White Paper

Semantic Container

Verein zur Förderung der selbstständigen Nutzung von Daten
2540 Bad Vöslau
ZVR: 789007092

Contact: semcon@ownyourdata.eu

Content

1 Introduction

Today, the economic potential of data as a cornerstone of modern economies as well as its important role as a driver for technologies such as machine learning and big data analytics is largely undisputed. Although data is often likened to scarce natural resources such as oil, it is in fact very different in that it can become more useful the more it is used. Despite open data initiatives and some efforts to create markets for data in particular domains, however, data is increasingly being monopolized in many domains today. Moreover, potential data users continue to face tedious and time-consuming hurdles when buying, accessing or working with data from external sources.

This is unfortunate, given that data can be replicated and moved at very low cost and can generate great economic value when shared. To date, however, only a few successful data markets have been established, which may be attributed partly to technical issues and partly to business rationale and questions over the viability (and sometimes legality) of business models built around the sales of data. A lot of these issues can be attributed to the problem that data providers selling data (rather than just access to data) largely have to give up control over the terms under which the data will be used. On the other hand, from a data consumer perspective, both technical challenges and limited trust in the quality, completeness, and origin of the available data have contributed to the restricted adoption of data markets today. This highlights the need for a light-weight, non-time-consuming, transparent, standardized, open source, clearly defined and user-friendly solution for data provisioning.

Semantic Containers tackle these challenges by developing a standardized infrastructure for data provisioning. The proposed concept allows data providers to efficiently distribute data without giving up control over its usage and monetization while providing data consumers with efficient and well-managed mechanisms to obtain and integrate data in a trustworthy and reproducible manner. By packaging data and processing capabilities into reusable containers, describing the semantics of the content and permissible usage, and providing uniform interfaces, a data set becomes a commodity with well-defined content, properties, quality and usage policy, as well as clear ownership and a price tag.

The Semantic Container approach leverages existing container technologies such as Docker, which already provide scalable mechanisms for deploying complex software assemblies and use them as a foundation for an infrastructure for data discovery, provisioning, and integration. To create a suitable environment for the emergence of a commodity market around data, we capture a set of rules for permissible usage of the data in semantic descriptions, provide cryptographic methods to prove ownership, and apply blockchain technology to guarantee immutability. Complete audit trails of data sources and processing steps provide gapless provenance and facilitate reproducibility.

2 Roles and Needs

This section outlines profiles the key actors involved in the Semantic Container ecosystem with a special focus on individual roles and needs addressed by the proposed solution.

2.1 Data Collector

A data collector provides services for collecting, storing and managing data. Data collectors need to setup data collection mechanisms, allowing individuals to donate data (data donation), while tracking the sources of collected data. Moreover, associated usage rights and, where appropriate, consent to the processing of personal data need to be made explicit.

Examples:

- research institutions (citizen science project)
- official institutions/associations (diabetes data collection in Denmark)
- data journalism
- pharmaceutical industry (medical studies and trials)

2.2 Data Providers

A data provider makes data available for others to use, sets prices and defines usage policies. Data providers need to find customers for selling data and generally aim at increasing the access to and usage of data while participating in a liquid data market. To do so, data providers need to package data, define associated usage policies and track the use of provided data (openness, transparency).

Examples:

- Central Institution for Meteorology and Geodynamics (ZAMG)
- Collaboration for Earth Observation (EODC)
- Open Data Portal
- Harptech (robotic sensor platform Verner)
- Scientific data

2.3 Data User

A data user handles data and services provided by Semantic Containers for commercial or non-commercial purposes, e.g., research as well as product or app development. Typical tasks of data users include data aggregation, statistical analysis, predictions and visualization of results. Data users need to find and compare relevant data sources. They are looking for a simple infrastructure to access data in a timely manner and to exchange data in a defined, reproducible, and automatically verifiable way. Automating manual tasks for distributing, cleaning and verifying data as well as keeping track of changes in data and code to reliably reproduce analyses are key functionalities in this regard. In addition, data users need to be able to combine semantic containers into data pipelines for data driven applications.

Examples:

- researchers
- students
- data scientists

2.4 Container Developer

Container developers implement solutions to specific problems and make them available as Semantic Containers, which in turn can be applied by data users. Container developers need a well-documented environment agnostic to programming language with defined input and output as well as development and automated testing tools.

Examples:

- data scientist
- software developer

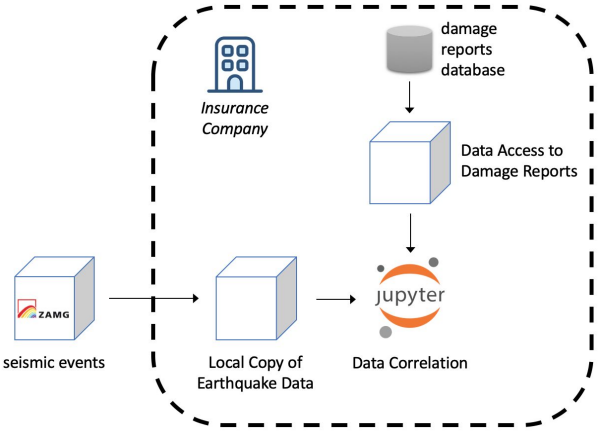
2.5 Example Scenario

In the following, the individual roles are further illustrated by means of an example.

Scenario Description

An insurance company wants to check if damage reports about earthquake damage match with actual seismic events in the vicinity of the reporter.

ZAMG (Central Institute for Meteorology and Geodynamics in Austria) provides information about seismic events as Open Data used in this scenario.



Data Collector: Insurance Company
Seeks earthquake data to match seismic events with damage reports.

Data Provider: ZAMG and insurance company
ZAMG provides seismic event data of the last month at the URL http://geoweb.zamg.ac.at/static/event/lastday.json; the insurance company provides an API to its internal database of damage reports.

Container Developer
A container developer sets up a static container to store a local copy of earthquake data. This container queries the ZAMG api on a daily basis and filters for relevant seismic events in the country the insurance company operates and stores it permanently.
Another dynamic container provides access to the damage reports database in a simplified data format.

Data User
A Data Scientist creates a Jupyter Notebook to correlate earthquake damage reports with actual seismic events in the vicinity of the reporter.

3 Semantic Containers

3.1 Characteristics

The needs of the individual roles in section 2 are addressed by the following characteristics of the proposed solution. Section 4 describes what technologies are used and how these characteristics are achieved.

Characteristics	Description
Provisioned	Semantic Containers provide a defined, reproducible, and automatically verifiable data status.
Distributed	Semantic Containers support data exchange between two parties without the need for a centralized intermediary.
Composable	Semantic Containers can be combined for creating a data processing pipeline.
Packaged	Semantic Containers combine data, semantic description and program logic in one distribution mechanism.
Discoverable	Semantic Containers make it easier and faster to find suitable data and allow to identify and compare similar data sets.
Tradeable	Semantic Containers provide an easy way to trade data in a secured way and unambiguously describe the usage rights for the data.
Standardized	Semantic Container offer a standardized infrastructure for data provisioning.

3.2 Data and Processing

Semantic Containers propose to define the following two entities in a data exchange environment:

Definition of Data
Data itself should be annotated with a Usage Policy (describing what is allowed for this data) and provenance information (providing a complete audit trail about the source, ownership, usage rights and processing steps performed so far). To make this information (content, usage policy and provenance) immutable a hash is stored in a distributed ledger and a trusted timestamp is generated.

Definition of Processing
An entity that processes data should have a unique identifier, provide a Format or Syntax Description of this data, the Usage Policy that applies to the results of the processed data and a description of the processing itself (to be added to the provenance).

Semantic Container automate the work to check data, match usage policies and update the provenance. Therefore, this environment enables a GDPR-compliant data processing by:

- explicit consent for (personal) data processing through a usage policy
- traceability of the data along a Semantic Container processing pipeline through the unique identifier of a Semantic Container and an immutable audit trail (provenance)

3.3 Interoperability Levels

Information about data to be processes can be available on different levels. Sometimes there is only metadata available, other times there is also a defined syntax for the records available, and in the ideal case it is possible to describe the exact semantics of the data at hand. Semantic Container take into account that data sources are not always described in a well-defined way and the proposed solution enables users to enrich the available data in a step-wise approach.

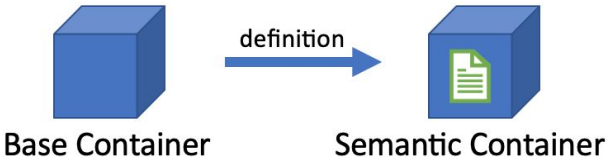
A Semantic Container describes the container input and output on three possible levels:

1. **Metadata Level:** every container includes the publisher of the data as well as a verbal description of the content; additionally, the allowed usage, provenance, and optional billing information is provided
2. **Syntax Level:** on this level it is possible to specify and validate the data format (syntax) of the data processed and stored by the container; this information can be used to check compatibility between container
3. **Semantic Level:** the highest level of interoperability includes metadata and syntax information as well as a semantic description about the individual data attributes; on this level it is possible to automatically validate data and ensure a defined level of data quality

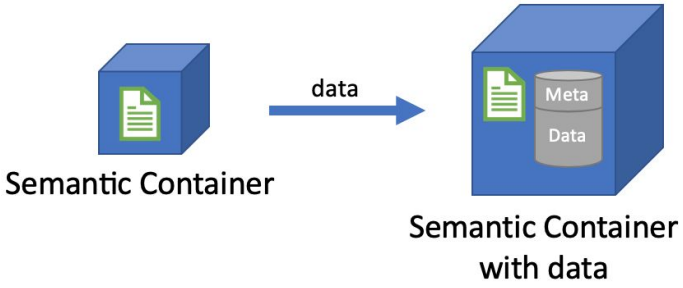
3.4 Workflow

The following steps describe the typical life cycle of a Semantic Container:

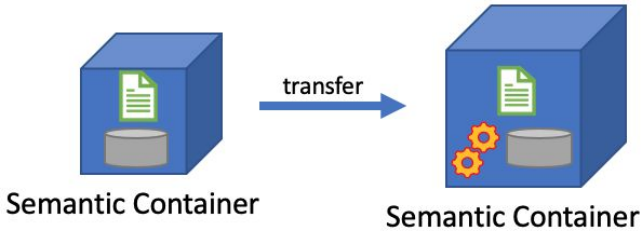
1) Setup a semantic container, based on a base container by providing a definition of data and a specific usage policy



2) Data is transferred into the container via a API endpoint and this method performs automatic data validation and generates a complete audit trail



3) Process data by chaining together semantic containers with specific functionality; get automatic verification for allowed usage and an updated audit trail



4) Share data in a well-defined way (i.e., including usage policy and full provenance) and document any access
    a) for operational containers data can be accessed through an API using the OAuth2 protocol; optionally, access can be subject to a charge and is billed automatically through a cryptocurrency
    b) containers can be shut down and distributed through images; for data access it is necessary to start the image (e.g., locally on your computer)

4 Technology

Based on the description of needs (section 2) and the previously described characteristics and design decisions (section 3), in the following the underlying technology used in Semantic Containers is outlined.

4.1 Cryptography

Semantic Containers use hash functions extensively to unambiguously identify digital information. Specifically SHA-256 is used to create a digest for data, policies, and containers. To explicitly assign data and metadata (like usage policy) to a user Semantic Containers provide mechanisms to digitally sign those hash values and provide automatic validation.

4.2 Blockchain

To make hash values or signed hash values immutable and verify this information independently those values are stored in a blockchain. An additional benefit of using a blockchain is to get a verified timestamp. For billing Semantic Containers use a cryptocurrency to transfer the monetary value between the two parties.

Ethereum was chosen as a distributed ledger to store data as well as cryptocurrency because of the widespread use and maturity of the technology.

4.3 Semantics

At the core of Semantic Containers is the semantic annotation of data and semantic description of metadata using W3C standards:

- the usage policy clearly specifies what is allowed and not allowed with regard to the use of data
- the provenance documents the complete process from the data source to the current state
- the API description describes available functions in a container

4.4 Container

The above mentioned technologies are combined in a single self-contained package. As underlying technology Semantic Containers use Docker to also benefit from the existing infrastructure to store and distributed images. Semantic Containers are built in a way to be derived from a base container that can be extend with necessary functionality.

5 Use Cases

This section lists use cases we currently explore and partly implement during the funding provided by the Austrian Research Promotion Agency (FFG).

5.1 Data Donation

Participants donating data provide data for free, but at the same time they can control how their donated data will be used through defining specific permissible usage rights and/or providing specific consent for the processing of personal data. An additional benefit is the possibility to track donated data as well as derived results.

Examples for data donations:

- collect GPS data from your phone and support a study from the Vienna University of Economics and Business to evaluate the quality of route recommendations provided by Google Maps
- download step count data from your iPhone through Apple HealthKit and aggregate this information to show average step counts per day for all participants

5.2 Providing Data

Organizations that want to make data available either to other businesses or to the general public can use Semantic Containers as light-weight and decentralized distribution platform. The provided functionalities include billing as well as tracking the use of the data.

Examples for organizations providing data through Semantic Containers:

- ZAMG (Austrian Central Institution for Meteorology and Geodynamics) will provide weather and seismic data through a Semantic Container API; requests to weather data are not free and subject to a fee
- EODC (Earth Observation Data Centre) will provide processed satellite images from the EU Copernicus program for given GPS coordinates with a 5-day update interval
- Semantic Containers can provide the technical infrastructure to fulfill a new “Payment Service Directive” for banks which are required to make account statements available to customers and transferable between institutions.

5.3 Selling Data

Establish a data-flow between private citizens and corporations to allow individuals monetizing data while keeping their privacy and businesses to access personal data in a GDPR-compliant way.

Example for a private-corporate partnership

a group of people with diabetes collects and anonymizes their blood sugar level; these data are in demand by the pharmaceutical industry to comply with new regulations regarding real world evidence (i.e., is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of real world data)

6 Final Remarks

In this white paper, we envision simple, efficient, and stable data exchange mechanism between multiple parties. We described the relevant roles and their needs and outlined how Semantic Containers can satisfy those requirements in an elegant way. Finally, some use cases were provided to give an outlook on the future development of data mobility with Semantic Containers.

did:sov:CYQLsccvwhMTowprMjGjQ6

You can find the latest version of this document on https://www.ownyourdata.eu/semcon/whitepaper.

Additionally, a reference is stored in the Sovrin blockchain at the address shown on the left. There you can always find the complete history and further information about this document.

Use https://uniresolver.io/#did=did:sov:CYQLsccvwhMTowprMjGjQ6 to resolve the address.

Semantic Containers is funded in the program “IKT der Zukunft” by the Federal Ministry for Transport, Innovation and Technology (bmvit) under grant number 869781.

Please don’t hesitate to contact us with any comments and feedback via semcon@ownyourdata.eu.