



# DOME 4.0

## Deliverable D2.5 - Tools for data analytics

<b>Responsible Partner:</b>	CMCL	29/11/2024
<b>Contributor(s):</b>	Chung Ting Lao (CMCL)	29/11/2024
<b>Reviewer(s):</b>	Adham Hashibon (UCL), Nicola Marzari (EPFL)	29/11/2024
<b>Coordinator:</b>	CMCL	29/11/2024
<b>Dissemination Level:</b>	Public	
<b>Due Date:</b>	M48 (November 2024)	
<b>Submission Date:</b>	29.Nov.2024	

## Project Profile

<b>Programme</b>	Horizon 2020
<b>Call</b>	H2020-NMBP-TO-IND-2020-twostage
<b>Topic</b>	DT-NMBP-40-2020 Creating an open marketplace for industrial data (RIA)
<b>Project number</b>	953163
<b>Acronym</b>	DOME 4.0
<b>Title</b>	Digital Open Marketplace Ecosystem 4.0
<b>Start Date</b>	December 1 <sup>st</sup> , 2020
<b>Duration</b>	48 months



This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 953163. It is the property of the DOME 4.0 consortium and do not necessarily reflect the views of the European Commission.

## Document History

Version	Date	Author	Remarks
V0.1	18/10/2024	Chung Ting Lao	First Draft
V0.2	04/11/2024	Chung Ting Lao	Incorporated details on visualisation
V0.3	06/11/2024	Martin Uhrin	Incorporated details on connector
V0.4	07/11/2024	Chung Ting Lao	Minor editing
V0.5	29/11/2024	Adham Hashibon	Review
V0.6	29/11/2024	Nicola Marzari	Review
Final	29/11/2024	Chung Ting Lao, Bijan Yadollahi and Amit Bhave	Final Version

## Publishable Summary

Deliverable D2.5 focused on enhancing the DOME 4.0 data workflow by developing tools and services for efficient data querying, mining, and translation. Key achievements include adopting and extending standard APIs, developing capabilities for data stream processing, and creating visualisation tool and data analysis tool. These tools were integrated into the DOME 4.0 platform, significantly improving data onboarding, sharing, and collaboration. The advancements empower users to explore, analyse, and integrate data effectively, positioning DOME 4.0 to support a wide range of materials science and engineering research activities. This deliverable is a public release, due on M48 of the project.

## Executive Summary

Deliverable D2.5 focused on enhancing the DOME 4.0 data workflow by developing tools and services for efficient data querying, mining, and translation. Key achievements include adopting and extending standard APIs (for example, with other digital services, databases and marketplace projects), developing capabilities for data stream processing, and creating visualisation tool and data analysis tool. These tools were integrated into the DOME 4.0 platform, significantly improving data onboarding, sharing, and collaboration. The advancements empower users to explore, analyse, and integrate data effectively, positioning DOME 4.0 to support a wide range of materials science and engineering research activities. This deliverable is a public release, due on M48 of the project.

## Table of Contents

Publishable Summary .....	2
Executive Summary.....	3
Table of Contents.....	4
List of Figures.....	5
List of Tables .....	5
1. Introduction .....	6
2. External data connection .....	7
2.1 External data connectors.....	7
2.1.1 Onboarding .....	7
2.1.2 Usage.....	7
2.2 Catalog data.....	7
2.2.1 Onboarding .....	7
2.2.2 Usage.....	8
2.3 Ontology.....	8
2.3.1 Onboarding .....	8
2.3.2 Usage.....	9
3. Data hand-off.....	12
3.1 Onboarding.....	12
3.2 Usage.....	14
4. Advanced tool for data analysis .....	16
4.1 Visualisation tool .....	16
4.2 Support for data analytic tools .....	17
5. Conclusions / Next steps.....	20
6. Lessons learnt .....	21
7. Deviations from Annex 1.....	22
8. References .....	22
9. Acknowledgement .....	22
10. Table of Abbreviations.....	23
Annex 1.....	<b>Error! Bookmark not defined.</b>

## List of Figures

Figure 1: overview of WP2. D2.5 focuses on connecting external data and hand-off to modelling workflows. ....	6
Figure 2: Process for registering new catalogue data with DOME .....	8
Figure 3: ontologies can be uploaded to DOME.....	9
Figure 4: Uploaded ontologies are discoverable from a list. ....	10
Figure 5: Advanced search for uploaded ontologies.....	11
Figure 6: An example of semantic link between data connector and tools. Here "MATERIALSPROJECT" is a data connector and "AIIIDALAB" is a tool. Both conform to the "OPTIMADE_API_SPECIFICATION". ....	12
Figure 7: registration form for API specification. ....	13
Figure 8: registered API specifications will appear as options for registration of connectors and tools. ...	14
Figure 9: An example of data hand-off between data provider and data consumer through DOME 4.0 platform. User may send data discovered from a connector (1) to a compatible tool (3) given that they both conform to a certain standard (2). ....	15
Figure 10: example output of MedusaVis. Properties are shown as rectangular boxes, values are represented as diamonds, and resources are displayed as ellipses. ....	16
Figure 11: multiple layout options are available to users, including "Network" (left) and "Hierarchical" (right).....	16
Figure 12: sample call to public search API. ....	17
Figure 13: the query URL of a connector is shown on its information page. This API call requires an API key to be included in the header.....	18
Figure 14: form to generate access key for connector API. The read scope must be included. ....	19

## List of Tables

No table of figures entries found.

# 1. Introduction

Task T2.5 focuses on developing tools for querying, mining, and translating data, as specified in Task T2.1. These tools are provided as services to enhance the DOME 4.0 data workflow. By bridging data sources and services DOME brings added value to both by enabling use cases that were not necessarily foreseen by their creators. Standard APIs such as OPTIMADE and VIMMP, along with standard ontologies like EMMO, have been adopted and further developed to facilitate efficient querying of both open and proprietary databases, some notable examples of the former being Materials Cloud and Materials Project.

This task has developed the capabilities to support joining, splitting, and aggregating data streams, as well as translating various formats and encodings into DOME 4.0 compliant data streams. This ensures data is prepared for simulation and modelling workflows developed in Task T2.6. Additionally, advanced tools for data clustering and visualisation are developed.

The data services and tools developed in this task have been integrated into the DOME 4.0 ontology through collaboration with Task T3.1 and incorporated into the platform in cooperation with Task T1.1. This integration will ensure a cohesive and efficient data management system within the DOME 4.0 framework.

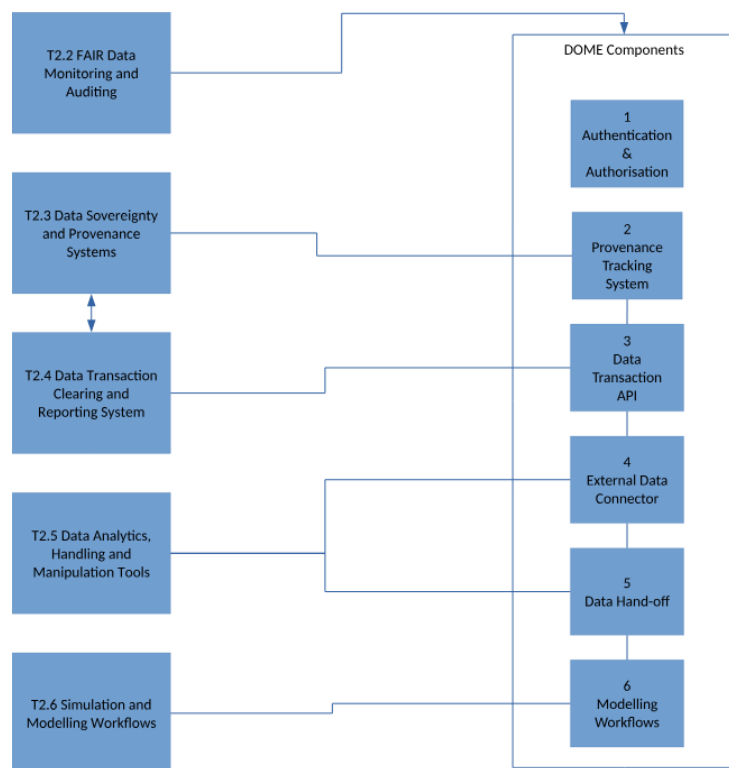


Figure 1: overview of WP2. D2.5 focuses on connecting external data and hand-off to modelling workflows.

## 2. External data connection

External data can be on-boarded to DOME in three ways. Firstly, data can be on-boarded using a connector, which acts as a wrapper around the API of external data platforms. Secondly, data can be registered individually through catalogue data, lastly semantic data can be uploaded directly via ontology.

### 2.1 External data connectors

#### 2.1.1 Onboarding

Connecting to an external data platform is accomplished using custom-built connectors, which are based on a reference connector initially developed in T3.5 and T3.7. This reference connector can be accessed through the Cookiecutter template found at [`DOME-4-0/reference-connector`](#) on GitHub.

After a connector is developed and deployed, users can register the data source via the DOME 4.0 platform's website using the [`Register DOME 4.0 Connector`](#) feature available at `dome40.io`.

To date, we have successfully on-boarded several external data connectors for OPTIMADE, an open standard designed for the exchange of atomistic data. This includes connectors for databases like Materials Cloud and The Materials Project, providing users with access to millions of crystal structures.

#### 2.1.2 Usage

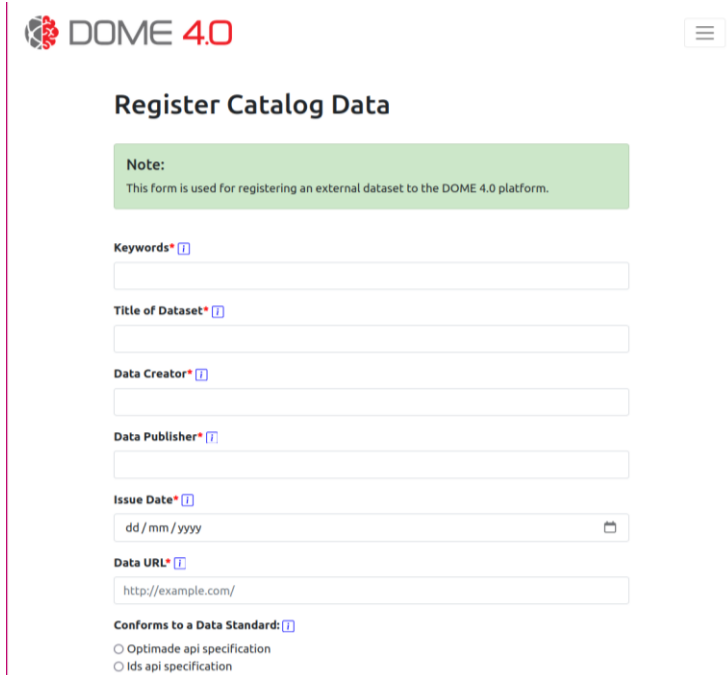
Users can perform free-text searches across all registered connectors through our semantic discovery service. The search results include both metadata and data, standardized according to the DOME dataset ontology, which is developed incrementally within the DOME 4.0 framework. Both the metadata and data are provided in JSON format, ensuring consistency and ease of use across datasets. The DOME dataset ontology can be accessed at [`DOME-4-0/data-set-ontology`](#) on GitHub, which provides further information on the standards applied for data exchange.

### 2.2 Catalog data

#### 2.2.1 Onboarding

A data catalog is a structured collection of metadata that describes data assets, such as datasets or data sources, making them searchable and easier to discover. The DOME 4.0 platform allows users to register individual data catalog entries, enabling broader access and easier searchability through the platform. When registering a data catalog, users can include keywords, data creator and publisher information, the issue date, a URL for data access, compatibility with data standards, licensing information, a description, and relevant topics. This information helps ensure datasets are easily accessible and can be integrated with other tools and services on DOME 4.0.





**DOME 4.0**

## Register Catalog Data

**Note:**  
This form is used for registering an external dataset to the DOME 4.0 platform.

**Keywords\***

**Title of Dataset\***

**Data Creator\***

**Data Publisher\***

**Issue Date\***

**Data URL\***

**Conforms to a Data Standard:**

- Optimade api specification
- Ids api specification

Figure 2: Process for registering new catalogue data with DOME

The URL provided in the data catalog registration serves as the access point for the dataset. This URL does not necessarily need to be publicly accessible; it can point to data that resides behind an external login or within an internal network. The URL ensures that users and tools on the DOME 4.0 platform have a designated link to locate the dataset, whether it is openly accessible or restricted. This setup supports secure data access, while still allowing datasets to be searchable and usable within the DOME 4.0 ecosystem.

## 2.2.2 Usage

Users can perform free-text searches across all registered data catalogs on the DOME platform through a semantic discovery service. This search functionality provides results that include only the metadata of the datasets, delivered in JSON format, allowing users to quickly review the essential details of each catalog entry.

## 2.3 Ontology

### 2.3.1 Onboarding

Ontologies can be uploaded to and stored on DOME 4.0 platform website [Upload \(dome40.io\)](https://dome40.io). At the time of writing, the ontology file has to be in Turtle format. The content of the file will be stored in an individual graph, identified by a graph URI supplied by the user.



## Upload An Ontology

### Note:

This is a pre-release, that only accepts .ttl files, a more general upload will be implemented shortly.

Choose File No file chosen

The data will reside in a separate graph (a name space) please provide a valid Graph URI:

For example: "https://your\_organisation\_uri/yourname/dataname" ,  
replace organisation etc as appropriate.

Upload

Figure 3: ontologies can be uploaded to DOME.

### 2.3.2 Usage

All uploaded ontologies can be discovered on DOME 4.0 platform website ([Upload \(dome40.io\)](https://dome40.io)) as shown in Figure 4. They can be visualised by the visualisation tool (See Section 4.1).



## Uploaded Ontologies

**Note:**

Choose which Ontology to Explore and Visualise!

<a href="http://dome40.io/dataset/data/dome40_core_dataset_trial0_reasoned">http://dome40.io/dataset/data/dome40_core_dataset_trial0_reasoned</a>	<a href="#">Visualise</a>
<a href="http://dome40.io/dataset/data/platforms_dome_core_reasoned_Hermit">http://dome40.io/dataset/data/platforms_dome_core_reasoned_Hermit</a>	<a href="#">Visualise</a>
<a href="https://w3id.org/function/ontology/1.0.0">https://w3id.org/function/ontology/1.0.0</a>	<a href="#">Visualise</a>
<a href="http://dome40.io/dataset/data/dome-all-data">http://dome40.io/dataset/data/dome-all-data</a>	<a href="#">Visualise</a>
<a href="https://imd.ucl.io/miso">https://imd.ucl.io/miso</a>	<a href="#">Visualise</a>
<a href="https://db1">https://db1</a>	<a href="#">Visualise</a>
<a href="https://simulation/test">https://simulation/test</a>	<a href="#">Visualise</a>

Figure 4: Uploaded ontologies are discoverable from a list.

In addition, the “advanced search” ([Upload \(dome40.io\)](#)) allows users to search through all uploaded ontologies. User may specify substrings to filter subjects, predicates or objects of uploaded ontologies. The result will be shown by the visualisation tool (See Section 4.1).



## Advanced Search (Simplified SPARQL Queries)

**Note:**

This allows triples to be searched through string matching of subject, predicate or object.

Subject

Predicate

Object

Submit Query

Figure 5: Advanced search for uploaded ontologies.

### 3. Data hand-off

The DOME 4.0 platform stores information about both data providers (described in Section 0) and tools and services which processes data. A mechanism is needed to enable seamless dataflow between compatible data providers and tools.

In the DOME 4.0 ecosystem, relationships between data sources and tools are defined semantically using the ontology developed in T3.2. Data connectors and catalogue data can conform to API specifications, while tools and services may also adhere to these standards. This alignment facilitates connections between data providers and consumers.

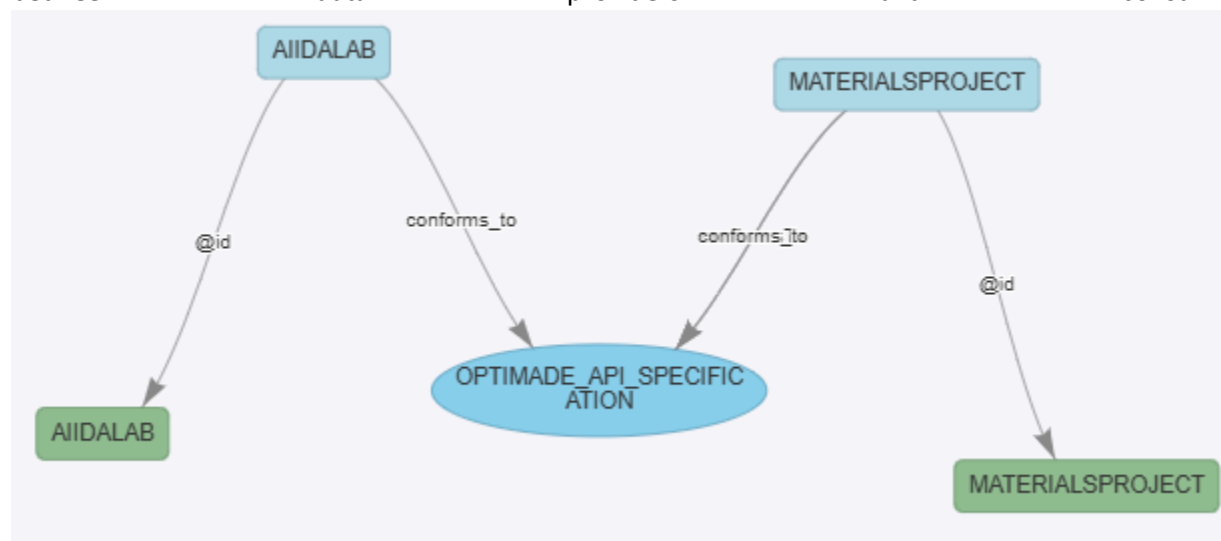
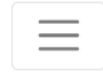


Figure 6: An example of semantic link between data connector and tools. Here "MATERIALSPROJECT" is a data connector and "AIIDALAB" is a tool. Both conform to the "OPTIMADE\_API\_SPECIFICATION".

When users search for data, compatible tools are suggested to them, enhancing the usability and efficiency of the platform. This feature ensures that users can easily find and utilise the tools that best match their data needs.

#### 3.1 Onboarding

User may register new standards to the DOME 4.0 platform via a registration form ([Register API specification \(dome40.io\)](https://dome40.io)):



## Register API specification

**Note:**

This form is used for registering API specification on DOME. Data and tools are linked if they both conform to the same API specification.

**Name of API specification\*** [i](#)

**API specification documentation URL** [i](#)

Register

Figure 7: registration form for API specification.

Registered standards will appear as options on the registration forms of data connector, catalogue data and tools. User may specify which standards a connector/tool complies with.



### Register DOME 4.0 Connector

**Note:**  
This form is used for registering DOME 4.0 connectors adhering to the Connector API (<https://github.com/DOME-4-0/reference-connector>). This allows external data platform to be discoverable on DOME.

Name of your connector as it will show up in the list of providers\* [?](#)

Human readable description of your connector [?](#)

Conforms to a Data Standard: [?](#)

- Optimade api specification
- Ids api specification



### Register DOME 4.0 Tool or Service

**Note:**  
This form is used for registering tools or services adhering to the DOME 4.0 tools and service API (<https://github.com/DOME-4-0/Tools-Services-Plugin-Template>). This allows data found through DOME to be processed by external tools or services.

Name of Tool or Service\* [?](#)

Tool/Service Description [?](#)

Conforms to Standard [?](#)

- Optimade api specification
- Ids api specification

Figure 8: registered API specifications will appear as options for registration of connectors and tools.

## 3.2 Usage

When user searches for data via the DOME 4.0 platform, the platform will perform an internal search for compatible tools based on the information stored in the triplestore. If any compatible tools are found, they will appear as buttons on the result page as shown in Figure 9. Clicking this button will redirect users to the tool with the metadata of the dataset being passed to the tool for further processing.

## Materials Project 1

The Materials Project provides open web-based access to computed information on known and predicted materials as well as powerful analysis tools to inspire and design novel materials.

### Metadata

```
{
  "Dataset": [
    "Dataset of C206Zn2 structures"
  ],
  "IssueDate": "2022-07-01T21:21:12Z",
  "License": "materialscloud.org",
  "Title": "C206Zn2",
  "URL": "https://aiida.materialscloud.org/mc3d/optimade/v1/structures/77806",
  "dataCreator": "materialscloud.org",
  "dataPublisher": "materialscloud.org",
  "keyword": "C206Zn2"
}
```

### Data

```
{
  "attributes": {
    "_mcloud_ctime": "2019-08-15T00:18:16Z",
    "assemblies": null,
    "cartesian_site_positions": [
      [
        2.356003244,
        1.3602391071,
        5.0009086005
      ],
      [
        0.0,
        0.0,
        7.6213630207
      ]
    ]
  }
}
```

**Free Platform**  
true

**Domain**  
NATURAL\_SCIENCES

**Offers**  
MODELLING\_DATA

**Home Page**  
<https://materialsproject.org/>

**Conforms to standard** 2  
OPTIMADE\_API\_SPECIFICATION

**FAIR score(s)**  
[FOOPS!](#) score: 4%

**Query URL** ?  
<https://nextgen.dome40.io/api/discover/results/MATERIALSPROJECT>

**DataInstance URL** ?  
[https://nextgen.dome40.io/api/discover/results/datum/MATERIALSPROJECT?search\\_string=carbon;zinc&keyword=C206Zn2](https://nextgen.dome40.io/api/discover/results/datum/MATERIALSPROJECT?search_string=carbon;zinc&keyword=C206Zn2)

Open in AIIDA LAB 3

Figure 9: An example of data hand-off between data provider and data consumer through DOME 4.0 platform. User may send data discovered from a connector (1) to a compatible tool (3) given that they both conform to a certain standard (2).



## 4. Advanced tool for data analysis

### 4.1 Visualisation tool

MedusaVis is an advanced visualisation tool in the DOME 4.0 platform for exploring ontologies and semantic data. It visualises ontologies as interactive network graphs, with nodes representing concepts and edges showing relationships.

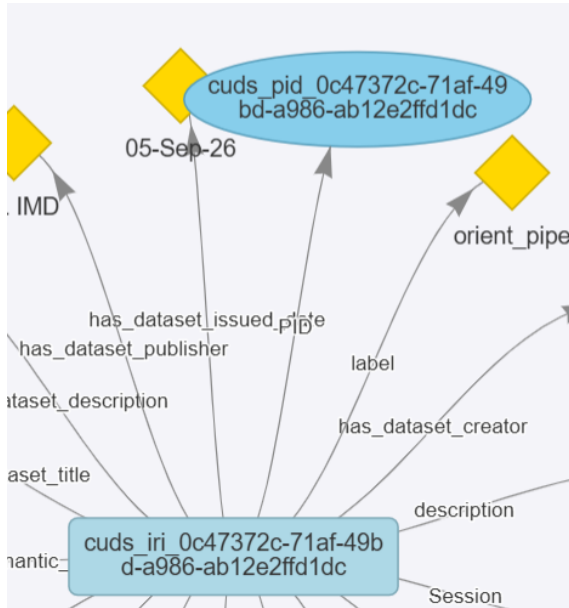


Figure 10: example output of MedusaVis. Properties are shown as rectangular boxes, values are represented as diamonds, and resources are displayed as ellipses.

Users can explore search results via multiple means, including zooming and panning, opening linked resources (by double-clicking), and changing the graph layout options.

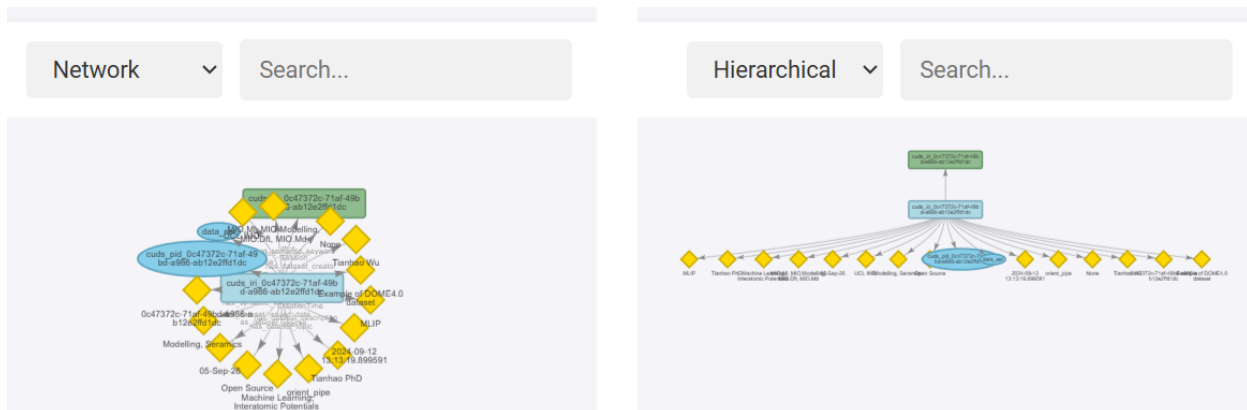


Figure 11: multiple layout options are available to users, including "Network" (left) and "Hierarchical" (right).

MedusaVis enhances DOME 4.0's semantic features by making complex ontological structures and search results more accessible and intuitive to users across various domains and expertise levels. Its combination

of semantic parsing and interactive visualisation enables users to gain insights that may not be apparent from text-based representations alone.

## 4.2 Support for data analytic tools

DOME 4.0 offers comprehensive support for bespoke data analytic tools through its public search APIs. Customised tools may be developed to search through the DOME 4.0 platform and analyse the gathered data.

There are two main APIs available to facilitate this process. The first API allows for a free-text search across all registered connectors and catalogue data, providing a broad search capability. This API takes a `search\_string` parameter. The result is returned in JSON format.

```
9 res_conn_query = requests.get(  
10     "https://nextgen.dome40.io/api/discover/results?search_string=carbon,zinc",  
11     timeout=10,  
12 )  
13  
14 print(res_conn_query.json())
```

PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS   GITLENS

Output of DOME search:  
=====

```
[{'data': [{'attributes': {'_cod_Rall': 0.062, '_cod_Z': '1', '_cod_Zprime': 0.25,  
. ; Zabukovec Logar, N.; Patarin, J.; Kaucic, V.', '_cod_b': 10.8477, '_cod_beta': 9  
P16 Zn8 -', '_cod_cellformula': '- C16 H96 N8 O68 P16 Zn8 -', '_cod_chemname': 'Oc  
hydrogenphosphato(V)octazincate tetrahydrate', '_cod_date': '2020-10-21', '_cod_do  
cod_firstpage': '373', '_cod_flags': 'has coordinates', '_cod_formula': '- C16 H96  
_journal': 'European Journal of Solid State and Inorganic Chemistry', '_cod_lastpag  
: 'C -2yc', '_cod_sgNumber': '9', '_cod_siga': 0.0007, '_cod_sighb': 0.0005, '_cod_s
```

Figure 12: sample call to public search API.

The second API enables a free-text search through a specific connector, allowing for more targeted searches. The specific API for a connector can be found on its own page as shown in Figure 13.



## Chemeo

Chemeo is an open, high quality chemical properties database.

**Free Platform**  
true

**Domain**  
NATURAL\_SCIENCES

**Offers**  
MATERIAL\_PROPERTY

**Home Page**  
<https://www.chemeo.com/>

**Query URL** ⓘ  
<https://nextgen.dome40.io/api/discover/results/CHEMEO>

### How to query

API keys need to be passed in the header along with the query url to access data. headers={"apikey": your-api-key } Please go to Access Keys under My Profile section and generate an API Key Also pass search\_string as a query parameter


Figure 13: the query URL of a connector is shown on its information page. This API call requires an API key to be included in the header.

The connector API requires an API key, which can be generated once user has logged in ([nextgen.dome40.io/access-keys](https://nextgen.dome40.io/access-keys)). The `Read` scope must be included when generating an API key for querying the connector API.



## Generate Access key

**Expiry Date\***



**Scopes\***

Read

Write

[Generate Key](#)

Figure 14: form to generate access key for connector API. The read scope must be included.

This dual API approach ensures that users can efficiently find and utilise the data they need for their specific analytic purposes.

## 5. Conclusions / Next steps

The DOME platform now supports onboarding data from various data sources, including creating connectors for external data platforms, registering the catalogue of generic external datasets, and uploading ontologies.

The platform also includes tools for data hand-off from providers to consumers, which include an extended ontology to semantically describe relationships between compatible providers and consumers. Users can now link providers and consumers during registration and trigger data hand-off from providers and consumers after searching for data.

Additionally, an advanced tool to support data analysis has been developed, including a visualization tool for exploration of semantic data and ontologies. A public search API enabling integration of bespoke data analytic tools and leverage DOME capabilities, secured through an API key, is also available.

These new features significantly enhance the capabilities of the DOME platform for data management and analysis. By providing connectors for external data platforms and a catalogue for generic external datasets, the platform facilitates efficient data onboarding from diverse sources. The extended ontology and data hand-off tools enable seamless data sharing and collaboration between providers and consumers. Furthermore, the advanced visualisation tool and public search API empower users to explore, analyse, and integrate data effectively.

## 6. Lessons learnt

One lesson was learnt on the necessity of facilitating the onboarding of data from diverse sources and formats. A balance was required between establishing common features and structures for efficient communication, exchange, and integration, while simultaneously accommodating the inherent heterogeneity of the data. The chosen solution adopts a hybrid approach. Metadata associated with data records are aligned with the DOME dataset ontology, while raw, free form, data is permitted in JSON format. This dual strategy fulfils two key objectives: data consumers can effectively search for and identify datasets from various DOME data providers through standardised metadata, and the interpretation of free-form raw data remains solely within the purview of individual data consumers, with DOME maintaining a neutral stance. This hybrid strategy fosters a flexible and scalable onboarding process. It provides a unified framework for data exchange and integration, while simultaneously embracing the diversity of data sources. This approach is crucial for ensuring that DOME can effectively meet the evolving needs of its user community.

Another important lesson learned was about the importance of user experience in data discovery and analysis. To address this need, a visualisation tool was developed, offering a clear and comprehensive view of the data. Various interactive features were incorporated to further enhance the user experience. One notable feature is the ability to open linked resources as URLs, providing a convenient way for users to search and access data through the DOME platform.

## 7. Deviations from Annex 1

N/A

## 8. References

N/A

## 9. Acknowledgement

The author(s) would like to thank the partners in the project for their valuable comments on previous drafts and for performing the review.

Project partners:

#	Type	Partner	Partner full name
1	SME	CMCL	Computational Modelling Cambridge Limited
2	Research	FHG	Fraunhofer Gesellschaft zur Förderung der Angewandten Forschung E.V.
3	Research	INTRA	Intrasoft International SA
4	University	UNIBO	Alma Mater Studiorum – Università di Bologna
5	University	EPFL	Ecole Polytechnique Federale de Lausanne
6	Research	UKRI	United Kingdom Research and Innovation
7	Large Industry	SISW	Siemens Industry Software NV
8	Large Industry	BOSCH	Robert Bosch GmbH
9	SME	UNR	Uniresearch B.V.
10	Research	SINTEF	SINTEF AS
11	SME	CNT	Cambridge Nanomaterials Technology LTD
12	University	UCL	University College London



*This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 953163. It is the property of the DOME 4.0 consortium and do not necessarily reflect the views of the European Commission.*

## 10. Table of Abbreviations

Abbreviation	Explanation
GA	Grant Agreement
SC	Showcase
TRL	Technology Readiness Level