tetrascience | The Scientific Data Cloud Company

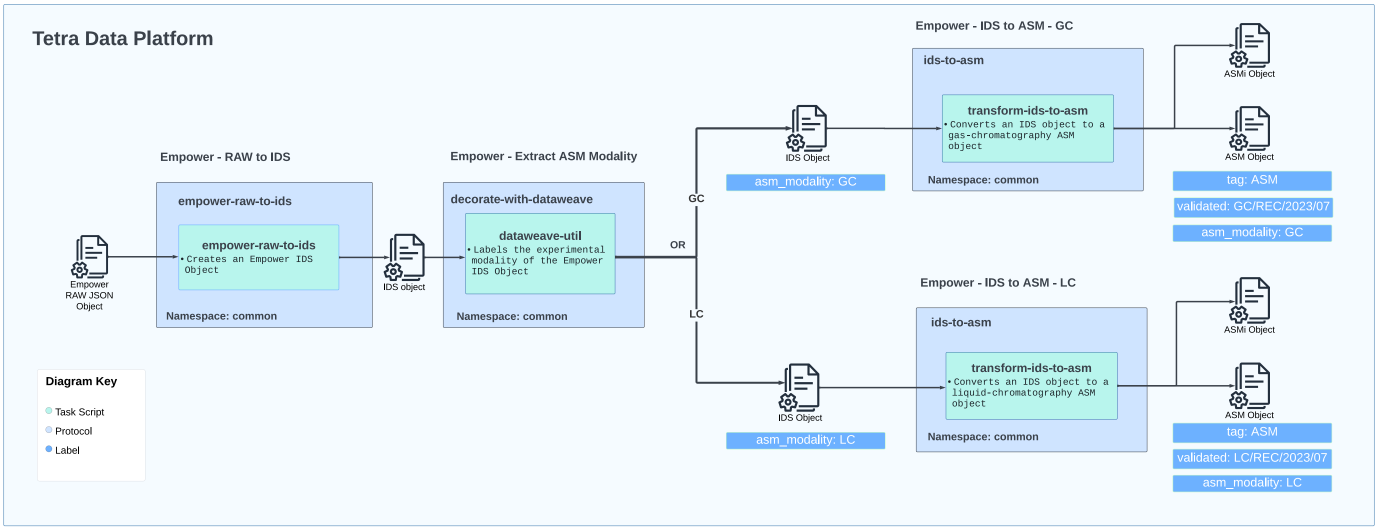
Creating data in the Allotrope Simple Model (ASM) at scale using the TetraScience Intermediate Data Schema (IDS)

FACT SHEET

A few years after TetraScience decided to build an Intermediate Data Schema (IDS) on JSON, Allotrope Foundation developed and released the Allotrope® Simple Model (ASM), which, like the IDS, also uses JSON.

TetraScience has advocated for this move in blog posts and within the Allotrope Community, since this approach can make Allotrope data models more compatible with data engineering efforts, the data science–related tooling ecosystem, tech trends, and TetraScience product offerings. ASM enables strong compatibility with the TetraScience IDS.

If your organization is interested in adopting the ASM, you can use the Tetra Scientific Data Cloud to transform your data into the ASM at scale. To fine-tune this process, TetraScience collaborated with a top 25 pharma company to integrate primary lab data into the cloud, engineer the data into our IDS, and then transform data into the ASM.



The end-to-end workflow for creating ASM objects at scale: Raw instrument data is engineered into the TetraScience IDS objects and further contextualized with a label that indicates the experimental modality of the file. The modality is used to route the file to the corresponding mapping logic that converts the data into the ASM object.

How does the ASM transformation work?

In the case of our collaboration with a pharma company, the journey to the ASM began with the raw instrument data emanating from Tetra Data acquisition components, such as agents or connectors. There was no easy to understand schema associated with the raw scientific data, and the content within was not searchable and actionable. We converted the file to our IDS object, which was indexed by Elasticsearch and searchable via a SQL-like interface in AWS Athena and data warehouses such as Redshift and Snowflake.

The IDS was designed to be an intermediate file for a plethora of scientific use cases outside of mapping-related work. It typically maps to the source system, for example, the particular Chromatography Data System (CDS) the raw data is extracted from. The IDS is designed such that it can preserve as much information from the source system as possible while using taxonomies and reusable schema building blocks for consistency.

Since sometimes the ASM is designed to map to an instrument technique, there is a one-to-many relationship between an IDS and ASM object(s). Consequently, you need to determine the instrument technique of the raw file. We used our DataWeave as the low-code scripting configuration to extract the instrument technique of the IDS object and assign a label to the file. For example, currently two ASMs are supported for Empower: liquid chromatography and gas chromatography. We used the label as a way to route to the appropriate IDS-to-ASM mapping, which converts the data into the ASM object. The instrument technique label can also be used to exclude instrument techniques that may be leveraged by the instrument for which there is no ASM schema or use case identified thus far.

The IDS maintains a common structure in the major components of the data. This common structure promotes consistency in the logic used for the IDS-to-ASM mapping scripts. It allows us to borrow existing logic when creating a new mapping, which saves time in the development cycle.

Keep your raw-to-IDS pipeline running

Working with the pharma company highlighted some key best practices for the ASM transformation. For example, it is important to keep an existing pipeline of raw data to IDS running. All a user needs to do is to add a new pipeline triggered on the IDS JSON to generate the ASM JSON.

As described above, this second pipeline will:

1. Determine which ASM the IDS object should be mapped to. This is important because TetraScience has IDS mapped to source systems, such as Thermo Fisher Chromeleon. And Allotrope has two relevant data schemas based on instrument technique or modality. As a result, it is important to determine how to map the modality to the data object.

2. Generate the ASM object in the “PROCESSED” or “TMP” category.

3. Generate the IDS-compatible ASM (“ASMi”) object (see definition below).

This approach has been tested on ASMs for more than 1,500 instruments, including some of the most complex ones, such as gas chromatography and liquid chromatography ASMs.

Step 1 in the pipeline is crucial. If you are interested in creating ASM objects at scale for thousands of instruments, there needs to be clear routing logic.

Enable queries with ASMi

Once the ASM object is generated in the platform as a file, it is available via REST API, Elasticsearch Query, and SQL. However, its content is not recognizable by the platform. That’s why TetraScience introduced the concept of ASMi to allow an ASM object to be available via Elasticsearch and SQL interfaces provided in the platform. This file is like a mirror of the ASM that allows users to query directly over the ASM terms.

There are some minor adaptations, however. As of August 2023, a field in ASM can either be a literal or an object and literals of different data types. For example, a field can be either a string, a number, or an object with subfields.

There is a great deal of structural flexibility in the semantic world: Objects can be many types. That flexibility can be challenging for data analytics in the whole (e.g., evaluating data for a specific schema) both in SQL and Python. When generating the ASMi object, the overall continuous tracking of data lineage by the Tetra Data Platform remediates any potential data loss.

Why IDS and then ASM?

As baked in the name of the “Intermediate” Data Schema, an IDS object is designed as an intermediate schema that can be easily transformed into another schema based on business needs. For example, users can transform the IDS schema into a final schema used in an ELN, LIMS, or data warehouse. In this case, it can be transformed into the ASM.

Why use IDS as an intermediate layer?

- **Access to more data.** The ASM is based on an instrument technique. It does not support all the information in the scientific data—especially information specific to the particular vendor's instrument. Using the IDS, the data engineering team can extract and structure as much data as possible. Even though only a subset goes into the ASM, users still have access to other data sets. Furthermore, they can preserve and audit data lineage and derivation information.
- **Efficient mapping**. The IDS makes the mapping exercise from heterogeneous scientific data produced by hundreds of vendors to ASM much more efficient. Since the data is already schematized via the IDS, data engineers working on creating the ASM just need to perform what is mostly a JSON-to-JSON mapping. Our intermediate schemas are designed using reusable components, providing a consistent starting point for data engineering and mapping logic design.
- **Immediate value from data.** Scientists can get immediate value from their data when the ASM is missing or evolving. When an ASM is not available, data consumers can use the IDS to schematize their scientific data and then later standardize to the ASM when the Allotrope Foundation releases the relevant ASM.
- **Cost savings and time acceleration.** For companies that are early adopters of the ASM, the IDS-to-ASM strategy provides significant cost savings and time acceleration. Without the IDS, they would need to map the heterogeneous (and often proprietary) scientific data from vendor format/schema to ASM objects from scratch—bearing the cost of design, development, testing, and maintenance. By contrast, the IDS is maintained and supported by TetraScience—and TetraScience is adding new schemas and improvements to the IDS library on a weekly or monthly basis.

In the case the IDS misses information that is needed in the ASM, TetraScience will work with customers to add the improvements to the TetraScience roadmap.

Summary

In our collaboration with a pharmaceutical company, we contributed valuable feedback for the ASM. That feedback exemplifies our commitment to contributing to the Allotrope Foundation, and fostering a continuous cycle of refinement and progression.

We will make similar contributions with each upcoming IDS-to-ASM conversion. This iterative, collaborative process aids the Allotrope Foundation, TetraScience, and biopharma companies in comprehending the mapping logic, schema design, and the ways scientific data drives scientific outcomes. At the same time, it also addresses the constraints inherent in data acquisition and data engineering. Consequently, this iterative collaboration leads to substantial enhancements to both IDS and ASM standards. It also fuels the creation of large-scale, liquid, contextualized, and compliant reengineered scientific data for analytics and AI.

TetraScience is The Scientific Data Cloud Company with a mission to transform life sciences, accelerate discovery, and improve and extend human life.

To learn more about how the Tetra Scientific Data Cloud can help scientific data engineers streamline their pipelines, visit tetrascience.com

Corporate Headquarters | 177 Huntington Avenue, Suite 1703, Boston, MA 02115
Atlanta | 3280 Peachtree Road NE, 7th Floor, Atlanta, GA 30305

tetrascience

tetrascience.com