# Research on Semantic Driven Urban Pipeline Dataspace Construction Method

Binghan Li [1] , Liang Huo [1], Yaodong Yang [2], Peng Bao [3], Maolin Zhang [1], Yucai Li [1]

[1] Beijing University of Civil Engineering and Architecture, Beijing, China – lbh292419@163.com, huoliang@bucea.edu.cn, (18769355718, liyucai1211) @163.com
[2] State Geospatial Information Center, Beijing, China – yangyd@sgic.net.cn
[3] CIGIS (CHINA) LIMITED, Beijing, China – baopeng0806@163.com

**Keywords:** Semantic-Driven, Dataspace, Semantic Feature Expression, Pipeline RDF Model, Neo4j.

**Abstract**

Urban pipeline data is heterogeneous in multiple sources and rich in data volume, and there are problems such as data conflict and difficult organization and management due to the heterogeneity of multiple sources when accessing the data in large-scale concurrently. To address this problem, this paper proposes a semantics-driven urban pipeline dataspace construction method, which aims to realize the efficient organization of pipeline data. Firstly, this method combines the classification and characteristics of urban pipelines, and expresses the semantic information of pipeline geographic entities from four dimensions: semantic description, spatial location, attribute characteristics and time evolution. Then, the four expression sets are embedded into the dataspace RDF model as predicates, and the associated description mechanism of pipeline geographic entities is established by means of genus classes and so on, so as to construct the pipeline dataspace RDF model. Finally, the model is stored and graphically visualized using neo4j to achieve fast retrieval of data within the pipeline dataspace. The research results show that this method provides a unified expression of pipeline entities, solves the problem of pipeline multi-source heterogeneous data conflict and organization difficulties, and improves the efficiency of multi-source heterogeneous pipeline data organization while ensuring the integrity of pipeline information to the maximum extent.

## 1. Introduction

As the "bloodline" and "nerve" of the city, underground pipelines play an important role in the health and sustainable operation of the city. The complexity of underground pipeline types, the variety of uses, and the different construction periods, as well as the lack of uniform data organization methods and standards, have led to problems such as redundancy, duplicate collection, and difficulty in sharing data (Wang, 2021). Efficiently managing and organizing these data is essential to improve the accessibility and standardization of pipeline data (Zhang et al., 2015). Existing data organization can be divided into two categories: management and maintenance and visualization and scheduling (Dai, 2018), but when dealing with urban pipeline data, which is rich in semantic information, high in granularity, and diverse in data sources, problems such as missing semantics, data conflicts, and imbalance in granularity may occur during large-scale concurrent access (Liu et al., 2018).

In November 2023, the Beijing International Data Lab and the International Dataspace Association jointly released the Dataspace Development Initiative, which aims to rapidly advance dataspace technology. Dataspace (Jeffery et al., 2008) unifies the expression of heterogeneous data from multiple sources by means of packing agents to achieve efficient organization and management. Applying the dataspace concept to urban pipelines not only can utilize the dataspace concept to show the linear topology of pipelines as much as possible and enhance the geographic presentation, but also helps to solve the problem of integrating geographic data from different sources and types and provides a unified framework for integrating geographic information (Su, 2019). The three core elements of dataspace are subject, dataset and service, and current research mainly focuses on model construction (Guo et al., 2023). Existing dataspace models include the iDM model (Franklin et al., 2005), the CoreSpace model (Li and Meng, 2009), the Triple model, and the RDF model (Cheng et al., 2016), each with different advantages and disadvantages. Among them, the RDF model represents data in the form of triple, a structure that can flexibly describe the relationships and attributes between pipeline data. For the pipeline dataspace, the RDF model can be used to accurately express the topological relationships between pipelines, geographic location information, pipeline attributes, etc., which improves the semantic expression of the data; in addition, the RDF model is a graph-based data model (Song, 2023), which has good scalability. In the pipeline dataspace, the pipeline network is often dynamically changing, and the adoption of the RDF model can conveniently extend and modify the data model to meet the needs of the continuous evolution of the pipeline dataspace(Rong, 2015); moreover, the RDF model is a W3C-recommended standard,which has good standardization and interoperability, and the adoption of the RDF model makes the data format of the pipeline dataspace in line with international standards and It facilitates data exchange and integration with other systems, which promotes data sharing and cooperation (Forresi et al., 2023). Therefore, this model is adopted in this paper for the construction of pipeline dataspace.

The semantic-driven urban pipeline dataspace construction method proposed in this paper establishes a semantic-driven pipeline geographic entity expression framework based on the data content and characteristics of urban pipelines, and then embeds this expression framework into the dataspace RDF model to construct the pipeline dataspace, which in turn realizes the multilevel data organization of urban pipelines.

## 2. Study Area and Data Source

In this study, Baigou New City, Baoding City, Hebei Province, China, was selected as the study area. This area has a rich variety of pipelines and a complex distribution pattern, which plays an important role in the management and maintenance of urban infrastructure. All kinds of urban pipeline data in the whole area of Baigou(Hebei,China) New City are selected as the research object, including seven categories of water supply, drainage, gas, heat, communication, electric power and industrial pipelines. It

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-189-2024 | © Author(s) 2024. CC BY 4.0 License.

includes vector data of all kinds of urban pipelines, whose attribute fields include pipeline type, interface form, number of holes, flow direction, burial method, design pressure, etc. It also includes vector data of all kinds of urban tube wells, whose attribute fields include water level, mixing type, depth setting method, well cover shape, well grade, well material, etc. It also includes 3Dtiles data of urban pipelines and pipe points, which can support large-scale data access, dynamic loading and rendering. There is also a map of the current status of the pipe network from 2020 to 2022, which includes the ancillary facilities of the pipe network, such as the water supply plant and the water supply planning scope. The detailed data information is shown in Table 1, and some of the data are shown in Figure 1.

| Name | Format | Volume | Content |
|---|---|---|---|
| Pipelines | SHP | 600MB | Pipeline space, attribute information |
| | 3D Tiles | 3.0GB | |
| Tube well | SHP | 528MB | Tube well space, attribute information |
| | 3D Tiles | 3.1GB | |
| Pipe network appurtenances | jpg | 3MB | Spatial distribution, trend |

Table 1. Status of data for this study

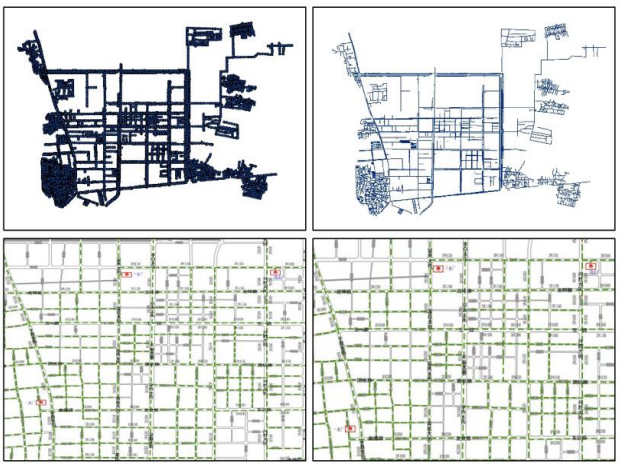

Figure 1. Partial data presentation

## 3. A Semantics-Driven Approach to Pipeline Dataspace Construction

### 3.1 Technological route

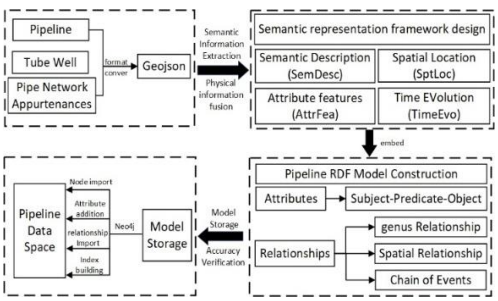


Figure 2. Technological route

The technical route of this research is shown in Figure 2, firstly, data preprocessing is carried out to convert all types of data into a unified format, then the semantic expression framework of pipeline geographic entities is designed by extracting semantic information and fusing entity information from the data, and the pipeline geographic entity information is expressed through four parts: semantic description set, spatial location set, attribute feature set and time evolution set, and then these four sets are The four sets are then embedded into the dataspace RDF model to express the attribute information of the pipeline RDF model, and at the same time, the association relationship of pipeline geographic entities is established from the aspects of genus class relationship and spatial relationship, and finally, the pipeline dataspace model is stored and visualized by neo4j to construct the association relationship between the entity data and the dataspace node to achieve the rapid query and retrieval of the data in the pipeline dataspace.

### 3.2 Semantics-driven RDF model construction for pipeline dataspace

RDF model is a kind of data model that describes and identifies information, in which subject, attribute and relationship are the three main elements for constructing pipeline RDF model. In this paper, combining the elemental fields and spatial information of pipeline geographic entities, a semantic-driven approach is adopted for the construction of spatial RDF model of pipeline data, which, at the attribute level, introduces the expression method of semantic features of geographic entities, and carries out the normalized representation of the attributes of pipeline geographic entities through the combination of semantic descriptions, spatial locations, attribute features and temporal evolution; and, at the relational level, combining the spatiotemporal information of GIS At the relational level, combining with the characteristics of GIS spatio-temporal information, spatial relationships and event chains are established to jointly express the associated relationships of pipeline geographic entities. The semantic-driven construction of pipeline data spatial model helps to improve the semantic expression ability of pipeline data, promote the integrity and interoperability of pipeline data, and effectively realize the construction and service application of pipeline dataspace.

| Field | Type | Description | Example |
|---|---|---|---|
| type | string | GeoJSON object types | "Feature" |
| id | number | Unique identifier of the feature | 1 |
| Geom | object | geometrical object information | [[lon, lat,...] |
| geom.type | string | geometric type | "LineString" |
| properties | object | Attribute information for features | {fid,id,zzms} |
| … | … | … | … |
| createTime | string | createTime | "18-Apr-2022" |

Table 2. GeoJSON data format

**3.2.1 Data preprocessing**: First of all, due to the variety of data formats needed in this study, and the different attribute information contained in different pipeline and pipeline point data, it is necessary to pre-process each type of data (Zhou et al., 2021), and to unify the format of the data.GeoJSON is a format for encoding various geographic data structures, and it is a geospatial information exchange format based on the JavaScript object representation method. Data exchange format, which can

effectively express the geometry, features or feature geometry of spatial data, and the use of JSON parsing library can be realized to parse and extract the semantic information of GeoJSON, so this study will be all kinds of pipeline data for the transformation of data format. For the vector data of pipelines and tube wells, the data format can be converted directly by ArcGIS software; for the data of pipeline network accessory facilities map, it is necessary to carry out the vectorization operation, which is firstly converted into vector data and then converted into GeoJSON data. The converted data contains information as shown in Table 2:

**3.2.2 Semantic feature representation of pipeline geographic entities**: In this paper, the canonical representation of the attributes of pipeline geographic entities is carried out in the form of semantic description, spatial location, attribute characterization, and temporal evolution combination. The semantic description of a geographic entity refers to the meaning of the concept to which the entity belongs, as well as the description of the uniqueness and essentiality of the entity (Ling et al., 2023). It is defined as.

$$SemDesc = \{EtyID, EtyType, EtyTypeID, EtyDH\}, \quad (1)$$

Where EtyID is the pipeline number, i.e., the unique identification of the pipeline; EtyType is the pipeline type; and EtyTypeID is the pipeline type layer name.

Spatial location describes the absolute spatial location information of the existence of the pipeline geographic entity, emphasizing the spatiality of the entity (Ke, 2016). For pipeline data, it is necessary to describe its start point coordinates as well as center point coordinates; for tube well data, it is necessary to describe its latitude, longitude and elevation information. Specifically defined as:

$$LocSpt = \{geomStart, geomLast\}, \quad (2)$$

$$LocSpt = \{geomX, geomY, geomH\}, \quad (3)$$

Where geomStart is the pipeline start coordinate; geomLast is the pipeline termination coordinate. geomX, geomY, and geomZ are the latitude, longitude, and elevation of the tubewell, respectively.

Attribute features are the portrayal of non-spatial information of pipeline geographic entities (Liu et al., 2017), which record the basic attributes contained in the pipeline geographic entities, such as laying mode, pipe diameter, pipe type, interface form, etc., and use the attribute abbreviations to indicate their meanings. Specifically defined as:

$$AttrFea = \{fsfs, gj, gxzl, jkxs \cdots\}, \quad (4)$$

Temporal evolution describes the time-related information contained in a pipeline geographic entity, which includes, on the one hand, temporal information related to the operational state of the pipeline, e.g., time of creation, state of use, time of closure, etc., and, on the other hand, records temporal information about the occurrence of certain events in the pipeline entity, which is specifically defined as:

$$TimeEve = \{pipeTime, EventTime\}, \quad (5)$$

**3.2.3 Pipeline dataspace RDF triples**: A collection of RDF triples consists of an RDF graph, where the subject and object (Han et al., 2022) correspond to the vertices and predicate to the edges in the graph, respectively. In this paper, we establish a spatial RDF model for pipeline data, and for a single geographic entity, the semantic description, spatial location, attribute features, and temporal evolution of the semantic feature expression of pipelines proposed in the previous section are embedded into the RDF model, which in turn standardizes the expression of pipeline information (Chen et al., 2019). The specific embedding form is: semantic description, spatial location, attribute features and time evolution as the first-level nodes, stored in the form of key-value pairs, such as the ternary group (water pipeline, rdf:type, SemDesc) indicates that "water pipeline contains semantic description set". And all kinds of features express specific information as a secondary node, also in the form of key-value pairs to store all kinds of specific information of the pipeline, such as ternary (AttrFea, rdf: gj,23) that (in the attribute feature set, the diameter of the pipe for the 23), by this way you can achieve the complete construction of the pipeline information, the construction of the program example is shown in the following figure 3 and figure 4:

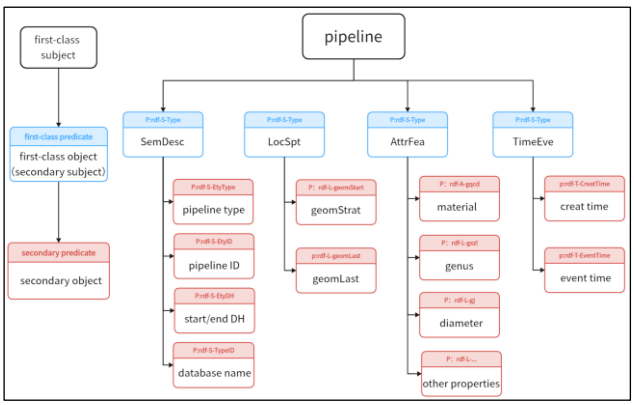


Figure 3. pipeline RDF.

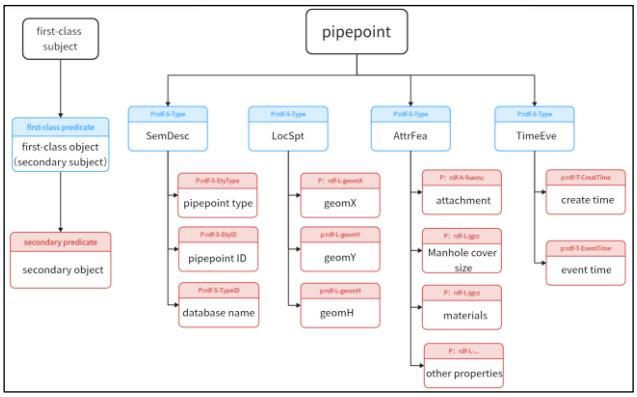


Figure 4. pipepoint RDF.

**3.2.4 Pipeline entity association relationship construction**: After completing the construction of pipeline RDF triad, the construction of association relationship between different pipeline geographic entities is needed. Therefore, this paper considers to use genus and spatial relationships to express the association relationship of RDF model. Element classification refers to the Underground Pipeline Element Data Dictionary (GB/T 41455-2022) and the National Standard for Engineering Surveying (GB 50026-2020) to categorize the various types of data of urban pipelines, which includes the major classes of pipelines (points) as well as the genera of pipelines (points) for categorization. The major categories include drainage, water supply, gas, heat, electricity, communication and industry, and the genera corresponding to the major categories are shown in

Table 3 and Table 4 below:

| Main class | Pipeline genus |
|---|---|
| drainage | Stormwater, sewage, combined flow, etc. |
| water supply | Raw water, water transfers, mid-water, etc. |
| gas | Gas, liquefied petroleum gas, natural gas, etc. |
| heat | Hot water, steam, etc. |
| electricity | Power supply, street lighting, traffic signals, etc. |
| communication | Telephone, limited television, information networks, etc. |
| industry | Hydrogen, oxygen, acetylene, etc. |

Table 3. Classification of pipeline elements

| Main class | Pipepoint genus |
|---|---|
| drainage | Manholes, hydrants, underground hydrants, fire wells, valves, etc. |
| water supply | Inspection wells, overflow wells, gate wells, drop wells, etc. |
| gas | Test wells, condensate tanks, compensators, etc. |
| heat | Manholes, service manholes, condensate tanks, etc. |
| electricity | Transformers, maintenance shafts, junction boxes, ventilation shafts, control cabinets, etc. |
| communication | Manholes, handholes, distribution boxes, transfer boxes, etc. |
| industry | Flow meters, compensators, boiler rooms, pumping stations, etc. |

Table 4. Classification of pipepoint elements

## 4. Validation of the Effect of Spatial Organization of Urban Pipeline Data

In this study, the graph database Neo4j was used to store and visualize the constructed urban pipeline dataspace. The RDF model of the pipeline dataspace was successfully mapped to the Neo4j database through operations such as the construction of genus-class relationships, the definition of pipeline geographic entity nodes, and the establishment of event chain association relationships. In this model, each node represents a subject or object in the RDF model of pipeline dataspace, and each association relation represents a predicate. The specific construction results are shown as follows.

Figure 5 shows the construction of pipeline genus relationship, which clearly presents the association relationship between various types of pipelines and their genus. Figure 6 illustrates the construction of pipe point genus relationships to visualize the association between various types of pipe points and their genera. It is worth noting that the construction of these two types of relationships does not involve specific instances of a single pipe point or pipe line. Figures 7 and 8, on the other hand, show the storage structure of a single pipe line and pipe point RDF model. In this figure, the instance of a single pipe point is built at the next level of the pipe point genus class relationship, and its nodes store a quaternion that includes semantic description, spatial location, attribute features, and temporal evolution to fully express the various types of information of each pipeline geographic entity.

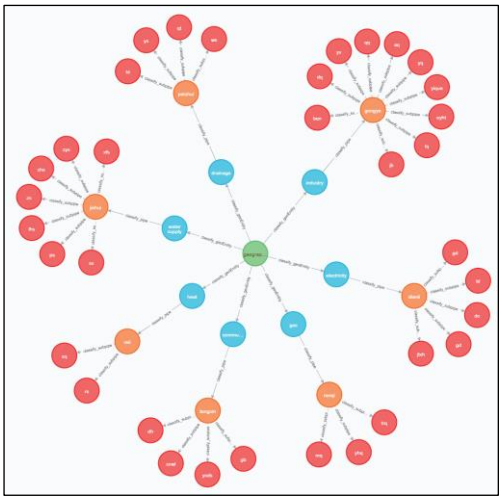

Figure 5. Pipeline Category Relationships.

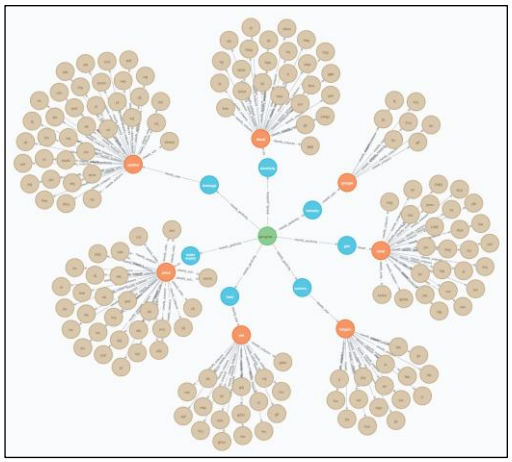

Figure 6. Pipepoint Category Relationships.


Figure 7. Single pipeline RDF.

Figure 9 illustrates the spatial association between pipelines and pipe points. Through the application of Neo4j database, effective storage, management and visualization of urban pipeline dataspace are successfully achieved, which provides important support for further analysis and utilization of pipeline data.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-189-2024 | © Author(s) 2024. CC BY 4.0 License.

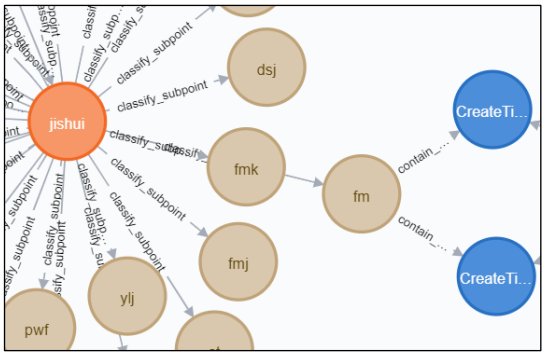


Figure 8. Single pipepoint RDF.

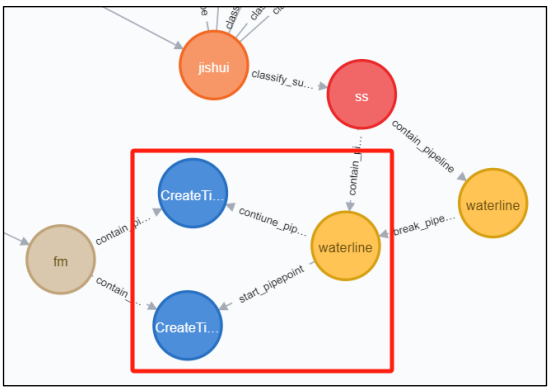


Figure 9. Spatial relationship building.

After completing the construction of the pipeline dataspace, the same data are stored in the relational database PostgreSQL, and the data organization effect of the pipeline dataspace is verified by comparing the retrieval efficiency of the two under different data volumes. In this study, experiments were conducted on five parts such as querying a single pipeline and entity range query, respectively. Figure 10 shows the results of the first time data retrieval time comparison between pipeline dataspace and PostgreSQL, and it can be concluded that the response time of the two is basically comparable when querying single entity query and entity attribute query. As the data volume continues to increase, the pipeline dataspace shows better performance performance. With the further increase of data volume, the query time of pipeline dataspace is significantly better than PostgreSQL, and the maximum query efficiency is improved by about 10%, which is more effective for pipeline data organization. At the end of the first retrieval data experiment, a second retrieval of the same data was performed, and the retrieval results are shown in Figure 11. Based on the comparison of the two retrieval times, it can be concluded that the time required for the secondary retrieval of the two methods is much faster than that of the first retrieval. As the data volume increases, the query time increases as well, but the time for the secondary retrieval using the pipeline dataspace is also significantly better than using the PostgreSQL database, and the maximum query efficiency is improved by about 15%, which indicates that the pipeline dataspace has a good data organization effect.

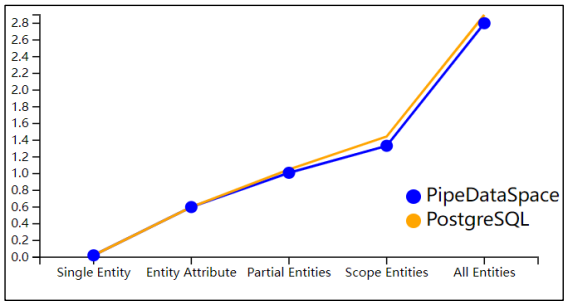


Figure 10. Comparison of first search time.

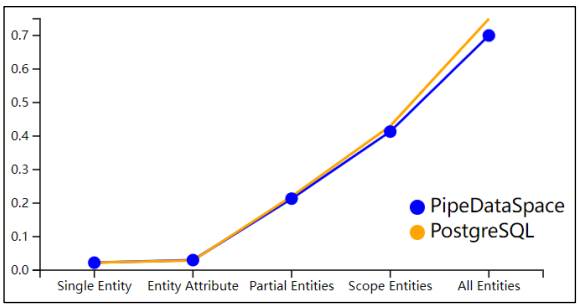


Figure 11. Comparison of secondary search times.

## 5. Conclusion

In this study, based on the classification and characteristics of urban pipelines, the semantic information of pipeline geographic entities is comprehensively expressed in four dimensions, namely, semantic description, spatial location, attribute characteristics and temporal evolution. Subsequently, these four expression sets are embedded as predicates into the RDF model of the dataspace, and the association description mechanism between the spatio-temporal elements of pipeline geographic entities is established through the genus and spatial relationships, and the RDF model of the pipeline dataspace is finally constructed and mapped to the Neo4j graph database for storage and visual expression, and the graph database Cypher query language is utilized to achieve the data retrieval of the pipeline data retrieval in the pipeline dataspace using the graph database Cypher query language, and the data organization results are better. The proposed method solves the data conflict and organizational management difficulties caused by the heterogeneity of pipeline data from multiple sources, and provides an effective method for efficiently organizing urban pipeline data.

## Acknowledgements

This work was supported by the Research Project of China Academy of Railway Sciences Corporation Limited (Project No. 2022YJ299).

## References

Chen, J., Liu, W.Z., Wu, H., Li, Z.L., Zhao, Y., Zhang, L., 2019: Basic Issues and Research Agenda of Geospatial Knowledge Service. Geomat. Inform. Sci. Wuhan Univ., 44(1), 38-47. doi:10.13203/j.whugis20180441.

Cheng, B., Guan, X.F., Xiang, L.G., Gao, M., Wu, H.Y., 2017: A Conceptual Data Model for Dynamic Changes Expression of

Spatio-temporal Object and Its Association. J. Geo-Inform. Sci., 19(11), 1415-1421. doi:10.3724/SP.J.1047.2017.01415.

Dai, Y.B., 2018: Data organization method of 3D city model based on object importance. MA thesis, Wuhan University.

Forresi, C., Gallinucci, E., Golfarelli, M. et al. A dataspace-based framework for OLAP analyses in a high-variety multistore. The VLDB Journal 30, 1017–1040 (2021). doi.org/10.1007/s00778-021-00682-5.

Franklin, M. , Halevy, A. , & Maier, D. . (2005). From databases to dataspaces: a new abstraction for information management. SIGMOD record: ACM SIGMOD (management of data)(4), 34.

Ministry of Natural Resources, People's Republic of China, 2022: Data dictionary for underground pipeline features, GB/T 41455-2022. China Standards Press, Beijing.

Ministry of Housing and Urban-Rural Development, 2020: Engineering survey standard, GB 50026-2020. China Standards Press, Beijing.

Han, X., Zhang, Z.Q., Yan, L., 2022: Temporal RDF Modeling Based on Relational Database. Comput. Sci., 49(11), 90-97. doi:10.11896/jsjkx.211100065.

Jingwei Guo, Ying Cheng, Dongxu Wang, Fei Tao & Stefan Pickl (2023) Industrial Dataspace for smart manufacturing: connotation, key technologies, and framework, International Journal of Production Research, 61:12, 3868-3883.doi.org/10.1080/00207543.2021.1955996.

Jeffery, S. R. , Franklin, M. J. , & Halevy, A. Y. . (2008). Pay-as-you-go user feedback for dataspace systems. ACM. doi.org/10.1145/1376616.1376701.

Ke, S., 2016: Adaptive map visualization for spatial data distribution features. PhD thesis, Wuhan University.

Ling, Z.Y., Li, R., Wu, H.Y., Li, J., Gui, Z.P., 2023: Semantic-driven construction of geographic entity association network and knowledge service. Acta Geod. Cartogr. Sini., 52(3), 478-489. doi:10.11947/j.AGCS.2023.20210349.

Li, Y. , & Meng, X. . (2009). Exploring Personal CoreSpace for DataSpace Management. Semantics, Knowledge and Grid. IEEE Computer Society.

Liu, C., Yin, H., Zhang, Y.T., Xie, X., Cao, Z.Y., 2018: WebGL-based 3D Pipeline Lightweight Visualization Method. J. Spatio Temp. Inform., 25(4), 48-51,57. doi:10.3969/j.issn.1672-1586.2018.04.009.

Liu, Z.H., Li, R., Wang, J.Q., 2017: A Dynamic Representation Method of Considering Semantic Scales of Attributes of spatio-temporal Object. J. Geo-Inform. Sci., 19(9), 1185-1194. doi:10.3724/SP.J.1047.2017.01185.

Rong, J.T., 2015: Research on In-Depth Ordering Mechanism of Knowledge Organization Based on Linked Data. Libr. Inform. Serv., 59(13), 134-141. doi:10.13266/j.issn.0252-3116.2015.13.019.

Song, X.Y., Zhang, W.M., Zhang, X.Q., 2023: Storage of Semantic Knowledge Hypergraph Based on a Resource Description Framework. J. China Soc. Sci. Tech. Inform., 42(8), 967-979. doi:10.3772/j.issn.1000-0135.2023.08.008.

Su, C., 2019: Research on entity parsing methods for dataspace. MA thesis, Harbin Engineering University.

Wang, J.J., 2021: Research on 3D GIS modeling method of urban underground integrated pipe corridor based on BIM. MA thesis, Nanjing Normal University. doi:10.13474/j.cnki.11-2246.2015.0603.

Zhang, B.G., Yang, B.G., Tao, Y.C., 2015: Research on fine management mode of urban underground pipeline in Beijing. Bull. Surv. Mapp., 11-15. doi:10.13474/j.cnki.11-2246.2015.0603.

Zhou, C.H., Wang, H., Wang, C.S., Hou, Z.Q., Zheng, Z.M., Shen, S.Z., Cheng, Q.M., Feng, Z.Q., Wang, X.B., Lv, H.R., Fan, J.X., Hu, X.M., Hou, M.C., Zhu, Y.Q., 2021: Research on Geoscience Knowledge Graph in the Era of Big Data. Sci. Sin., 51(7), 1070-1079. doi:10.1360/SSTe-2020-0337.