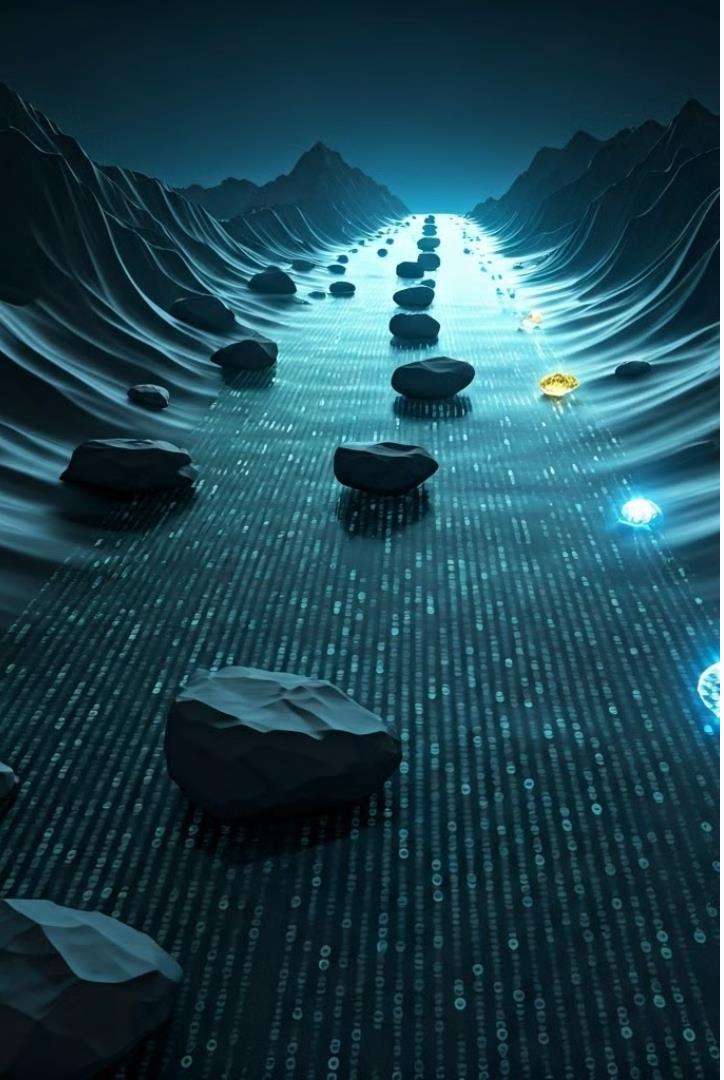
Evolution of ETL Processes Towards Data Product Pipelines

Gorka Zárate, Maria Jose Lopez, Benjamin Navarro, Antonio Gimeno, Jordi Arjona, Urtza Iturraspe, Ana Isabel Torre

22/10/24 Mainz



Evolution of ETL Processes Towards Data Product Pipelines

The rise of data as a first-class asset has led to the creation of infrastructures and tools designed to enhance organizations' abilities to monetize data internally and externally. Traditional Extract-Transform-Load (ETL) processes are evolving into sophisticated data pipelines aimed at creating data products. This article examines how current ETL tools are prepared to address this new concept within the framework of the DATAMITE project, which provides real scenarios and use cases to demonstrate benefits and applicability.

The Rise of Data-Driven Organizations

Over the past 15 years, data has emerged as a critical asset for organizations. Factors like industry digitization, Big Data analytics, and cloud computing have enabled massive data creation and processing. Companies that properly monetize their data can increase income by 10-30%. Beyond internal monetization, new opportunities for external data monetization are arising through initiatives like International Data Spaces, Gaia-X, and SIMPL, which enable safe and reliable data exchange between different actors.

1. **Industry Digitization**
   Enabling creation of massive amounts of data
2. **Big Data Analytics**
   Providing tools to work with large data volumes
3. **Cloud Computing**
   Democratizing access to computing resources
4. External Data **Monetization**
   New initiatives enabling safe data exchange

Challenges in Data Monetization

To unlock the value of data, organizations must ensure data relevance and adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable). Internal catalogues provide a global view of datasets, supported by automated pipelines extracting information from various sources. This allows data to be listed, enriched, and exploited by non-technical users. The goal of data pipelines now extends beyond overseeing operations to correctly defining data services or products, from preparation to exposure or sharing.

Findable

Ensuring data can be easily discovered

Accessible

Making data retrievable by humans and machines

Interoperable

Enabling data integration with other datasets

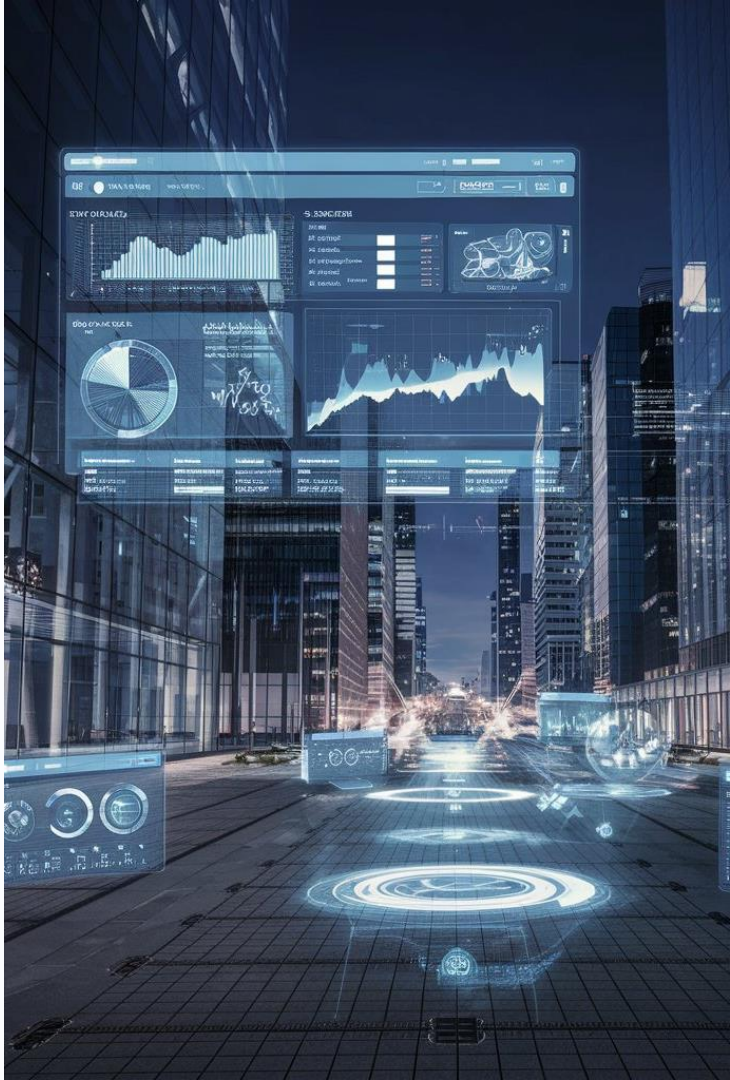
Reusable

Allowing data to be repurposed for various applications

State of art

Currently, many open source and private projects are working in the direction of improving data management and monetization. Initiatives such as DataSpace, Eclipse, GaiaX and SIMPL are developing tools to create policies, contracts and other solutions on data sets in order to monetize them.
Traditional ETL tools, while useful, are primarily geared toward users working with structured data and SQL. Their strength lies in visual interfaces that simplify the creation of transformation logic, but their biggest drawback is that they are designed for a limited number of predefined data sources, which limits their flexibility in today's data environment.

1. **Data Silos**
   Initial focus on isolated data storage.
2. **Data warehouses and lakes**
   Evolution towards more integrated structures to improve analysis.
3. **ETL Tools**
   Development of processes to extract, transform and load data.
4. **Data Products**
   Current trend towards creating data-driven services.



Comparative analysis of data tools

In the context of the evolution of data processes, two main groups of tools have been identified: those for data discovery and orchestration, and those for cataloging data products using appropriate metadata. To analyze both groups, comparative frameworks have been created based on the functionalities and requirements extracted from the DATAMITE project.

Table 1: Comparison of Data Pipelines Orchestration tools and frameworks

Feature	NiFi	Airflow	Mage
Data Flow	Visual flow design	Workflow DAG	Workflow DAG
Language	Java	Python	Java
Real-time Processing	Yes	No	Yes
Batch Processing	Yes	Yes	Yes
Scalability	High	High	High
Community Support	Large	Large	Growing
Extensibility	Processor API	Plugins	Plugin API
Monitoring	Built-in	External tooling	Built-in
Fault Tolerance	Yes	Yes	Yes
Workflow Scheduling	Limited	Yes	Yes
Analytic Oriented	Yes	Yes	Yes
Publication Functionalities	Yes	Yes	Yes
Standards	Compliance support	Compliance support	Compliance support

Table 1 compares three popular orchestration tools, which provide capabilities for automated data access, transformation, and storage.

Comparative analysis of data tools

Table 2: Comparative Analysis of OpenMetadata, DataHub, and Apache Atlas as data catalog tools

Feature	OpenMetadata	DataHub	Apache Atlas
Metadata Capture	Graph-based model	Scalable architecture	Extensible architecture
Metadata Discovery	Yes	Yes	Yes
Data Lineage	Yes	Yes	Yes
Governance Policies	Yes	Yes	Yes
Integration	API-driven	LinkedIn ecosystem	Various platforms
Extensibility	Yes	Yes	Yes
Community Support	Growing	Active	Hadoop ecosystem

Table 2 compares open source data cataloging and metadata management tools, such as OpenMetaData, LinkedIn DataHub, and Apache Atlas, which are critical for metadata storage, discovery, and governance.

Why DATAMITE?

ETLs

- These tools fall short on their own, but they serve as a
- tool within an overall architecture that provides
- us with the ability to manage and **monetize** these data sets

- Now the goal is to generate **Data Products**
- or data-driven services on their own

DATAMITE Approach to Data Ingestion and Discovery

DATAMITE proposes a modular, open-source framework to empower organizations in monetizing their data internally and externally. It provides data ingestion and storage components, along with data discovery connectors. The framework considers three types of data sources: bulk uploads, streaming connections, and external databases or repositories. Bulk ingestion leverages the S3 Multipart library, while streaming data adopts a plugin-based approach using protocols like Kafka, MQTT, and OPC-UA.

1

Data Sources

Bulk uploads, streaming connections, external repositories

2

Ingestion

S3 Multipart for bulk, plugin-based for streaming

3

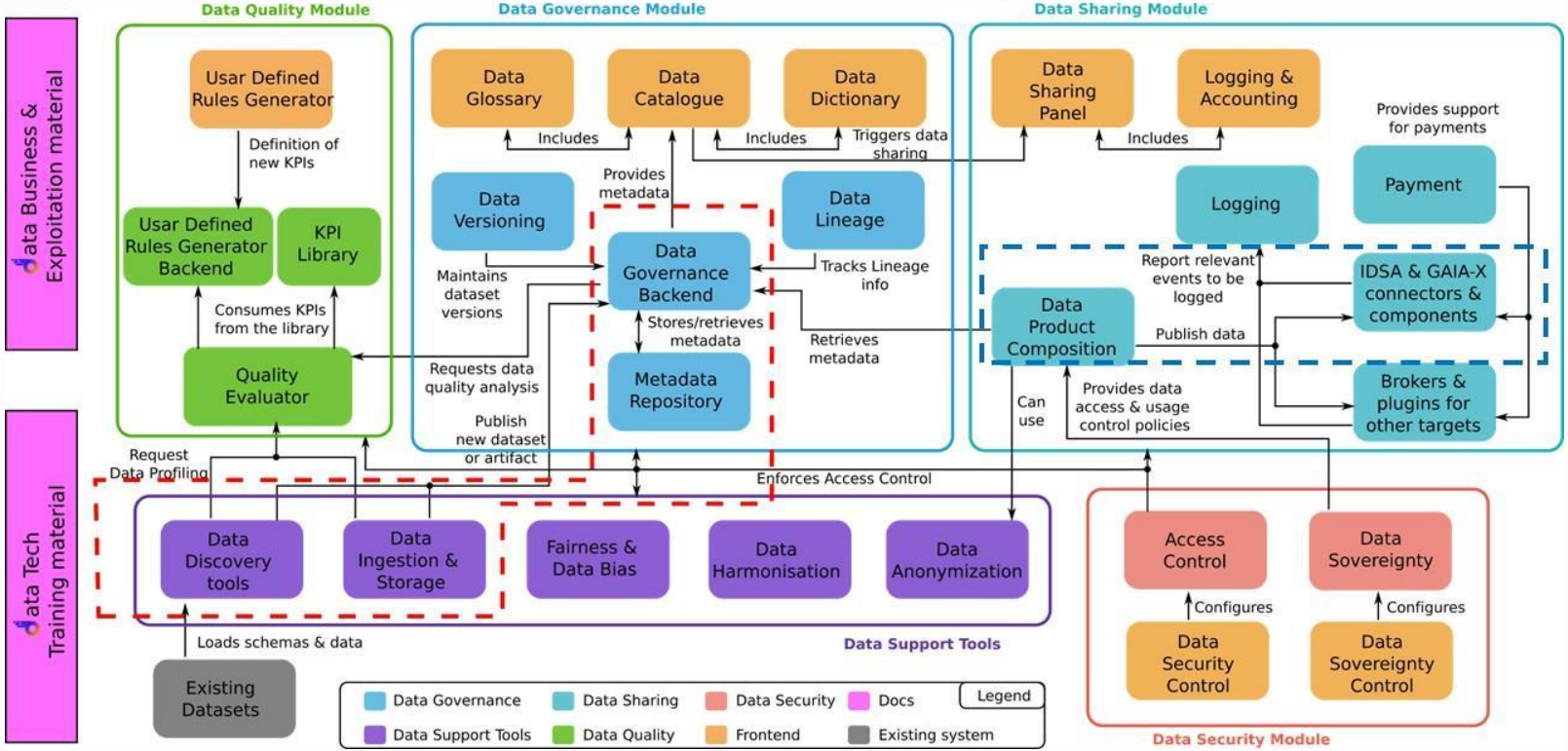
Storage

Internal MinIO instance for temporary and persistent storage

4

Discovery

Connectors to extract and map metadata to DATAMITE schema


Data Business & Exploitation material
Data Tech Training material
Data Quality Module
Data Governance Module
Data Sharing Module
Usar Defined Rules Generator
Definition of new KPIs
Usar Defined Rules Generator Backend
KPI Library
Consumes KPIs from the library
Quality Evaluator
Data Glossary
Data Catalogue
Data Dictionary
Includes
Includes
Triggers data sharing
Provides metadata
Data Versioning
Data Lineage
Data Governance Backend
Maintains dataset versions
Tracks Lineage info
Stores/retrieves metadata
Metadata Repository
Requests data quality analysis
Retrieves metadata
Publish new dataset or artifact
Request Data Profiling
Enforces Access Control
Data Sharing Panel
Logging & Accounting
Includes
Provides support for payments
Logging
Payment
Report relevant events to be logged
IDSA & GAIA-X connectors & components
Data Product Composition
Publish data
Brokers & plugins for other targets
Can use
Provides data access & usage control policies
Data Discovery tools
Data Ingestion & Storage
Fairness & Data Bias
Data Harmonisation
Data Anonymization
Loads schemas & data
Existing Datasets
Data Support Tools
Access Control
Data Sovereignty
Configures
Configures
Data Security Control
Data Sovereignty Control
Data Security Module
Legend
Data Governance
Data Sharing
Data Security
Docs
Data Support Tools
Data Quality
Frontend
Existing system

Metadata Extraction and Management

DATAMITE uses a plugin-based approach for metadata extraction, connecting to different storage technologies to extract structural information and map it to the DATAMITE metadata schema. The Data Governance Backend (DGB) acts as an interface to the metadata repository, using Apache Atlas for its flexibility. The metadata model is based on DCAT but extended to consider complex data assets. It accurately represents objects displayed in the catalogue, including technical and business aspects, as well as quality measurements following the DQV specification.

1

Metadata Extraction

Connection to data sources and extraction of structural information.

2

Metadata Mapping

Adaptation of information to the DATAMITE schema.

3

Send to ATLAS

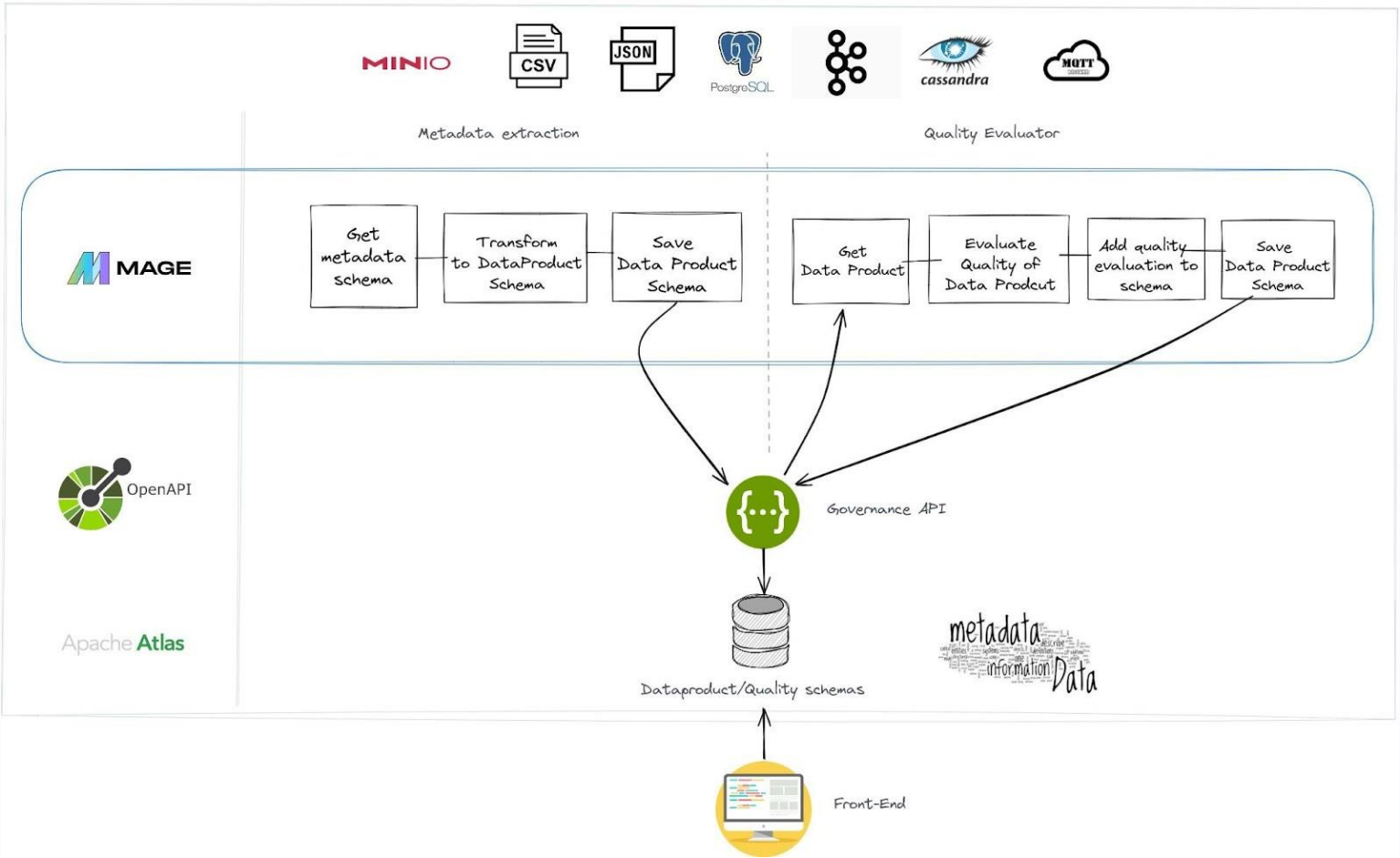
Transferring metadata to the governance backend.

4

Data Profiling

Generation of initial quality metadata.

DATAMITE WORKFLOWS



Data Products and GAIA-X Compatibility

A data product is a standardized unit packaging relevant data resources and services into a consumable form. DATAMITE offers a framework to define and create GAIA-X compliant data products. The DATAMITE GAIA-X onboarding process creates verifiable credentials for all participants: the legal entity, data product provider, and the product itself. The product is transformed into a GAIA-X service offering, ensuring precise definition and documentation. DATAMITE's data product Verifiable Presentation (VP) specification adheres to the GAIA-X Trust Framework, including the Legal Participant VP within the Data Product VP.

Verifiable Credentials

Ensures trust and authenticity of participants and products

Standardized Packaging

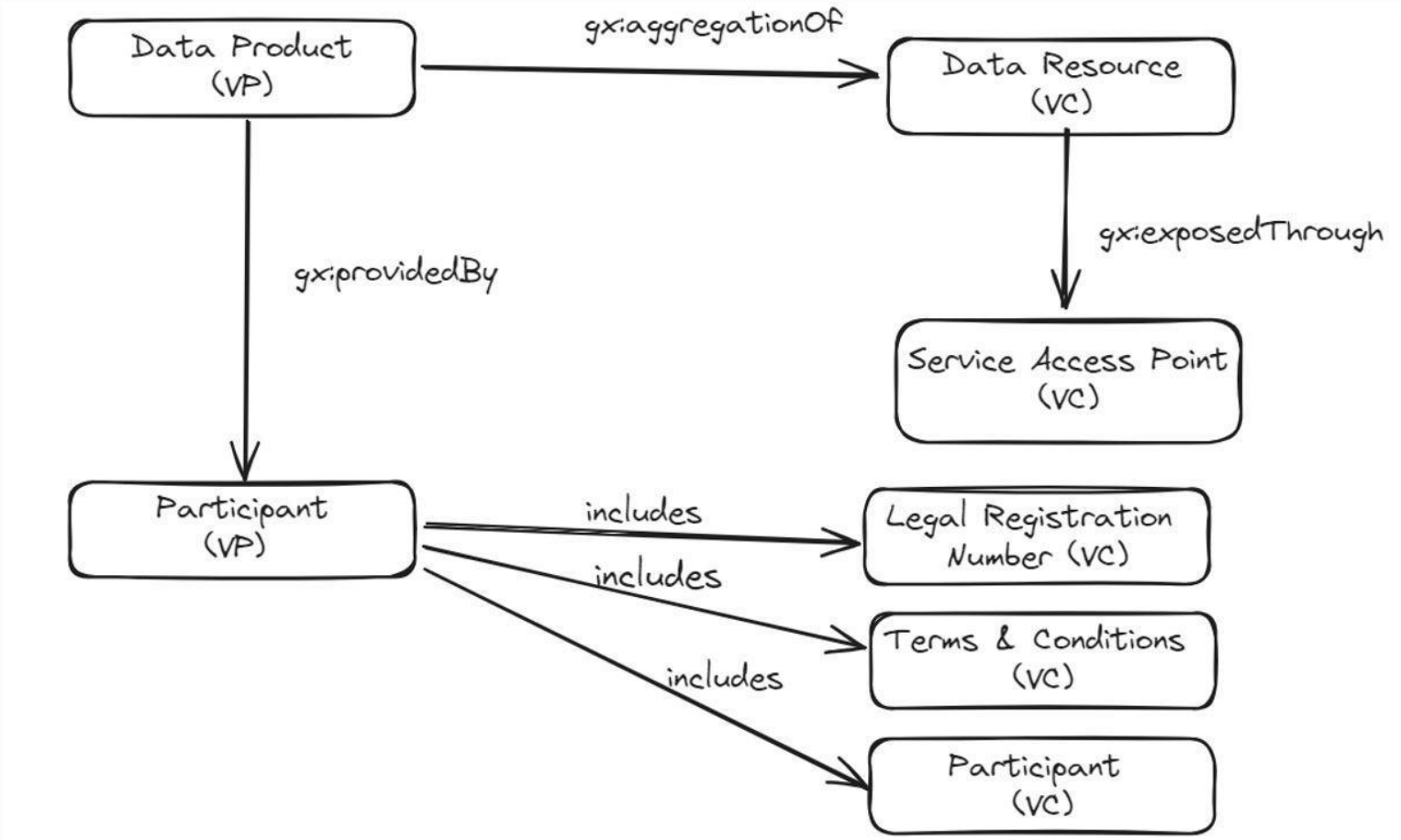
Combines data resources and services in a consumable form

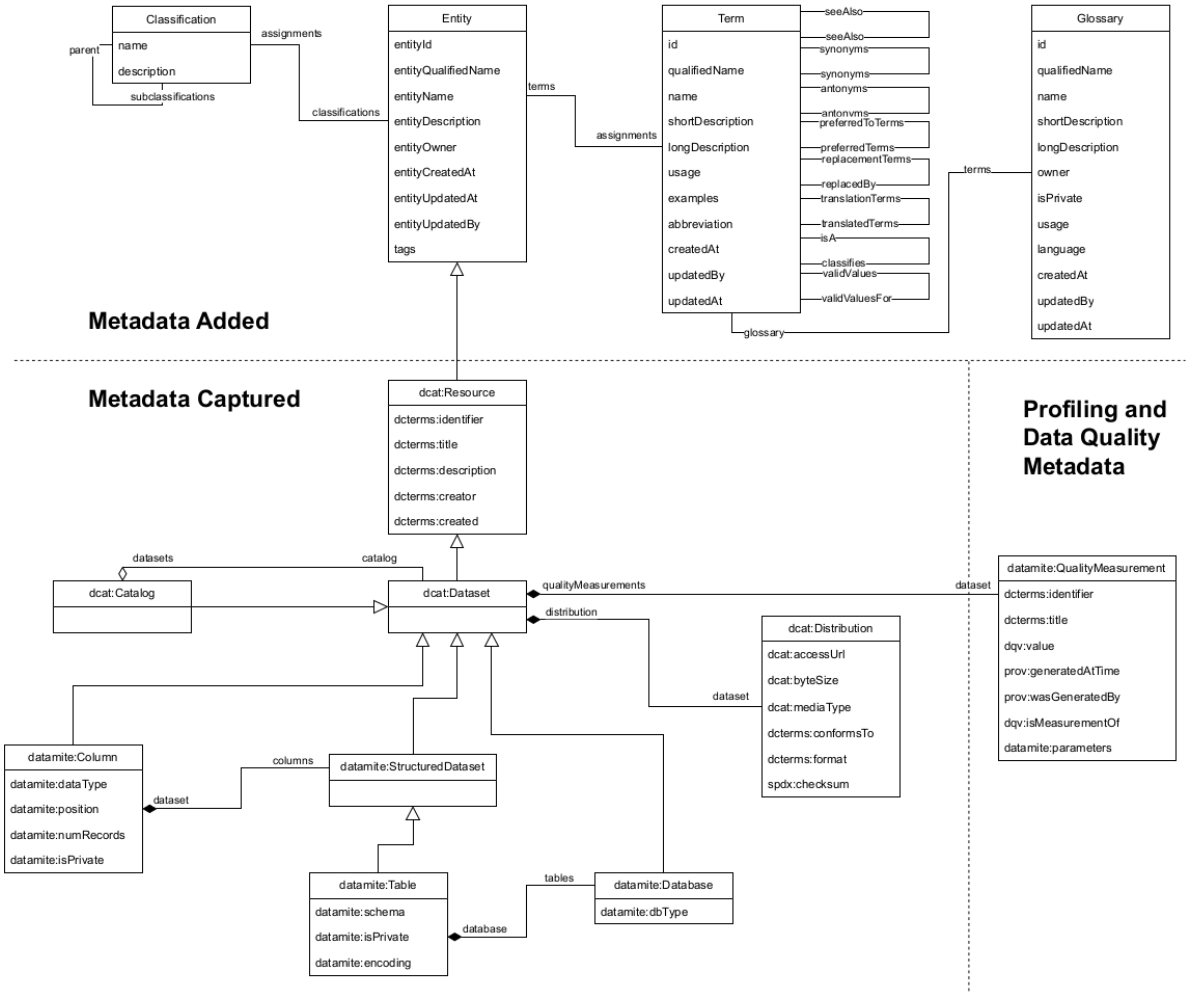
GAIA-X Compliance

Adheres to the GAIA-X Trust Framework specifications

Interoperability

Enables seamless integration within data spaces


Data Product
(VP)
gx:aggregationOf
Data Resource
(VC)
gx:exposedThrough
gx:providedBy
Service Access Point
(VC)
Participant
(VP)
includes
Legal Registration
Number (VC)
includes
Terms & Conditions
(VC)
includes
Participant
(VC)


Classification
name
description
parent
subclassifications
assignments
classifications
Entity
entityId
entityQualifiedName
entityName
entityDescription
entityOwner
entityCreatedAt
entityUpdatedAt
entityUpdatedBy
tags
terms
assignments
Term
id
qualifiedName
name
shortDescription
longDescription
usage
examples
abbreviation
createdAt
updatedBy
updatedAt
seeAlso
seeAlso
synonyms
synonyms
antonyms
antonyms
preferredToTerms
preferredTerms
replacementTerms
replacedBy
translationTerms
translatedTerms
isA
classifies
validValues
validValuesFor
glossary
terms
Glossary
id
qualifiedName
name
shortDescription
longDescription
owner
isPrivate
usage
language
createdAt
updatedBy
updatedAt
Metadata Added
Metadata Captured
dcat:Resource
dcterms:identifier
dcterms:title
dcterms:description
dcterms:creator
dcterms:created
Profiling and Data Quality Metadata
datasets
catalog
dcat:Catalog
dcat:Dataset
qualityMeasurements
distribution
dataset
datamite:QualityMeasurement
dcterms:identifier
dcterms:title
dqv:value
prov:generatedAtTime
prov:wasGeneratedBy
dqv:isMeasurementOf
datamite:parameters
dcat:Distribution
dcat:accessUrl
dcat:byteSize
dcat:mediaType
dcterms:conformsTo
dcterms:format
spdx:checksum
dataset
datamite:Column
datamite:dataType
datamite:position
datamite:numRecords
datamite:isPrivate
dataset
columns
datamite:StructuredDataset
datamite:Table
datamite:schema
datamite:isPrivate
datamite:encoding
database
tables
datamite:Database
datamite:dbType

Future Challenges and Conclusion

The DATAMITE project aims to contribute to a modular, open-source framework providing European organizations with data governance, quality, security, and sharing tools. While progress has been made in processing data to constitute data products and publish them in data spaces, future challenges remain. Organizations struggle to determine effective and profitable strategies for monetizing their data products. Pricing data products is complex due to the intangible nature of data and absence of standardized models. Data consumers' reluctance to pay for data further complicates profit optimization strategies.

1 Monetization Strategies

Developing effective approaches for data product pricing and profit generation

2 Standardization

Creating industry-wide models for data valuation and exchange

3 Consumer Education

Addressing reluctance to pay for data through demonstrating value

4 Ethical Considerations

Balancing profit motives with responsible data use and privacy concerns

Thank you