NFDI4Energy Conference
1. NFDI4Energy Conference
10.5281/zenodo.10625406

Published: 6.2.2024

Metadata Extraction from a Huge Time Series Database

Andreas Schmidt[1,2] [https://orcid.org/0000-0002-9911-5881],
Mohamed Anis Koubaa[1] [https://orcid.org/0000-0001-8552-2008],
Nan Liu[1] [https://orcid.org/0009-0005-8768-7072],
Karl-Uwe Stucky[1] [https://orcid.org/0000-0002-0065-0762], and
Wolfgang Süß[1] [https://orcid.org/0000-0003-2785-7736]

[1]Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Department of Computer Science and Business Information Systems, University of Applied Sciences, Karlsruhe, Germany
*Correspondance: Andreas Schmidt, andreas.schmidt@kit.edu

Keywords: Technical Metadata, Semantic Metadata, Time Series Database, Zeitgeist

1 Introduction

The Energy Lab 2.0 [1] runs a big cluster of an Influx time series database, where a wide variety of energy-related data are stored in a large number of individual databases for up to 15 years. These data form the basis for a number of research projects like SEKO (Sector Coupling) [2], Living Lab Energy Campus [3], Kopernikus 2X [4], and others. In order to make the experiments performed at KIT reproducible for further research, it is necessary to make these data available in an appropriate manner. As a first step in this direction we developed *Zeitgeist* [5], a tool to facilitate and automate the publication of FAIR time series data sets. The general idea behind *Zeitgeist* is the interactive specification of a time series to be published, enriching it automatically and manually with meta information and make it available in an RO-Crate [6]. In a next version of *Zeitgeist*, we plan to enhance the tool to not only contain a single time series, but a bunch of semantically related time series data sets. However, it has turned out that finding semantically related data is anything but trivial. One of the reasons for this is the schema free characteristic of the Influx database, so that a priori no information about the data structures is available. This is particularly advantageous at the beginning of a project, when many objectives are not yet clear and it is difficult to establish an initial structure, before first data sets can be stored. On the other hand, the lack of an explicit schema also has several disadvantages, as no assumptions can be made about the data set structure. So, we decided to develop a tool that automatically analyzes and visualizes the metadata provided, in order to get an overview of the scope and structure of the Influx databases.

2 Technical and Semantic Metadata

In the following, we refer to *technical metadata* when extraction is possible independently of the underlying domain. In contrast, we speak of *semantic metadata* when

knowledge of the specific domain is required for extraction. Figure 1 shows the different levels of metadata we already extract or plan to extract and how they can be derived.

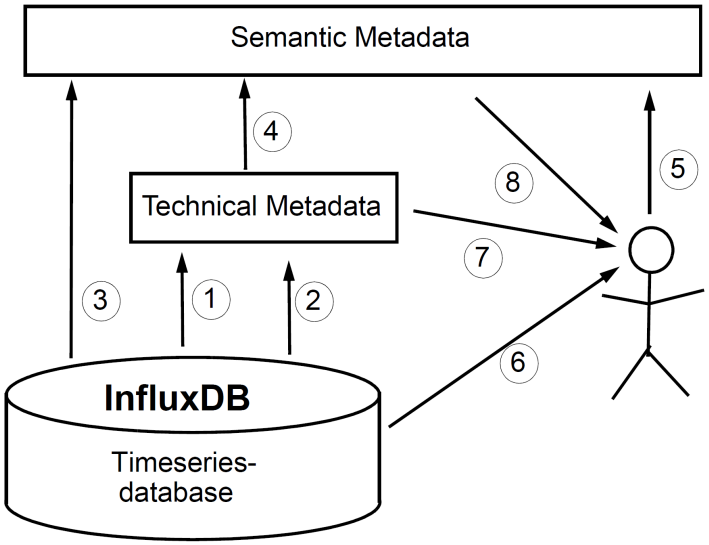

Figure 1. *Technical & and semantic metadata.*

Database supported Metadata extraction (1): Technical metadata is extracted using the Influx Metadata API [7]. This includes the database names and the *measurements* (a table-like structure inside an Influx database), attributes with their data types and so on.

Metadata extraction from regular database queries (2): This can, for example, be the minimum or maximum value of a certain field. Another example is the earliest and latest time stamp of a *measurement*.

Extraction of semantic metadata with regular database queries (3): This includes the identification of time-limited record sequences within a *measurement* and the associated sample rate.

Semantic metadata generation from already extracted metadata (4): Based on the extracted attributes (1), data structures can be reconstructed and similarity measures between structures in different *measurements* can be determined.

Human generated metadata (5): A human with in-depth knowledge of the system and the measurement setup can, for example, classify the measurement series of several sensors as semantically related and relate them to other information, such as another time series with weather data. This will form a semantically comprehensive package of time series data. In this cognitive process, the human accesses actual data (6), extracted technical metadata (7), and already available semantic metadata (8) to generate further semantic metadata.

3 Architecture

Figure 2 shows the architecture of our current system. The extensions we will discuss in Section 4 are shown in red. The *InfluxMetadata Extractor* uses the REST API provided by InfluxDB to obtain information about the characteristics of the data records via both the metadata API and the normal query API. It uses configuration files as input, which contain the access information, as well as information about the databases and *measurements* to be analyzed. In addition, output templates are describing the

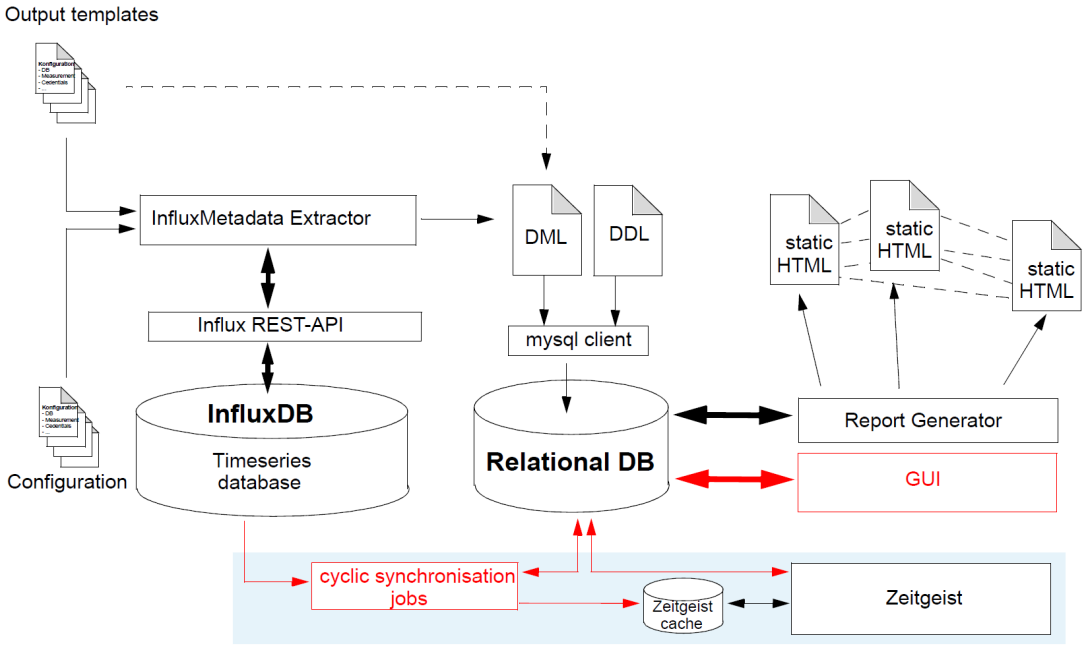

Figure 2. *Architecture of Influx MetadataExtractor.*

output format. We have developed a template that realizes the output in the form of SQL insert and update statements for a relational database. In the current version, a report generator displays the results as static, interlinked HTML pages. In the next version, we will replace the static HTML pages with an interactive GUI that also allows manual metadata to be added ((5) in Figure 1). The results of the metadata extraction tool are intended to support users of the *Zeitgeist* tool in defining RO-Crate exports, for example by displaying semantically related time series. How this will be implemented in practice is currently still the subject of ongoing research.

4 Next Steps

In the actual version of our metadata extraction tool we mainly extract technical metadata (see (1), (2) in Figure 1). Initial examinations of the extracted meta information have shown that the absence of a schema causes a real proliferation of data structures, which must be disentangled in order to efficiently use the data. Our system already has a function that visualizes similar data structures at the measurement level (using heat maps). This approach only works if the individual measurements contain homogeneous data sets. As first examinations have shown, this is often not the case. Accordingly, this approach fails and the different data structures must be identified at data set level. This can be achieved by identifying the attributes that occur together. In this way, a temporal evolution of the data structures can also be determined.

In order to identify semantically related time series, it is beneficial to first identify the recording intervals inside the *measurements*. Afterwards, the system can automatically search for intervals of measurements that have start and end times that are close together. By representing these findings graphically, they can act as the base for a human-assisted grouping.

When we extract the meta information with our metadata extraction tool, it is initially only a snapshot of the status at a certain point in time. In order to always have the current status of the system, it is necessary to expand it so that changes to structures

(databases, measurements) and data sets are detected. This can be done by cyclically checking the cluster for new databases, the databases for new measurements, and the measurements for new data sets. These processes can communicate directly with the relational database that holds the meta information (Figure 2).

Our tool *Zeitgeist* also benefits from this expansion. It has turned out that with very large data sets, such as those found in the KIT energy lab, reading out some meta information, such as the fields in a measurement, can take quite a long time. For this reason, *Zeitgeist* manages its own small cache, which stores this information. The *Zeitgeist* user can then decide to load this meta information from the cache (fast) or to access the database (probably slow), This cache can now also be filled by the cyclical process that checks the Influx database for changes and therefore remains up to date. Alternatively, *Zeitgeist* could also access the database managed by the extractor tool directly. The interaction between the tool presented here in *Zeitgeist* is shown in Figure 2 in the area highlighted in gray.

For the human provided metadata, we need to extend the structure of our relational database and especially find a mechanism that this metadata ”survives” further runs of our metadata generator tool.

Author contributions

Andreas Schmidt: Conceptualization, Investigation, Methodology, Software, Writing – original draft

Anis Koubaa, Nan Liu, Karl-Uwe Stucky, Wolfgang Süß: Writing – review & editing

Competing interests

The authors declare that they have no competing interests.

Funding

The authors' work is conducted as a contribution to the project NFDI4Energy in the German National Research Data Infrastructure, a special funding initiative of the Deutsche Forschungsgemeinschaft (DFG).

This publication was also supported within the Hub Energy of the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

References

[1] *Welcome to the Energy Lab 2.0*, https://www.elab2.kit.edu/english/index.php, 2023.

[2] KIT, *Seko (sector coupling)*, https://www.esd.kit.edu/85.php, 2021. (visited on 02/05/2024).

[3] Forschungszentrum Jülich, *Living lab energy campus*, https://www.fz-juelich.de/de/llec. (visited on 02/05/2024).

[4] BMBF, *P2x - how the kopernikus project p2x converts renewable electricity into plastics and fuels, gases and heat*, https://www.kopernikus-projekte.de/en/projects/p2x. (visited on 02/05/2024).

[5] A. Schmidt, M. A. Koubaa, J. Schweikert, K.-U. Stucky, W. Süß, and V. Hagenmeyer, "Zeitgeist - A Generic Tool Supporting the Dissemination of Time Series Data following FAIR Principles," in *Proceedings of the International Conference on Knowledge Management and Information Systems*, Insticc, SCITEPRESS, Nov. 2023. DOI: doi:10.5220/0012254300003598.

[6] S. Soiland-Reyes, P. Sefton, M. Crosas, *et al.*, "Packaging research artefacts with RO-Crate," *Data Science*, vol. 5, no. 2, pp. 97–138, 2022. DOI: doi:10.3233/DS-210053.

[7] *Explore your schema using InfluxQL*, https://docs.influxdata.com/influxdb/v1.8/query_language/explore-schema/, 2022.