College of Computer Science and Technology, Jilin University

– NULL Group –

Knowledge Graph for financial analysis

Document Data:	Reference Persons:
- 2023-11-10 -	- Yongjun Fan, Jiahui Hou, Yuze Li -

BaH (internal) reports are for internal only use within the NULL Group. They describe preliminary or instrumental work which should not be disclosed outside the group. BaH reports cannot be mentioned or cited by documents which are not BaH reports. BaH reports are the result of the collaborative work of members of the NULL group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the NULL group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.

Contents

Revision History:

Revision	**Date**	**Author**	**Description of Changes**

1 Introduction

A-shares, Hong Kong stocks (H-shares), and the New Third Board are three different stock trading markets within Mainland China, each with its distinct characteristics and regulations. The A-share market specifically refers to Mainland China's primary stock trading market, used for the trading of stocks of companies registered within Mainland China. The Hong Kong stock market operates in Hong Kong.It is globally recognized as an international financial hub, permitting international investors to buy and sell shares of companies registered in Hong Kong, which may include numerous Mainland Chinese companies. The New Third Board, officially known as the National Equities Exchange and Quotations (NEEQ), is an over-the-counter market situated within Mainland China. It provides financing and equity trading opportunities, primarily for small and medium-sized enterprises. These markets offer diversified investment opportunities for both Mainland China and international investors.

In this project, we collect data from three platforms and construct a financial knowledge graph by extracting equity information between companies, thereby providing knowledge services in the financial field.

The report is organized as follows.Section 2 defines the purpose and specific domain of interest of the project. It provides a high-level overview of low quality resources that will be created by producer, and high quality resources that will be identified and composed by consumer.Section 3 formalize our research purpose by defining a set of scenarios, personas, competency questions (CQs) to identify a set of entities, which are modeled in an Enity-Relationship (ER) model.Section 4 delineates the sources of data and their utilization.In Section 5, we extract entity types, data properties, and object properties from the aforementioned data. Section 6 is devoted to the formalization of teleologies for the three sources, which are then integrated into a comprehensive equity knowledge graph.Section 7 outlines the formalization of data based on the equity knowledge graph. Section 8 evaluate the quality of our data and knowledge results and show the Knowledge Graph (KG) exploitation. Finally, we draw the conclusion and give open issues in 9.

2 Project Description

2.1 Objective

Project purpose:The objective of this project is to create a financial equity knowledge graph by extracting ownership information between companies and individuals. This knowledge graph aims to provide end-to-end knowledge research and services throughout the knowledge lifecycle in financial contexts.

Project domain of interest (DoI):Our research focuses on the equity relationships among companies, financial institutions, and individuals across three trading platforms: A-share, Hong Kong stock,

and the New Third Board.

2.2 Project development

2.2.1 Data Production

In this phase, producer aims to produce datasets that meet our project purpose, these resources need to be created if they that don't exist yet or exist in a bad quality.In our project, we need to create and formalize resources about securities trader, A share and Hong Kong stocks. The data about securities trader should include trading statistics and important indicators, such as net capital, risk coverage ratio, and financial leverage ratio. A-share and Hong Kong stock markets represent stock information, primarily consisting of the names of listed companies and their corresponding securities firms.

2.2.2 Data Composition

Consumer identifies the existing high quality resources to satisfy our project purpose,i.e., Financial shares OSM dataset which contains a large number of data collected from different platforms to construct entity categories such as the New Third Board, securities companies, individuals, A-shares, and Hong Kong stocks. These data are cleaned and pre-processed to support the construction of a complete financial knowledge map.All the selected resources from consumer will be composed with the formalized resources from producer.Here we group the same listed companies, brokerages, individuals into the same set. For example, brokerages match according to name, and those with exactly the same name are matched to the same set.

3 Purpose Formalization

The purpose of our project is to collect basic information and data of NEEQ listed companies, major securities brokerages, A-shares and Hong Kong stocks from three different financing platforms.After that, we used neo4j to carry out knowledge storage, knowledge completion and knowledge fusion on the knowledge graph to construct a high-quality financial knowledge graph.

To describe multiple aspects considered by the project purpose, we list a set of usage scenarios as follows:

- Scenario 1. In financial markets, there exist relationships of control and being controlled between companies, known as supervisory relationship,shareholding and being held shares relationships.
- Scenario 2. In the financial markets, a person can own shares in multiple companies.

In the scenarios defined above, we represents a set of real companies and person with specific features included in the project purpose, which are listed as follows:

- Company 1. Ping An Bank Co., Ltd.,founded in 1987, is a cross-regional joint-stock commercial bank controlled by China Ping An Insurance (Group) Co., Ltd. It is one of the twelve national joint-stock commercial banks in mainland China.

- Company 2. China Vanke Co., Ltd.,founded in 1984, has emerged as a leading domestic provider of urban and rural construction and lifestyle services after nearly four decades of development.The company's operations are centered on the three most dynamic economic circles and key cities in the central and western regions of the country.
- Person 1. Shi Ying, female, born in August 1975, master degree, Chinese nationality. Director or executive director of several companies.

Taking into account the personas in the scenarios defined, we create Competency Questions(CQs):

- CQs1.What is the equity ralationship between Ping An Bank Co., Ltd. and China Vanke Co., Ltd.?
- CQs2.Does Ping An Bank Co., Ltd. hold China Vanke Co., Ltd.'s stock?
- CQs3.Which companies does Shi Ying hold shares in?

From the CQs, referring to Personas and Scenarios, we extract Entities with properties. The details of this work are outlined in Table 1.

Table 1 Entities extraction and classification

Scenarios	companies or Personas	CQs	Entities	Properties
1	c1	1,2	Ping An Bank Co.	name:string date:string Controller:string Net capitalstring Risk coverage:float ratiofloat Financial leverage multiplefloat Transaction statisticslist
1	c2	1,2	China Vanke Co.	name:string date:string Actual Controller:string Legal representative:string Org$_{f}orm : string$ Company profile:string Financial politician:list Securities informationlist
1,2	c1,c2	1,2,3	security	stock code:int stock abbreviation:string company name:string main business income:float net profit:float Listing date:string
2	p1	3	Shi Ying	name :string job:string gender:string education background:string age:int holding shares Num:int Curvitae:string
1	c1,c2	1,2	Holding company	name :string Nature of shareholder:string holding shares Num:int Share holding ratio:int

Based on the entities in Table1.we design an Entity–relationship (ER) model as Figure 1.

4 Information Gathering

In this section, we will introduce that by using crawler technology, I obtained rich data sets from multiple financial-related websites, covering information in multiple fields such as the stock market and company information. These data include important information such as stock prices, trading volumes, and financial statements, as well as basic information and business scope of each company. I collected and organized these data and built a comprehensive

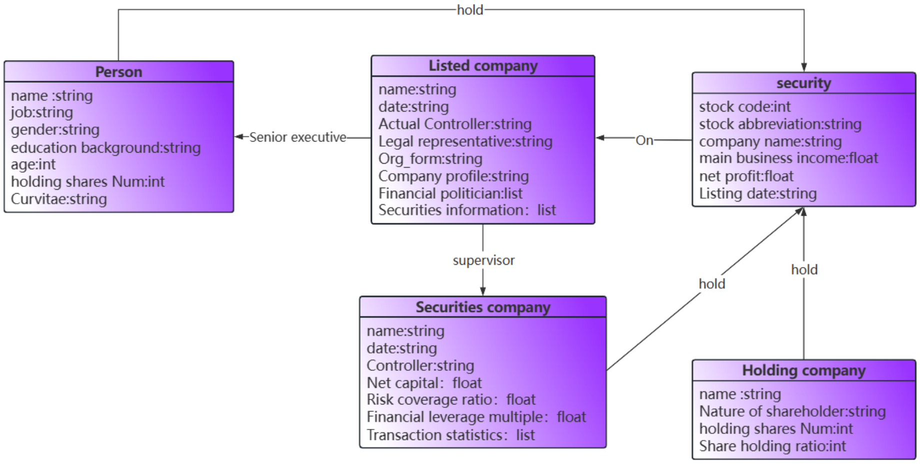


Figure 1: ER model

financial database to provide users with rich financial information resources. Next, I will describe the process of how I collect, process, and manage these data sources.

4.1 Data and Knowledge Source

We have provided four data sources along with their descriptions, including A-share market information, Hong Kong stock market information, brokerage firm information, and listed company information. These data sources are provided by the China Industrial Research Institute, covering comprehensive information in various fields. The data format is primarily in JSON files and can be accessed through the respective URLs. These data sources offer rich information, including basic company details, business scope, financial indicators, and more. Additionally, these data sources also include descriptions of knowledge related to their respective fields. This information will provide users with in-depth insights into areas such as finance, the Hong Kong stock market, brokerage firms, and listed companies.

4.1.1 Informal Data and Knowledge Source from Producer

Table 1: A-Share Catalog

Field	A Finance
Keywords	Finance, Bank, Commercial Bank
Language	Chinese
Provider	China Industrial Research Institute
Data URL	https://s.askci.com/stock/h/?reportTime=2018-06-30
Data Format	JSON file
Data Description	Contains various information about Ping An Bank Co., Ltd., such as company name, English name, region, industry, company website, etc. It also includes detailed information such as main business, product names, controlling shareholders, and more.
Knowledge URL	N/A
Knowledge Description	N/A

Table 2: Hong Kong Stock Catalog

Field	Hong Kong Stock Information
Keywords	Hong Kong stock comprehensive information
Language	Chinese
Provider	China Industrial Research Institute
Data URL	https://s.askci.com/stock/h/?reportTime=2018-06-30
Data Format	JSON
Data Description	Contains various information about companies listed in the Hong Kong stock market, such as company name, industry, number of employees, registered address, main business, etc. It also includes detailed information like company headquarters, phone number, fax, website, email, etc.
Knowledge URL	N/A
Knowledge Description	Contains knowledge about companies listed in the Hong Kong stock market, such as business scope, history, etc.

Table 3: Brokerage Catalog

Field	Brokerage Firm
Keywords	Comprehensive Information on Brokerage Firms
Language	Chinese
Provider	China Industry Research Institute
Data URL	http://xinsanban.eastmoney.com/api/Organization/Broker/qslb?
Data Format	JSON
Data Description	Contains various information about brokerage firms, such as company name, establishment date, legal representative, company website, business scope, etc. It also includes important indicators such as net capital, return on net capital, risk coverage ratio, financial leverage multiple, and more detailed information.
Knowledge URL	N/A
Knowledge Description	Contains knowledge about brokerage firms, including basic information, important indicators, transaction statistics, and more.

Table 4: Listed Company Catalog

Listed Company	
Keywords	Comprehensive Information of Listed Companies
Language	Chinese
Provider	China Industry Research Institute
Data URL	http://xinsanban.eastmoney.com/api/DataCenter/JGCG/GetSSGSCG?
Data Format	JSON
Data Description	Contains various information about listed companies, such as company profiles, securities information, financial summaries, key indicators, shareholder lists, executive lists, etc.
Knowledge URL	N/A
Knowledge Description	Contains knowledge about listed companies.

4.1.2 Formal Data and Knowledge Source from Consumer

table 5: Data and knowledge catalog.

Resource Name	Financial Equity
Domain	Financial Equity
Keywords	Stocks, Securities, Listed Companies
Language	Chinese
Provider	Knowledge Graph and Knowledge Fusion
Data Description	Based on Neo4j graph database, it stores basic information data of New Third Board listed companies, major securities firms, A-shares, and Hong Kong shares.
Knowledge Description	The knowledge is stored based on the Neo4j graph database. It involves tasks such as knowledge completion and knowledge fusion in the knowledge graph. Relationships such as ”Supervising Brokerage” are established to connect company nodes with brokerage nodes. The processing of shareholder lists connects company nodes with shareholder nodes, and holds the attributes of shareholding quantity and shareholding proportion as relationship properties. The processing of executive lists creates a new node for each executive, marked as ”Individual,” with the name as the node name. The ”Executive” relationship is established to connect company nodes with executive nodes, and holds the position and shareholding quantity (if available) as relationship properties.

4.2 Resource Collection, Processing and Scraping

4.2.1 Informal Resource Collection, Processing and Scraping from Producer

For the retrieval and processing of brokerage firm information:
1. Initially, requests were made to the Eastmoney API to obtain a dataset containing a list of brokerage firms.
2. Each page's data was iterated through, and the code and detail page link for each brokerage firm were saved in respective lists.
3. For each brokerage firm's detail page, web scraping and parsing were conducted. This included extracting basic company information, important indicators, and transaction statistics.
4. All acquired data was saved in JSON format. The company names were used as keys, with corresponding company information as values.

For the retrieval and processing of listed company information:
1. Similarly, requests were made to the Eastmoney API to obtain a dataset containing a list of listed companies.
2. Each page's data was iterated through, and the code and detail page link for each listed company were saved in respective lists.
3. For each listed company's detail page, web scraping and parsing were conducted. This encompassed gathering basic company information, securities details, financial highlights, important indicators, shareholder lists, and executive teams.
4. All acquired data was saved in JSON format. The company names were used as keys, with corresponding company information as values.

4.2.2 Formal Resource Collection and Scraping from Consumer

For each company, we initiate by creating a new node labeled as "New Third Board", with the node's name set as the full name of the company. Following this, we add the basic company information as node attributes. Subsequently, we incorporate details such as securities data, financial summaries, and key indicators as strings in the node attributes.
Afterwards, a broker node is generated and labeled as "Broker", with its name being the continuous supervising broker. A "Supervising Broker" relationship is then established, connecting the company node and the broker node. Moving on, we manage the list of shareholders for the company. For each shareholder, a new node is formed, with the node type denoting the shareholder's nature and their name serving as the label. We proceed to establish a "Shareholder" relationship, linking the company node with the shareholder node. Concurrently, we append the shareholding quantity and percentage as relationship attributes.
To conclude, we address the list of executives within the company. For each executive, a new node of type "Individual" is generated, with the name set as the executive's name. The relevant attributes are then added to the node. Subsequently, an "Executive" relationship is forged, uniting the company node with the executive node, and the position is added as a relationship attribute. In instances where the executive holds a quantity of shares (measured in ten thousand shares), this is also integrated as an attribute in the relationship.

5 Language Definition

In this section, we will extract entity types (lexical types), data properties, and object properties from these four sources. Next, by mapping these concepts to global identifiers in the Common Knowledge Core, we formalize the language of these concepts for each source code.

5.1 Formalize Etypes with Properties from resources by Producer

Based on the provided A-share market data, we conducted extraction, resulting in attributes, property types, and entity types.

Property	Property Type	Entity Type
Company Name, English Name, Former Names, Region, Industry, Company Website, Main Business, Product Names, Controlling Shareholder, Actual Controller, Ultimate Controller, Chairman, Secretary of the Board, Legal Representative, General Manager, Registered Capital, Number of Employees, Telephone, Fax, Zip Code, Office Address, Company Introduction	Data Property	Company

Based on the provided data from the Hong Kong stock market, we conducted extraction, resulting in attributes, property types, and entity types.

Property	Property Type	Entity Type
Company Name, Industry, Chairman of the Board, Securities Affairs Representative, Number of Employees, Year-End Date, Registered Address, Company Headquarters, Telephone, Fax, Website, Email, Auditor, Legal Advisor, Main Business, Company Introduction	Data Property	Company

Based on the provided brokerage-related data, we conducted extraction, resulting in attributes, property types, and entity types.

Basic Information:

Property	Property Type	Entity Type
Company Name, Establishment Date, Legal Representative, General Manager, Registered Capital (RMB '0000), Net Capital (RMB '0000), Company Website, Email, Business License Number, Registered Address, Office Address, Main Business, Business Scope, Company Introduction	Data Property	Company ,Brokerage firm

Key Indicators:

Property	Property Type	Entity Type
Net Capital (RMB '0000), Net Capital ROI (%), Risk Coverage Ratio (%), Financial Leverage Ratio	Data Property	Company ,Brokerage firm

Trading Statistics:

Property	Property Type	Entity Type
Transaction Amount (RMB '0000), Percentage of Total Volume (%), Transaction Volume (10,000 Shares), Number of Transactions, Number of Traders	Data Property	Company ,Brokerage firm

Based on the provided listed company data, we conducted extraction, resulting in attributes, property types, and entity types.

Securities Information Table:

Property	Property Type	Entity Type
Securities Code, Listing Date, Transfer Method, Total Capital (10,000 shares), Recommended Listing Broker, Securities Abbreviation, First Trading Day, Market Segmentation, Circulation Capital (10,000 shares), Continuous Supervision Broker	Data Property	Securities

Financial Indicators Table:

Property	Property Type	Entity Type
Total Operating Revenue, YoY Operating Revenue, Attributable Net Profit, YoY Attributable Net Profit, Deducted Net Profit, YoY Deducted Net Profit, Earnings Per Share, Gross Margin, Weighted ROE, Net Assets	Data Property	Company

Market Performance Table:

Property	Property Type	Entity Type
Total Capital (10,000 shares), Total Market Value (10,000 RMB), Circulation Capital (10,000 shares), Circulation Market Value (10,000 RMB), PE Ratio TTM, PB Ratio MRQ, Financing Issuance Times, Cumulative Financing Amount (10,000 RMB), Proportion of Actual Trading Days in the Last 3 Months , Number of Violations Since Listing	Data Property	Company

Shareholder Table:

Property	Property Type	Entity Type
Shareholder Name, Shareholder Nature, Holding Quantity, Holding Ratio	Data Property	Company

Executive Table:

Property	Property Type	Entity Type
Name, Position, Gender, Education, Age, Holding Quantity, Resume	Data Property	Individual

5.2 Formalize Object Properties for Composing Resources from Consumer

Consumers generate two object properties, namely 'consultation,' where it can combine source types from these four resources. Object properties are associated from source lexical types to target lexical types.

Attribute	Attribute Type	Entity Type
Is	Object Property	Company Name, English Name, Former Names, Region, Industry, Company Website, Main Business, Product Names, Controlling Shareholder, Actual Controller, Ultimate Controller, Chairman, Secretary of the Board, Legal Representative, General Manager, Registered Capital, Number of Employees, Telephone, Fax, Zip Code, Office Address, Company Introduction, Securities, Shareholders, Executives, etc.

6 Knowledge Definition

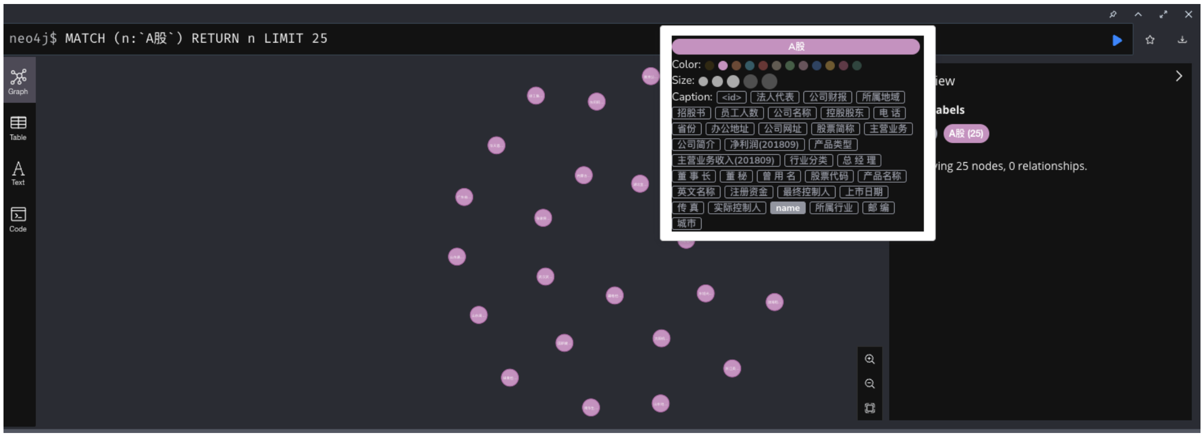
In this section, the producer's goal is to generate ontologies for each dataset, always considering the reusability of knowledge and enhancement. At the same time, the consumer's goal is to integrate ontologies from the producer.

In both the Hong Kong and A-share markets, the company names are defined in the basic information of securities. We have also defined information about securities, shareholder companies, and executives in listed companies. The definition of company names is an entity type that we aim to share or reuse across different data entities. This means that attributes like Company Name, English Name, Former Names, Region, Industry, Company Website, Main Business, Product Names, Controlling Shareholder, Actual Controller, Ultimate Controller, Chairman, Secretary of the Board, Legal Representative, General Manager, Registered Capital, Number of Employees, Telephone, Fax, Zip Code, Office Address, and Company Introduction are consistent across A-shares, Hong Kong stocks, and the basic information of securities. Additionally, attributes such as Company Name, Establishment Date, Legal Representative, General Manager, Registered Capital (RMB '0000), Net Capital (RMB '0000), Company Website, Email, Business License Number, Registered Address, Office Address, Main Business, Business Scope, and Company Introduction are also aligned with the company information in the securities information and executives in listed companies.

As shown in the Figure 2,3,4, we can observe that the company name information is present across different entities. We initiate the KG by using the listed companies dataset in the Figure 5,6.

7 Data Definition

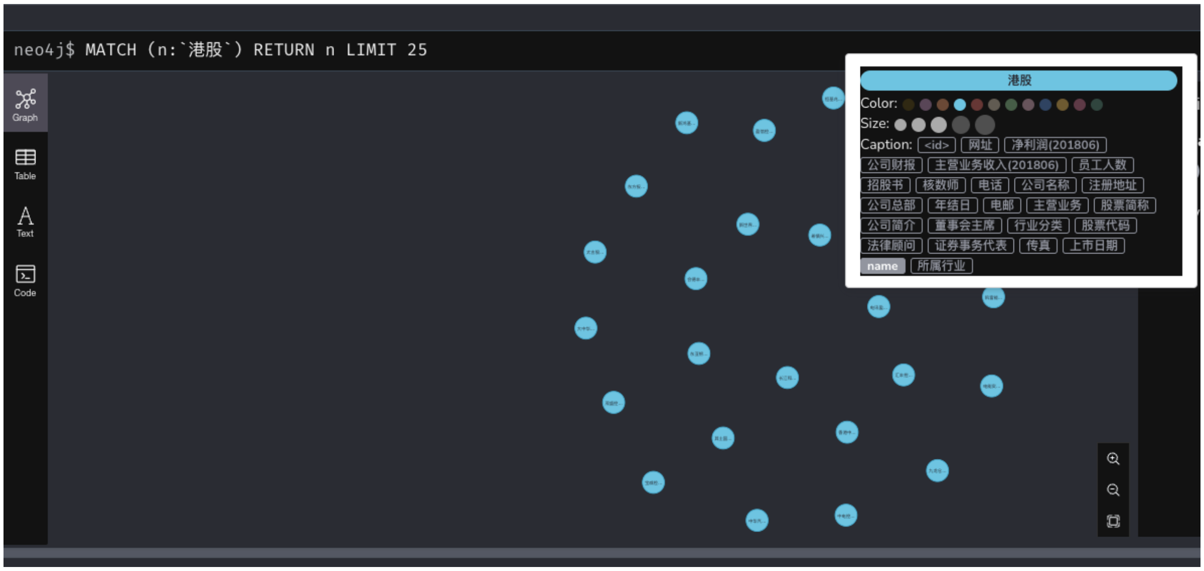
In this section, the goal of the producer is to formalize each dataset and map each dataset to their respective schema. At the same time, the goal of the consumer is to combine all data sets and merge the composed data with the final remote ontology.



```
neo4j$ MATCH (n:`A股`) RETURN n LIMIT 25
```

Graph
Table
Text
Code
A股
Color:
Size:
Caption: <id> 法人代表 公司财报 所属地域
招股书 员工人数 公司名称 控股股东 电 话
省份 办公地址 公司网址 股票简称 主营业务
公司简介 净利润(201809) 产品类型
主营业务收入(201809) 行业分类 总 经 理
董 事 长 董 秘 曾 用 名 股票代码 产品名称
英文名称 注册资金 最终控制人 上市日期
传 真 实际控制人 name 所属行业 邮 编
城市
iew
abels
A股 (25)
ing 25 nodes, 0 relationships.

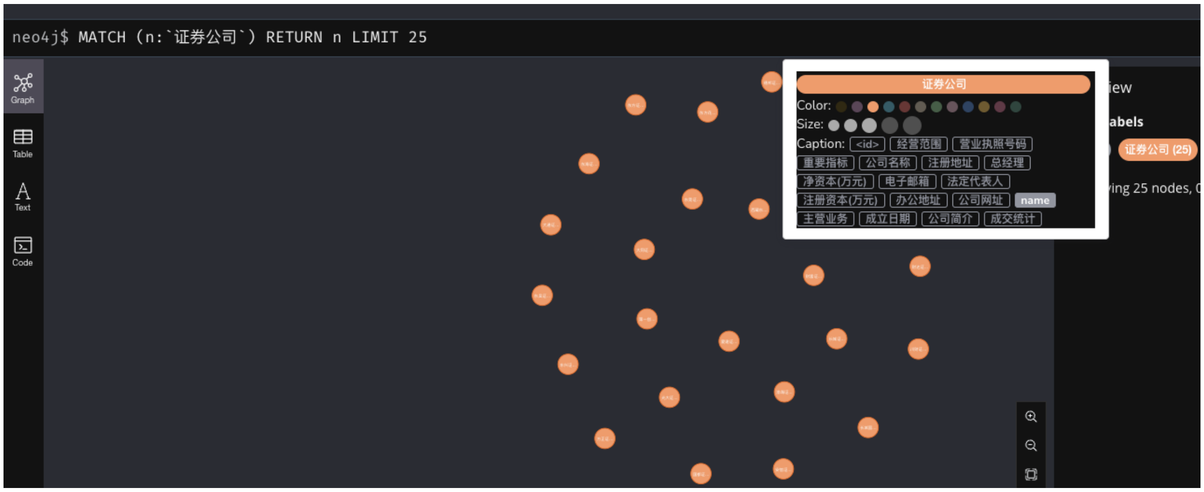
Figure 2: A-share Company Information



```
neo4j$ MATCH (n:`港股`) RETURN n LIMIT 25
```

Graph
Table
Text
Code
港股
Color:
Size:
Caption: <id> 网址 净利润(201806)
公司财报 主营业务收入(201806) 员工人数
招股书 核数师 电话 公司名称 注册地址
公司总部 年结日 电邮 主营业务 股票简称
公司简介 董事会主席 行业分类 股票代码
法律顾问 证券事务代表 传真 上市日期
name 所属行业

Figure 3: Hong Kong Stock Company Information



```
neo4j$ MATCH (n:`证券公司`) RETURN n LIMIT 25
```

Graph
Table
Text
Code
证券公司
Color:
Size:
Caption: <id> 经营范围 营业执照号码
重要指标 公司名称 注册地址 总经理
净资本(万元) 电子邮箱 法定代表人
注册资本(万元) 办公地址 公司网址 name
主营业务 成立日期 公司简介 成交统计
iew
abels
证券公司 (25)
ing 25 nodes, (

Figure 4: Securities Company Information

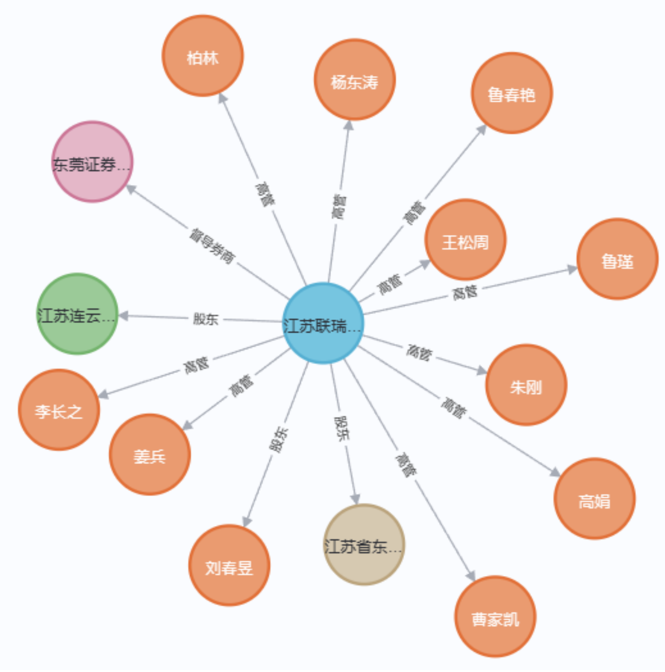


Figure 5: KG initialization

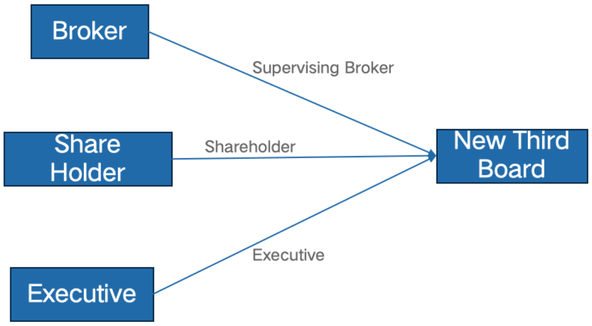


Figure 6: KG initialization

7.1 Dataset formatting by consumer

Consumers formalize data to instantiate entity types, data attributes, and object attributes defined in A-shares, Hong Kong stocks, securities, and listed company ontologies. To distinguish data files from different sources, we merged source information for each file. This applies to a variety of files, such as A-share data files, Hong Kong stock files and securities files. If the files contain instances of the same entity type, they are integrated. The text contains entities of the bus station entity type. During this integration process, data attributes from the integration files are also merged.

Dir	Property Type
A-Share	Company Name, English Name, Former Names, Region, Industry, Company Website, Main Business, Product Names, Controlling Shareholder, Actual Controller, Ultimate Controller, Chairman, Secretary of the Board, Legal Representative, General Manager, Registered Capital, Number of Employees, Telephone, Fax, Zip Code, Office Address, Company Introduction
H-Share	Company Name, Industry, Chairman of the Board, Securities Affairs Representative, Number of Employees, Year-End Date, Registered Address, Company Headquarters, Telephone, Fax, Website, Email, Auditor, Legal Advisor, Main Business, Company Introduction

7.2 Entity identifier provided by the manufacturer

In each file, the producer identifies identical entities by comparing their names and coordinates and eliminates any duplications. Essentially, for each file, we have a set of entity identifiers, including the data attribute ”Company Name”, which is used to determine whether two entities in the file are the same. We assume that two entities are considered the same entity if they have the same name.

7.3 The object attribute ”is” is identified by the consumer

Data definition is a crucial step in information management and database design, providing a clear and accurate framework for the relationships between various entities. In the process, we recognize that the connections between entities such as individuals, companies, and securities are connected through the attribute of ”is.”

For example, when we define a person, we can introduce the attribute: ”what company”. This attribute describes the affiliation between an individual and a company and can cover aspects such as employment, ownership, etc. Not only does this allow us to understand someone’s current occupation, but it also allows us to trace their historical relationship with the company.

Similarly, for the definition of listed companies and securities, we can introduce attributes: ”What company does the listed company’s securities belong to?” This attribute demonstrates the connection between a public company and its securities, providing investors and analysts with important information to help them understand a company’s financial condition and business performance.

Through such data definition, we have established an organic information network, making the relationships between individuals, companies, securities and other entities more transparent and traceable. This is critical for data management, corporate decision-making and financial analysis. Such a definition not only provides structured data, but also facilitates system query and analysis, promoting deeper data understanding and utilization.

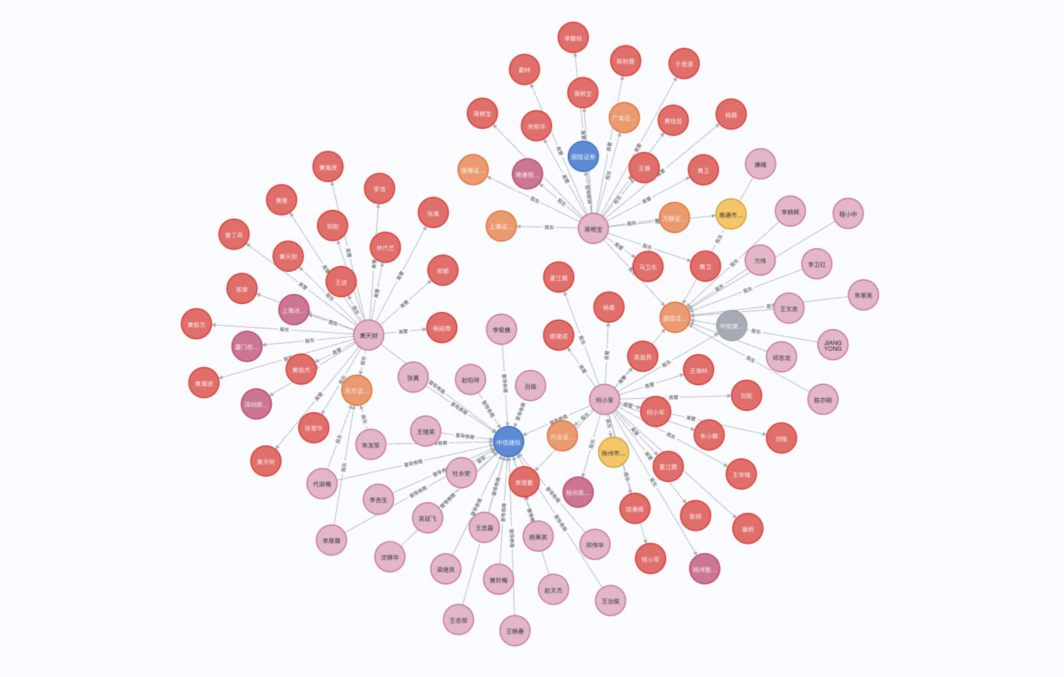


Figure 7: Enter Caption

In summary, through carefully designed data definitions, we are able to capture the relationships between entities and provide a solid foundation for the construction of information systems and data-driven decision-making. This method not only makes the data more practical, but also provides more reliable decision support for decision makers in various fields.

7.4 data mapping

Data mapping refers to matching existing data with predefined ontology or models and establishing a mapping relationship between data. In our scenario, by connecting the integrated data with the individual entities in the entity graph, we achieve an efficient mapping of the relationships between these entities. This mapping not only helps understand the correlation between data, but also provides greater efficiency and accuracy for data query and analysis.

The generation of entity diagrams makes the organizational structure of data clearer and the relationships between entities more intuitive. Through data mapping, we correspond the integrated data to actual business entities, so that the data is no longer an isolated piece of information, but is organically integrated into the entire business logic. As shown in the figure, we can clearly see that we have equivalence between entities and can connect different data sets in Figure 5.

8 Conclusion and Open Issues

This document follows the iTelos method process and consists of three spatial resources. Throughout this process, we identified a number of relevant open issues. First, the resources we selected record stock information at different times, and there may be company shareholder information with the same name, leading to errors in matching.

Therefore, for such issues we must ensure that the information of shareholders with the same name should be differentiated, which we have not yet completed. Despite the above challenges, iTelos' methodological process does provide a comprehensive framework for integrating resources from various sources. This approach not only allowed us to effectively achieve the goals of the project, but also provided us with a set of knowledge and data integration tools to work collaboratively on