

What is the difference between the "stock_code" in LSE database and the ISIN code identifying a company?

The ISIN (International Securities Identification Numbers) is an international code to uniquely identify a financial asset, a company in this case. In the same way, the "stock_code" from LSE is a unique code used to distinguish the companies registered on the base, but also to distinguish the compartment where the title is traded.

5. BEDOFIH - Questions about Eurex and Xetra (Deutsche Boërse)

What is the depth of order books available on Eurex?

All order books from the Eurex data have a depth of 20. This is due to the data registration system of the German exchange platform that saves only the 20 best bids and best asks.

How many different types of messages do exist on Eurex?

Altogether there are five different types of messages to build the order book. We find a message "New", which adds a new order in the order book and shifts all orders already in it, a message "Change" that only changes the values of the book for one price and three messages "Delete" that cancel one or more orders in the book.

What are the frequencies of Eurex order books?

The frequency of the order book depends on the data. On Eurex, we have data with an accuracy of around one nanosecond, which is both an advantage and a disadvantage. The data is more accurate, but it requires more time to be managed and compiled for the calculations.

6. IT Infrastructure (Data Storage and Computing) Questions

What are the Virtual Machines' specifications?

The available operating systems are Microsoft's Windows 7 Enterprise and Linux' Ubuntu 14.04, the CPU limit is 8 Cores and the RAM limit is 64 Go.

What is the available software?

We can install SAS, R, Matlab and PGadmin.