Poster: Sustainable Data Management Flow for Spatio-Temporal Datasets

Yoshiteru Nagata
Nagoya University
Nagoya, Aichi, Japan
teru@ucl.nuee.nagoya-u.ac.jp

Daiki Kohama
Nagoya University
Nagoya, Aichi, Japan
kohama@ucl.nuee.nagoya-u.ac.jp

Yoshiki Watanabe
Nagoya University
Nagoya, Aichi, Japan
yoshiki@ucl.nuee.nagoya-u.ac.jp

Shin Katayama
Nagoya University
Nagoya, Aichi, Japan
shinsan@ucl.nuee.nagoya-u.ac.jp

Kenta Urano
Nagoya University
Nagoya, Aichi, Japan
vrano@ucl.nuee.nagoya-u.ac.jp

Takuro Yonezawa
Nagoya University
Nagoya, Aichi, Japan
takuro@nagoya-u.jp

Nobuo Kawaguchi
Nagoya University
Nagoya, Aichi, Japan
kawaguti@nagoya-u.jp

ABSTRACT

Spatio-temporal data is utilized in various fields, but its scale is generally vast, leading to significant labor and costs in storage and processing. Therefore, the value that can be derived from spatio-temporal data is diluted due to management costs. We propose a new data management flow using various metadata and common programs for spatio-temporal data utilization. Traditionally, various spatio-temporal data processing have been implemented and processed according to each spatio-temporal data. We defined spatio-temporal data structure metadata and performed data processing based on metadata using a common data processing program. Furthermore, we automated the generation of data structure metadata by combining our data skeleton recognition method and generative AI model. Using this flow, we expect to improve the sustainability of utilizing spatio-temporal data.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Spatio-Temporal data, Big data, Semantic web

ACM Reference Format:
Yoshiteru Nagata, Daiki Kohama, Yoshiki Watanabe, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2024. Poster: Sustainable Data Management Flow for Spatio-Temporal Datasets. In *ACM International Conference on Mobile Systems, Applications, and Services (Mobisys '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3643832.3661423

Mobisys '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06
https://doi.org/10.1145/3643832.3661423

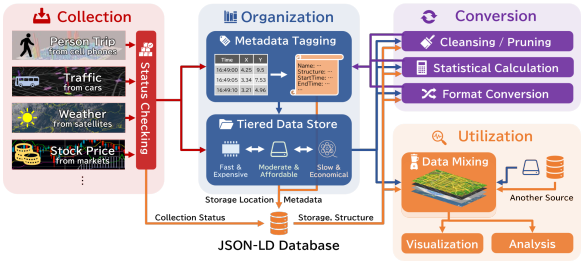

Figure 1: Spatio-Temporal Data Flow

1 INTRODUCTION

Spatio-temporal data is utilized in various fields such as transportation, weather, disaster management, and marketing, providing some "value" within these fields. Various studies have also been conducted on the management of spatio-temporal data [4]. However, many spatio-temporal datasets are so large that the effort and cost involved in management can dilute the value that can be derived from the data. According to a survey by Anaconda [1], 37.75% of data analysts' efforts are devoted just to data preparation and cleansing. As a result, there are cases where spatio-temporal data analysis is abandoned, and the data is left unutilized or discarded.

We propose a new data management flow for automating spatio-temporal data processing for data analysis, and a metadata format along with a method for generating metadata for spatio-temporal data structure, as depicted in Figure 1. We define metadata representing the structure of spatio-temporal data, data collection methods, and storage locations. In addition, we define a template for the data processing program, enabling the implementation of various data processing tasks. We also introduce a method for automatically generating the structural metadata for automating data processing, by combining our data skeleton recognition method and generative

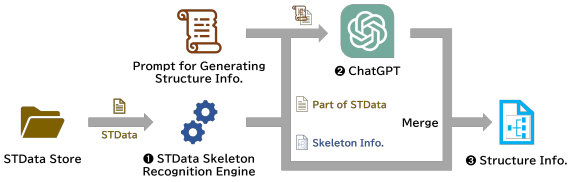

Figure 2: Structure Information Generation

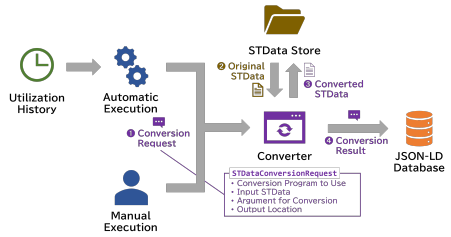

Figure 3: Spatio-Temporal Data Conversion Schema

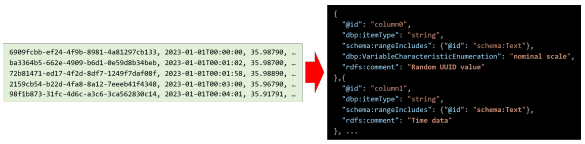

Figure 4: Example Output of Metadata Tagging

AI model. Using this flow, we expect to reduce the cost of utilizing spatio-temporal data, leading to sustainable data utilization.

Our contributions are follows: 1) defined metadata format and common program template for spatio-temporal data, 2) developed automated metadata generation system, and 3) introduced data management flow using them.

2 DATA MANAGEMENT FLOW DESIGN

Overview: This flow (Figure 1) consists of multiple modules used in the stages of collection, organization, conversion, and utilization of spatio-temporal data. In addition, the JSON-LD database manages the metadata generated at each stage. The following describes the details of each module.

JSON-LD Metadata Database: In the domain of the semantic web and Linked Open Data, studies have been undertaken to attach semantic metadata to spatio-temporal data using ontologies and to describe data relationships with RDF and/or JSON-LD [2, 3, 5]. We have expanded these ideas and developed a format that represents various metadata for spatio-temporal data (e.g., sources of spatio-temporal data, storage locations, data structures, and conversion programs used) in JSON-LD. In addition, we have developed a database and user interface to manage these metadata easily.

Collection: A lot of spatio-temporal data is collected from a large number of devices such as smartphones and IoT sensors. Therefore, data loss can occur when individual devices encounter problems during data collection, posing challenges for spatio-temporal data analysis. In this flow, communication from each device is proxied, and the data collection status is automatically recorded in a JSON-LD database.

Organization: In this flow, we organize spatio-temporal data through 1) metadata tagging and 2) tiering of data. As shown in Figure 2, spatio-temporal data is first input to the skeleton recognition engine, which outputs the skeleton information. The spatio-temporal data is then input together with the skeleton information to a generative AI model (ChatGPT), which assigns meta-information about the data itself (e.g., what values it represents, units, etc.). In addition, we define storage metadata for various storage destinations and perfome data tiering according to usage situations.

Conversion and Utilization: In this flow, parameters required for data conversion are generalized, and data conversion is performed using data conversion programs with metadata related to functionality, as shown in Figure 3. This improves the traceability of spatio-temporal data and enables the reproduction and scaling of conversion processes.

3 EXPERIMENT: METADATA TAGGING

We conducted experiments on metadata tagging for several spatio-temporal datasets using the structural metadata tagging system we developed. The results showed that, in many cases, the system could recognize the overall structure of the data and appropriate labels and qualitative characteristics for each variable in well-structured JSON/CSV data, as shown in Figure 4. However, it was also observed that the system could not fully capture the data structure and variable characteristics in spatio-temporal data where the number of CSV columns varied by row or when the variables themselves had few distinctive features. We are now working to address these issue by updating the skeleton recognition engine algorithm and the prompt for ChatGPT.

ACKNOWLEDGMENTS

This research was supported in part by commissioned research (No.22609) by NICT, and JSPS KAKENHI (Grant Number 22H03696) in Japan.

REFERENCES

[1] Anaconda. 2022. State of Data Science Report 2022. https://www.anaconda.com/resources/whitepapers/state-of-data-science-report-2022. (Accessed on 3/27/2024).

[2] Trupti Padiya, Minal Bhise, and Prashant Rajkotiya. 2015. Data Management for Internet of Things. In *2015 IEEE Region 10 Symposium*. 62–65. https://doi.org/10.1109/TENSYMP.2015.26

[3] Sudha Ram and Jun Liu. 2009. A New Perspective on Semantics of Data Provenance. In *Proceedings of the First International Conference on Semantic Web in Provenance Management - Volume 526* (Washington DC) *(SWPM'09)*. CEUR-WS.org, Aachen, DEU, 35–40.

[4] Naser Shirvanian, Maryam Shams, and Amir Masoud Rahmani. 2022. Internet of Things data management: A systematic literature review, vision, and future trends. *International Journal of Communication Systems* 35, 14 (2022), e5267. https://doi.org/10.1002/dac.5267

[5] Viktor Zayakin, Lyudmila Lyadova, Mikhail Smirnov, Viacheslav Lanin, Nada Matta, and Elena Zamyatina. 2022. Event Series Generation and Analysis Based on Multifaceted Ontology. In *2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT)*. 1–6. https://doi.org/10.1109/AICT55583.2022.10013573

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009