Semantic sensor data federation in dynamic knowledge graphs using RDF-star

Yingying Zhang[1] and Jakob Beetz[1]
[1]RWTH Aachen University, Germany
yzhang@dc.rwth-aachen.de

Abstract.

The ability of Linked Data (LD) to handle heterogeneous datasets has been demonstrated, with promising initial results in the Architecture, Engineering, and Construction (AEC) domain for linking unstructured or semi-structured data. However, data in the AEC sector often assumes that the underlying knowledge graph is static, whereas in reality, events are composed of a series of temporal evolutions. Especially in sensor networks, a large number of dynamic events can lead to insights into time. In this research, we propose a method for dynamic knowledge graphs based on RDF-star syntax, which can integrate diverse joint resources generated by sensor network systems and support efficient storage and querying. Through the case study, performance based on RDF-reification modelling approach, RDF-star approach, and a hybrid modeling approach of RDF-star plus times series database were tested. Finally, through a query case, the flexibility of SPARQL federated queries across RDF resources and non-RDF resources was demonstrated.

1. Introduction

As the Internet of Things (IoT) and wireless sensor technologies continue to proliferate, distributed sensor networks in buildings are becoming an important source of highly accurate data for facilitating building energy retrofit. In recent years, Semantic Web (SW) technology has emerged as a promising complement to Building Information Modeling (BIM) technology in the Architecture, Engineering, and Construction (AEC) domain. This is due to SW's ability to promote interoperability between different data sets and establish links across multiple domains Pauwels et al. (2017). Despite previous efforts to integrate heterogeneous building-related resources into static knowledge graphs through Linked Data, a significant amount of sensor network data remains stored in native formats (e.g., CSV, relational database and time-series database) independent of sensor contextual information. Consequently, there is an urgent need to develop efficient methods for integrating sensor network information in heterogeneous smart buildings and establishing a common access protocol for retrieving information in a timely and efficient manner.

The dynamic and private nature of the data generated by the Internet of Things (IoT) presents obstacles to data federation efforts. In order to effectively coordinate and integrate building context data with sensor network datasets within a network context, the ideal scenario would involve an open framework in which any user can publish information on the network. In this scenario, a mechanism for identifying changes and declaring authorship would be required, and access to data would be governed by user-defined data disclosure preferences.

In this research, an innovative approach that utilizes dynamic knowledge graphs to federate contextualized time series data in buildings was proposed. Our approach takes into account the temporal changes, provenance information, and contextual environment information in sensor networks. Firstly, a trade-off is made between performance and interoperability to address the heterogeneous nature of building environment data. The data is stored in corresponding databases based

on the characteristics of the data, while heterogeneous static information is mapped and stored in RDF-star. The RDF-star's powerful metadata annotation capabilities allow capturing and annotating different types of metadata related to sensor networks, including temporal evolution and authorization information. As a result, data is federated in the dynamic knowledge graph. From efficiency considerations, time-series data are stored based on a retention policy in the time-series database, while a common access protocol based on SPARQL can be used to perform federated queries on the graph database and TSDB. Finally, with the support of a more flexible SPARQL federated query syntax, complex queries across heterogeneous resources with specified patterns can be performed.

2. Related Work

The interoperable ecosystem in Linked Building Data (LBD) is becoming a topic of interest, and the process of heterogeneous information evolving over time is an indispensable part of building a complete knowledge graph. Supported by existing research, this paper introduces dynamic knowledge graphs to the building sensor network domain to interpret state information. Here we review the existing heterogeneous data integration approaches, decentralized data ecosystem for building sensor networks, and how to describe state information in dynamic knowledge graphs.

2.1. Heterogeneous Semantic Sensor Data Interpretation in Knowledge Graph

In the context of IoT-assisted AEC industry, low-power and wide-area sensor systems enable real-time monitoring and feedback of intelligent buildings. The integration of heterogeneous data in BIM and IoT has gained a lot of attention Isikdag (2015), and the native IFC schema is extended to support IoT external modules Ruiz-Zafra et al. (2022). Although the extended IFC schema can only provide limited and case-level support for IoT solutions due to the weak support for sensor entities in IFC schema and the constraints of geometry-centric features.

The Semantic Web community's research in recent years has provided a new paradigm for the interoperability of heterogeneous systems. W3C Standards such as RDF, OWL, SPARQL, etc Hayes (2004). enable linking heterogeneous datasets on the web, providing query services, and performing rule-based reasoning.

A knowledge graph is a structured representation of information by nodes, edges and labels, by extracting the implicit document information structure and organizing it into usable knowledge Auer et al. (2018). In order to exchange information through common protocol or standard, generic domain ontology has been created in multi-domains based on consensus. For example, the IFC EXPRESS schema was interpreted as ifcOWL and became an open standard for semantic data exchange in building and construction sectors Beetz et al. (2009). The Semantic Sensor Network (SSN) ontology and the SOSA (Sensor, Observation, Sample, and Actuator) ontology are proposed to describe sensor and actuator observations as well as contextual information Compton et al. (2012) Haller et al. (2019). The Semantic Web and Linked Data are rapidly growing in the industry with their advantages in heterogeneous dataset compatibility, and big data around the RDF schema have shown performance in large-scale reasoning and complementary role for artificial intelligence systems.

2.2. Decentralized Data Ecosystem for Building Sensor Networks

As previously mentioned, the Semantic Web and linked data are also in a process of continuous evolution, and one of the major challenges is how to federate data in a secure and efficient way in the interoperable ecosystem. The main focus in the building and construction sector is currently on the feasibility of applying Linked Data, while the hierarchical autonomy of information has not been widely addressed. Meanwhile, the management of sensor data in buildings has become a challenge due to its privacy characteristics in data ownership and authentication. As the development of self-sovereign identity (SSI) methods for securely transmitting information over networks has become a consensus Tobin & Reed (2016), and data ownership management for Linked Data technologies is no exception. The Solid protocol which led by Tim Berners-Lee, provides a decentralized platform for linking data Capadisli (2020), aiming to improve privacy issues in the Semantic Web by allowing users to take ownership of the data. Oraskari et al. (2022) proposed a distributed CDE platform approach to exchange BIM Collaboration Format (BCF) information in the building and construction domain, and the solid protocol was used to decentralize the data ownership mechanism. The solid protocol was used in conjunction with the RDF-star syntax to self-validate resources on the web Braun & Käfer (2022). Each person in the decentralized data ecosystem has a private database and is able to selectively share the updated information in a timely manner, thus avoiding the inefficiency of repeated dissemination of stale and erroneous information, which is the significance of the existence of dynamic knowledge graphs.

A potential challenge in building decentralized data ecosystems for sensor networks, as opposed to statically stated data in the AEC domain, is ensuring that the network can scale to accommodate the volume of data generated by sensors. The management of high-volume sensor data in decentralized system will be addressed in this paper. Overall, the decentralized data ecosystem provides an efficient, secure, and interoperable way for building sensor networks to federate data organically.

2.3. Describe Contextual Information in RDF

In RDF models Hayes (2004) Manola et al. (2004), data and metadata are typically not distinguished, and data are represented as triples consisting of subject, predicate, and object. The RDF model lacks annotation support at the metadata level. There are different annotation requirements in different use scenarios, such as the need to add provenance information, certainty, location or temporal information to the data. The standard RDF Reification provides a simple and intuitive way to write the subject, predicate, and object along with metadata as attributes in a triple. However, it has low semantic standardization and can only add statement-level data, which inevitably leads to inefficiency due to the addition of extra triples Nguyen et al. (2014). While the Named Graph Carroll et al. (2005) approach adds metadata to the $< s, p, o >$ triples, it is more concise compared to RDF Reification, allows for arbitrary combinations of RDF statements, and has gained broader support from standardized tools. However, to some extent, data redundancy remains high, and related data may be duplicated in multiple graphs. The Singleton Properties approach allows statements to be added to predicates to form new triples, but it has high query complexity and low efficiency. Other practices involve time annotations on RDF graphs, such as Temporal RDF, which Zhang & Beetz (2023) applied in the AEC domain to capture the temporal evolution of data. In general, different data schemas have their own application scenarios, and the appropriate syntax should be chosen based on practical considerations such as the clarity of semantics, data volume requirements, or query performance requirements.

With the proposal and improvement of RDF-star, it extends the RDF syntax to allow nested triple annotations in both subject and object positions. Compared to other methods, RDF-star provides more concrete syntaxes, and it has been shown to be more efficient than other RDF methods in expressing statement-level metadata. Considering the high volume and complexity of sensor network information, RDF-star was chosen as the model for the application in this paper.

3. Methodology

Through literature review and survey, a dynamic knowledge graph solution for contextualized time series data is proposed to log metadata from multi-data sources such as authenticity, versioning, and authorship. The iterative process of knowledge graphs is captured through statement-level annotations. Specifically, the following steps are used to conduct the study:

1) SSN ontology and extended SSN-log ontology module are used to describe sensor and observation data. Weather data from external sources, building context data, and information uploaded by different collaborators all work in the same graph through RDF format.

2) Information about version updates during collaborative work is brought into RDF graph through annotations. The RDF-star syntax will be used here.

3) Performance benchmarking was performed to compare three modeling approaches: RDF-reification, RDF-star, and RDF-star plus TSDB modeling approaches. Also, a query example of building sensor network was used to test federated queries across multiple resources. Some experiences and findings in the application process are summarized.

3.1. Verifiable Log Extension Module for SSN Ontology

In this work, the PROV ontology Khalid Belhajjame (2013) was reused to describe provenance information and the Verifiable Credentials Data Model Manu Sporny (2013) was referred to extend it into a verifiable log extension module for SSN. In the SSN model sosa:Sensor, sosa:Actuator, sosa:Sampler, and other entities are aligned as prov:Agent, but they are restricted to describe the log information and attributes of the device, while in an ideal verifiable decentralized sensor network ecosystem, different types of agents such as person, device, and organization business processes are logged and data validation is accomplished by separating the holder from the issuer to perform different roles.

The preferred prefix for the extended module is ssn-log, a glossary that includes 3 classes additions and 3 property additions. As shown in Figure1.

`ssn-log:agent`: agent class is the owner of the activities defined by ssn-log, an agent can be person, group, or physical artifact, example holders include students, employees, and software bots.

`ssn-log:publisher`: publisher is responsible for publishing events and verifying that credentials and schemas are compliant to ensure that the published content is legal and compliant.

`ssn-log: versionInfo`: it generates version identifiers and stores records in a machine-readable data format

`ssn-log:hasStartTime`: the start time of an event.

`ssn-log:hasEndTime`: the end time of an event.

`ssn-log:hasLocation`: spatial information associated with a sensor or an observation.

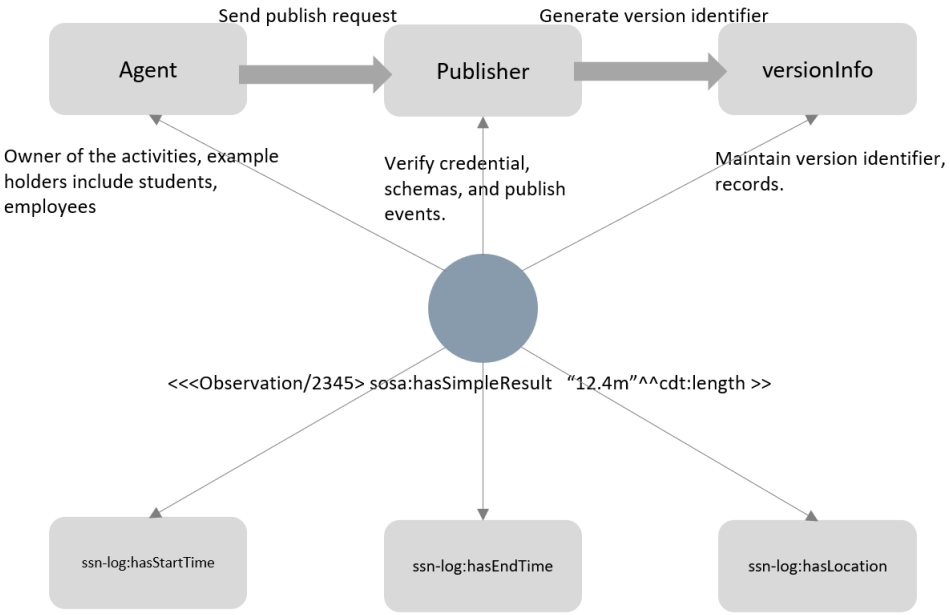

Figure 1: SSN-log extension vocabularies

3.2. Mapping Heterogeneous Metadata to RDF-star Representation

In building sensor networks, there is a large variability of data. Traditional linked building data is mainly hierarchical, ranging from geometric data to business process exchanges, and the volume of data is not particularly large. While sensor network information is time-series data and is often characterized by high data noise and high data volume. The building data is usually categorized as the contextual information of the sensor network. In this study, the sensor networks in buildings are the main subject of research, and a reasonable storage structure is required to effectively federate the large volume of sensor data with building information in dynamic knowledge graphs. The hybrid storage architecture can be used as a flexible approach to cope with the large volume of data and hierarchical contextual information,i.e., time series data is stored in the native database, and the linked data approach can be used to link heterogeneous contextual information from multiple sources and index the values in the time series database by identifier. This hybrid storage structure has been proven to be highly efficient Hu et al. (2016).

In the proposed approach, data are categorized and stored by corresponding data features, see in Figure2. In general, data in sensor networks can be classified into four categories: static data, dynamic data, state data, and contextual data, which are stored in Graph Database and Time Series database according to the data storage cost, scalability, and openness principles. As shown in Figure2, the sensor metadata is stored in time series databases such as Influx DB to ensure efficient query and storage due to its high volume and redundant nature. While heterogeneous static data, state data, sensor context data, and building context data are stored in graph databases to be linked on the web.

Since time series data are constantly updated and prone to being redundant, we divide the sensor read values into different influxDB instances according to the database retention policy, and then use the HTTP API to create a URL for each instance and attach the necessary authentication and

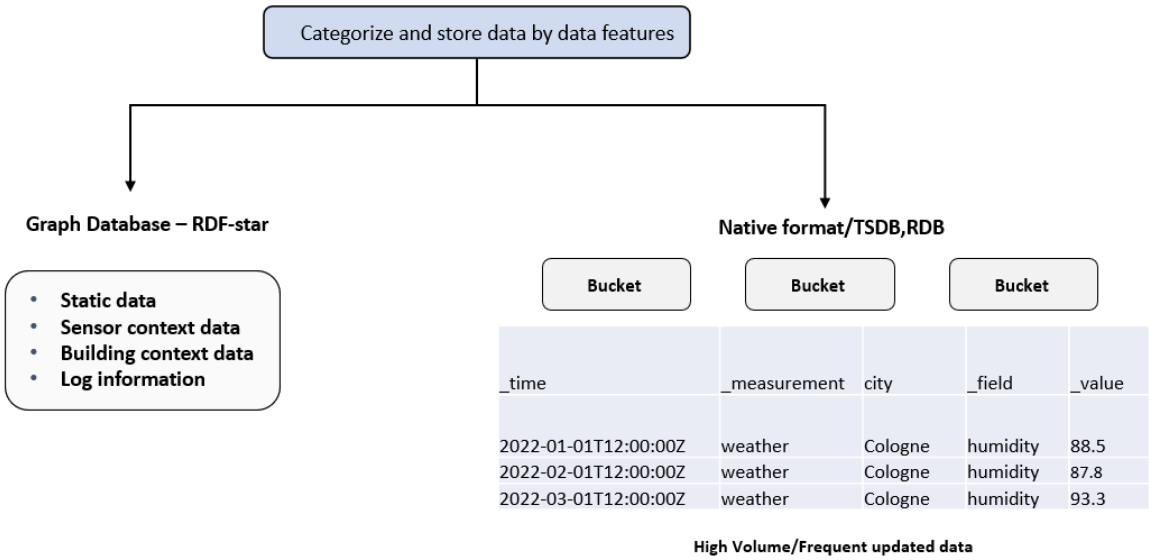

Figure 2: Hybrid data storage structure for contextualized time series data

configuration information. As shown in Figure 3 SPARQL federated query syntax can execute queries on data distributed over different endpoints, whether it is graph databases or RDF triples transformed through middleware. The common SPARQL endpoint server, Apache Jena Fuseki, can be configured to support federated queries across multiple data sources. Given the flexible requirements of specific scenarios, either the Time-series data instances are linked to the RDF graph for querying via URLs as indexes, or the data can be retrieved from the time-series database using the REST API and converted to Linked Data format for loading into the HTTP repository, and then the corresponding information in the HTTP repository and the original graph can be queried federated.

Meanwhile, the W3C community has developed a language for mapping RDB to RDF, R2RML, which supports mapping heterogeneous datasets to RDF based on a custom structure and enables Linked Data. As RDF-star was proposed and supports providing statement annotations on RDF, an extended RML-star Delva et al. (2021) syntax was introduced which allows defining data from any other source format as an RDF dataset. In this study, RML-star is used to customize the conversion of data in time-series DB to RDF graphs.

3.3. Represent Semantic Sensor System State Information in Knowledge Graph

The Semantic Web enables a standardized approach to facilitate the integration, retrieval, and reasoning of heterogeneous data in a machine-readable language. Although Semantic sensor networks (SSN) ontology follows a standardized glossary to describe sensor networks and contextual information. However, the current approach does not support state information well and therefore there is a need to investigate the use of RDF-star to document iteration records and data provenance information in knowledge graphs.

In the previous section, we proposed the verifiable SSN-log extension module, since data in sensor networks are autonomous and highly self-governed information, the SSN-log vocabulary allows to describe verifiable provenance and state information in the knowledge graph. The emer-

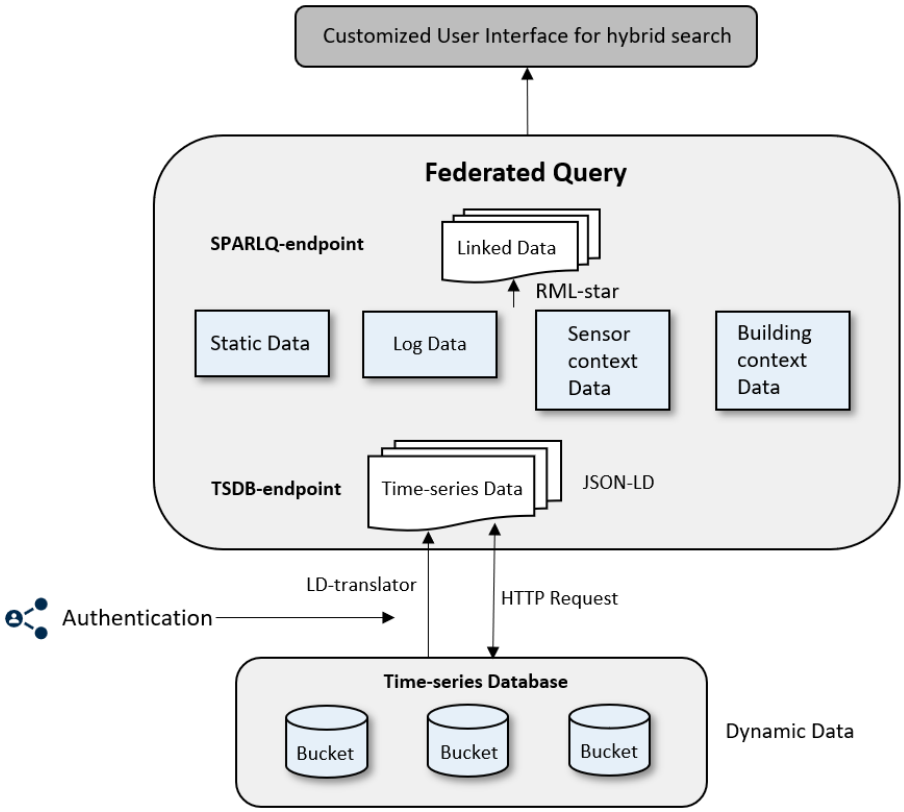

Figure 3: Hybrid architecture for efficient retrieval of heterogeneous information

gence of RDF-star provides the possibility to add statement-level annotation to traditional triples, in the following listing 1, the instance of observation is extended by the SSN-log module using the rdf-star syntax.

```
<< <Observation/235714> rdf:type sosa:Observation >> ssn-log:agent   "
   YingZhang"^^xsd:string;
    ssn-log:publisher "DC chair, Uni.";
    ssn-log:versionInfo "m18022023" ;
    rdf:type time:Interval;
    ssn-log:hasStartTime "2023-02-18T00:00:00+00:00"^^xsd:dateTimeStamp;
    ssn-log:hasEndTime "2023-02-19T00:00:00+00:00"^^xsd:dateTimeStamp.
<< <Observation/235714> sosa:hasSimpleResult   "26.5"^^xsd:double >> rdf:type
   time:Instant;
sosa:resultTime "2023-02-18T00:00:12+00:00"^^xsd:dateTimeStamp .
```

Listing 1: Representing Contextual Data in Semantic Sensor Using RDF-star

4. Evaluation

In this section, we evaluate the modeling approach proposed in this paper in terms of performance benchmarks and federal query applications. We build an evaluation scenario using a real case of a university office and compare the modelling performance of RDF-reification, RDF-star, and RDF-

star + TSDB hybrid method. We also discuss flexible queries based on RDF-star modeling from the perspective of query applications.

4.1. Performance Benchmark

The evaluation scenario uses a built environment monitoring system set up in the university office where the data is stored in a time series database, see in Figure 4. Approximately 1000 sets of monitoring key-value from one of the monitoring nodes were extracted for benchmarking. We evaluated RDF reification modeling, RDF-star modeling, and hybrid RDF-star and TSDB modeling approaches in expressing contextualized time series data, respectively. In Figure 5, we have differentiated the portions of time-series data and contextual data in the dataset using colors. It can be seen from the first bar that the standard RDF reification method generates relatively large amounts of data, with particularly abundant contextual data. The data redundancy caused by the repetition of subject, predicate, and object in the reification syntax. In contrast, the RDF-star method produces nearly half the total triples compared to the RDF-reification modelling approach. In addition, we also tested the hybrid modeling approach using both RDF-star and TSDB, where the time series data are all stored in the TSDB, so this graph only contains contextual information and a small number of observation identifiers. As can be seen from the figure, the number of statements is reduced sharply because a large amount of contextual information about the observations is repeated, so we only need to describe the data blocks according to the data chunking principle (e.g., according to event-based segmentation or retention policy), thus saving a lot of memory in expressing contextual information, the hybrid method is both efficient and scalable.

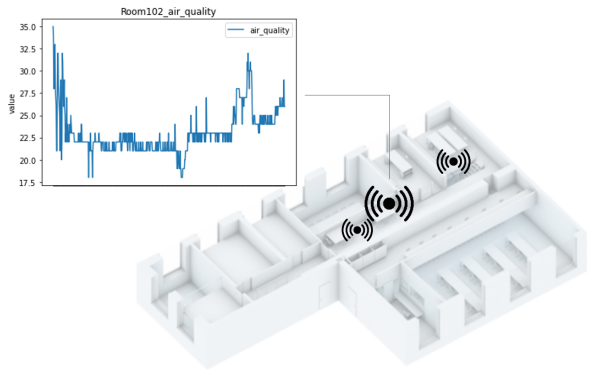

Figure 4: Contextualized time-series data in university building

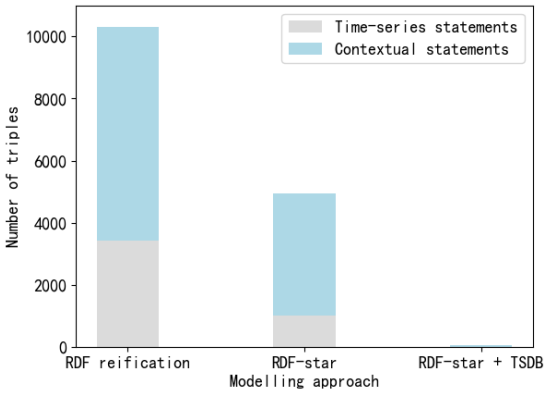

Figure 5: Performance of different modeling approaches

4.2. Federated Query Implementation

SPARQL Federated Query extends the standard SPARQL query syntax, which allows for more flexible queries across multiple data sources. In this case, we use an IoT-enabled building system in university as a motivated scenario. Where the context data is stored in RDF and the time series data is stored in a time series database in its native format. We performed the query in Apache Fuseki (see in Figure 6), where the contextual data and time series data are linked in separate endpoints.

Facade-X is used here which provides a unified way to access heterogeneous data sources Asprino et al. (2023). The sensor values, agent, sensor type, observed property, and publisher are retrieved from RDF and JSON by the federated query.

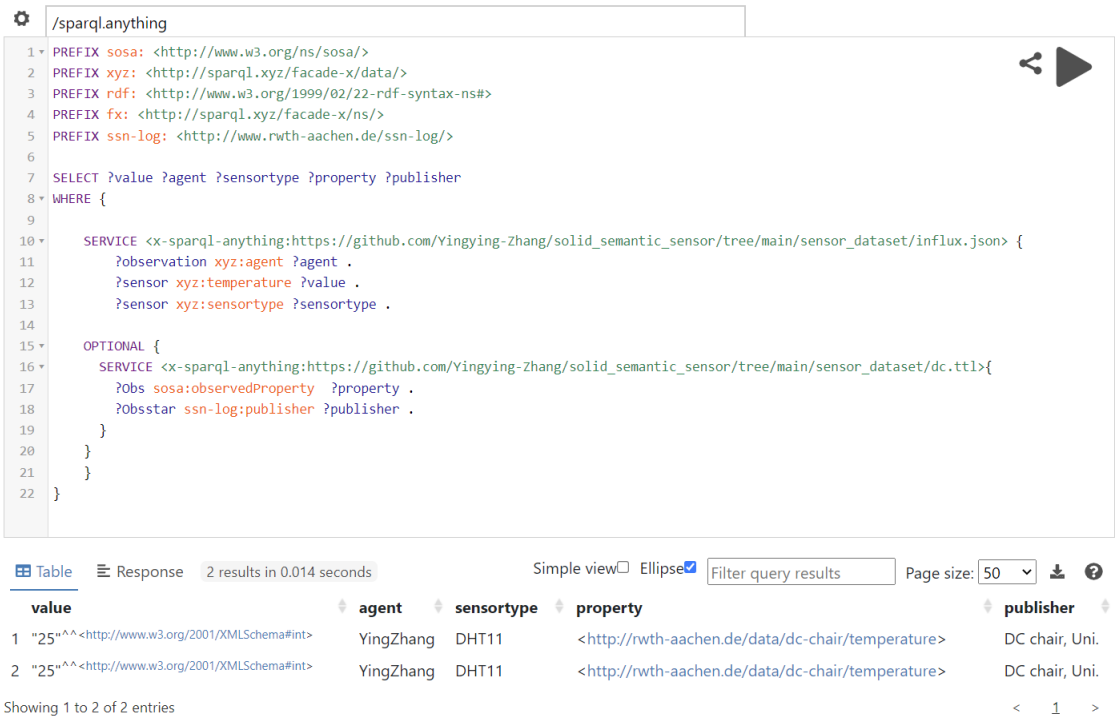

Figure 6: Federated query to access heterogeneous contextualized time series data

5. Conclusion

In this research, we present a method to federate heterogeneous sensor network data in a dynamic knowledge graph using RDF-star. The dynamic evolution of knowledge graphs is noticed as IoT-assisted building systems are no longer static models but rather dynamic systems with continuous iterations involving multiple parties. The Semantic Web demonstrates its immense potential in data exchange and interoperability. In our envisioned scenario, different users decentralized controlling different resources, authorization information and status information are attached to the data as metadata and formalized in RDF-star graphs. Complex queries based on SPARQL-star enable retrieval across multiple named graphs, and decentralized data systems can be federated for queries. This study explores the potential of the semantic web in expressing the temporal evolution of data in the AEC field, without being limited by the representation constraints of static knowledge graphs. Future research in dynamic knowledge graph reasoning capabilities is also an interesting challenge.

References

Asprino, L., Daga, E., Gangemi, A. & Mulholland, P. (2023), 'Knowledge graph construction with a façade: a unified method to access heterogeneous data sources on the web', *ACM Transactions on Internet Tech-*

nology **23**(1), 1–31.

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M. & Vidal, M. E. (2018), Towards a knowledge graph for science, *in* 'Proceedings of the 8th international conference on web intelligence, mining and semantics', pp. 1–6.

Beetz, J., Van Leeuwen, J. & De Vries, B. (2009), 'Ifcowl: A case of transforming express schemas into ontologies', *Ai Edam* **23**(1), 89–101.

Braun, C. H.-J. & Käfer, T. (2022), Self-verifying web resource representations using solid, rdf-star and signed uris, *in* 'The Semantic Web: ESWC 2022 Satellite Events: Hersonissos, Crete, Greece, May 29–June 2, 2022, Proceedings', Springer, pp. 138–142.

Capadisli, S., B.-L. T. V. R. K. K. (2020), 'SPARQL Query Language for RDF'.
URL: *https://solidproject.org/TR/protocol*

Carroll, J. J., Bizer, C., Hayes, P. & Stickler, P. (2005), 'Named graphs', *Journal of Web Semantics* **3**(4), 247–267.

Compton, M., Barnaghi, P., Bermudez, L., Garcia-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A. et al. (2012), 'The ssn ontology of the w3c semantic sensor network incubator group', *Journal of Web Semantics* **17**, 25–32.

Delva, T., Arenas-Guerrero, J., Iglesias-Molina, A., Corcho, O., Chaves-Fraga, D. & Dimou, A. (2021), Rml-star: A declarative mapping language for rdf-star generation, *in* 'ISWC2021, the International Semantic Web Conference', Vol. 2980, CEUR.

Haller, A., Janowicz, K., Cox, S. J., Lefrançois, M., Taylor, K., Le Phuoc, D., Lieberman, J., García-Castro, R., Atkinson, R. & Stadler, C. (2019), 'The modular ssn ontology: A joint w3c and ogc standard specifying the semantics of sensors, observations, sampling, and actuation', *Semantic Web* **10**(1), 9–32.

Hayes, P. (2004), 'Rdf semantics', *http://www. w3. org/TR/rdf-mt* .

Hu, S., Corry, E., Curry, E., Turner, W. J. & O'Donnell, J. (2016), 'Building performance optimisation: A hybrid architecture for the integration of contextual information and time-series data', *Automation in Construction* **70**, 51–61.

Isikdag, U. (2015), 'Bim and iot: A synopsis from gis perspective.', *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* **40**.

Khalid Belhajjame, J. C. (2013), 'Prov-o: The prov ontology', *https://www.w3.org/TR/prov-o/* .

Manola, F., Miller, E., McBride, B. et al. (2004), 'Rdf primer', *W3C recommendation* **10**(1-107), 6.

Manu Sporny, D. L. (2013), 'Verifiable credentials data model v1.1', *https://www.w3.org/TR/prov-o/* .

Nguyen, V., Bodenreider, O. & Sheth, A. (2014), Don't like rdf reification? making statements about statements using singleton property, *in* 'Proceedings of the 23rd international conference on World wide web', pp. 759–770.

Oraskari, J., Schulz, O., Werbrouck, J. & Beetz, J. (2022), Enabling federated interoperable issue management in a building and construction sector, *in* 'Proceedings of the 29th EG-ICE International Workshop on Intelligent Computing in Engineering : Aarhus, Denmark, July 6-8, 2022', EG-ICE.

Pauwels, P., Zhang, S. & Lee, Y.-C. (2017), 'Semantic web technologies in aec industry: A literature overview', *Automation in construction* **73**, 145–165.

Ruiz-Zafra, A., Benghazi, K. & Noguera, M. (2022), 'Ifc+: Towards the integration of iot into early stages of building design', *Automation in Construction* **136**, 104129.

Tobin, A. & Reed, D. (2016), 'The inevitable rise of self-sovereign identity', *The Sovrin Foundation* **29**(2016), 18.

Zhang, Y. & Beetz, J. (2023), Describe and query semantic building digital twin data in temporal knowledge graph, *in* 'Ready for publication', CIB-W78.