noblis
For the best of reasons

Comprehensive Knowledge Mesh for Emergency Management and Response in Coastline Disasters

Final Report

Report Authors	Achuna Ofonedu; Amogha Sinha; Dawson Miller
Project Contributors	Achuna Ofonedu; Amogha Sinha; Archana Donavalli; Chip Steiner; Dawson Miller; Kyle Arbide

Version 1.0 – December 15, 2025

Acknowledgement: *This project was funded through the Joint Venture Partnerships Program, Office of Systems Architecture and Engineering (SAE), NOAA/NESDIS.*

Table of Contents

List of Figures

List of Tables

1. Executive Summary

The National Oceanic and Atmospheric Administration (NOAA) Knowledge Mesh (KM) for Emergency Management and Response in Coastline Disasters is a research project aligned with the National Environmental Satellite, Data, and Information Service (NESDIS) and NOAA mission to conserve and manage coastal and marine ecosystems and resources by improving discovery, access, use, reproducibility, and trust of environmental intelligence. Noblis' knowledge graph solution leverages open standard taxonomies and ontologies for maximizing interoperability and scalability and connects with a digital twin for simulating predicted observations of potential oil spills. The NOAA KM creates a centralized entry point to access siloed, cross-domain structured and unstructured data via an Application Program Interface (API), providing a reliable foundation for advanced analytics and Artificial Intelligence (AI) applications.

1.1. Objective

The primary objective of this project is to develop a comprehensive and robust prototype for tracking and managing oil and chemical spills along coastlines, leveraging advanced knowledge graph capabilities to enhance NOAA's environmental intelligence framework. The prototype can be used to significantly improve emergency response and environmental protection efforts by integrating real-time, multi-modal, and historical data into a unified, semantically enriched NOAA KM. Multiple user personas benefit from the design and testing of the Knowledge Mesh, ensuring it is able to provide accurate and contextually relevant information in response to questions from users across various fields.

1.2. Key Findings and Outcomes

Noblis successfully demonstrated the significance of integrating diverse environmental intelligence datasets through advanced knowledge graph capabilities that leverage open semantic web standards for interoperability and scalability. By deploying open standards, containerized and serverless architecture, and digital twin integration, Noblis has developed a knowledge mesh prototype that serves as a scalable framework to transform previously siloed environmental intelligence datasets into an interconnected, API-accessible knowledge mesh resource that provides a basis for improved decision-making during coastline disasters such as oil spills.

The NOAA KM focused specifically on the Chesapeake Bay region for developing the prototype, linking together structured geospatial data, unstructured documents, and real-time satellite observations into a coherent knowledge mesh framework. Through collaboration with Sedaro, Noblis successfully integrated digital twin capabilities that enable simulation of satellite observations over potential oil spills for monitoring and emergency response planning. Multiple personas are beneficiaries of the NOAA KM. For example, emergency response managers benefit from real-time disaster monitoring and predictive scenario modeling to coordinate effective responses to coastal disasters, while coastal scientists gain unified access to integrated weather forecasts from NOAA’s hydrodynamic models. Ecologists can leverage the system's real-time environmental monitoring capabilities and cross-domain ecological datasets, climatologists gain consistent variable definitions and unified access to climate, ocean, and ecological data, and urban planners obtain integrated geospatial information and historical case studies on oil spill responses for infrastructure planning.

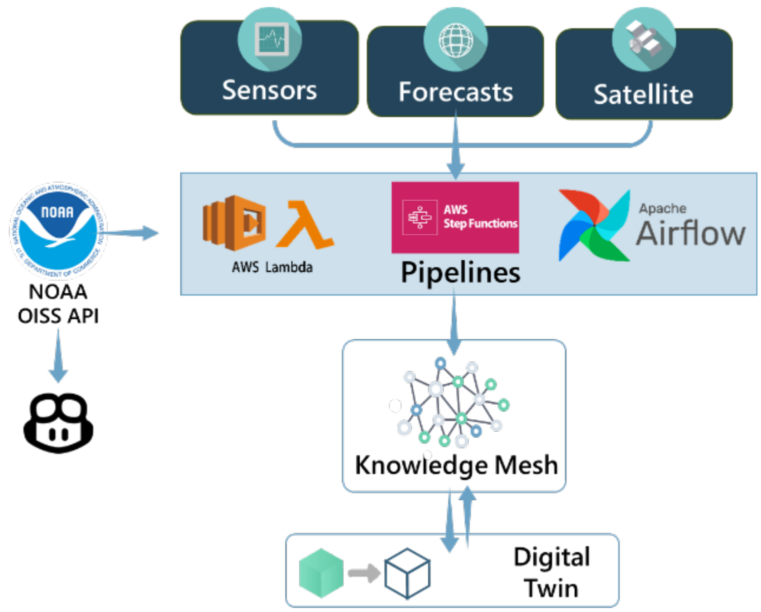


Figure 1. Overview of NOAA KM Integrated Solution

While developing the Knowledge Mesh, Noblis integrated several of the associated data pipelines with NOAA's Open Information Stewardship Service (OISS) python package, and in collaboration with NOAA stakeholders, the team at Noblis developed an interoperable AI solution using the Model Context Protocol (MCP)[1] to save development time, answer questions from the documentation, and perform supervised code generation. The MCP server can deploy alongside Large Language Model (LLM) providers like GitHub CoPilot, where developers can utilize the MCP server's capabilities directly in their Visual Studio Code workspace. Rather than manually searching OISS documentation, developers can directly ask questions and perform supervised code generation based on the OISS text, architecture diagrams, and source code.

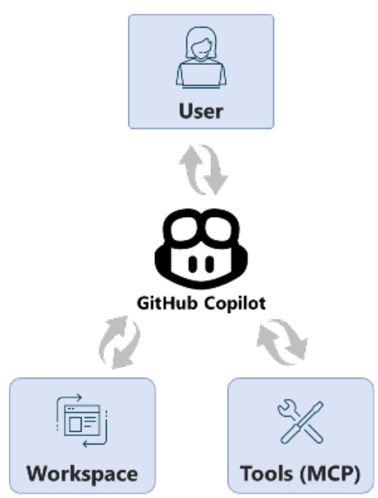


Figure 2. General Architecture for Model Context Protocol (MCP) Server GitHub CoPilot Deployment

[1] MCP servers allow standardized access to resources like tools and data sources across LLMs, enabling interoperability across teams and use cases.

Additionally, while not a primary focus of the project, in collaboration with NOAA stakeholders, Noblis developed a proof-of-concept R2RML (RDB to RDF Mapping Language) integration with OISS, allowing applications and knowledge graphs to query existing databases using SPARQL as though they were native Resource Description Framework (RDF) stores, providing virtualization capabilities and preserving data provenance. In an R2RML demonstration, Noblis established a foundational workflow showing how relational data can be virtually transformed into RDF using the python package Morph-KGC and integrated within the OISS framework. This proof-of-concept validated the end-to-end pathway for future virtualization pipelines, laying the groundwork for a scalable, Findable, Accessible, Interoperable, and Reusable (FAIR)-compliant approach to transforming and querying external datasets on-demand.

Based on the findings from this project, Noblis recommends the following high-level next steps. These next steps are discussed in detail in Section 5.

- Expansion of the NOAA KM to additional geographic regions beyond the Chesapeake Bay.
- Integrate AI-powered information retrieval with the knowledge mesh as a semantically rich and consistent data foundation.
- Expand the digital twin simulation capabilities to additional satellites, missions, and instrumentation.
- Productionize the MCP server to assist with the broad developer adoption of OISS.
- Implement architecture diagram tools from the MCP server as a user business logic (UBL) template that can be leveraged across teams through OISS.
- Further research and validation for deploying R2RML virtualization via OISS such as accommodating additional data formats like Zarr, and supporting federated virtualization
- Within OISS, expand the range of Archival Information Collection (AIC) and Archival Information Unit (AIU) templates to accommodate emerging environmental datasets and analytical models.

2. Introduction

NOAA has identified a need to improve and expand the open discovery, access, use, reproducibility, and trust of its environmental intelligence – both in terms of process and information. In support of this initiative, the primary objective of this project is to develop a comprehensive and robust prototype for tracking and managing oil and chemical spills along coastlines, leveraging advanced knowledge graph capabilities to enhance NOAA's environmental intelligence framework. The prototype can be used to significantly improve emergency response and environmental protection efforts by integrating real-time, multi-modal, and historical data into a unified, semantically enriched knowledge mesh. Multiple personas have guided and benefited from the design and testing of the Knowledge Mesh, ensuring the knowledge mesh solution is able to provide accurate and contextually relevant information in response to questions from users across various fields.

2.1. Background and Motivation

NOAA and its partner institutions produce vast amounts of environmental intelligence. However, the data sets and information repositories are often designed for specific users and not connected with one another, resulting in information with limited scope and opportunity for cross-domain discovery.

No single interface exists to bring the data together and support the needs of a wide variety of NOAA users, such as scientists, engineers, fishermen, planners, and other agencies. Currently, to bring together cross-domain datasets for environmental intelligence workflows, stakeholders need to discover a variety of data access locations, familiarize themselves with unique portals and access methods, and understand the formats and contents before manually making connections between them.

This effort is designed to improve discovery and accessibility across NOAA and other agencies, integrate with digital twin technology, and establish a foundation for future agentic AI integration. Importantly, the solution must support key personas, including emergency managers, urban planners, ecologists, and coastal scientists, by providing a structured and scalable approach to integrating and semantically enriching NOAA datasets, models, and tools.

Noblis' solution establishes a NOAA KM centered around oil spill response in the Chesapeake Bay. The NOAA KM employs open semantic web standards along with open, standard taxonomies and ontologies to enhance data linkage, query efficiency, and machine-readable interoperability. Since the knowledge mesh leverages these open and standard taxonomies and ontologies, it is interoperable with other knowledge graph resources. This interoperability makes scaling the solution more feasible, as the same standards can be reused across project teams and data products. With the NOAA KM, Noblis is creating semantic relationships across a range of environmental intelligence datasets, transforming structured and unstructured data into a unified, interoperable, and API-accessible entry-point across domains.

The NOAA KM supports multiple key personas including emergency response managers, urban planners, ecologists, and coastal scientists. In practical demonstration of relevance to these personas, the NOAA KM also integrates with a digital twin for simulating forecasted observability of potential oil spills. The digital twin and knowledge mesh are connected via a python API client.

2.2. Project Scope and Stakeholders

The NOAA KM project encompasses the development of a comprehensive prototype integrating, enriching, and exploiting environmental intelligence data related to coastline disasters, focusing on oil and chemical spills in the Chesapeake Bay region. The scope of the project involves demonstration of several key milestones across two primary phases: KM Construction and KM Grooming. Importantly, the Knowledge Mesh is intended to benefit multiple user personas including coastal scientists, ecologists, climatologists, and emergency managers. The construction of the mesh lays the foundation for future KM exploitation. The primary phases addressed in this project involve several key milestones given below. Note that the project's scope is limited to a demonstration of each milestone, not deployment to end-users.

2.2.1. KM Construction

The Knowledge Mesh construction phase includes: (1) conducting user persona interviews for data and user story identification, (2) creation of data pipelines for ingesting data into the Knowledge Mesh with demonstration of alignment with NOAA's OISS initiative, and (3) the creation of custom taxonomies for data normalization. In support of NOAA's commitment to open-source technologies and software development, the data pipelines employ open-source and industry-standard data engineering and Natural Language Processing (NLP) tools and frameworks to ingest, enrich incoming data with relevant metadata, and normalize entities.

2.2.2. KM Grooming

The Knowledge Mesh grooming phase includes: (1) the development of ontologies that leverage predefined namespaces such as Simple Knowledge Organization System (SKOS), Dublin Core (DC), Schema.org (SDO), and XML Schema Definition (XSD), (2) API development for seamless and open data access to the Knowledge Mesh according to FAIR (Findable, Accessible, Interoperable, and Reusable) principles, ensuring the system can answer highly contextual queries and retrieve precise, relevant information for a range of NOAA users, (3) quality monitoring to ensure consistency of the data in the Knowledge Mesh, (4) the development of a digital twin that integrates with the Knowledge Mesh for coastline disaster scenario analysis and planning, and (5) creation of a data catalog.

3. Methodology

3.1. Use Case and Dataset Identification

3.1.1. Use Case Evaluation Framework

The initial stage of the project centered on identifying and prioritizing use cases most relevant to NOAA's Knowledge Mesh objectives. To ensure alignment with agency goals and technical feasibility, Noblis established a structured evaluation criteria framework comprising three key dimensions: applicability, feasibility, and stakeholder value.

1. Applicability: Is the use case relevant to the objectives of the NOAA KM project?

This criterion measures how well each proposed use case supports the overarching goals of the NOAA Knowledge Mesh, improving data discoverability, interoperability, and accessibility for coastal and environmental decision-making. Applicability also considered alignment with NOAA's identified priorities, including Digital Twin integration, satellite data fusion, and natural language data interoperability.

2. Feasibility: Does the use case contain challenges that are impractical?

Feasibility assesses the technical complexity, data availability, and subject matter expertise required to implement the use case within the project timeline. Considerations include the maturity of relevant data sources, ease of integration into the Knowledge Mesh architecture, and expected dependencies on NOAA subject matter experts (SMEs).

3. Stakeholder Value: Which user personas benefit?

This dimension focuses on the end-user impact of each use case. It evaluates the potential of each use case to address the needs of NOAA's core user personas, including ecologists, coastal scientists, climatologists, emergency managers, and urban planners. Key questions included:

- Does the use case solve a primary operational need?
- Does it serve multiple stakeholder groups?
- Are similar solutions already available, or would this provide novel value?

These evaluation criteria provided a common framework for discussion and transparent prioritization of proposed scenarios.

3.1.2. Persona-Centered Design and Alignment

The use case identification process was deeply informed by five primary NOAA user personas, derived from interviews, field research, and structured evaluation (see Figure 3). Each persona represents a stakeholder group directly impacted by coastal hazard management and environmental response operations. By grounding the selection process in the personas listed below, the project ensures that chosen use cases address tangible operational challenges while representing diverse NOAA stakeholder perspectives.

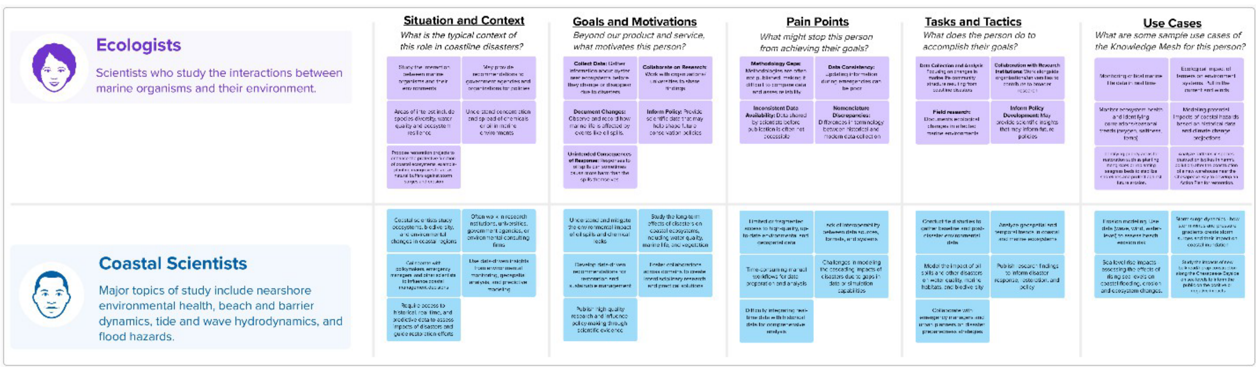


Figure 3. Example of Noblis' Structured Evaluation of User-Personas

1. Ecologists

Ecologists focus on understanding interactions between marine organisms and their environment. Their motivations center on data collection for conservation policy, with challenges around methodological consistency, fragmented datasets, and delayed information during emergencies.

Key Needs: Real-time environmental monitoring, access to cross-domain ecological datasets, and integration of observational and model-based data.

2. Coastal Scientists

These users study nearshore dynamics including tides, currents, and storm surges. They require integrated datasets combining hydrodynamic models, bathymetry, and pollution data for predictive analysis.

Key Needs: Access to historical and forecast model outputs, multi-source integration for scenario planning, and visualization tools for communicating results.

3. Climatologists

Climatologists assess long-term climate impacts, variability, and trends affecting coastal resilience. Their pain points include limited temporal continuity across datasets and difficulty comparing model results across scales.

Key Needs: Unified access to climate, ocean, and ecological data; consistent variable definitions for cross-model comparison; and analytic support for climate trend analysis.

4. Emergency Managers

Emergency managers require timely, reliable data to plan and coordinate responses to coastal disasters. Fragmented data systems and delayed model outputs hinder decision-making during crises.

Key Needs: Real-time disaster monitoring, predictive scenario modeling, and integration of social vulnerability data for evacuation planning.

5. Urban Planners

Urban planners leverage environmental intelligence for zoning, infrastructure resilience, and community protection. They face data access challenges across disparate portals and inconsistent metadata standards.

Key Needs: Integrated geospatial data, historical hazard trends, and predictive analytics to guide development and emergency planning.

3.1.3. Selected Use Cases

The Noblis team conducted a comprehensive evaluation to select three uses cases as top priorities. The team gathered and analyzed potential use cases from user persona interviews and research to understand stakeholder needs and participated in a structured voting process to prioritize the use cases best aligned with the project. Based on the evaluation and voting outcomes, five use cases emerged as top priorities, from which three were selected for implementation under this phase of the project:

1. Real-Time Disaster Monitoring and Early Warning Systems

The objective of this use case is to detect and forecast coastal hazards such as oil spills, floods, and storm surges by integrating near real-time satellite observations from the Marine Pollution Surveillance Reports (MPSR) with data from the Operational Forecast System (OFS). Through the Knowledge Mesh, these datasets are semantically linked to enable automated alerting, cross-domain event correlation, and improved situational awareness. This integration allows analysts to trace pollution events in relation to forecasted ocean dynamics and coastal impact zones, supporting faster and more informed decision-making. The architecture also establishes a foundation for future digital twin simulations that fuse live observations with predictive models, providing NOAA with a scalable early warning capability for environmental response and planning.

2. Identifying Socially Vulnerable Areas at Risk Near the Chesapeake Bay

Identification of vulnerable areas and habitats is critical in coastline disaster response, which this use-case aims to support by understanding the intersection of environmental hazards, at-risk coastal resources, and human exposure. Both the Environmental Sensitivity Index (ESI) and Natural Resource Damage Assessment and Restoration (NRDAR) datasets are relevant to this use case, as they characterize human and ecological exposure to coastal disasters, supporting urban planners and emergency managers in preparedness planning.

3. Understanding the Spread of Oil Spills

Provided OFS data can offer correlations to oil dispersion rates. This use case directly supports ecologists and coastal scientists through simulation-based predictive insights. By incorporating this use case into the construction of the Knowledge Mesh, ecologists and coastal scientists benefit from integrated access to datasets such as the OFS, which provides nowcasts and forecasts of coastline hydrodynamics including temperature, salinity, and current velocity. These model outputs form the foundation for simulating oil spill dispersion and behavior under varying environmental conditions. The digital twin integration connects these data to predictive modeling workflows, enabling scientists to generate satellite overpass simulations and align them with forecasted spill movement. This combination of observed and modeled data enhances NOAA's capacity to evaluate the spread and impact of spills in near real time, improving readiness and coordination during response operations.

Together, these selected scenarios collectively address NOAA's mission goals across environmental monitoring, climate resilience, and emergency management ensuring broad applicability and measurable operational value.

3.1.4. Dataset Selection

In selecting the datasets for the Knowledge Mesh, Noblis utilized the user persona interviews, analysis on user and data challenges and benefits, and discussion with NOAA stakeholders in the monthly meetings. The datasets listed below vary in sourcing, format, and use cases to demonstrate how the Knowledge Mesh can benefit multiple user personas. Additional technical details for developers and technical users are provided in the Technical Data Package.

1. Operational Forecast System (OFS)

OFS (NOAA Center for Operational Oceanographic Products and Services, 2025) contains time series nowcasts and forecasts with 48-120 hours horizons on parameters from the National Ocean Service's (NOS) hydrodynamic models such as water levels, temperature, salinity, and currents. These forecasts are updated four times per day with 6-hour nowcasts. The Chesapeake Bay region forecasts contained in the OFS data provide key, real-time context for oil spill response, enabling users to query core weather expectations over a possible oil spill.

2. National Resource Damage Assessment and Restoration (NRDAR)

The NRDAR (U.S. Department of the Interior, 2025) Case Document Library is managed by the Department of the Interior, and compiles data and documents on various environmental incidents to support the assessment of natural resources. Similar to NOAA's Data Integration, Visualization, Exploration, and Reporting tool (DIVER, 2020), the NRDAR Case Document Library provides a public, map-based library of restoration cases that cover the full process from assessment through monitoring and closure. It is used to support Natural Language Processing (NLP) and semantic tagging for extracting geospatial and environmental insights. For the Vulnerable Areas use case, NRDAR adds context on affected habitats, species, and restoration plans, providing insight into environmental damage and recovery.

3. Environmental Sensitivity Index (ESI)

NOAA's ESI data (NOAA Office of Response and Restoration, 2025) is a multi-layered spatial dataset that maps shoreline types, habitats, species occurrences, and human-use areas. Each layer ties biological or human-use features to geometry and source metadata and includes breeding and abundance data that are crucial for response and planning. ESI supports the Vulnerable Areas use case as the main structured dataset for assessing ecological exposure and cleanup priority.

It provides spatially consistent data for understanding which sensitive habitats and resources are most at risk and where they are located.

4. Marine Pollution Surveillance Reports (MPSR)

The MPSR dataset (NOAA Office of Satellite and Product Operations, 2025), provides near real-time detection of potential oil spills and pollution events derived from satellite imagery. Each report includes detailed metadata such as image acquisition time, satellite sensor, spatial resolution, estimated slick area, confidence level, and analytical remarks, along with derived map products and shapefiles that delineate observed spill extents. These geospatial components enable direct transformation into RDF within the Knowledge Mesh, supporting semantic and spatial reasoning for incident tracking. As the primary dataset for the Early Warning use case, MPSR links satellite observations with model forecasts and environmental sensitivity layers, offering rapid, data-driven insights for coastal monitoring and emergency response.

3.2. Taxonomies and Ontologies

To support consistent semantic alignment across NOAA datasets, the Knowledge Mesh relies on two types of semantic resources: taxonomies and ontologies. Taxonomies provide standardized vocabularies for labeling entities, and ontologies define the formal structures used to represent relationships, geospatial features, provenance, and metadata. Taxonomies ensure consistent naming and classification, while ontologies define the schema that allows datasets to interoperate and be queried semantically.

3.2.1. Taxonomies

The NOAA KM prototype uses a combination of existing taxonomies and a custom taxonomy created to support both structured datasets and natural language sources processed by the pipelines. These taxonomies provide standardized vocabularies for identifying species, geographic features, and other key entities, ensuring that extracted information is normalized and linked reliably across the mesh.

1. **National Center for Biotechnology Information (NCBI) Taxonomy**

   The NCBI Taxonomy provides a standardized scientific classification for species, allowing all organism names extracted from datasets or text to be normalized to consistent identifiers.

2. **Geographic Features Taxonomy**

   The Geographic Features Taxonomy is a custom vocabulary built from Geographic Names Information System (GNIS) feature types and mapped to Environment Ontology (ENVO) feature classes. This allows natural language references to places like rivers, creeks, and bays to be consistently normalized and connected to the environmental feature ontology described in the following section.

In the NRDAR natural language pipeline, entities are tagged from unstructured case documents using the geographic features taxonomy and represented as nrdar:TaggedTerm instances. The documents themselves are modeled as dcat:Distribution records and linked to their extracted entities using the nrdar:hasTaggedTerm attribute. Each tagged term captures the matched text span, the spatial context, and the feature type. It also references the corresponding entry in the geographic features taxonomy, allowing extracted information to be semantically normalized and linked across datasets.

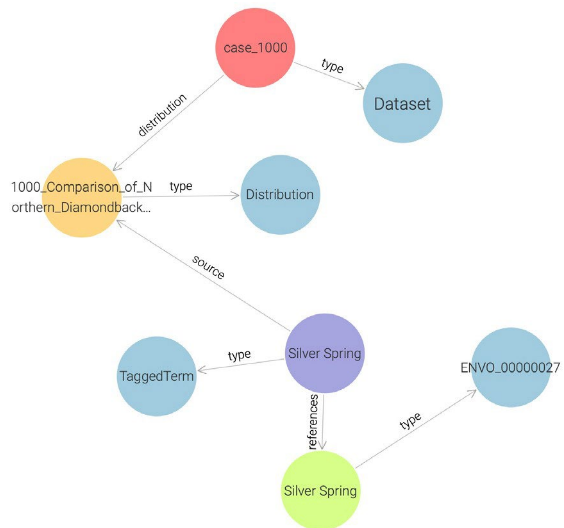


Figure 4. NRDAR Tagged Term Visualization

For example, the term Silver Spring is tagged in the "Estimate of Total Injury to Diamondback Terrapins" study report from the Potomac Electric Power Co. Chalk Point oil spill case in the NRDAR Case Document Library. This tagged term captures the text span, spatial information ("West Virginia"), and feature type ("Spring"). Its reference in the Geographic Features Taxonomy includes the official GNIS label ("Silver Spring"), the geographic coordinates (POINT(39.2903794 -78.0199985)), the type of feature (Spring), and a mapping to the ENVO class envo:00000027, which represents a spring. This structure ensures that natural language references to geographic features are normalized, semantically linked, and interoperable across the Knowledge Mesh.

3.2.2. Ontologies

Table 1. List of Ontologies Used in the Mesh

Category	Ontologies	Description
General Semantic Web	RDF (W3C), RDFS (W3C), OWL (W3C), SKOS (W3C), XSD (W3C)	Core semantic-web standards for structuring and typing semantic data
Cataloging/ Metadata	DCAT (W3C), PROV-O (W3C), DCTERMS (DCMI), Schema.org, FOAF (FOAF Project)	Metadata, dataset descriptions, provenance, and agents
Environment & Sensors	SOSA (W3C), SSN (W3C), TIME (W3C), QUDT (QUDT.org)	Sensor observations, measurements, units, and time
Geospatial	GEO/WGS84 (W3C), GEOSPARQL (OGC), WKT Literals (OGC), ENVO (OBO), PATO (OBO)	Geospatial features, environmental terms, and geometry
Biological	DWC (TDWG), NCBI Taxonomy (NCBI)	Species classification and biological occurrence data

Network Common Data Form (NetCDF) & Scientific Data	NetCDF (Unidata), CF-Metadata (CF), OISS (NOAA)	Scientific data formats and climate conventions

The Knowledge Mesh draws on a broad suite of ontologies to provide semantic meaning, consistent typing, and interoperable structure across all the datasets. Core semantic-web standards such as RDF, Web Ontology Language (OWL), and SKOS define the basic framework for representing knowledge (W3C RDF Working Group, 2014; W3C OWL Working Group, 2012; Simple Knowledge Organization System Working Group, 2009). Metadata and provenance are described using Data Catalog Vocabulary (DCAT), Dublin Core Metadata Terms (DCTERMS), and Provenance Ontology (PROV-O) (W3C Data Catalog Vocabulary Working Group, 2020; Dublin Core Metadata Initiative, 2020; W3C Provenance Working Group, 2013). Environmental and sensor models include ENVO, Sensor-Observation-Sampling-Actuator ontology (SOSA), Semantic Sensor Network ontology (SSN), and Quantities, Units, Dimensions, and Types (QUDT)(Buttigieg, 2013; W3C Semantic Sensor Network Incubator Group / OGC, 2017; QUDT.org, 2025), while geospatial ontologies such as GeoSPARQL and WGS84 provide spatial context (Open Geospatial Consortium, 2022, 2019). Biological vocabularies like Darwin Core and NCBI Taxonomy support species and habitat data (Darwin Core Task Group, 2021; National Center for Biotechnology Information, 2023), and scientific data standards such as CF-Metadata and OISS enable consistent measurement and observation reporting (CF Conventions Committee, 2023; Open Geospatial Consortium, 2024). Across these, ENVO, GeoSPARQL, DCAT/PROV-O, and OISS serve as the mesh's core semantic pillars, providing the environmental, spatial, metadata, and scientific observation frameworks that anchor interoperability across all datasets.

In addition to these standard ontologies, the ESI, NRDAR, and MPSR datasets define their own custom namespaces to capture attributes unique to their domains. The ESI namespace captures Environmental Sensitivity Index layers, shoreline types, breeding occurrences, and habitat-specific metrics. The NRDAR custom namespace represents case documents, incidents, trustees, and tagged environmental features, enabling more detailed resource representation in the mesh. Likewise, the MPSR namespace supports marine pollution surveillance reports, including attributes supporting imagery, spill locations, spill types, confidence scores, and observational metadata. These custom ontologies ensure that domain specific properties and classifications can be represented in the mesh consistently while still integrating seamlessly with the broader semantic framework.

Together, these ontologies create a unified semantic layer that supports consistent representation of datasets, entities, observations, and provenance across the Knowledge Mesh.

3.3. Software Development Environment and Standards

The Noblis development team utilized a robust software development environment built around containerization with Docker, which enabled consistent builds and deployment of data pipelines and API. The pipelines are packaged using Poetry for package and dependency management into the noaa-cd-km (NOAA Coastline Disaster Knowledge Mesh) python package. [2] The containerization and dependency management accommodate both on-prem and cloud deployments, streamlining the use and future development of the Knowledge Mesh's pipelines.

In developing the python package, Noblis utilized consistent styling by adhering to Google's python style guide, which benefits readability and future maintaining developers in quickly becoming familiar with the codebase. For API development, the team followed FAIR principles, making their services more discoverable and usable by other systems. Furthermore, security was integrated into the software development workflow through automated scanning via GitHub Enterprise Security, which helps identify vulnerabilities on each push to the git-controlled repository.

[2] The repository containing the resulting codebase is included in the Technical Data Package.

This resulted in security bugs and dependency vulnerabilities being discoverable as early as possible in the development process.

3.4. Data Pipeline Construction and Deployment

The NOAA KM utilizes advanced data engineering solutions to transform heterogenous data sources into a cohesive, semantically enriched knowledge graph. A dual architecture for demonstrating the data pipelines provided in the noaa-cd-km package: (1) an on-prem deployment of the OFS pipeline with Apache Airflow and Ontotext GraphDB that allows for a cloud-agnostic option with industry-standard, robust scheduling capabilities, and (2) an Amazon Web Service (AWS) deployment of the geospatial pipelines with AWS Lambda, Step Functions, Event Bridge, Elastic Container Registry, S3, and Neptune that illustrates how the pipelines can integrate with NOAA's OISS package.

There are two categories of data pipelines in the NOAA KM: geospatial and natural language. The geospatial pipelines include OFS, MPSR, and ESI, while the natural language pipeline includes the NRDAR dataset. Each pipeline follows the pattern of extraction, optional cleaning, transformation into RDF, and export into the Knowledge Mesh. In AWS, the geospatial pipelines are deployed in containerized and serverless architecture to ensure cost efficiency and scalability. Dedicated AWS Step Functions listen to EventBridge rules for triggering the Extract, Transform, Load (ETL) AWS Lambda functions. See Figure 5 for a cloud architecture diagram of the deployment in AWS, and the discussion below for further technical details.

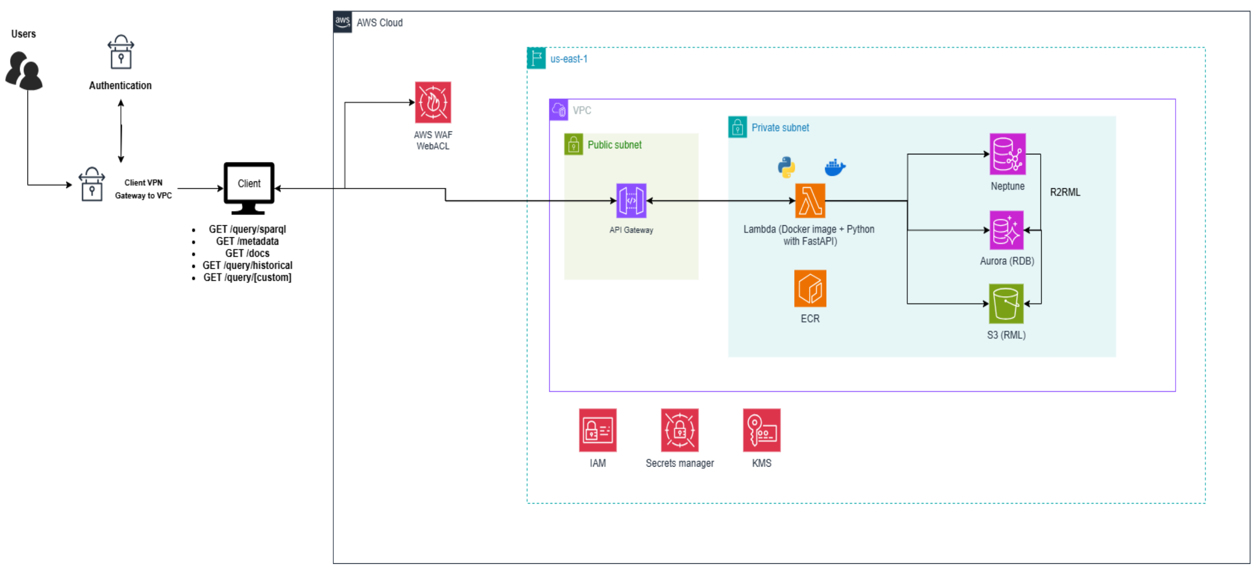


Figure 5. AWS Architecture

3.4.1. Geospatial Pipelines

The geospatial pipelines in the NOAA KM handle structured spatial data from multiple sources, transforming this data into semantically enriched RDF that is compatible with GeoSPARQL standards. Each pipeline employs consistent patterns for coordinate reference system standardization, conversion of spatial features to Well-Known Text (WKT) format, and semantic annotation with domain-specific vocabularies. Quality validation occurs throughout the pipeline execution, ensuring spatial accuracy, metadata completeness, and ontological consistency.

Additionally, global Shapes Constraint Language (SHACL) validation is possible through the provided shapes files and external, open-source SHACL software such as pySHACL (Sommer, 2024) and Apache Jena SHACL (Apache Software Foundation, 2021).

In AWS the geospatial pipelines are deployed in serverless and containerized architecture. Dedicated AWS Step Functions manage sequential and parallel tasks for data acquisition, transformation, and publication. The noaa-cd-km pipelines are containerized inside dedicated AWS Lambda functions to be orchestrated by the associated Step Functions. When new data such as new OFS model estimates are detected, an AWS EventBridge rule triggers the associated Step Function, which initiates the retrieval, processing, and registering of the latest data. See Figure 6 for an example of this deployment architecture for the OFS pipeline.

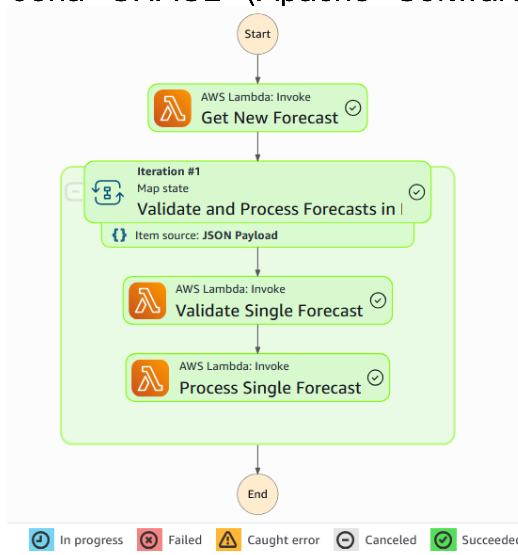


Figure 6. AWS Step Function for OFS Pipeline Deployment

Operational Forecast System (OFS)

The OFS pipeline automatically ingests, validates, and transforms NOAA's OFS data for the Chesapeake Bay region into semantically enriched RDF suitable for integration into the knowledge mesh. As the OFS data is hosted on the Registry of Open Data on AWS Simple Storage Service (S3), the OFS pipeline can be deployed on a recurring schedule to pull fresh nowcasts and forecasts as they become available. The pipeline implements a comprehensive workflow that begins with forecast identification against user-supplied constraints such as specification of specific forecasts. Then the pipeline validates the incoming data against expected variables and metadata structures, ensuring incoming data matches the expected structure by comparing variables against those contained in the existing graph.

During transformation, the OFS pipeline extracts key temporal features including model run datetime, forecast time, and associated time intervals. It processes oceanographic variables such as salinity, temperature, oxygen levels, and wind vectors, converting them into standardized observation patterns. Each observation is semantically enriched with geospatial coordinates expressed as WKT literals, temporal references, and links to variable metadata. The pipeline supports both surface-level and depth-specific measurements through appropriate coordinate dimensionality.

The OFS pipeline implements the OISS framework through Archival Information Collection (AIC) patterns, member descriptions, and semantic relationships that ensure data is FAIR-compliant. Dataset-level metadata is organized hierarchically with catalog, dataset, and distribution levels, while variable metadata preserves Climate and Forecast (CF) metadata convention attributes. By aligning with GeoSPARQL standards, the pipeline enables advanced spatial queries through standardized representation of geospatial data.

Overall, the OFS pipeline demonstrates a comprehensive approach to transforming complex oceanographic forecast data into semantically rich, spatially accurate RDF that supports advanced querying and analysis. By aligning with the OISS framework through the implementation of AIC patterns and member descriptions, the pipeline ensures that all observations and metadata are properly contextualized within the broader NOAA Knowledge Mesh architecture, enabling seamless integration with other environmental data sources.

Marine Pollution Surveillance Reports (MPSR)

The MPSR pipeline automatically ingests, validates, and transforms pollution observation data into semantically enriched RDF suitable for integration into the Knowledge Mesh. This pipeline is deployed as a serverless AWS workflow that connects data ingestion from NOAA repositories to RDF graph construction in Neptune, ensuring each report is traceable, spatially accurate, and FAIR-compliant.

A dedicated AWS Step Function orchestrates the MPSR pipeline, managing sequential tasks for data acquisition, transformation, and publication. When new reports are detected, an EventBridge rule triggers the Step Function, which initiates the process to retrieve, process, and register the latest data. The ingestion phase identifies new MPSR reports, validates their geospatial coverage, and filters entries based on a defined area of interest. During transformation, the pipeline extracts metadata fields such as date, time, region, confidence, and remarks, and converts geometric data into standardized WKT formats.

Each report is then semantically enriched with domain-specific vocabularies and linked to relevant ontologies through RDF properties. Custom predicates such as hasSourceLocation and hasAreaLocation extend the GeoSPARQL vocabulary to describe incident geometries with high spatial fidelity. Quantitative attributes such as area and oil type are encoded using QUDT and UNIT ontologies to ensure that all data values are machine interpretable. The system also preserves data provenance by associating each RDF node with its source URLs, including text summaries, images, and downloadable report archives.

Once RDF construction is complete, the Step Function writes the resulting triples to the Knowledge Mesh hosted in Amazon Neptune. Validation and consistency checks are performed during this phase to confirm geometry types, coordinate reference systems, and metadata completeness. The pipeline's modular design allows each phase to be executed independently and record its output as a discrete, queryable artifact in the graph.

Overall, the MPSR data pipeline demonstrates an end-to-end automated process that ingests raw marine pollution observations, transforms them into semantically annotated geospatial RDF, and integrates them into the Knowledge Mesh. By aligning this workflow with the OISS framework and the broader NOAA Knowledge Mesh architecture, Noblis established a repeatable, transparent, and FAIR-aligned model for future environmental data pipelines.

Environmental Sensitivity Index (ESI)

The ESI pipeline was designed to streamline the ingestion and semantic transformation of NOAA's ESI geodatabases into RDF graphs for integration into the NOAA Knowledge Mesh. The workflow converts multi-layer data— covering shoreline types, species occurrences, habitats, and human-use areas— into graphs that preserve source relationships and spatial accuracy. By aligning with

NOAA's OISS framework, the ESI pipeline makes environmental sensitivity data more discoverable and usable for vulnerability assessments.

The ESI pipeline is managed through an AWS Step Function triggered by EventBridge when new ESI geodatabases are uploaded to S3. The first Lambda function validates the incoming data. The second Lambda function processes the data and transforms it into RDF. During this step, layer attributes are standardized and organized into triples.

The resulting RDF graph captures the ESI data using a mix of standard and custom vocabularies for spatial, ecological, and provenance details. DCAT describes datasets, sources, and distributions; Darwin Core defines biological occurrences and species information; and GeoSPARQL handles geometry in WKT format, including area and perimeter metrics. Furthermore, a custom ESI ontology extends these models to represent shoreline types, sensitivity codes, and habitat attributes.

In the graph, biological and human-use layers are linked to three core components— Habitat occurrence, ESI lines and polygons, and Resource Areas. Each connects to Area and Source nodes to retain spatial and provenance detail. Habitat Occurrences include details such as location, species, abundance, breeding activity, and observation months. ESI Line and Polygon features detail shoreline and intertidal environments, including environment type, line type, and shore classifications tied to ENVO features. Resource Areas describe managed or protected sites and record attributes like use type, inundation depth, and sensitivity codes.

Overall, the ESI pipeline automates the full workflow from ingesting NOAA's ESI geodatabase to producing structured RDF graphs in Neptune. It captures shoreline types, habitats, species occurrences, and human-use areas, linking them to spatial and provenance details while standardizing key attributes. By embedding this data into the Knowledge Mesh under the OISS framework, the pipeline enables users to explore environmental sensitivity information in a queryable way— supporting vulnerability assessments, planning, and ecological decision-making.

3.4.2. Natural Language Pipelines

The NRDAR pipeline automatically processes unstructured text documents from the NRDAR Case Document Library into semantically enriched RDF suitable for integration into the Knowledge Mesh. The pipeline extracts valuable environmental insights by identifying and tagging key features such as geographical locations, habitats, species, and restoration activities mentioned within assessment documents. Through a two-phase approach, the pipeline first processes document metadata to establish case-level relationships, then applies natural language processing techniques to extract and semantically tag environmental features within the document text.

During metadata processing, the NRDAR pipeline creates a hierarchical structure linking cases to their associated documents, capturing essential information such as case identifiers, descriptions, spatial data, and document dates. The feature extraction phase employs a taxonomy-based tagging system that identifies environmental features within document text, enriching them with contextual information and geospatial coordinates when available. Each tagged term is linked to its source document and referenced feature in the knowledge graph, enabling traceability from extracted insights back to source materials.

By transforming unstructured NRDAR case documents into structured, semantically rich RDF, the pipeline enables advanced queries across environmental incidents, affected habitats, species impacts, and restoration activities. This semantic enrichment aligns with GeoSPARQL standards for spatial representation and leverages domain-specific vocabularies to ensure proper contextualization within the broader Knowledge Mesh architecture. The resulting knowledge graph enhances the discoverability and utility of NRDAR case information, providing valuable insights for vulnerability assessments, restoration planning, and environmental impact analysis across diverse ecosystems and incident types.

3.5. API Development

The development of robust APIs is critical to providing seamless access to the Knowledge Mesh, ensuring that the system can answer highly contextual queries and retrieve precise, relevant information for a range of NOAA users. These APIs need to use SPARQL to grab data and metadata directly from the graph database and return structured responses to the users. Noblis followed OpenAPI specifications, ensuring that all endpoints are well-documented and adhere to industry standards. The APIs created for access to the knowledge graph leverage OpenAPI specifications to create industry standard documentation for all endpoints. APIs are readily available for integration into other tools or software.

The development of robust APIs was a central focus of the Knowledge Mesh architecture, ensuring that users could access environmental intelligence through a secure, scalable, and standards-based interface. Each endpoint was designed to provide structured access to the underlying graph database through parameterized SPARQL queries, allowing users to dynamically retrieve contextual data and metadata across NOAA's key datasets.

The API layer was implemented using a modern, serverless design within the AWS Cloud environment. Core services included Amazon API Gateway for routing and authentication, AWS Lambda for execution of API logic, Amazon Elastic Container Registry (ECR) for container image storage, and Amazon Neptune as the graph database. Amazon S3 provided storage for R2RML mapping files and metadata configurations, while AWS Secrets Manager and Key Management Service handled secure credentials and encryption.

The API was developed in Python using FastAPI, containerized with Docker, and deployed to AWS Lambda through ECR. Each endpoint follows the OpenAPI 3.1 specification, ensuring discoverability and interoperability with other systems. Clear documentation and a live interface were created to allow developers and analysts to interact with the Knowledge Mesh API directly, testing parameters and viewing structured outputs.

Each endpoint corresponds to a predefined SPARQL query template that can be dynamically customized through user parameters such as geographic region, date range, or dataset type. This approach simplifies query creation while maintaining flexibility and reproducibility.

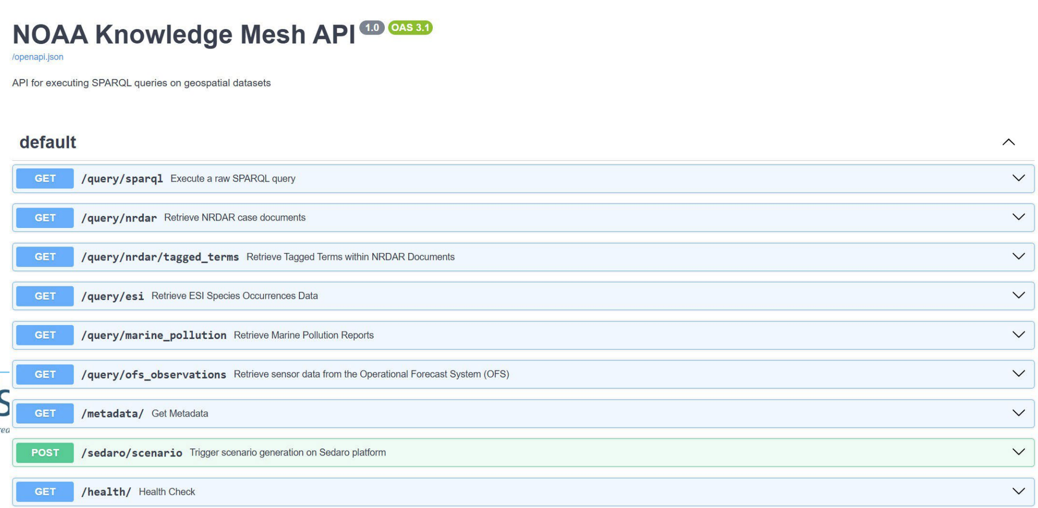


Figure 7. Pre-defined Endpoints

Key endpoints include:

- GET /query/sparql: Executes user-defined SPARQL queries on the graph database.
- GET /query/esi: Retrieves Environmental Sensitivity Index data on species occurrences and shoreline types.
- GET /query/marine_pollution: Provides access to Marine Pollution Surveillance Reports filtered by date, region, or event type.
- GET /query/ofs_observations: Returns Operational Forecast System outputs such as temperature, salinity, and current velocity.
- GET /query/nrdar and /query/nrdar/tagged_terms: Retrieve Natural Disaster and Restoration case documents and tagged entities extracted from text.
- POST /sedaro/scenario: Initiates a simulation request within the Sedaro digital twin environment.

These endpoints provide flexible and standardized access to NOAA's environmental intelligence data through the Knowledge Mesh, enabling both interactive analysis and automated integration into downstream applications.

3.6. Digital Twin Integration

One of the outstanding questions from NOAA was how digital twin teams would be able to integrate with the Knowledge Mesh. In support of this initiative, Sedaro leveraged its digital twin capabilities to develop digital twin simulations that connect with the Knowledge Mesh, allowing users to retrieve synthetic data of predicted satellite overpasses of possible oil spill locations that are observable by satellites from NOAA, the National Aeronautics and Space Administration (NASA), the European Space Agency (ESA), the Canadian Space Agency (CSA), as well as other commercial Earth orbit satellites. This predicted data can be used to inform the MPSR by identifying platforms that may have imagery of affected locations to better characterize the severity of the spill.

Additionally, the Knowledge Mesh was designed to interface with NOAA's broader digital twin initiatives, enabling dynamic simulation and forecasting capabilities based on real-world and model-generated data. Integration with the Sedaro platform provides a bridge between observed data and predictive modeling workflows. Through the **/sedaro/scenario** endpoint, users can initiate and manage digital twin scenarios directly through the API. This capability allows analysts to explore hypothetical or forecasted events such as oil spill dispersion, temperature anomalies, and flooding scenarios. Scenario metadata and outputs are automatically captured and registered within the Knowledge Mesh, ensuring full traceability and compliance with FAIR data principles.

By integrating via python APIs, it's possible for users of the Knowledge Mesh to have a continuous feedback loop between environmental observation and predictive modeling. Data from operational systems such as OFS, MPSR, and ESI can be used to drive simulations, while outputs from the digital twin can be ingested back into the Knowledge Mesh for comparative analysis and validation. This creates a powerful decision support environment where real-time observations and simulated forecasts are semantically linked, improving situational awareness for emergency managers, planners, and researchers.

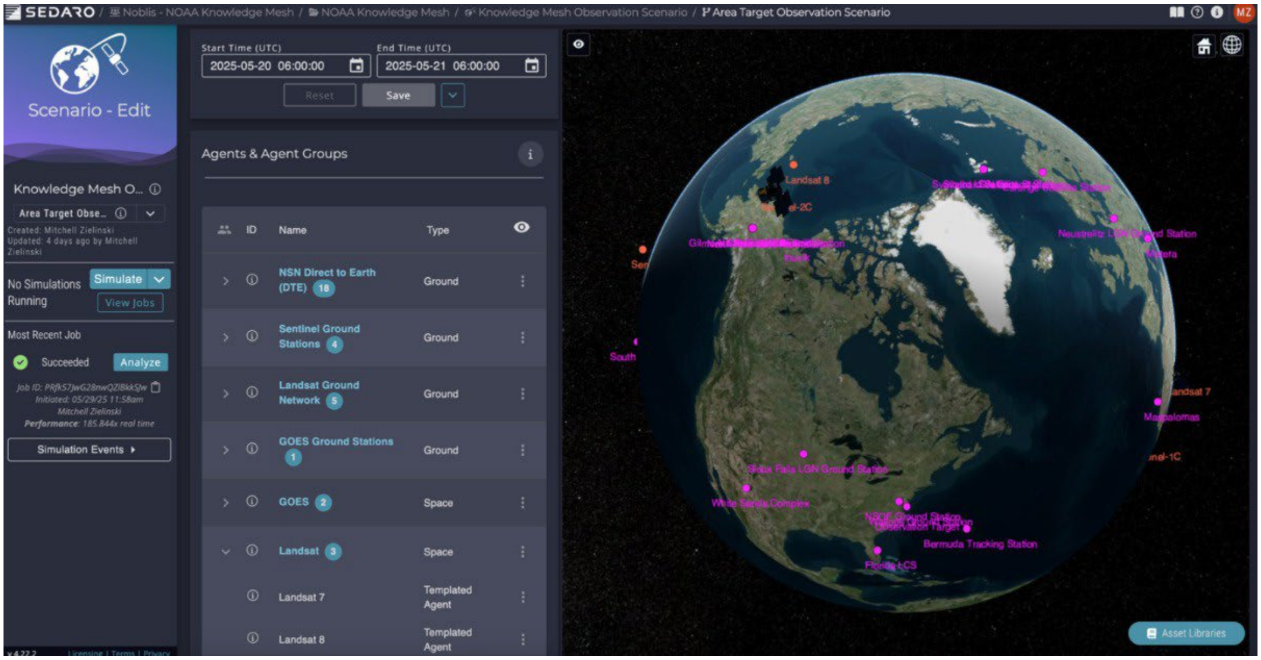


Figure 8. Sedaro Digital Twin Platform

3.7. Open Information Stewardship Service (OISS) Integration

OISS provides the conceptual and technical framework for defining how data is acquired, enriched, and published in a consistent and traceable manner. The Knowledge Mesh integrates with OISS to demonstrate how data pipelines can be deployed with transparency, reusability, and semantic governance. By leveraging OISS, Noblis was able to build a system that models each stage of the MPSR, OFS, and ESI ingestion pipelines as semantically described processes with explicit provenance.

3.7.1. OISS Conceptual Framework

OISS organizes workflows using four core components: Pattern, Template, Task Pattern, Task Template. Each element contributes to a structured, FAIR-aligned process lifecycle that governs how data moves from acquisition to publication.

A **Pattern** defines the abstract structure of an object or process, describing what the schema or data model should look like. For example, a storage pattern may define fields such as title, date, and region for an environmental report.

A **Template** represents a concrete configuration of a pattern. It includes metadata about the creator and organization, pointers to execution logic such as Lambda or Step Function Amazon Resource Names (ARNs), and contextual information about inputs and outputs. In practice, a template is a ready-to-deploy implementation of a pattern.

A **Task Pattern** describes the structure and expected behavior of a computational task. It defines inputs, outputs, and logical operations in an abstract form. For instance, an AIU task pattern may specify inputs such as year or mode and outputs such as the Uniform Resource Identifier (URI) of retrieved data.

A **Task Template** is a fully realized version of the task pattern that binds it to executable logic. This includes references to specific AWS Lambda functions or Step Functions and a corresponding UBL template that defines data field types and transformations.

3.7.2. Application in Knowledge Mesh Pipelines

Noblis implemented the OISS structure consistently across all geospatial pipelines, OFS, MPSR, and ESI, using the same layered logic. Each dataset follows a standard OISS lifecycle consisting of three major stages:

- The **Archival Information Unit** (AIU) is responsible for acquisition. It retrieves raw data from source systems such as NOAA's S3 repositories or web services and captures essential metadata describing the acquisition event.
- The **Archival Information Collection** (AIC) transforms and enriches the outputs of the AIU. This layer standardizes the data, applies relevant taxonomies and ontologies, and converts it into a semantically annotated RDF graph that can be queried through SPARQL in Neptune.
- The **Dissemination Information Package** (DIP) is a layer used for distribution or registration of results. It packages AIC outputs for downstream systems, enabling data sharing or formal registration within other FAIR-aligned repositories.

The AIU and AIC layers are defined in terms of Patterns and Templates within Neptune and implemented through containerized Lambda functions orchestrated by AWS Step Functions.

In the MPSR pipeline, for example, the AIU retrieves the latest pollution surveillance reports through a containerized Lambda function called *get_new_reports*. The AIC then processes these reports in a second container, *process_incoming_reports*, enriching them with geographic and temporal metadata and structuring the information as an RDF graph. Both stages are linked through a Step Function that orchestrates their sequence of execution.

AWS EventBridge monitors for incoming process IRIs and automatically triggers the Step Function when new data is available. Identify and Access Management (IAM) policies ensure appropriate permissions, and AWS Secrets Manager manages credentials securely. Once the process is completed, the resulting data and provenance metadata are written to Neptune as RDF triples.

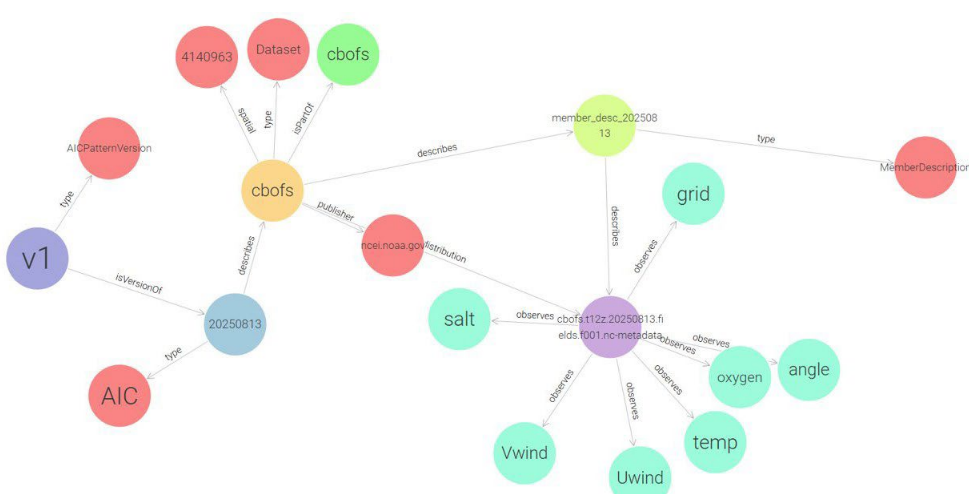


Figure 9. Ontotext GraphDB – OFS & OISS AIC Integration

3.7.3. Semantic Governance and FAIR Compliance

Before deployment, each pipeline defines its AIU and AIC Patterns and Templates through Python scripts that use the OISS SDK. These scripts register the metadata and logical structure directly into Neptune, creating a semantic model of the pipeline itself. When a process runs, an Instance and corresponding Record are generated, capturing input parameters, execution time, and success or failure status.

This system creates a transparent and semantically rich layer of governance. Every pipeline, from ingestion to publication, is described using open ontologies and linked metadata, enabling users to trace not only the data but also the logic that produced it.

By integrating OISS concepts into the AWS infrastructure, Noblis established a repeatable, FAIR-compliant workflow that combines the flexibility of serverless computing with the traceability and semantic rigor required for NOAA's Knowledge Mesh. This integration ensures that all geospatial data pipelines, OFS, MPSR, and ESI, are not only automated and scalable but also fully documented, version-controlled, and discoverable through Neptune's SPARQL endpoint.

3.7.4. Data Pipelines

The data pipelines developed for the Knowledge Mesh were constructed alongside the OISS framework to ensure that each dataset ingested was semantically described, reproducible, and governed according to FAIR principles. Each pipeline was defined through a hierarchy of OISS components, including Patterns, Templates, and Member Descriptions (MDs), which collectively capture both the structure and provenance of data throughout its lifecycle.

For each pipeline, Noblis implemented an end-to-end process that utilized the AIU and AIC layers to handle data acquisition and transformation, respectively. The AIU defined the schema and metadata requirements for the raw input data, capturing essential descriptors such as report date, image acquisition time, satellite source, and file URLs for associated shapefiles and imagery. The AIC layer transformed these inputs into semantically enriched RDF, linking observed features.

A Process Template was created to combine the AIU and AIC into a single executable workflow, governed by OISS metadata. This template registered its parameters, AWS Step Function ARN, process IRIs stored in secrets manager, and input-output relationships within the Knowledge Mesh as an RDF representation of the process itself. An AWS EventBridge rule and OISS trigger agent was configured to handle execution of the step function tied to the pipeline. Through this architecture, each OISS Member Description serves as both a semantic contract and a governance object that ensures every dataset and transformation is transparently linked to its origin, ontology, and computational context.

3.7.5. Model Context Protocol (MCP)

Model Context Protocol (MCP) standardizes tool and resource access for LLMs, streamlining adoption of AI capabilities across teams and use cases (Anthropic, 2024). The OISS MCP server was developed to assist software developers and data analysts working with NOAA's OISS API. This specialized tool provides three primary capabilities: general question answering against OISS documentation, intelligent code generation, and architecture diagram understanding.

General QA functionality employs vector retrieval through Retrieval Augmented Generation (RAG) against OISS documentation PDFs, helping developers quickly find answers to conceptual questions about OISS implementations and integrations. For code generation, the MCP server implements a custom code-chunking approach that intelligently segments Python imports, methods, and classes, while generating summaries of code chunks to enhance vector retrieval effectiveness.

The MCP server further enhances developer productivity through diagram extraction and interpretation capabilities, which identify architecture diagrams within OISS documentation and translate them into structured, natural language information using vision-enabled LLMs. This structured information is then leveraged for vector retrieval and question answering about architectural components. Importantly, as MCP is inherently interoperable, the server integrates seamlessly with GitHub Copilot and potentially other model providers, enabling developers to access these specialized OISS capabilities through their preferred development environments. This proof-of-concept application was developed on an ad-hoc basis within a few weeks yet demonstrates how AI-assisted tools can significantly improve developer efficiency when working with complex systems like OISS.

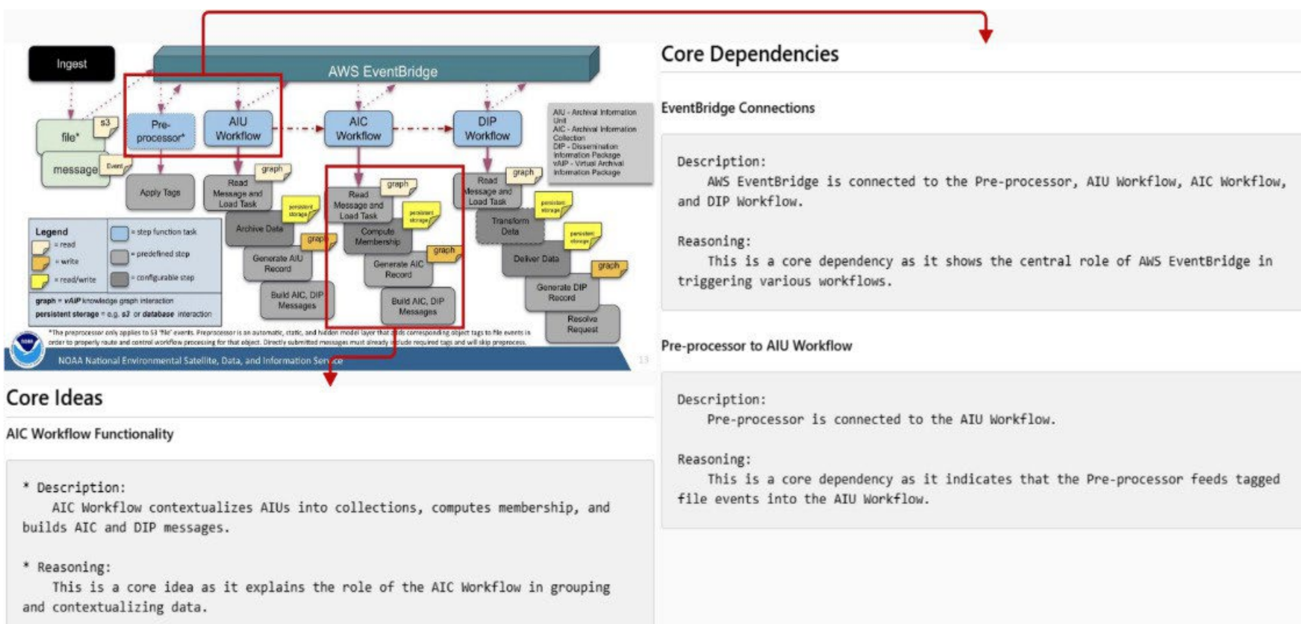


Figure 10. Example Conversion of OISS Cloud Architecture Diagram to Core Ideas & Dependencies This Diagram shows partial output of pre-processing an OISS cloud architecture diagram to core ideas and dependencies. Transitive and geometrically implicit relationships are highlighted.

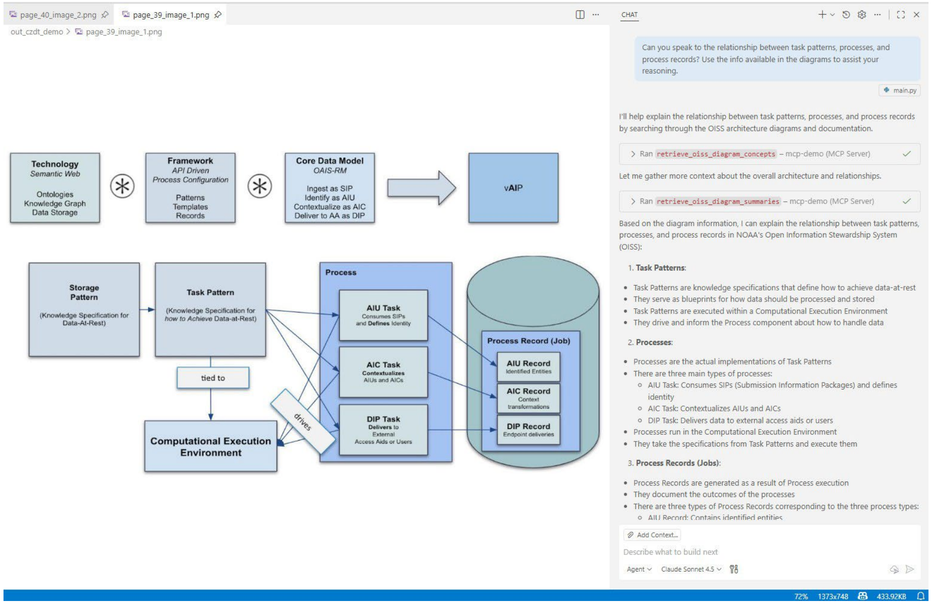


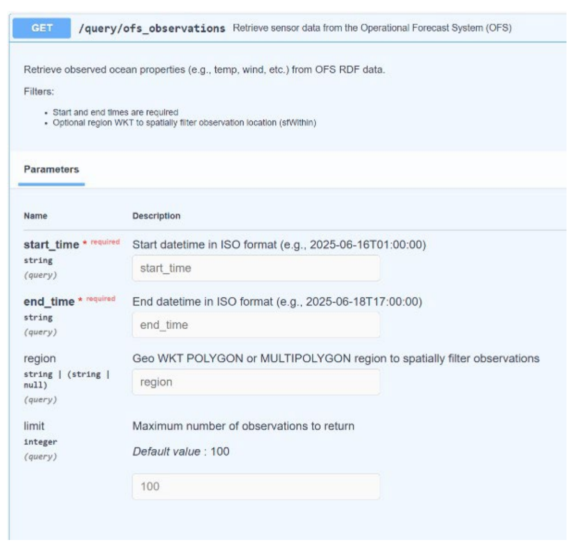
Figure 11. GitHub CoPilot Agent Response with OISS MCP Server

This Diagram shows a response from GitHub CoPilot using the OISS MCP server for responding to the query "Can you speak to the relationship between task patterns, processes, and process records? Use the info available in the diagrams to assist your reasoning." The response leverages the architecture diagrams in its response, pulling concepts extracted from the included OISS diagram to the left of the chat output.

3.8. Quality Monitoring

In support of quality monitoring for the Knowledge Mesh, Noblis provides multiple ways to validate the data and API including programmatic methods that can be automated as well as visualization for manual inspection. These methods help ensure correctness and consistency with the various data sources ingested into the knowledge mesh.

3.8.1. API



The Knowledge Mesh API implements comprehensive quality monitoring through standards-based documentation following OpenAPI 3.1, providing consistent schemas, response formats, and live test interfaces that enhance developer experience. All API inputs undergo strict validation for parameters like region, dataset, and date, ensuring that responses maintain predictable structure and traceability for reliable downstream consumption. The implementation leverages FastAPI's built-in capabilities for request validation, response type checking, and automated documentation generation, strengthening interface consistency and reliability. Robust monitoring and logging systems track API usage patterns, enabling effective debugging and valuable analytics on system

Figure 12. API Documentation for the /query/ofs_observations Endpoint

utilization. The API employs clear versioning practices that maintain backward compatibility while allowing for evolution, giving consumers confidence that their integrations will remain stable over time.

3.8.2. Data Validation

The python package noaa-cd-km provides core abstract methods for establishing data validation practices for incoming data, requiring implementation of metadata and variable validation prior to completing the conversion into RDF. The general operational flow is given below in Figure 13.

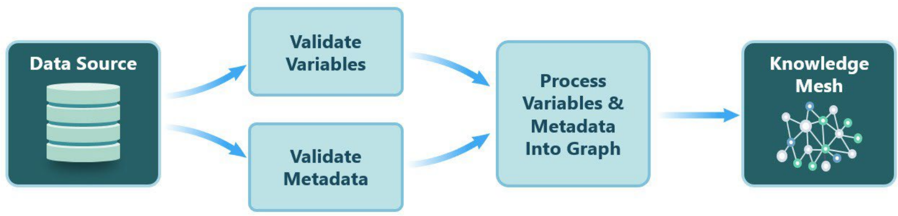


Figure 13. General Architecture for Data Ingestion and Validation

As the Knowledge Mesh scales, these data validation requirements establish good practice for efficient and robust development.

In addition to validating data within the ingestion pipelines, the codebase provides containerized SHACL validation, which enables the specification of constraints against which RDF graphs must validate. SHACL validation is provided through the Java tool Apache Jena SHACL, which demonstrated significant performance over the commonly used PySHACL python package. Each data source has an associated SHACL file which can be used to validate output graphs incrementally or perform periodic checks on the global graph. For example, a portion of the shape file for MPSR data is shown in Figure 14, where both reportDate and reportTime are required fields (minCount set to 1), while the description is optional. By containerizing the SHACL utility, this validation method can be deployed for automation and API access, such as in AWS Batch or a standalone API endpoint.

```
# Marine Pollution Report Shape
[] a sh:NodeShape ;
  sh:targetClass mpsr:MarinePollutionReport ;
  sh:property [
    sh:path dct:description ;
    sh:datatype xsd:string ;
    sh:minCount 0 ;
  ] ;
  sh:property [
    sh:path mpsr:reportDate ;
    sh:datatype xsd:date ;
    sh:minCount 1 ;
  ] ;
  sh:property [
    sh:path mpsr:reportTime ;
    sh:datatype xsd:time ;
    sh:minCount 1 ;
  ] ;
...
```

Figure 14. Example Portion of SHACL File for MPSR

3.9. Data Catalog

A comprehensive data catalog was developed and maintained in dual formats: a JSON representation enabling programmatic access and automated processing, and a Mural-based visual representation facilitating intuitive inspection of data relationships and dependencies. For visual inspection of the Knowledge Mesh, interactive graph visualization of data and process semantics was demonstrated using Ontotext GraphDB's visualization tools (Figure 15) as well as OISS' built-in process hierarchy visualization. These visualizations are valuable for developers who need to manually inspect data or OISS processes in the Knowledge Mesh, and they provide an efficient means of communicating the semantics of the Knowledge Mesh to stakeholders as the complexity of the Knowledge Mesh grows.

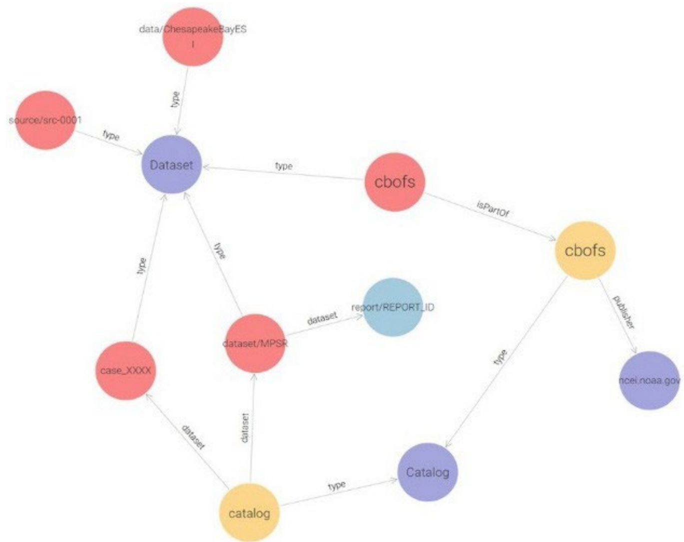


Figure 15. Visual of Data Catalog Ontotext GraphDB

3.10. R2RML Exploration

3.10.1. R2RML Definition

R2RML (RDB to RDF Mapping Language) is the W3C standard for expressing customized mappings from relational databases to RDF datasets. It enables relational data (tables, rows, columns) to be represented as RDF triples without requiring data replication. Instead, a mapping file describes how elements in the relational schema correspond to RDF classes and properties. This allows applications and knowledge graphs to query existing databases using SPARQL as though they were native RDF stores, providing virtualization capabilities and preserving data provenance.

3.10.2. Storage Options

When implementing OISS process workflows with R2RML, the storage layer becomes critical because it determines whether data is to be materialized (copied) or virtualized (accessed on demand). Options include:

- **Relational Databases (SQL Server, PostgreSQL, MariaDB, Athena)**
  Data resides in a traditional Relational Database Management System (RDBMS) or query engine. R2RML mappings enable SPARQL queries to be translated into SQL and executed directly against the database.
- **File-Based Formats (CSV, JSON, Parquet, NetCDF, Zarr)**
  Data remains in its native file format, often in S3 or other object stores. R2RML/RML mappings can be defined to convert tabular or multidimensional data into RDF on demand. Parquet and Zarr are particularly suitable for large vector or array datasets.
- **Streaming and Messaging (Kafka Topics)**

Data is accessed through message queues or streaming feeds. Virtualization through RML allows events to be exposed as RDF triples without permanent storage in OISS.

- **Materialized RDF Stores (e.g., Neptune, Ontotext GraphDB)**
  Data is transformed once into RDF triples and stored persistently in a triplestore. This approach favors fast query execution but requires duplication and ongoing synchronization.

The choice depends on performance, scale, and whether real-time access (virtualization) or fast repeat querying (materialization) is prioritized.

3.10.3. Member Descriptions Definition and Use

Within OISS, **Member Descriptions** are pattern-defined, context-isolated RDF graphs that encapsulate metadata, provenance, mappings, and structural information for a dataset or process. Each MD links fields (object, property, literal, or statement) to the AIC, ensuring that datasets and transformations are semantically described and interoperable.

Key roles of Member Descriptions in this effort:

- **Encapsulation of Virtualization Logic**: MDs will store references to R2RML/RML mapping files, enabling datasets to be queried virtually without replication.
- **Storage Flexibility**: MDs define whether data is external (in S3, SQL, or Kafka) or materialized in a triple store, supporting both strategies transparently.
- **Provenance and Semantics**: MDs provide provenance narratives that capture how mappings were applied, whether data was accessed virtually or materialized, and what ontologies (DCAT, SOSA, ENVO) were used.
- **API Integration**: MDs can serve as the semantic contract between data and APIs. An API stub can dynamically read an MD, locate the mapping file, and expose the dataset as an access-oriented service.

In short, **R2RML provides the mapping language, storage options define where and how data resides, and Member Descriptions serve as the semantic layer that binds the two together in the OISS framework**.

3.10.4. R2RML Outlined Approach

NOAA has identified the value of implementing and R2RML approach with the OISS framework. Although this is out of scope of the NOAA BAA, Noblis has outlined an approach and mock example of how R2R can be utilized and its benefits. The approach extends OISS Member Descriptions, so they include not only provenance and metadata but also mapping logic using R2RML or RML. This allows datasets to remain in their original storage locations (S3, SQL, Parquet, Kafka) while being accessed virtually instead of fully materialized. AIC templates and Member Descriptions serve as the semantic contract, storing DCAT metadata and mapping files that define how external data is exposed as RDF. An AWS-based API layer with Lambda, API Gateway, and the python package Morph-KGC (Arenas-Guerrero, 2024; Corcho, 2024) uses these mappings to let users query and integrate external datasets as part of the OISS knowledge mesh without duplication.

In future phases, the Knowledge Mesh can serve as the semantic backbone for federated digital twin ecosystems, enabling interoperability across multiple simulation platforms beyond Sedaro. NOAA's long-term digital twin strategy involves connecting operational models, hazard simulators, atmospheric models, and satellite tasking simulators into an integrated decision-support network.

By grounding each dataset and simulation input/output in common ontologies such as SSN/SOSA for observations, ENVO for environmental features, and DCAT/PROV for metadata, these digital twins can exchange information in a consistent, ontology-driven manner. This will allow analysts to chain models together, such as using OFS forecasts to trigger Sedaro satellite visibility predictions or routing ESI habitat layers into impact-assessment simulations.

Each digital twin remains independently managed while semantically linked through the Knowledge Mesh. This avoids monolithic architecture and instead enables a mesh of modular, discoverable simulation capabilities. As NOAA and NASA expand their digital twin portfolios, this mesh-based integration ensures that synthetic outputs—like orbital intersects, forecasted spill trajectories, or simulated plume dispersion, can be compared, validated, or combined with real-world observations in a uniform semantic context.

Additionally, integrating digital twin outputs through OISS Member Descriptions (MD) offers a clear path toward materialized and virtualized simulation products. Some simulation results may be stored and materialized for long-term reuse (e.g., weekly OFS analyses), while others, such as Sedaro scenario runs, can remain virtualized and generated on demand. This aligns with NOAA's interest in "simulation-as-data" and ensures that predictive products can be surfaced to analysts without incurring unnecessary storage costs. Together, these strategies build toward a future where digital twin systems function as live, federated extensions of the Knowledge Mesh.

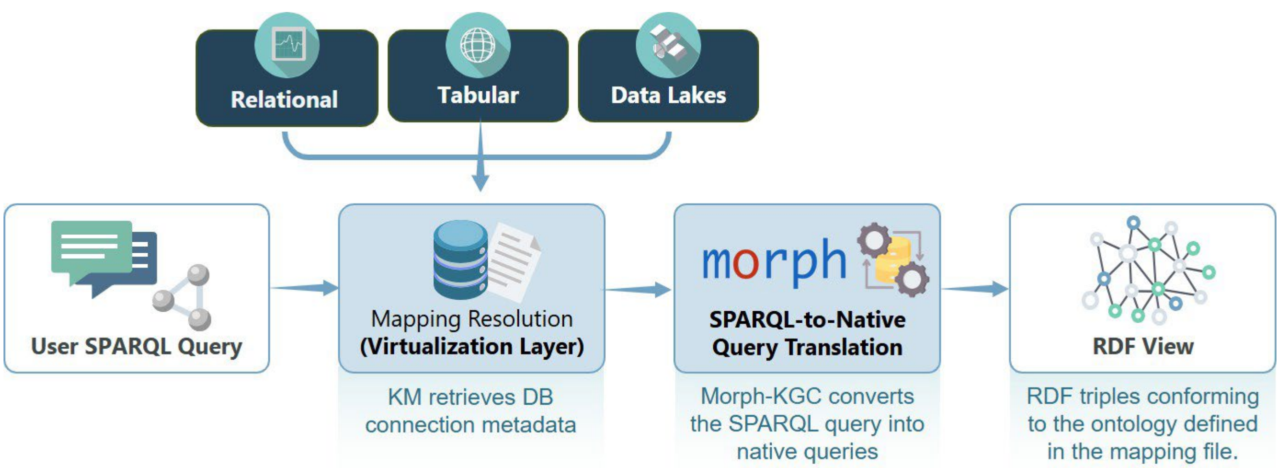


Figure 16. AIC Patterns and MDs

This Diagram shows the workflow for leveraging mappings to translate SPARQL queries to the native query language, enhancing access to external data sources

1. Background and Current Progress

OISS MDs are the foundational units within the OISS framework that encapsulate metadata, provenance, and mappings to data and processes. By working with AICs, templates, and tasks, we have established a baseline for modeling NOAA datasets in a structured, interoperable way. Member descriptions provide semantic scaffolding for connecting NOAA's diverse data formats (NetCDF, tabular, geospatial) to higher-level knowledge mesh architectures.

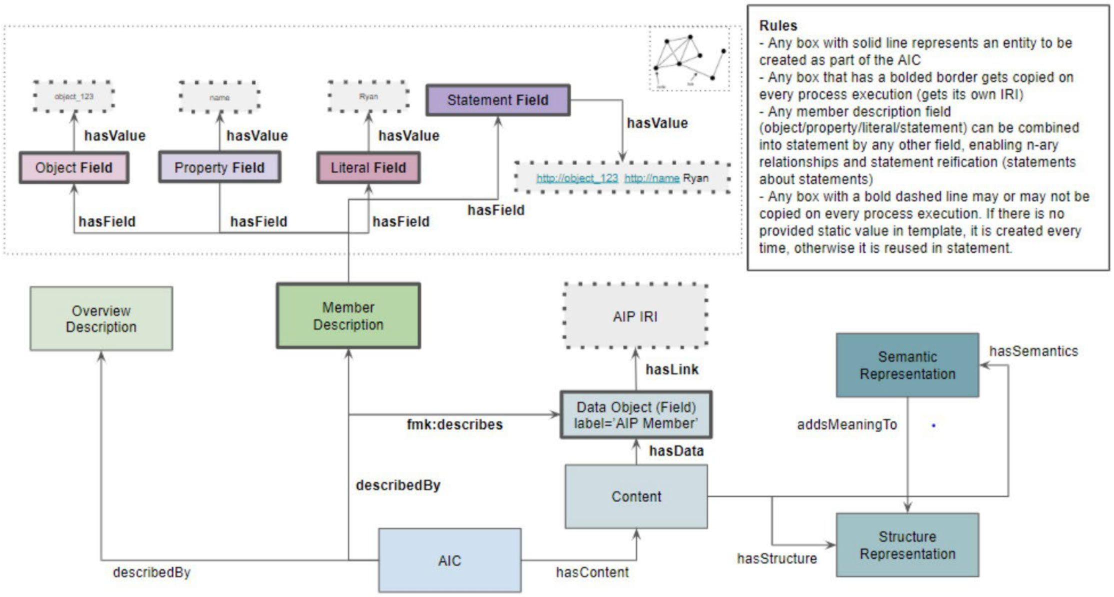


Figure 17. AIC Patterns and MDs

This Diagram shows the implementable design of AIC patterns, specifically calling out the Membership Description(s), its (their) containment, content, and links to AIP IRIs contained in a unique Data Object that is automatically added at runtime to the Content Object of the AIC.

The process workflows created so far demonstrate how AIU, AIC, and DIP layers interact to ingest, describe, and deliver data. However, a critical gap has emerged: the need to explicitly model materialization and virtualization strategies for data access and integration. NOAA has highlighted the importance of bridging this gap using R2RML/RML mapping strategies, API stubs, and virtualization approaches to provide an access-oriented view of data without necessarily materializing it in OISS storage.

2. Identified Gap

- Current MDs define provenance and mappings but do not yet account for virtualization strategies (i.e., pointing to external sources without replication).
- R2RML mappings and related tools (Morph-KGC, Ontotext Refine, RDF2RML) provide a pathway for dynamic translation between relational/tabular data and RDF-based OISS descriptions.
- There is no consistent API-level demonstration of how AIC MDs can store mapping headers (e.g., R2RML mappings, DCAT records) and expose them as service endpoints.

3. Proposed Approach

Step 1: Extend OISS Member Descriptions

- Define an **AIC template** that includes slots for:
  - dcat:Dataset references (location of data, storage type such as Parquet, Kafka, SQL).
  - r2rml:Mapping file references stored in MD headers.

  - Provenance narratives documenting whether the data has been materialized (copied) or virtualized (queried live).

- Implement this as a **storage template + task template** pair to ensure repeatable deployment.

Step 2: Develop R2RML Integration

- Create a pilot mapping for one NOAA dataset (e.g., Chesapeake Bay Operational Forecast System [CBOFS] stations file) utilizing Morph-KGC. The mapping reference will reside in the MD header and define how to expose the dataset as RDF/OS language. Morph-KGC is used for dynamic RDF graph generation from tabular and RDB sources. The mapping file is stored within the cloud environment

Step 3: API Stub and Virtualization Layer

- A **lightweight API stub** in FastAPI is defined that reads the MD, retrieves the mapping, and dynamically exposes SPARQL queries via Morph-KGC translation.
- Only the MD, mapping file, and DCAT metadata are stored in Neptune; the raw dataset remains external (e.g., in S3 or web resource).
- **API Gateway + Lambda** was used to expose a read-only virtualization endpoint that demonstrates access-oriented retrieval without duplicating the data.

Step 4: AWS Prototype Architecture

- **PostgresSQL**: stores a mock tabular dataset.
- **Lambda (FastAPI in Docker)**:
  - Reads AIC MDs.
  - Loads R2RML mapping (Morph-KGC).
  - Executes queries against external sources (RDB).
- **API Gateway**: provides secure REST/SPARQL endpoint for accessing virtualized data.
- **Neptune**: stores MDs, templates, and DCAT metadata.

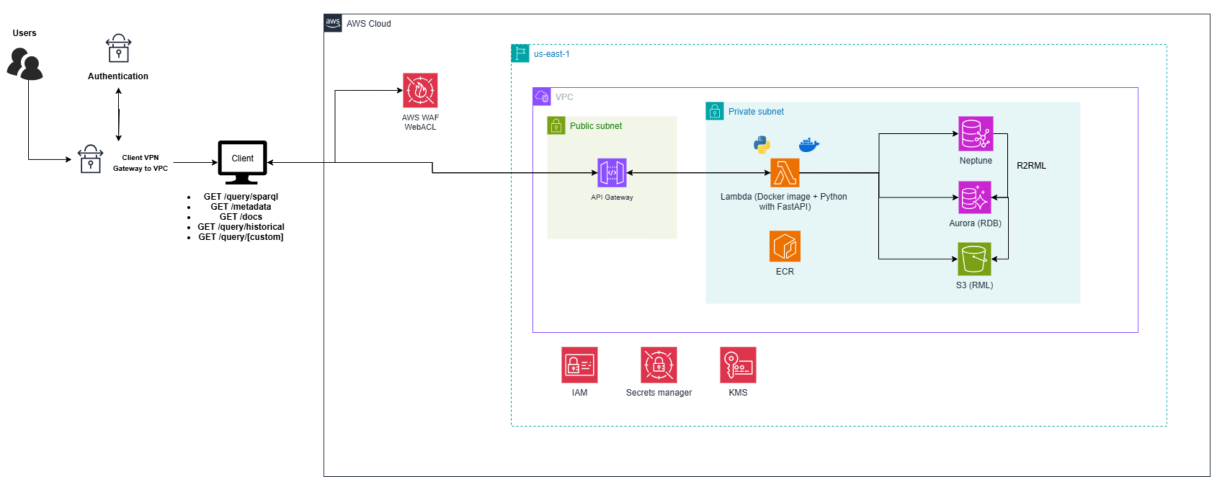


Figure 18. AWS Architecture

3.10.5. Process Flow

- **User Input**: A **SPARQL query** that is sent to the OISS endpoint (the query is executed against the Member Description).
- **System Action**:
  1. The Member Description provides metadata and points to the relevant R2RML mapping file.
  2. The mapping engine (e.g., Morph-KGC) uses that mapping file to translate the SPARQL query into the native query language of the external data source (SQL, Parquet scan, API call, etc.).
  3. The external data source is queried live.
- **System Output**: The engine returns the results as **RDF triples** that match the ontology defined in the Member Description, so the user sees a consistent RDF view no matter where the data was stored.

Noblis successfully implemented this approach locally against a mock PostgresSQL table within a container. By extending OISS member descriptions to include R2RML mappings, DCAT metadata, and explicit materialization/virtualization strategies, NOAA can close the identified gap and demonstrate a fully access oriented approach to knowledge mesh integration. The proposed AWS prototype architecture, paired with Morph-KGC and lightweight API stubs, provides a scalable and standards-based path forward. This work will generate study evidence, improve interoperability, and set a foundation for future digital replica and R2RML initiatives.

4. Findings and Insights

The development and deployment of the NOAA Knowledge Mesh demonstrates that semantic data integration across multiple, disparate data domains is both technically feasible and operationally valuable for coastal monitoring, emergency management, and environmental research. The implementation establishes a repeatable and scalable methodology for transforming, linking, and exposing NOAA datasets through a unified, ontology-driven architecture aligned with the FAIR data principles. Importantly, by combining containerized AWS-based workflows, semantic web standards, and dynamic API integration, the project delivered a flexible foundation that can scale to new data domains and emerging NOAA mission requirements.

4.1. Semantic Integration and Data Pipeline Insights

The Knowledge Mesh successfully linked heterogeneous datasets from four major NOAA sources, the OFS, MPSR, ESI, and NRDAR case library, into a semantically enriched graph. Through this integration, both structured (geospatial, time-series) and unstructured (textual, report-based) data were harmonized using open ontologies such as DCAT, GeoSPARQL, and ENVO. Each pipeline aligned to the OISS framework, which provided the standardized lifecycle management of data through AIU, AIC, and DIP stages.

For geospatial pipelines such as MPSR and ESI, RDF transformations were performed directly from raw geospatial files using automated scripts that generated standardized geometry encodings (WKT) and metadata annotations. Custom properties such as hasSourceLocation and hasAreaLocation extended GeoSPARQL to provide fine-grained spatial relationships between observed pollution events and impacted regions. The use of dedicated ontologies allowed consistent measurement encoding for variables such as spill area, oil type, and forecast parameters, ensuring machine-readable interoperability across data models.

Pipeline orchestration using AWS Step Functions and EventBridge provided full automation for ingestion, validation, and publication. Each pipeline included SHACL-based validation to ensure that geometries, coordinate systems, and units conformed to schema definitions before integration into the Knowledge Mesh stored in Amazon Neptune. This modular, serverless approach proved both cost-efficient and resilient, confirming that future NOAA workflows can scale horizontally across new datasets and geographies.

4.2. OISS Framework Implementation

OISS integration demonstrated the strength of semantically governed data management. By encoding the structure and logic of every pipeline as Patterns and Templates within the OISS schema, Noblis ensured that each ingestion process was transparent, discoverable, and fully documented as part of the graph itself. Each pipeline instantiation generated its own OISS record, containing detailed metadata on input parameters, timestamps, and provenance. This provided NOAA with the capability to track not only what data was produced, but also how and when it was produced, satisfying provenance and reproducibility requirements essential to FAIR data stewardship.

Through containerized Lambda executions and Neptune-stored metadata, OISS enabled pipelines to function as first-class semantic objects, linking processing logic directly to data outputs. This approach laid the groundwork for NOAA to expand OISS usage beyond the initial four datasets, potentially automating metadata capture and quality control across future ingestion pipelines for marine debris, fisheries, and permitting datasets.

4.3. MCP Developer Enablement and AI Integration

The proof-of-concept MCP server for OISS development provides value as a time-saving tool for developers working with the OISS documentation and implementation and can be extended to additional domains. Developers can rapidly extract relevant information from both textual documentation as well as architecture diagrams, removing the burden of manually searching the documentation and images to answer their questions. While automated code generation would require further development, the initial research conducted resulted in novel chunking of the python code that establishes a valuable, extensible foundation for code generation. By combining RAG-based vector retrieval against the OISS documentation, custom conversion of architecture diagrams into contextualized units of information, and intelligent python code-chunking approaches, the MCP server enables developers to quickly access context-specific information relevant to their immediate tasks.

4.4. R2RML Virtualization and Knowledge Mesh Extensibility

The proof-of-concept R2RML implementation illustrated the potential of query virtualization for NOAA's data ecosystem. Using Morph-KGC, Noblis demonstrated how relational and tabular data can be dynamically exposed as RDF without being copied or pre-transformed, using mapping files that define logical correspondences between source fields and ontology terms.

By storing R2RML mapping references within OISS Member Descriptions, datasets residing in SQL, S3, or Parquet can be queried as if they were natively part of the Knowledge Mesh.

This approach decouples storage from access, providing a flexible and cost-efficient mechanism to connect disparate systems across NOAA without redundant data replication. It establishes a model where the Knowledge Mesh can act as a virtual integration layer, capable of querying live data streams while maintaining provenance and semantic consistency. This strategy aligns closely with NOAA's vision for distributed, FAIR-aligned data ecosystems and offers a pathway toward dynamic, on-demand data federation.

4.5. Cloud Architecture and API Insights

From a systems perspective, the AWS cloud architecture provided a robust foundation for operationalizing the Knowledge Mesh. Core components included API Gateway for routing and authentication, Lambda for scalable compute, ECR for container storage, and Neptune for RDF graph storage and persistence. Together, these services supported a serverless architecture that minimizes operational overhead while maintaining scalability and security.

Each Knowledge Mesh endpoint was implemented in FastAPI with OpenAPI documentation and configurable SPARQL templates, enabling developers and analysts to execute parameterized queries over datasets such as OFS, MPSR, ESI, and NRDAR. The modular endpoint design supports extension to new datasets with minimal configuration, ensuring that future NOAA users can integrate new sources or models into the API layer with consistent governance and documentation.

4.6. Digital Twins

The integration with Sedaro's digital twin demonstrates the opportunity for digital twin teams to interact with the Knowledge Mesh for increased insights by teams leveraging those outputs as well as workflow improvements. For the Coastal Zone Digital Twin (CZ-DT) team a key value add of this capability is eliminating the dependence on external vendors or sources for capturing the time to observation of a potential spill. By leveraging Python API access, teams can seamlessly trigger simulations and retrieve data through a unified interface, creating substantial time savings in operational workflows. This successful implementation suggests the Knowledge Mesh architecture can be extended to incorporate multiple digital twins such as the Earth Observations Digital Twin, creating a more comprehensive ecosystem for environmental monitoring and prediction capabilities.

5. Future Work and Scalability

The prototype developed under this effort establishes a strong foundation for

NOAA's long-term vision of interoperable, AI-ready environmental data management. Building on this framework, several areas of future work have been identified to enhance scalability, developer usability, and intelligent data interaction within the Knowledge Mesh (see Table 2).

Table 2. Categories of Future Work

Geographic & Thematic Expansion	Expand the Knowledge Mesh beyond Chesapeake Bay and coastal regions.
AI Search Interface & Knowledge Mesh Operations	The Knowledge Mesh provides a reliable data foundation for AI-enabled search that enhances accessibility across users. For example, generating complex SPARQL queries through natural language. AI models can also help with the construction and operation of the Knowledge Mesh (i.e., automated ontology alignment).
Additional Digital Twin Simulation Capabilities	Expand the digital twin simulation capabilities to additional satellites, missions, and instrumentation.
Productionize MCP Server & Apply to UBL Templates	Operationalize the MCP server as an interactive developer toolkit for real-time documentation assistance and development assistance for OISS templates and workflows. Embedding MCP's diagram extraction and structured representation capabilities as a reusable UBL within OISS would further standardize pipeline documentation and automation.
R2RML Integration with Support of Additional Data Formats	Extend the PoC to additional data formats and operationalize within Knowledge Mesh, enabling access through API endpoints. Additionally, provide semi-automated creation of mapping files to reduce time to integration.
OISS Expansion	Expand OISS AIC and AIU templates to accommodate emerging environmental datasets and analytical models.

5.1. Geographic & Thematic Expansion

While the Knowledge Mesh focused on Coastline disaster in the Chesapeake Bay, additional scalability opportunities include expanding the geographic and thematic coverage of the Knowledge Mesh beyond the Chesapeake Bay region. The OISS-based pipelines already demonstrated can be reused and adapted for these future datasets, ensuring consistent semantic modeling and data provenance. As additional data types are incorporated, performance optimizations such as parallel pipeline execution, intelligent caching, and graph partitioning should be explored to support larger-scale deployments.

Examples of potential regions to consider for continued development include the Gulf of America, Gulf Coast of Florida, and the Pacific Northwest. Potential themes include oil spill monitoring, hurricane impact assessments such as debris monitoring, socio-economic effects of these disasters for vulnerable coastline communities, ecosystem monitoring such as coral reef health and fishery monitoring, and natural phenomena forecasting and monitoring such as monitoring and forecasting red tides.

The geospatial and natural language datasets used in this work have coverage over the Gulf of America and the Gulf Coast of Florida, making this region a good choice for incremental expansion of the Knowledge Mesh within the theme of oil spill monitoring, planning, and response. Additional datasets across these themes include but are not limited to:

- Ocean acidity levels from the Pacific Marine Environmental Laboratory (PMEL) Carbon Program data portals data portal (NOAA PMEL Carbon Group, 2025)
- U.S. EPA Disaster Debris Recovery Tool (United States Environmental Protection Agency, 2025) for incorporating debris cleanup facilities in hurricane response planning
- NOAA's NCCOS harmful algal bloom forecasts for Florida, Texas, and the Pacific Northwest (NOAA NCCOS, 2025)
- CDC/ATSDR SVI databases (Centers for Disease Control and Prevention, 2024) for incorporating socio-economic factors in disaster response planning

When incorporating new data sources and regions, a similar development approach of user-persona value identification through interaction with SMEs, identification of relevant taxonomies and ontologies, identification of gaps that require custom taxonomies and ontologies, and the construction of pipelines that convert the incoming data into RDF-compatible graphs for the graph database.

5.2. AI Search and Knowledge Mesh Operations

The Knowledge Mesh's data standardization provides optimal conditions as a data foundation to AI applications. As part of future Knowledge Mesh exploitation, LLMs can integrate with the Knowledge Mesh as a natural language interface that takes natural language queries from users and assists in SPARQL generation for complex queries. Utilizing LLMs for text-to-SPARQL can leverage open models and datasets for fine-tuning to the Knowledge Mesh. At the time of writing, several datasets exist for general fine-tuning that could be adapted to dataset specific to the Knowledge Mesh. Examples include LC-QuAD (Trivedi, 2017), QALD-9-plus (Perevalov, 2022) and Instruct-to-SPARQL (Mehdi Ben Amor, 2025). Text-to-SPARQL is an active area of research and contemporary models should be assessed at the time of implementation to determine if the costs that are typically associated with fine-tuning present compelling value. This should be done both against the Knowledge Mesh as well as open benchmarks such as the Spider4SPARQL benchmark (Kosten, 2023) to account for data domain and user growth.

Additionally, LLMs can assist developers in data ingestion operations such as generating schema recommendations during ontology development, automated ontology alignment, and interacting with MCP servers for extending existing OISS UBL templates. This would open pathways for more intuitive and intelligent interaction with NOAA's data ecosystem, improving usability for both technical and non-technical users.

5.3. Digital Twin Expansion

In consultation with Sedaro (Sedaro, 2025) and NOAA, Noblis recommends the following primary areas for future development of combining digital twins with the Knowledge Mesh: expansion of instrumentation and environmental conditions modeling and the inclusion of additional satellites. Additionally, in a federated context, Noblis sees the opportunity for digital twin teams leveraging the Knowledge

Mesh API both for accessing simulation data from other digital twins and other modeling input/output data included in the Knowledge Mesh.

5.3.1. Modeling

The following items were identified in consultation with Sedaro's digital twin team:

- **Modeling Individual Instruments:** Currently, all Sedaro models include instrument names; however, modeling individual instruments on a satellite (for example, a Landsat) would require a significant step up in model fidelity and development efforts.
- **Modeling Scanning Patterns:** While identifying which kind of imaging mode a given satellite would use is simple, accurately modeling scanning patterns would require in-depth research on the satellites in question, as well as significant model development time.
- **Imaging Mode:** Refers to the configuration or operational setting of a satellite's sensor system during data collection. It determines how the sensor captures imagery, including factors such as resolution, swath width, number of spectral bands, and viewing geometry (e.g., angle). This information could be useful if agents have the capability to model scanning patterns as described in the item above.
- **Imaging Angle:** The Sedaro team believes that it is possible to model imaging angles based on the existing data about the satellites used in this project; however, this would require a moderate amount of script development work.
- **Satellite Revisit Frequency:** The Sedaro team is currently outputting predicted satellite overpasses for the duration of the simulation but likely cannot determine satellite revisit rate without constructing another scenario focused on long-term prediction. Note: All the satellites used in this project are either geostationary or sun-synchronous, so the revisit frequency would be either "always" or "once per day at approximately the same time."
- **Cloud Cover Percentage:** An estimate of how much of the imaged area is obscured by clouds, expressed as a percentage (ex: 0% = clear sky, no cloud obstruction; 100% = fully cloud-covered image). High cloud cover reduces the usability of optical satellite imagery.

5.3.2. Missions

On May 22,2025, Noblis and Sedaro met with members of NOAA's CZ-DT Community of Practice and identified the following satellites and missions that could be incorporated in future work:

- **ASA Plankton, Aerosol, Cloud, Ocean Ecosystem (PACE):** NASA's PACE is an Earth-observing mission that launched in February 2024. It studies the distribution of phytoplankton in the ocean, aerosol particles and clouds in the atmosphere, and their impact on Earth's climate. The satellite orbits at an altitude of 676.5 km in a sun-synchronous polar orbit.
- **Suomi National Polar-orbiting Partnership (NPP):** NOAA's NNP is a key satellite in the Joint Polar Satellite System (JPSS). It was launched on October 28, 2011, and its primary mission is to collect and distribute remotely sensed data for weather and climate research.
- **NASA-ISRO Synthetic Aperture Radar (NISAR):** NISAR is a joint mission between NASA and the Indian Space Research Organization (ISRO) to map Earth's surface with radar. It is designed to measure changes in the planet's surface, including movements as small as centimeters, and will scan nearly all of

Earth's land and ice surfaces twice every 12 days. NISAR data will be free and publicly available once commissioned (estimated).

- **Surface Biology and Geology (SBG):** The SBG mission, a NASA Earth-observing initiative, aims to study the Earth's surface, including its biology, geology, and interactions. It involves two main satellite concepts: SBG Thermal Infrared (TIR) and SBG Visible and Short-Wave Infrared (VSWIR). These satellites will use advanced imaging technology to provide global, high-resolution data for a range of scientific and application needs.

- **Geosynchronous Littoral Imaging and Monitoring Radiometer (GLIMR):** The GLIMR is a NASA-funded, space-based instrument that will observe and monitor ocean biology, chemistry, and ecology in the Gulf of Mexico, the southeastern U.S. coastline, and the Amazon River plume. GLIMR will be uniquely positioned in a geosynchronous orbit, allowing for frequent and high-resolution observations of these

5.4. MCP Server

The MCP server developed during this project serves as an important step toward making OISS more accessible for developers, and operationalizing the tool for broad usage across NOAA, Noblis identifies the following areas for future work.

- **Interactive Developer Toolkit:** The OISS MCP server can be extended to provide real-time documentation assistance both on OISS and NOAA's broader software repositories. In following this path, the automatic code generation should be enhanced. As the MCP server was not an objective of the Knowledge Mesh project, the code generation capabilities remain in initial, functional development and require further testing.
- **Additional Platform Integration:** The MCP server was demoed under GitHub CoPilot, but other platforms such as additional integrated development environments (IDEs) would streamline onboarding and adoption of OISS across NOAA teams and partner organizations**.**
- **UBL Templates:** In operationalizing the MCP server, semantic integration with the Knowledge Mesh through OISS' UBL would aid discoverability, documentation, and use by other teams.

5.5. R2RML

A major opportunity for future enhancement lies in expanding the use of R2RML and related virtualization technologies across NOAA data systems. The current prototype shows how mapping files can translate relational or tabular data into RDF without materializing the underlying data in the knowledge mesh. However, creating these mappings still requires substantial manual effort. Developers must understand the structure of each source system, interpret how fields relate to environmental concepts, and manually select appropriate ontologies and vocabularies. As NOAA scales to include datasets from new systems, missions, and scientific models, this manual lift will become a bottleneck. Expanding R2RML support therefore requires investment in automation, code generation tools, and streamlined template-driven workflows that reduce the burden on developers.

Future work should explore the use of advanced language models and AI assisted schema reasoning to automatically generate candidate mapping files. These models can analyze dataset metadata, read database schemas, and cross reference known ontologies such as DCAT, SOSA, ENVO, QUDT, and SKOS to recommend mappings that align with NOAA's semantic standards. Automated tools can also identify reusable patterns across similar datasets, dramatically reducing the need to build mappings from scratch. These capabilities can be integrated directly into OISS templates so that mapping generation becomes a guided process within the pipeline creation workflow. Complementary validation mechanisms using SHACL or OWL reasoning can ensure that automatically produced mappings remain consistent with NOAA's ontology governance requirements.

Over time, this approach enables a library of reusable mapping components that reduce the learning curve for teams adopting semantic workflows.

Additional research should expand virtualization capabilities beyond traditional relational databases. Many NOAA datasets exist in scientific storage formats such as NetCDF, Parquet, Zarr, Cloud Optimized Geographic Tagged Image File Format (GeoTIFFs), or API driven services like ERDDAP and THREDDS. Future R2RML integration should focus on virtualizing these sources so that users can query multi-dimensional model outputs, gridded time series, satellite products, and forecast fields directly through SPARQL without duplicating the files into a triplestore. This work can extend to federated querying, allowing the knowledge mesh to pull live data from distributed systems while preserving provenance and performance. These advancements position NOAA to build a truly distributed, access-oriented architecture where the knowledge mesh can describe, link, and query any data source regardless of its physical location or format. Together, automated mapping generation, expanded format support, and federated virtualization will allow NOAA to scale the knowledge mesh to new missions, new domains, and new scientific workflows with minimal friction for developers and analysts.

5.6. OISS

Several enhancements to the OISS framework would benefit future scaling of the Knowledge Mesh. Further work should expand the range of AIC and AIU templates to accommodate emerging environmental datasets and analytical models. Integrating advanced validation workflows and automated FAIR compliance checks directly into OISS will help NOAA and its partners deploy standardized pipelines faster and with greater confidence. The extension of the DIP layer to register published outputs and downstream consumption events would also enhance traceability and transparency across agency-level data sharing.

Future development should focus on improving automation within OISS to reduce the overhead associated with designing and deploying new data pipelines. AI driven code generation tools, potentially integrated into the MCP environment, could guide developers through the creation of templates, task definitions, and semantic metadata without requiring deep familiarity with RDF or ontology engineering. Automated discovery of dataset structures, provenance capture, and lineage tracking would also streamline onboarding of new datasets into the OISS ecosystem. These improvements will help ensure that environmental intelligence workflows remain consistent across NOAA programs while significantly reducing manual configuration time.

Another promising area for expansion involves broadening OISS interoperability with external data systems, cloud services, and modeling workflows. For example, OISS templates could be extended to describe complex streaming pipelines, multi-step modeling workflows, digital twin simulations, or large-scale geospatial processing tasks. Integrating OISS with NOAA's cloud and high-performance computing (HPC) environments would allow the framework to orchestrate more computation intensive processes while maintaining semantic traceability across all inputs and outputs. Over time this will enable OISS to function not only as a metadata and workflow governance system but as a central coordination layer connecting NOAA’s environmental models, observational pipelines, and decision support tools into a unified, semantically governed ecosystem.

6. References

Anthropic. (2024, November 25). *Introducing the Model Context Protocol*. Retrieved 12 3, 2025, from www.anthropic.com: https://www.anthropic.com/news/model-context-protocol

Apache Software Foundation. (2021). Apache Jena SHACL. Retrieved from

https://jena.apache.org/ Arenas-Guerrero, J. a.-F. (2024). Morph-KGC: Scalable

knowledge graph materialization with mapping partitions. *Semantic Web, 15*, 1-20. doi:10.3233/SW-223135

Brickley, D. &. (2014). *FOAF Vocabulary Specification*. Retrieved from W3C FOAF Project: http://xmlns.com/foaf/spec/

Buttigieg, P. L. (2013). *ENVO: The Environment Ontology*. Retrieved from OBO Foundry : https://environmentontology.org/

Centers for Disease Control and Prevention. (2024, 12 16). *SVI Data & Documentation Download*. Retrieved from ATSDR Place and Health - Geospatial Researchh, Analysis, and Services Program (GRASP): https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html

CF Conventions Committee. (2023). *CF Metadata Conventions (v1.11)*. Retrieved from CF Conventions: https://cfconventions.org

Corcho, J. A.-G.-A.-D.-F. (2024). An RML-FNML module for Python user-defined functions in Morph-KGC. *SoftwareX, 26*, 101709. doi:10.1016/j.softx.2024.101709

Darwin Core Task Group (TDWG). (2021). *Darwin Core (DwC) — Biodiversity data standard*. Retrieved from TDWG: https://dwc.tdwg.org

DIVER. (2020). *Web Application: Data Integration Visualization Exploration and Reporting Application, National Oceanic and Atmospheric Administration*. Retrieved 12 8, 2025, from www.diver.orr.noaa.gov: https://www.diver.orr.noaa.gov/

Dublin Core Metadata Initiative. (2020). *Dublin Core Terms (DCTERMS)*. Retrieved from Dublin Core Metadata Initiative: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Kosten, C. a.-M. (2023). Spider4SPARQL: A Complex Benchmark for Evaluating Knowledge Graph Question Answering Systems. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 5272-5281). IEEE.

Mehdi Ben Amor, A. S.-Z. (2025). Instruct-to-SPARQL: A text-to-SPARQL dataset for training SPARQL Agents. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 390-395). New York, NY, USA: Association for Computing Machinery.

National Center for Biotechnology Information. (2023). *NCBI Taxonomy database*.

Retrieved from National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/taxonomy

National Oceanic and Atmospheric Administration . (2024). *Open Information Stewardship Service (OISS)*. Retrieved from National Oceanic and Atmospheric Administration: https://www.noaa.gov

NOAA Center for Operational Oceanographic Products and Services. (2025, 12 8). *Operational Forecast System*. Retrieved 12 8, 2025, from https://tidesandcurrents.noaa.gov/: https://tidesandcurrents.noaa.gov/models.html

NOAA GOA-ON. (2025). *GOA-ON*. Retrieved 12 8, 2025, from NOAA PMEL Carbon Program: https://www.pmel.noaa.gov/co2/story/GOA-ON

NOAA NCCOS. (2025). *Harmful Algal Bloom Forecasting*. Retrieved 12 8, 2025, from National Centers for Coastal Ocean Science: https://coastalscience.noaa.gov/science-areas/habs/hab-forecasts/

NOAA Office of Response and Restoration. (2025, 12 3). *Environmental Sensitivity Index (ESI) Maps and Data*. Retrieved 12 8, 2025, from www.response.restoration.noaa.gov: https://response.restoration.noaa.gov/oil-and-chemical-spills/oil-spills/environmental-sensitivity-index-esi-maps-and-data

NOAA Office of Satellite and Product Operations. (2025, 12 8). *Marine Pollution Products*. Retrieved from www.ospo.noaa.gov: https://www.ospo.noaa.gov/products/ocean/marinepollution/

NOAA PMEL Carbon Group. (2025). *PMEL Carbon Program*. Retrieved 12 8, 2025, from PMEL CO2 - Carbon Dioxide Program: https://www.pmel.noaa.gov/co2/

NOAA PMEL Carbon Program. (n.d.). *GOA-AN*.

OBO Foundry. (2023). *PATO: The Phenotype and Trait Ontology*. Retrieved from OBO Foundry : https://obofoundry.org/ontology/pato.html

Open Geospatial Consortium (OGC). (2012). *WKT — Well-known text representation of geometry (WKT) specification*. Retrieved from OGC: https://www.ogc.org/standards/wkt

Open Geospatial Consortium (OGC). (2015). *NetCDF (Network Common Data Form)*. Retrieved from Open Geospatial Consortium: https://www.unidata.ucar.edu/software/netcdf/

Open Geospatial Consortium (OGC). (2019). *WGS84 Geo Positioning (WGS84 vocabulary / geo namespace)*. Retrieved from W3C: https://www.w3.org/2003/01/geo/

Open Geospatial Consortium (OGC). (2022). *OGC GeoSPARQL — A geographic query language for RDF data*. Retrieved from OGC: https://www.ogc.org/standards/geosparql

Open Geospatial Consortium. (2015). *NetCDF (Network Common Data Form)*. Retrieved from Open Geospatial Consortium: https://www.unidata.ucar.edu/software/netcdf/

Perevalov, A. a. (2022). QALD-9-plus: A Multilingual Dataset for Question Answering over

DBpedia and Wikidata Translated by Native Speakers. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)* (pp. 229-234). IEEE.

QUDT.org. (2025). *QUDT: Quantities, Units, Dimensions and Data Types Ontologies*. Retrieved from QUDT.org: https://qudt.org

Schema.org. (2025). *Schema.org vocabulary*. Retrieved from Schema.org:

https://schema.org Sedaro. (2025). *Broad Agency Announcement Knowledge*

Graphs/Natural Language Processing (BAA-KM/NLP) Final Report.

Simple Knowledge Organization System (SKOS) Working Group, W3C. (2009). *SKOS Simple Knowledge Organization System — Reference*. Retrieved from W3C: https://www.w3.org/TR/skos-reference/

Sommer, A. &. (2024). pySHACL (v0.26.0). Zenodo. doi:10.5281/zenodo.4750840

Time Ontology in OWL Working Group, W3C. (2017). *Time Ontology in OWL*. Retrieved from W3C: https://www.w3.org/TR/owl-time/

Trivedi, P. a. (2017). Lc-quad: A corpus for complex question answering over knowledge graphs. In *International Semantic Web Conference* (pp. 210-2018). Springer.

U.S. Department of the Interior. (2025). *Natural Resource Damage Assessment and Restoration Program*. Retrieved 12 8, 2025, from www.cerc.usgs.gov: https://www.cerc.usgs.gov/orda_docs/

U.S. Geological Survey. (2022). *Geographic Names Information System (GNIS)*. Retrieved from U.S. Geological Survey: https://www.usgs.gov/us-board-on-geographic-names

United States Environmental Protection Agency. (2025). *Disaster Debris Recover Tool*. Retrieved 12 8, 2025, from Disaster Debris Mitigation and Planning: https://www.epa.gov/disaster-debris/disaster-debris-recovery-tool

W3C Data Catalog Vocabulary Working Group. (2020). *Data Catalog Vocabulary (DCAT) – Version 2*. Retrieved from W3C: https://www.w3.org/TR/vocab-dcat-2/

W3C OWL Working Group. (2012). *OWL 2 Web Ontology Language — Overview*. Retrieved from W3C: https://www.w3.org/TR/owl2-overview/

W3C Provenance Working Group. (2013). *PROV-O: The PROV Ontology*. Retrieved from W3C: https://www.w3.org/TR/prov-o/

W3C RDF Schema Working Group. (2014). *RDF Schema 1.1*. Retrieved from W3C: https://www.w3.org/TR/rdf-schema/

W3C RDF Working Group. (2014). *RDF 1.1 Concepts and Abstract Syntax*. Retrieved from W3C: https://www.w3.org/TR/rdf11-concepts/

W3C Semantic Sensor Network Incubator Group / OGC. (2017). *SOSA/SSN: Sensor, observation, sample, and actuator & Semantic Sensor Network Ontology*. Retrieved from W3C: https://www.w3.org/TR/vocab-ssn/

W3C XML Schema Working Group. (2012). *XML Schema Definition Language (XSD) 1.1 — Part 1: Structures*. Retrieved from W3C: https://www.w3.org/TR/xmlschema11-1/

7. Appendix

7.1. Terms

Term	Full Form
AI	Artificial Intelligence
AIC	Archival Information Collection
AIU	Archival Information Unit
API	Application Program Interface
ARN	Amazon Resource Name
AWS	Amazon Web Service
AWS ECR	AWS Elastic Container Registry
AWS KMS	AWS Key Management Service
AWS VPC	AWS Virtual Private Cloud
CBOFS	Chesapeake Bay Operational Forecast System
CF	Climate and Forecast
CSA	Canadian Space Agency
CZ-TD	Coastal Zone Digital Twin
DC	Dublin Core
DCAT	Data Catalog Vocabulary
DCTERMS	Dublin Core Metadata Terms
DIP	Dissemination Information Package
DPROD	Data Products Ontology
DWC	Darwin Core
ENVO	Environment Ontology
ESA	European Space Agency
ESI	Environmental Sensitivity Index
ETL	Extract, Transform, Load
FAIR	Findable, Accessible, Interoperable, Reusable

Term	Full Form
FOAF	Friend of a Friend
GEO/WGS84	World Geodetic System 1984
GeoTIFF	Geographic Tagged Image File Format
GLIMR	Geosynchronous Littoral Imaging and Monitoring Radiometer
GNIS	Geographic Names Information System
IaC	Infrastructure as Code
IAM	Identity and Access Management
ISRO	Indian Space Research Organization
JPSS	Joint Polar Satellite System
KM	Knowledge Mesh
LLM	Large Language Model
MCP	Model Context Protocol
MD	Member Description
MPSR	Marine Pollution Surveillance Reports
NASA	National Aeronautics and Space Administration
NCBI	National Center for Biotechnology Information
NESDIS	National Environmental Satellite, Data, and Information Service
NetCDF	Network Common Data Form
NISAR	NASA-ISRO Synthetic Aperture Radar
NLP	Natural Language Processing
NOAA	National Oceanic and Atmospheric Administration
NOS	National Ocean Service
NPP	National Polar-orbiting Partnership
NRDAR	National Resource Damage Assessment and Restoration
OFS	Operational Forecast System
OISS	Open Information Stewardship Service
OWL	Web Ontology Language

Term	Full Form
PACE	Plankton, Aerosol, Cloud, Ocean Ecosystem
PATO	Phenotype and Trait Ontology
PMEL	Pacific Marine Environmental Laboratory
PROV-O	Provenance Ontology
QUDT	Quantities, Units, Dimensions, and Types
R2RML	RDB to RDF Mapping Language
RDB	Relational Database
RDF	Resource Description Framework
RDFS	RDF Schema
S3	Simple Storage Service
SBG	Surface Biology and Geology
SDK	Software Development Kit
SDO	Schema.org
SHACL	Shapes Constraint Language
SKOS	Simple Knowledge Organization System
SME	Subject matter experts
SOSA	Sensor-Observation-Sampling-Actuator ontology
SSN	Semantic Sensor Network ontology
TIME	Ontology for time
TIR	Thermal Infrared
UBL	User Business Logic
URI	Uniform Resource Identifier
VSWIR	Visible and Short-Wave Infrared
WKT	Well-Known Text
XML	Extensible Markup Language
XSD	XML Schema Definition

7.2. Challenges Encountered

Challenge: Irregular update cycles and patterns across datasets.

Except for the OFS dataset, the geospatial and natural language datasets exhibit irregular updates. The AWS deployment of the pipelines is capable of handling repeated updates via scheduled calls to AWS Eventbridge. To ensure that only new data is imported into the graph, the Noblis team implemented metadata tracking and metadata checking within the pipelines.

Challenge: Combining geospatial and natural language features within data sources.

The Knowledge Mesh needs to be capable of bringing together both geospatial features and natural language features to effectively address the needs of multiple user-personas. Noblis used the MPSR data to establish a reusable solution pattern for semantically bringing together these feature categories. For example, for a prospective spill incident both geospatial boundaries as well as natural language reports are provided. Noblis' approach utilized a central node for the MPSR report and attaching that central node semantically to its geospatial and textual properties. This established a reusable foundation for additional datasets that have both geospatial and natural language features useful to knowledge mesh users.

Challenge: Integrating OISS with the Knowledge Mesh

Integrating the OISS framework with the Knowledge Mesh introduced significant conceptual and technical complexity. While OISS provides a powerful structure for describing data acquisition, transformation, and dissemination, aligning its pattern and template based architecture with the Knowledge Mesh's ontology driven data model required careful coordination. Each dataset pipeline needed to be semantically modeled not only in terms of its outputs but also the processes that generated those outputs, which created additional layers of metadata, provenance tracking, and configuration that had to be manually authored and validated. This integration was further complicated by the early stage maturity of the OISS tooling, requiring adjustments to both the KM pipelines and the OISS member descriptions. Despite these challenges, the integration revealed clear pathways for improving automation, introducing developer tooling through MCP assisted code generation, and expanding OISS templates so that future pipelines can be deployed with significantly less manual effort and greater semantic consistency.

Challenge: Manual standup of R2RML mapping files

Establishing R2RML mappings for virtualization proved to be a labor intensive process that demanded deep understanding of both the source database schema and the target semantic model. Developers were required to manually interpret relational structures, determine appropriate ontology alignments, and craft mapping files that governed how data would appear as RDF. This manual lift slowed experimentation and prevented rapid onboarding of new datasets or iterative refinement of semantic models. Moreover, ensuring that mappings were accurate, complete, and semantically coherent often required repeated validation across multiple tools and systems. These challenges highlight the need for automated mapping generation, ontology recommendation engines, and validation workflows supported by large language models and OISS integrated metadata checks. As described in the Future Work section, advancing these capabilities will enable NOAA to scale virtualization across many more data sources and formats while reducing the technical burden on developers and analysts.

7.3. Sedaro's Final Report

“Broad Agency Announcement Knowledge Graphs/Natural Language Processing (BAA-KM/NLP)”

Final Report

Subcontract Number: 25SNOAA1SC

Date: June 11, 2025

Revision 3: December 12, 2025

Project Manager: Sarah Moyers

sarah.moyers@sedaro.com

Technical Lead: Mitch Zielinski

mitch.zielinski@sedaro.com

1 Overview

The National Oceanic and Atmospheric Administration (NOAA) has identified a need to improve and expand the open discovery, access, use, reproducibility, and trust of its environmental intelligence – both in terms of process and information – in order to facilitate key strategic initiatives. Noblis is on contract to develop a foundational model capable of dynamically leveraging application programming interfaces (APIs) using a Knowledge Mesh capability for the purpose of answering highly contextual queries from a diverse swath of NOAA users. This Knowledge Mesh is intended to enable users across multiple personas, including ecologists, emergency managers, coastal scientists, and urban planners, to have improved data discoverability and access for coastline and coastline disaster datasets and use cases, as well as develop a scalable approach that can be iterated upon by adding additional datasets, geographic regions, and domains. The Knowledge Mesh integrates real-time and historical data and provides semantically enriched and queryable information related to oil and chemical spills.

In support of Noblis' work in this initiative, Sedaro has leveraged its digital twin capabilities to develop simulations that connect with the Knowledge Mesh, allowing users to retrieve synthetic data of predicted satellite overpasses of oil spill locations that are observable by satellites from NOAA, the National Aeronautics and Space Administration (NASA), the European Space Agency (ESA), the Canadian Space Agency (CSA), as well as other commercial Earth orbit satellites. This predicted data can be used to inform the Marine Pollution Surveillance Report (MPSR) by identifying the platforms that may have imagery of affected locations to better characterize the severity of the spill.

This final report captures accomplishments under Subcontract No. 25SNOAA1SC. The contents are provided for use by Noblis on Prime Agreement No: 1332KP24C0025. Although the full period of performance (PoP) runs from 1 April 2025 through 31 Dec 2025, this report only documents the deliverables due in the month of June. The final deliverable, an instance of the Sedaro platform hosted in Azure Government, will be provided to Noblis on October 1, 2025 per the terms of the Subcontract.

2 Scope

This report captures the work performed under Subcontract No. 25SNOAA1SC and serves as the deliverable required under "Documentation – Final Report." Sedaro's objectives for this effort included developing the ability to:

1. Read data from Knowledge Mesh Data Streams
2. Create basic simulations and return a set of forecasted values
3. Ensure that simulation results are accessible via the Knowledge Mesh

2.1 Deliverables

- Scenario Model
- System Models
- Python Scripts
- Documentation

- Other Direct Costs (ODCs)
- Additional Deliverables

2.2 Final Demonstration

Sedaro has provided a summary of the solution and accompanying workflow developed under this effort.

- Final Demo

2.3 Future Development Efforts

Sedaro has included data about the following subject matter to offer Noblis and NOAA additional insight about areas identified for further development in this effort:

- Noblis Topics
- NOAA Topics

2.4 Sedaro Architecture

Sedaro has compiled documentation on the platform's open architecture and capabilities including modeling standards and software interfaces. Sedaro has addressed these topics in the following sections:

- Links to "Sedaro Protocol" Standards and Software

3 Personnel

The primary Sedaro points of contact for this effort are:

- Dr. Robbie Robertson, CEO
- Sebastian "Bas" Welsh, CTO
- Daniel Martin, COO — Contracts Manager
    - Daniel.Martin@sedaro.com
- Sarah Moyers, PMP — Project Manager
    - Sarah.Moyers@sedaro.com
- Mitchell Zielinski – Technical Lead
- Dr. Brad Sease – Technical Team

4 Deliverables

4.1 Scenario Model

4.1.1 Criteria

Per the table included on page 14 of Subcontract No. 25SNOAA1SC, Sedaro supplied a scenario model that supported the following criteria:

- "User provides an input containing the location of the incident"
- "Scenario model generates a synthetic report of the incident"

Additional criteria provided on page 53 of Subcontract No. 25SNOAA1SC were also met in the development of Sedaro's scenario model:

- "Build a Digital Twin for Chesapeake Bay Oil Spill Response Analysis that queries and processes RDF-compliant data from the Knowledge Mesh"
- "Develop and deliver simulation scenarios in the Sedaro platform that allow users to retrieve synthetic data of satellite overpasses of oil spill locations for various Earth-observation satellite systems"
- "Design and implement a web-based or GIS-integrated dashboard for DT visualization"

Sedaro's scenario model was delivered on May 30, 2025 and was formally accepted by the Noblis Project Manager (PM)

4.1.2 Scenario Model

Sedaro's technical team developed a scenario model in the Sedaro simulation platform designed to simulate predicted observations of a potential oil spill. The scenario is populated by models of relevant spacecraft, specified by Noblis, and their respective ground segments. It is designed to be called from a Python script that provides the scenario time bounds, oil spill location (defined either as a single point or as a polygonal region), and up-to-date spacecraft states. The scenario is used to model predicted opportunities for observation of the oil spill, as well as predicted opportunities for that data to be downlinked to a ground station. The Sedaro platform enables users to easily add additional models to the scenario, and the platform's version control capabilities enable multiple modified versions of the scenario to be maintained. The scenario model for this effort can be viewed below in Figure 1.

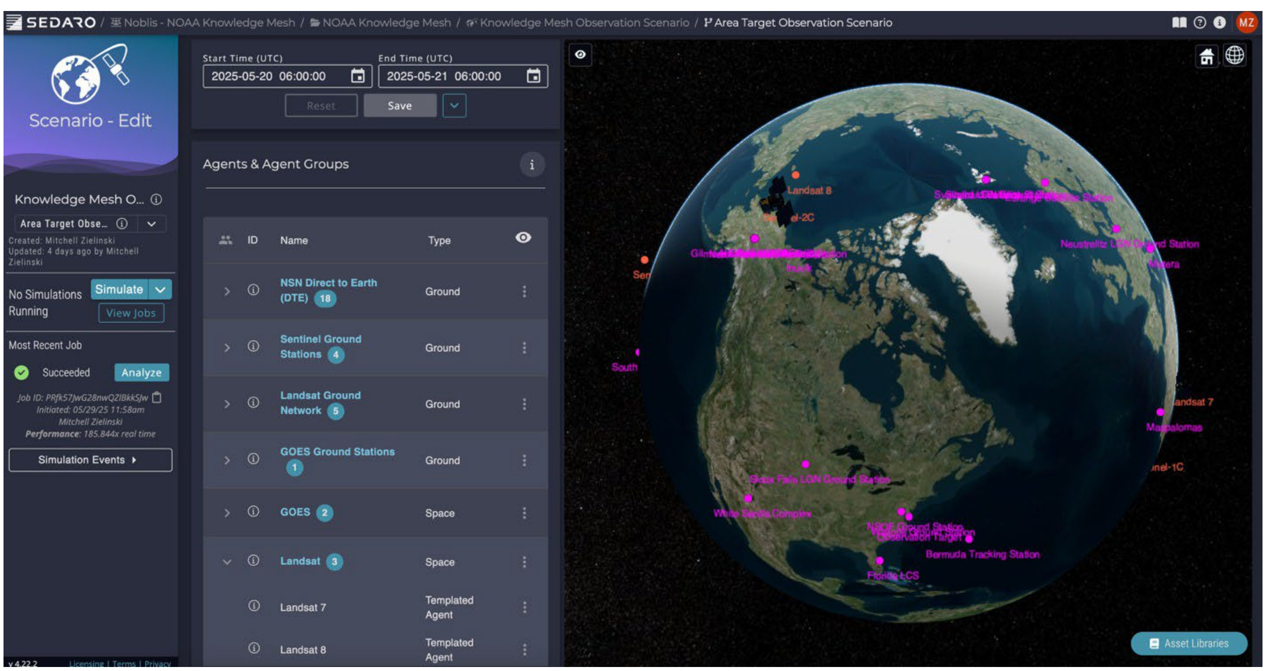


Figure 1: Sedaro scenario model containing various satellites and ground stations.

The dashboard visualizations referenced in the criteria for the scenario model include a number of dashboard options available within the Sedaro platform, which provide various analytics and applicable data for this use case. While these dashboards are not visible to end-users of the Knowledge Mesh, they will be available on Noblis' dedicated instance, which is scheduled to deploy on October 1, 2025. An example of a plot dashboard can be seen in Figure 2 below.

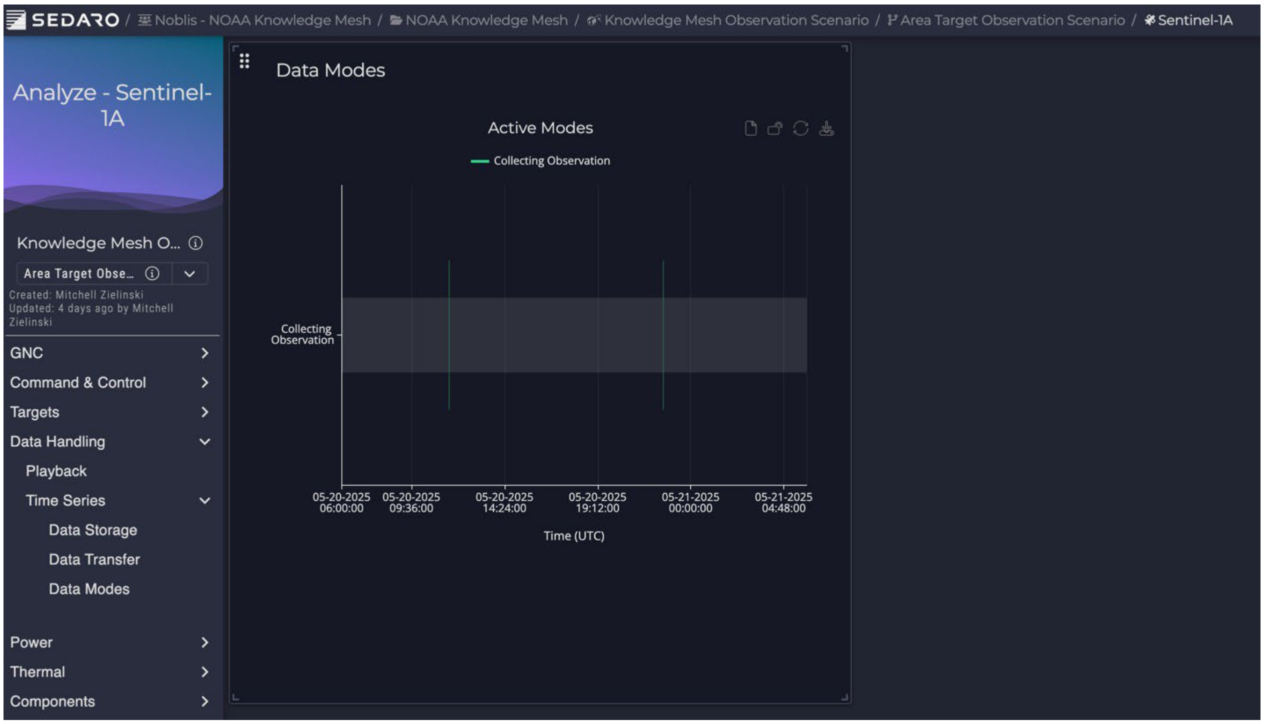


Figure 2: A Sedaro dashboard view displaying the timeframes in which the Sentinel A1 can view a designated set of coordinates.

As an additional way to facilitate feedback from the Noblis team and allow for transparency during the design process, Sedaro housed the scenario model in a dedicated Beta workspace that was shared with stakeholders from both Noblis and NOAA. Users have the ability to monitor updates to the models in real time and explore platform capabilities throughout development efforts.

4.2 System Models

4.2.1 Criteria

Per the table included on page 14 of Subcontract No. 25SNOAA1SC, Sedaro supplied system models that supported the following criteria:

- "Access to the underlying Sedaro models powering the Scenario Modeling tool"
- "Ability to write custom python scripts that interact with the underlying model"
- "Ability to spin up/down the instance running the model in Azure Government"

Additional criteria provided on page 53 of Subcontract No. 25SNOAA1SC were also met in the development of Sedaro's system models:

- "Develop and deliver system models of space vehicles for simulation. These shall include 4x NOAA GOES, 3x NASA Landsat, and 1x ESA Sentinel"
- "Develop and deliver Ground Segments models in the Sedaro Platform for the NASA Near Space Network"

Sedaro's system models were delivered on May 30, 2025 and were formally accepted by the Noblis PM on June 11, 2025. All models and scripts have been designed to be fully functional on Noblis' dedicated instance on Azure Government, which is scheduled to deploy on October 1, 2025.

4.2.2 Satellites

Sedaro developed models of the following satellites:

- **ESA Sentinel**. The Sentinel-1 is a polar-orbiting, all-weather, day-and-night radar imaging mission for land and ocean services. Sentinel-2 is a polar-orbiting, multispectral high-resolution imaging mission for land monitoring to provide, for example, imagery of vegetation, soil and water cover, inland waterways and coastal areas. Sentinel-2 can also deliver information for emergency services.
  - This scenario includes **Sentinel units 1A, 1C, 2A, and 2B.**
- **NASA Landsat**. The Landsat program consists of a series of Earth-observing satellite missions jointly managed by NASA and the U.S. Geological Survey (USGS). Since 1972, Landsat satellites have continuously acquired images of the Earth's land surface and provided an uninterrupted data archive to assist land managers, planners, and policymakers in making more informed decisions about natural resources and the environment.
  - This scenario includes **Landsat units 6, 7, 8, and 9.**
- **NOAA GOES**. The Geostationary Operational Environmental Satellite (GOES), operated by NOAA's NESDIS division, supports weather forecasting, severe storm tracking, and meteorology research. Spacecraft and ground-based elements of the system work together to provide a continuous stream of environmental data. The National Weather Service (NWS) and the Meteorological Service of Canada use the GOES system for their North American weather monitoring and forecasting operations, and scientific researchers use the data to better understand land, atmosphere, ocean, and climate dynamics.
  - This scenario includes **GOES units 16, 17, 18, and 19.**

4.2.3 Ground Stations

Sedaro developed models of the following ground station network:

- **NASA NSN**. The Near Space Network provides missions within one million miles of Earth with robust communications services. Using a blend of government and commercial assets, the network supports science, human spaceflight, and technology demonstration missions exploring the planet and solar system. This data is gathered through global direct-to-Earth antenna systems and a fleet of relay satellites.

4.3 Python Scripts

4.3.1 Criteria

Per the table included on page 14 of Subcontract No. 25SNOAA1SC, Sedaro supplied system models that supported the following criteria:

- "Scripts are responsible for taking the user's input and passing it into the scenario model"
- "Scripts are responsible for transforming the output of the scenario model into the desired format"

Additional criteria provided on page 53 of Subcontract No. 25SNOAA1SC were also met in the development of Sedaro's Python scripts:

- "Provide documentation, including data model, application processing interface (API) specifications, and scalability recommendations"
- "Enable SPARQL Protocol and RDF Query Language (SPARQL) queries for retrieving data from the Digital Twin by implementing a simple Python client that reads relevant data out of the Sedaro platform and makes it available for the Noblis team to publish into their Resource Description Framework (RDF) store in order to be queried."

Sedaro's Python scripts were delivered on May 30, 2025 and were formally accepted by the Noblis PM on June 11, 2025.

4.3.2 Python Scripts

Sedaro's technical team developed a lightweight Python package designed to act as middleware between the Knowledge Mesh and the Sedaro simulation platform. The package centers around a script function designed to be called when a call is made to the Knowledge Mesh API to kick off a simulation. The function accepts an observation region definition of one or more pairs of latitude / longitude coordinates, as well as start and end times for the simulation. A helper function reaches out to Space-Track.org, the USG portal for space object tracking data, via an API call to obtain the most up-to-date spacecraft orbital state information possible for the given scenario start epoch. The main script function uses the Sedaro Python client to access the oil spill observation scenario and updates the scenario time bounds, the oil spill observation target model, and the orbital state of each spacecraft model before kicking off the scenario. Simulation progress is communicated to the terminal by means of an updating progress bar.

Once the scenario simulation is complete, the full simulation results are downloaded and passed to another helper function for processing. This function uses the Sedaro Python client to query the simulation results and construct JSON-formatted sets of output data, which are passed back to the calling code.

4.3.3 Data from Output Sets

The list below provides detailed descriptions of each data field in the output set from a given query:

- **Downlink Start**. The timestamp when a satellite begins transmitting collected data (e.g., images, sensor readings) to a ground station. This occurs when the satellite is within communication range of a ground station and ready to send stored data.

- **Imaging Start**. The timestamp when a satellite begins capturing images or sensor data of a specific area of interest on Earth. This is often linked to the satellite's position and orientation over the target location.
- **Pass Direction**. The direction the satellite travels during an orbital pass over a target area. Common directions include ascending (moving from south to north) and descending (moving from north to south). This affects lighting, viewing angle, and time of day for image capture.
- **Spacecraft**. The name of the satellite with visibility of a given set of coordinates.
- **Imaging Time**. The total duration during which the satellite collects imagery of a specific area. This is typically the time between **Imaging Start** and **Imaging End.**
- **Time to Downlink**. The time interval between the end of image capture (**Imaging End**) and the **Downlink Start**. It represents the delay between data acquisition and the beginning of transmission to the ground station.

An example of a synthetic data report of a given oil spill location can be seen below in Figure 3.


```
"2025-06-01T22:34:00.000086+00:00": {
  "downlink_start": "2025-06-01T22:43:30.000086+00:00",
  "imaging_start": "2025-06-01T22:34:00.000086+00:00",
  "imaging_time": 50.0,
  "pass_direction": "ASCENDING",
  "spacecraft": "Sentinel-1A",
  "time_to_downlink": 520.0
},
"2025-06-02T02:49:40.000086+00:00": {
  "downlink_start": "2025-06-02T02:51:10.000087+00:00",
  "imaging_start": "2025-06-02T02:49:40.000086+00:00",
  "imaging_time": 20.0,
  "pass_direction": "ASCENDING",
  "spacecraft": "Landsat 9",
  "time_to_downlink": 70.0
}
```


Figure 3: Synthetic report of an oil spill observation.

4.4 Documentation

4.4.1 Criteria

Per the table included on page 14 of Subcontract No. 25SNOAA1SC, Sedaro supplied system models that supported the following criteria:

- "Thorough documentation on the solution created by Sedaro"
- "This should include a final report on the current functionality, limitations, and potential future state of the tool"

Additional criteria provided on page 53 of Subcontract No. 25SNOAA1SC:

- "Subcontractor shall submit monthly progress reports via email to the Noblis Project Manager and Subcontracts Administrator listed in this Agreement. Additional details will be provided upon award."

This requirement was voided by the Noblis PM as all of Sedaro's technical development work took place within the month of May. Information that would have gone into the monthly report has been included in this Final Report.

Sedaro's Final Report was delivered on June 11, 2025, and was accepted by the Noblis PM.

4.4.2 Final Report

This document serves as the deliverable for "Documentation – Final Report."

4.5 Other Direct Costs (ODCs)

Per the terms of Subcontract No. 25SNOAA1SC, Sedaro is obligated to deliver Noblis a license package with a start date of October 1, 2025. This license package includes the following:

- Multi-Tenant Sedaro Platform Licenses
- Sedaro Cloud Monthly 6-Month Lease
- Cybersecurity Maturity Model Certification (CMMC) Compliant Instance of the Sedaro

Platform Noblis' instance of the Sedaro Platform is scheduled to deploy on Azure Government

on October 1, 2025.

4.6 Additional Deliverables

Per page 53 of Subcontract No. 25SNOAA1SC, Sedaro completed the following additional deliverables:

- "Inputs to architecture and prototype development per request of the Project Manager"
- "Define the Digital Twin's system architecture and project plan"
  - "This should be a cloud-based solution in Azure Government and include cost estimates for development and hosting"

4.6.1 Inputs to Architecture and Development

Sedaro facilitated inputs to prototype architecture and development by establishing a weekly meeting cadence with the Noblis team. This allowed for frequent feedback and requests, which included adjustments to the data sets generated by Sedaro’s Python client and the generation of a metadata glossary to aid the Noblis team in their own API development.

4.6.2 Project Plan

Sedaro provided a detailed Project Plan, which outlined the constituent parts of each deliverable, a proposed interface between Sedaro and Noblis, and a timeline of estimated completion. Cost estimates were not required for Noblis’ instance on Azure Government as these expenses were already accounted for in the cost of this firm fixed price (FFP) contract. Sedaro’s Project Plan was presented on May 7, 2025. Written acceptance from the Noblis PM was received on May 9, 2025.

5 Final Demo

Sedaro’s solution interfaces with the Knowledge Mesh, allowing NOAA end-users to make inquiries about oil spill location targets and receive synthetic reports about the incidents. To begin, a user queries the Knowledge Mesh by inputting a set of latitude and longitude coordinates corresponding to an area of interest. The Knowledge Mesh then makes an API call to Sedaro, using a custom Python script to translate the Knowledge Mesh’s RDF-formatted data into a format that Sedaro can read. The Sedaro platform then runs simulations to determine which satellites can provide data about the given set of coordinates. In the example in Figure 4, a polygonal observation target derived from input coordinates can be seen in the Chesapeake Bay area. In Figure 5, the Sentinel 2B captures the observation target in its FOV as it orbits.

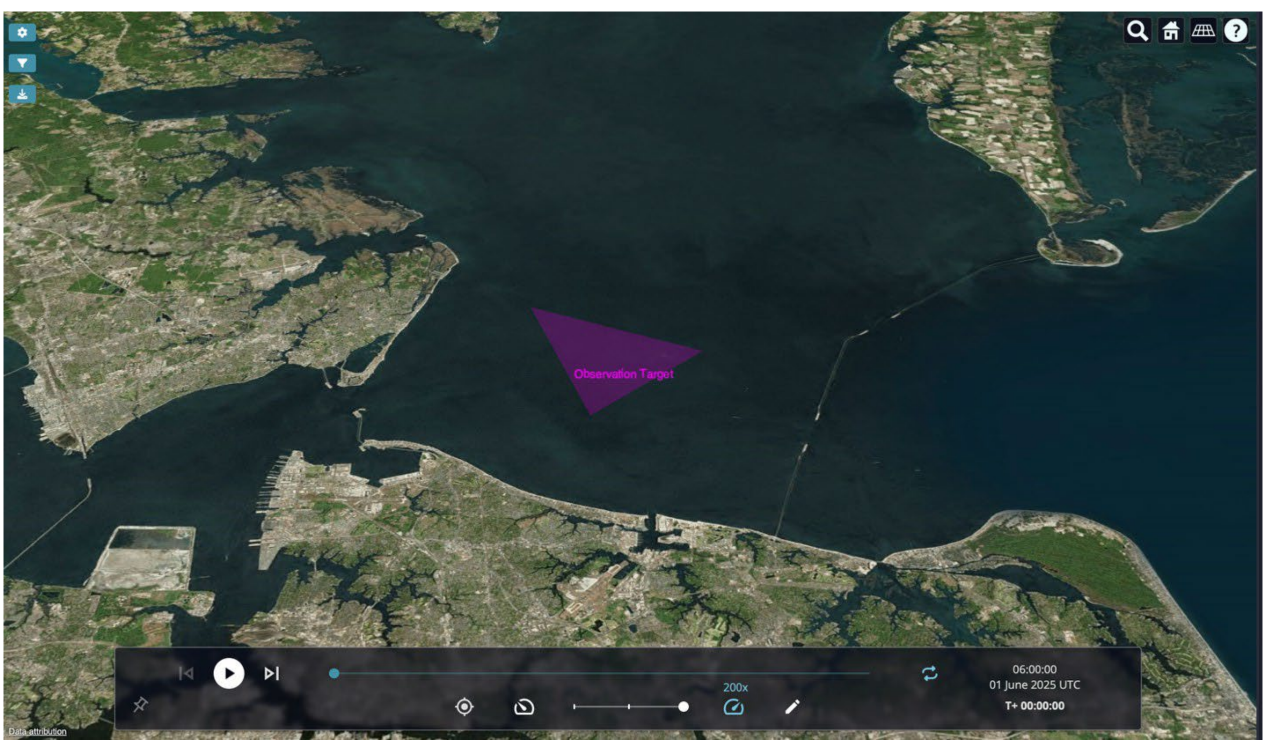


Figure 4: Polygonal observation target in the Chesapeake Bay area.

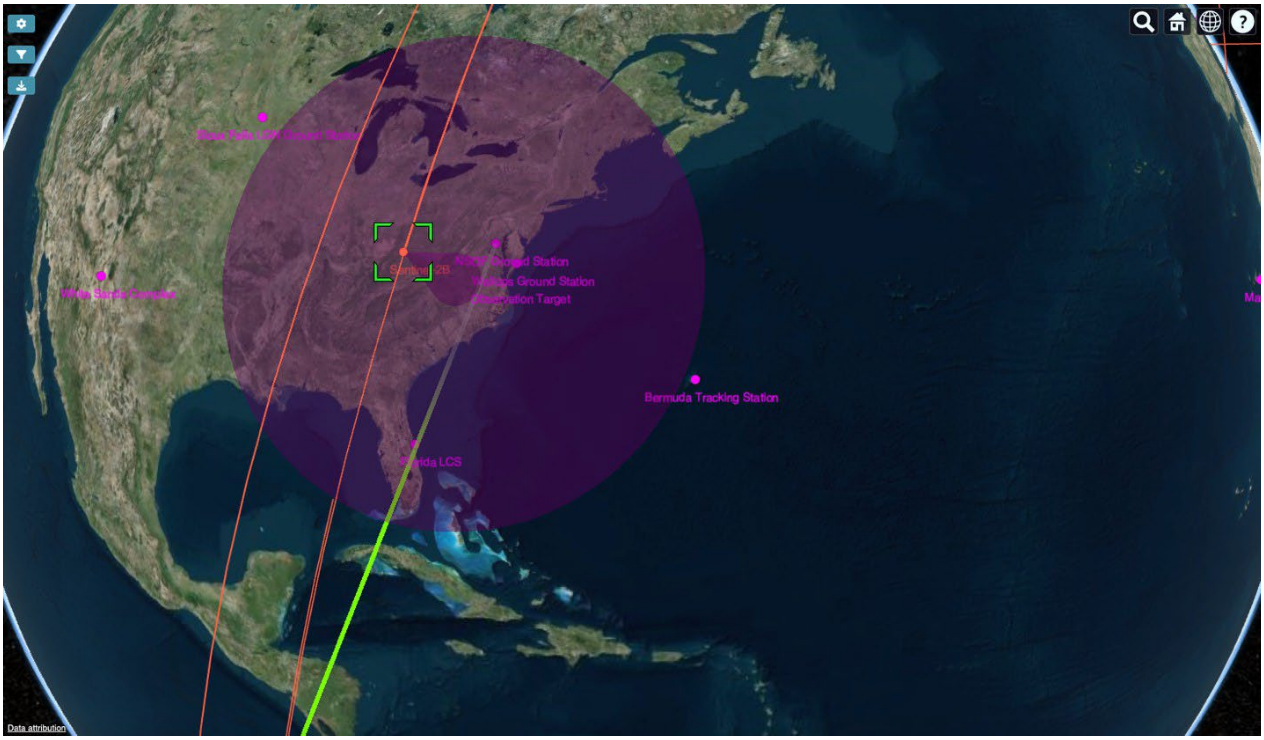


Figure 5: As Sentinel 2B travels along its orbit, the observation target passes beneath the satellite's FOV.

After running simulations, the Sedaro platform outputs a set of observations including imaging start, time from image to downlink, satellite information, and other useful data seen earlier in Figure 3. This output set is then translated back into RDF format and is passed back through the Knowledge Mesh to the NOAA end-user, providing them with knowledge about what satellites and systems to consult in order to learn more about a given oil spill or location. This process flow is captured in Figure 6.

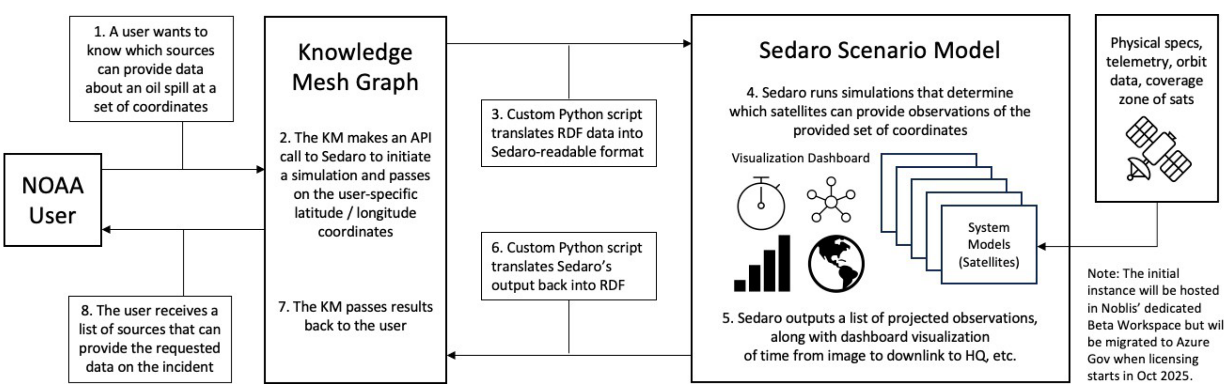


Figure 6: Diagram outlining the process flow between NOAA users, the Knowledge Mesh, and Sedaro.

Sedaro's Final Demonstration was presented on June 11, 2025, and was accepted by the Noblis PM.

6 Future Development Efforts

Sedaro has captured the following topics identified by NOAA and Noblis for future development, should NOAA decide to pursue follow on work for this experimental project.

6.1 Noblis Topics

While these items were determined to be out of scope for this preliminary effort, the Sedaro team has documented a record of them here for future consideration:

- **Modeling Individual Instruments**. Currently, all Sedaro models include instrument names; however, modeling individual instruments on a satellite (for example, a Landsat) would require a significant step up in model fidelity and development efforts.
- **Modeling Scanning Patterns**. While identifying which kind of imaging mode a given satellite would use is fairly simple, accurately modeling scanning patterns would require in-depth research on the satellites in question, as well as significant model development time.
- **Imaging Mode**. Refers to the configuration or operational setting of a satellite's sensor system during data collection. It determines how the sensor captures imagery, including factors such as resolution, swath width, number of spectral bands, and viewing geometry (e.g., angle). This information could be useful if agents have the capability to model scanning patterns as described in the item above.
- **Imaging Angle**. The Sedaro team believes that it is possible to model imaging angle based on the existing data about the satellites used in this project; however; this would require a moderate amount of script development work.
- **Satellite Revisit Frequency**. The Sedaro team is currently outputting predicted satellite overpasses for the duration of the simulation but likely cannot determine satellite revisit rate without constructing another scenario focused on long-term prediction. Note: All of the satellites used in this project are either geostationary or sun-synchronous, so the revisit frequency would be either "always" or "once per day at approximately the same time."
- **Cloud Cover Percentage**. An estimate of how much of the imaged area is obscured by clouds, expressed as a percentage (ex: 0% = clear sky, no cloud obstruction; 100% = fully cloud-covered image). High cloud cover reduces the usability of optical satellite imagery.

6.2 NOAA Topics

The Sedaro team met with representatives from NOAA's Coastal Zone Digital Twin (CZ-DT) Community of Practice on May 22, 2025 to garner initial feedback on development efforts. The CZ-DT team provided several suggestions for future satellites and missions to incorporate into the Knowledge Mesh model:

- **NASA Plankton, Aerosol, Cloud, Ocean Ecosystem (PACE)**. NASA's PACE is an Earth-observing mission that launched in February 2024. It studies the distribution of phytoplankton in the ocean, aerosol particles and clouds in the atmosphere, and their impact on Earth's climate. The satellite orbits at an altitude of 676.5 km in a sun-synchronous polar orbit.

- **Suomi National Polar-orbiting Partnership (NPP)**. NOAA’s NNP is a key satellite in the Joint Polar Satellite System (JPSS). It was launched on October 28, 2011, and its primary mission is to collect and distribute remotely sensed data for weather and climate research.
- **NASA-ISRO Synthetic Aperture Radar (NISAR)**. NISAR is a joint mission between NASA and the Indian Space Research Organization (ISRO) to map Earth's surface with radar. It is designed to measure changes in the planet's surface, including movements as small as centimeters, and will scan nearly all of Earth's land and ice surfaces twice every 12 days. NISAR data will be free and publicly available once commissioned (estimated 3-5 months from the date of this report).
- **Surface Biology and Geology (SBG)**. The SBG mission, a NASA Earth-observing initiative, aims to study the Earth's surface, including its biology, geology, and their interactions. It involves two main satellite concepts: SBG Thermal Infrared (TIR) and SBG Visible and Short-Wave Infrared (VSWIR). These satellites will use advanced imaging technology to provide global, high-resolution data for a range of scientific and application needs.
- **Geosynchronous Littoral Imaging and Monitoring Radiometer (GLIMR)**. The GLIMR is a NASA-funded, space-based instrument that will observe and monitor ocean biology, chemistry, and ecology in the Gulf of Mexico, the southeastern U.S. coastline, and the Amazon River plume. GLIMR will be uniquely positioned in a geosynchronous orbit, allowing for frequent and high-resolution observations of these coastal waters.

7 Sedaro Platform and Protocol

Sedaro is a digital twin software company focused on leveraging modern software frameworks and cloud computing to deliver a leap forward in the evolution of system Modeling, Simulation, and Analysis (MS&A). Sedaro is delivering this leap forward with the cloud-native software architecture illustrated in Figure 7. This digital twin software is more scalable, collaborative, and interoperable than alternative technologies for MS&A.

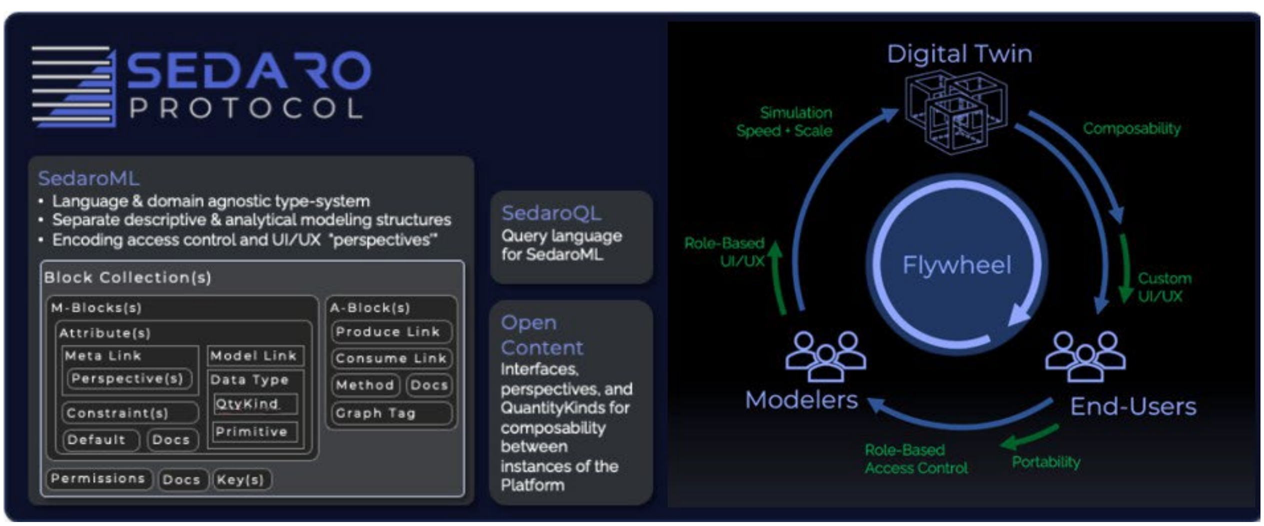


Figure 7: Sedaro Commercial Platform and Open Protocol (standards, software, and open content). The integrated platform delivers differentiated capabilities for collaborative modeling, parallelized hybrid-cloud

simulation, and digital integration. The modern architecture consists of several discrete microservices, all of which auto-scale as *needed to support a given number of concurrent collaborators, level of system complexity, or quantity of parallel simulations.*

Key components of the Sedaro platform that form the foundation of the proposed digital twin solution are illustrated in Figure 7, in addition to external applications and associated integration software. These external applications can be heritage software tools running on local machines or in containers co-hosted in the cloud for turnkey scalability and a faster connection to Sedaro's distributed runtime.

7.1 Links to "Sedaro Protocol" Standards and Software

Sedaro is the process of formalizing the community-facing documentation and semantics for "Sedaro Protocol." Included below is a list of links to open-source Sedaro documentation, system and simulation scenario ontologies, and digital integration software.

- **Introduction to Sedaro**: https://docs.sedaro.com/
- **Sample Spacecraft Metamodel**: https://docs.sedaro.com/ontologies#Spacecraft
- **Sample GroundSegment Metamodel**: https://docs.sedaro.com/ontologies#GroundSegment
- **Sedaro Terrestrial Vehicle Metamodel**: https://docs.sedaro.com/ontologies#TerrestrialVehicle
- **Sedaro Scenario Metamodel**: https://docs.sedaro.com/ontologies#Scenario
- **Sedaro API**: https://sedaro.github.io/openapi/
- **Sedaro Python Client**: https://github.com/sedaro/sedaro-python
- **Sedaro Modeling & Simulation Jupyter Notebooks**: https://github.com/sedaro/modsim-notebooks/tree/main
- **Sedaro Model Exchange (ModEx)**: https://github.com/sedaro/model-exchange
- **Sedaro MATLAB Cosimulation Client**: https://github.com/sedaro/sedaro-matlab

8 Summary

In support of Noblis' initiative to enhance the discovery, accessibility, reproducibility, and trustworthiness of NOAA's environmental intelligence, Sedaro has delivered a digital twin solution that simulates satellite overpasses and connects seamlessly to the Knowledge Mesh. This successful demonstration provides a valuable proof-of-concept for how digital twin simulations can be integrated with federated data environments to enhance situational awareness and decision support. By validating end-to-end interoperability—from data stream ingestion to real-time simulation output and semantic reintegration—this effort lays the groundwork for future development. Specifically, it illustrates a repeatable and extensible architecture that can support a wide range of environmental monitoring and disaster response scenarios, while promoting modular integration of additional simulation services and data domains over time.

9 Glossary

API – Application Processing Interface

CMMC – Cybersecurity Maturity Model

CZ-DT – Coastal Zone

Digital Twin ESA –

European Space Agency

FFP – Firm Fixed Price

GOES – Geostationary Operational Environmental

Satellite GUI – Graphical User Interface

ISRO – Indian Space Research

Organization JPSS – Joint Polar

Satellite System

MS&A – Modeling, Simulation, and

Analysis MSC – Meteorological

Service of Canada MPSR –Marine

Pollution Surveillance Report

NASA – National Aeronautics and Space Administration

NESDIS – National Environmental Satellite, Data, and Information

Service NOAA – National Oceanic and Atmospheric

Administration

NNP - Suomi National Polar-orbiting

Partnership NSN – Near Space Network

NWS – National Weather

Service ODC – Other Direct

Cost

PACE – Plankton, Aerosol, Cloud, Ocean

Ecosystem PM – Project Manager

PoP – Period of Performance

RDF – Resource Description

Framework SBG – Surface Biology

and Geology

SPARQL – SPARQL Protocol and RDF

Query Language TIR – Thermal Infrared

USGS – U.S. Geological Survey

VSWIR – Visible and Short-Wave Infrared