Deliverable D4.1 - Report on data availability, data structure, and data requirements for each showcase

Responsible Partner:	UKRI	26 May 2022
Contributor(s):	Silvia Chiacchiera (UKRI)	26 May 2022
Reviewer(s):	Adham Hashibon (UCL), Amit Bhave (CMCL)	25 May 2022
Coordinator:	CMCL Innovations	28 May 2022
Dissemination Level:	Public	
Due Date:	M18 (May, 2022)	
Submission Date:	31.May.2022	

Project Profile

Programme	**Horizon 2020**
Call	**H2020-NMBP-TO-IND-2020-twostage**
Topic	**DT-NMBP-40-2020** **Creating an open marketplace for industrial data (RIA)**
Project number	**953163**
Acronym	**DOME 4.0**
Title	**Digital Open Marketplace Ecosystem 4.0**
Start Date	**December 1st, 2020**
Duration	**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	5 April 2022	Silvia Chiacchiera	Initial structure
V0.6	13 May 2022	Silvia Chiacchiera, Kok Foong Lee, Bijan Yadollahi, Natalja Schafet, Jörg Hohe, Evgeny Kharlamov, Michael Hoffmann, Vinicius Carrillo Beber, Kristof Vanclooster, Paula Martinez, Lazlo Farkas, Konstantinos Sipsas, Martin Uhrin, Noel Vizcaino, Vasily Bunakov, Martin Horsch, Michael Seaton, Ilian Todorov	Info on I/O data, file formats and platforms from showcases #1, 2, 3, 4, 5, 6, 7, 9 transferred to document. Abbreviations added.
V0.7	16 May 2022	Silvia Chiacchiera	Introduction, Conclusions added. Formatting, small additions. Elucidations added in Appendix.
V0.8	17 May 2022	Silvia Chiacchiera	Additions and cleaning up (all sections). Added Appendix 2, with example queries and requirements from partners.
V0.9	25 May 2022	Adham Hashibon	Review and proof reading
V0.95	26 May 2022	Silvia Chiacchiera, Kok Foong Lee, Natalja Schafet, Jörg Hohe, Martin Uhrin, Michael Hoffmann, Baifan Zhou	Addressed reviewer comments, added table for showcase #8, minor updates/additions to other tables
V1.0	28 May 2022	Silvia Chiacchiera, Michael Seaton, Baifan Zhou	Cleaned up caption formatting. Row added in Tables 15 and 16, cleaned up Table 13. Other minor additions/updates.
V1.1	28 May 2022	Amit Bhave	Review of the text and proof reading. Minor revisions.

Executive Summary

In this deliverable we focus on the nine DOME 4.0 showcases to collect information on data, metadata and platforms involved, aimed at informing the DOME 4.0 platform development. This activity is supported through individual interviews and accompanied by collation of example files. Additionally, example queries are gathered.

Table of Contents

List of Tables

1. Introduction

This document reports on the activities of Task 4.1 of DOME 4.0 project: "Metadata, Data Acquisition, Curation, and Communication". It summarises the data aspects for DOME 4.0 showcases, to support the development of the DOME 4.0 platform, with a focus on data, metadata and platforms. The material presented here has been iteratively collected during the period from M3 to M18 and collated in the form of a large spreadsheet, that is here broken down into smaller parts for presentation purposes. Realistic example files from the showcases have also been gathered within task 4.1 and have been made available internally, notably for usage within the core work packages (WP1-DOME 4.0 Platform, WP2- Data Tools and Services and WP3- Ontology-driven Interfaces). For example, they have been used to support the implementation of tools assessing data FAIRness (T2.2) and as example datasets for the Ecosystem information model (T3.2).

From the very beginning of T4.1, we decided to collect information in a relatively "dry" (i.e. in the form of schematics and tables, avoiding verbosity) format, which allows to have a good overview of the various data types to be handled and is suitable for the aim of this document. Based on the close connection with the partners and the core WPs, we have gathered feedback and iteratively adjusted the spreadsheet to the form presented here. We note in passing that, to describe datasets and platforms, we do not adopt here any specific schema or ontology (as e.g., EMMO, Dublin Core - DC, DCAT), but most of the concepts used are covered there too. We also note that specific user epics and stories, personas and workflows are not the focus of this document and will be described in other project deliverables (e.g., D1.2). For brief introductions to all DOME 4.0 use cases, we point the reader to the dedicated DOME 4.0 webpage[1].

This document is structured as follows: in Sections 2 to 10 we detail the data inputs and outputs of the individual showcases. In Sec. 11 we collect information on the *platforms* (intended as web-based databases or services) and in Sec. 12 the main file formats related to the showcases. Finally, we draw our Conclusions in Sec. 13 and, in Annex 1 – Elucidations and Annex 2 – Example of queries and additional requirements respectively, we give elucidations and examples of both queries and data consumer requirements.

[1] https://dome40.eu/dome-40-showcases

2. Showcase #1. Chemical kinetics Knowledge Graph (KG) – marine, air quality

Showcase #1 is developed by CMCL and addresses air quality in relation to ship trajectories and weather conditions, involving a combination of observations and modelling. Below, in Table 1, we show the data that are produced, shared and consumed in Showcase #1, grouped into four main categories by "data type": Ship location in real time, Weather data, Topography data and Dispersion model. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 1: Data involved in Showcase #1

1	Data type (high level)	Ship location in real time	Weather data	Topography data	Dispersion model
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I	I	I	O
3	Probability of usage in the showcases (sure, high, low)	sure	Sure	sure	sure
4	In which showcase(s)?	1	1	1	1
5	We have access already	No	Yes	Yes	Yes
6	Structure (File format(s) etc)	JSON	JSON	.HGT ("height")	Tab-separated text file
7	Intermediate structure within the use case (File formats we convert to etc)	RDF	RDF	GeoTIFF	RDF
8	Folder structure (yes/no + notes)	No	No	No (albeit more likely than the others on the left to be in a bundle with each filetype with a function, typically layers)	No
9	File size (order of magnitude, including all types)	kb	kb	80GB for entire world	kb-Mb
10	Open-source format?	yes	Yes	Should be	
11	File identifier				
12	Name of file (e.g., it contains relevant info/metadata)			Related to location coordinates	Location, timestamp

13	**Changes in time (NA=not applicable)**	Hourly (our use)	Hourly (our use)	No	NA
14	**Binary, human readable? (B, H)**	H	H	B	H
15	**Concrete example files (paths)**	Stored internally in DOME4.0	Stored internally in DOME4.0	Stored internally in DOME4.0	Stored internally in DOME4.0
16	**Data type in more detail: keywords**	latitude, longitude, ship ID, speed, size	temperature, humidity, precipitation (mm/h)	height profile	Concentration profiles of hydrocarbons (XYZ coordinates + concentration)
17	**Is the structure documented?**	Yes	Yes	Maybe (not user friendly)	No (own format)
18	**Link to documentation**	[URL][2]	[URL][3]	[URL][4]	None
19	**How is the metadata given?**	JSON keys	JSON keys	Not sure yet	column headers
20	**Does it follow a schema?**	No	No	No	No
21	**Name/link of schema followed**				
22	**Where is the data now? (Data source)**	Need subscription data, dummy data available locally, stored in RDF	Queried from API (openweather), then converted to RDF	Hard drive, downloaded from source	Hard drive, file location stored in RDF
23	**On a platform? (yes/no)**	Yes	yes	yes	Not yet
24	**Platform name and acronym**	Vessel finder, AIS hub	OpenWeather		
25	**Platform home URL**	[URLs][5]	openweathermap.org		
26	**Data provider (name or type)**	Vessel finder, AIS hub	openweather	NASA	CMCL
27	**Data owner (name or type)**	Same as provider	Same as provider	Same as provider	CMCL + client
28	**Data access rights (license, registration or fee needed)**	fee	free (usage dependent)	free	Probably: view data (not raw file)

[2] https://www.vesselfinder.com/realtime-ais-data, https://www.aishub.net/api
[3] https://openweathermap.org/api
[4] http://www.viewfinderpanoramas.org/Coverage%20map%20viewfinderpanoramas_org3.htm
[5] www.vesselfinder.com, https://www.aishub.net/

29	**Interface (connector, etc)**	API (HTTP)	API (HTTP)	Webpage	Webpage, UI
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	Yes	Yes	No	Yes
31	**Name+acronyms of the ontology(ies)**	OntoShip	OntoStation	N/A	
32	**Documentation on the ontologies (paper URL)**				
33	**URLs of the ontologies**	[URL][6]	[URL][7]	N/A	N/A
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**	none	none	none	none
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	long/lat points, Bounding Boxes/GeoJSON Applicable Coordinate Reference System - CRS (e.g., EPSG:4326)	long/lat points, Bounding Boxes/GeoJSON Applicable CRS (e.g., EPSG:4326)	Likely to have associated geospatial metadata. ==> long/lat points, Bounding Boxes/GeoJSON Applicable CRS (e.g., EPSG:4326)	long/lat points, Bounding Boxes/GeoJSON Applicable CRS (e.g., EPSG:4326)
36	**Notes**	Gets coordinates of ships within a coordinate box. Raw data from JSON; within the use case, converted to RDF with own (CMCL) Ontology for Ships		Downloaded all files, that are named according to coordinates	Data produced “on-demand”
37	**Specific libraries and/or tools required to read data**			GDAL[8]	
38	**What we use to read it**	Python script, Java script	Java		

[6] https://github.com/cambridge-cares/TheWorldAvatar/blob/main/JPS_Ontology/ontology/ontoship/OntoShip.owl
[7] https://github.com/cambridge-cares/TheWorldAvatar/blob/main/JPS_Ontology/ontology/ontostation/OntoStation.owl

[8] https://gdal.org/

3. Showcase #2. Light weight construction – fibre reinforced plastics

Showcase #2 is developed by Fraunhofer IWM and BOSCH and addresses fibre-reinforced plastics in a combination of experiments and modelling. Below, in Table 2, we show the data that are produced and consumed in Showcase #2, grouped into six main categories by "data type": Experimental Creep data, Numerical Creep data, Material cards, Finite Elements (FE) simulation results and Experimental data at product level. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 2: Data involved in Showcase #2

1	Data type (high level)	Experimental Creep data (at level of composite recyclate material samples)	Numerical Creep data (microstructure + creep data for matrix recyclate material)	Material card for anisotropic recyclate material (available in Ansys)	Finite Elements (FE) simulation results	Experimental data at product level for FE-results validation
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I/O	I/O	O	O	O
3	Probability of usage in the showcases (sure, high, low)	sure	sure/high	sure	sure	sure/high
4	In which showcase(s)?	2	2	2	2	2
5	We have access already	No	No	No	no	No
6	Structure (File format(s) etc)	Excel or ascii	Excel or ascii	ascii	Calibrated material parameters: ascii Glass Fiber (GF)-orientation: .cof, .elem CAD-Model of product: .STEP FE-Model: .cdb	Ascii
7	Intermediate structure within the use case (File	csv or xlsx	csv or xlsx	ASCII	ASCII	csv or xlsx

	formats we convert to etc)					
8	**Folder structure (yes/no + notes)**	no	no	no	no	No
9	**File size (order of magnitude, including all types)**	1 GB (300MB x 32 measurements)	0.5 GB	<1KB	0.5-2TB	1GB
10	**Open-source format?**	Depends	Office License	Ansys[9] license Excel license IWM-internal calibration software (SW) for anisotropic materials	Excel license Converse, calibration SW for anisotropic materials Ansys license	Zwick-SW Shaker/Vibration-SW
11	**File identifier**	No	No	different	different	Different
12	**Name of file (e.g., it contains relevant info/metadata)**	Could contain specimen #, test conditions	Could contain material name, environmental conditions, math. Settings	Could contain material name, environmental conditions, math. Settings	Could contain product name, environmental conditions, math. Settings	Could contain product name, environmental conditions, math. Settings
13	**Changes in time (NA=not applicable)**	No	yes, can be extended by additional computations	yes, depending on the choice of material card type	yes, depending on PhD-student capacity as well as on availability of required experimental data	yes, depending on PhD-student capacity as well as on availability of required experimental data
14	**Binary, human readable? (B, H)**	H	H	H	H	H
15	**Concrete example files (paths)**	Stored internally in DOME 4.0			Stored internally in DOME 4.0	

[9] In particular, Ansys Mechanical finite element analysis program is used.

16	**Data type in more detail: keywords**	Creep curves, creep strain vs time, rapture time, rapture strain	For composite material: Creep curves, creep strain vs time, rapture time, rapture strain	Material card available in Ansys calibrated using IWM-internal SW for Mapping in Converse with Hill-potentials from Converse (which will be created by Bosch based on experimental data)	- injection simulation for product -> GF-orientation - FE-model build-up - Mapping of Material Card in Converse -Write out the APDL-File with anisotropic material data for Ansys	- Tests of clamping forces in the plastic part - Vibration test
17	**Is the structure documented?**	No for ascii	in some cases yes, in other cases not, structure might be different	manual of Converse can be received from PartGmbH after the registration	yes	in some cases yes, in other cases not, structure might be different
18	**Link to documentation**					
19	**How is the metadata given?**	Column/row headers	Column/row headers	Column/row headers	Column/row headers	Column/row headers
20	**Does it follow a schema?**	No, but standard concepts: DIN EN ISO 899-1 - 2 directions (0°, 90°); - 4 temperatures (Tmin/Tmax and 2 x around the Tg) - 2 loads (%of_UTS); - 2 Repetitions;	No, but standard concepts: '-%, dimensions, orientation of GF (Glass Fiber) - Creep material data for matrix based on experimental data - The cell-size depends on specific properties of the final composite material	No (see notes for workflow-related info)	No (see notes for workflow-related info)	Bosch-internal, depends on target of experiment on product level

21	**Name/link of schema followed**	N/A	N/A	N/A	N/A	N/A
22	**Where is the data now? (Data source)**	Local machines at IWM	Local machines at IWM	Local machines at IWM	local machines at Bosch	local machines at Bosch
23	**On a platform? (yes/no)**	No	No	No	No	No
26	**Data provider (name or type)**	IWM	IWM	IWM	Bosch	Bosch
27	**Data owner (name or type)**	IWM	IWM	IWM / Bosch	Bosch	Bosch
28	**Data access rights (license, registration or fee needed)**	Confidential	Confidential	Confidential	Confidential	Confidential
29	**Interface (connector, etc)**	N/A	N/A	N/A	N/A	N/A
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	No	No	No	No
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**	not used	not used	not used	not used	not used
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	Column/row headers	Column/row headers	material name/acronym	material name/acronym	Column/row headers
37	**Specific libraries and/or tools required to read data**	Excel license	Excel license	Excel license	Ansys license	Excel license
38	**What we use to read it**	Excel license	Excel license	Excel license	Ansys license	Excel license

4. Showcase #3. Polymeric additives for coatings: anti-corrosion

Showcase #3 is developed by Fraunhofer (FHG) IFAM and SIEMENS (SISW), and addresses anti-corrosion coatings, in a combination of different publicly available data sources, experiments and Machine Learning (ML) modelling. Below, in Table 3 and Table 4 we show the data that are produced and consumed in Showcase #3, grouped into nine (five in the first table, four in the second) main categories by "data type": Chemical Identifiers, Training data, Hansen Solubility Parameter, Chemical information, Toxicology information, Corrosion data, Spectroscopic data, and Material Safety Data Sheet (MSDS). To avoid repetitions, data concerning the ML modelling that are relevant to multiple showcases will be shown only once, when Showcase #9 is addressed (see Section 11). For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 3: Data involved in Showcase #3 (first set)

1	Data type (high level)	Chemical Identifiers - SMILES string	Model - Training Data	Hansen Solubility Parameter (HSP)	Chemical information, e.g. GHS (Globally Harmonized System) Classification	Toxicology Information
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I	I	I/O	O	O
3	Probability of usage in the showcases (sure, high, low)	high	high	high	high	high
4	In which showcase(s)?	3	3	3	3	3
5	We have access already	Yes	Yes	Yes	Yes	Yes
6	Structure (File format(s) etc)	ASCII-string	csv	csv	json	json
7	Intermediate structure within the use case (File formats we convert to etc)					
8	Folder structure (yes/no + notes)	No	No	No	No	No
9	File size (order of magnitude, including all types)	kB	MB	few MB	kB	kB

10	**Open-source format?**	Yes	Yes	Yes	Yes	Yes
11	**File identifier**	NA	csv	csv	NA	NA
12	**Name of file (e.g., it contains relevant info/metadata)**	NA	custom	custom	NA	NA
13	**Changes in time (NA=not applicable)**	No	NA	No	NA	NA
14	**Binary, human readable? (B, H)**	H	H	H	H	H
15	**Concrete example files (paths)**	CC(=O)OC1=CC=CC=C1C(=O)O	[URL][10]	Stored internally in DOME 4.0	[URL][11]	[URL][12]
16	**Data type in more detail: keywords**	chemical identifier	Data which is needed to train the model	data from HSP experiments (output case) / HSP data used for models (input case)	GHS hazard statement - code and phrase	Predicted toxicity
17	**Is the structure documented?**	NA	No	custom	Yes	Yes
18	**Link to documentation**	[URL][13]			[URL][14]	[URL][15]
19	**How is the metadata given?**	NA	column header	column header	NA	NA
20	**Does it follow a schema?**					
21	**Name/link of schema followed**					
22	**Where is the data now? (Data source)**	user input either as list or ASCII-file	public	local	public database	public service
23	**On a platform? (yes/no)**	No	Yes	No	Yes	Yes
24	**Platform name and acronym**		ACS.org		Pubchem	ProTox-II
25	**Platform home URL**		[URL][16]		[URL][17]	[URL][18]

[10] https://pubs.acs.org/doi/suppl/10.1021/ci034243x/suppl_file/ci034243xsi20040112_053635.txt

[11] https://pubchem.ncbi.nlm.nih.gov/compound/240#section=Safety-and-Hazards
[12] https://tox-new.charite.de/protox_II/index.php?site=faq#API

[13] https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system
[14] https://en.wikipedia.org/wiki/GHS_hazard_statements
[15] https://tox-new.charite.de/protox_II/index.php?site=faq#API
[16] https://pubs.acs.org/
[17] https://pubchem.ncbi.nlm.nih.gov/
[18] https://tox-new.charite.de/protox_II/

26	**Data provider (name or type)**		public	IFAM	Pubchem	Charite University of Medicine
27	**Data owner (name or type)**		public	IFAM	public	public
28	**Data access rights (license, registration or fee needed)**			confidential	Free	Free
29	**Interface (connector, etc)**	GUI or API	GUI		REST API	REST API
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No		No	would be possible	No
31	**Name+acronyms of the ontology(ies)**				PubChemRDF	
32	**Documentation on the ontologies (paper URL)**				[URL][19]	
33	**URLs of the ontologies**				[URL][20]	
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**					
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	NA	chemical identifier, physico-chemical property(s)	chemical identifier, HSP values	chemical identifier, GHS hazard statement	chemical identifier, parts of the toxicity data

[19] https://jcheminf.biomedcentral.com/articles/10.1186/s13321-015-0084-4

[20] https://pubchemdocs.ncbi.nlm.nih.gov/rdf

Table 4: Data involved in Showcase #3 (second set)

1	Data type (high level)	Corrosion data (Filiform, NSST-neutral salt spray test)	Spectroscopic data (e.g. X-ray photoelectron spectroscopy)	Chemical and Physical Properties (e.g., molecular weight)	Material Safety Data Sheet (MSDS)
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I	I	O	I
3	Probability of usage in the showcases (sure, high, low)	medium	low	medium	low
4	In which showcase(s)?	3	3	3	3
5	We have access already	Yes	Yes	Yes	Yes
6	Structure (File format(s) etc)	picture format (jpg, png); results of corrosion evaluation as csv or xlsx	xlsx, csv	json	pdf
7	Intermediate structure within the use case (File formats we convert to etc)				
8	Folder structure (yes/no + notes)	No	No	No	No
9	File size (order of magnitude, including all types)	10 MB per picture; KB for aggregated data	few MB per file	kB	few MB
10	Open-source format?	depends	Yes	Yes	depends
11	File identifier	NA	xlsx/csv	NA	NA
12	Name of file (e.g., it contains relevant information/metadata)	custom	custom	NA	custom
13	Changes in time (NA=not applicable)	No	No	NA	NA
14	Binary, human readable? (B, H)	B (H for aggregated data)	H	H	B
15	Concrete example files (paths)	Stored internally in DOME 4.0	Stored internally in DOME 4.0	[URL][21]	Stored internally in DOME 4.0

[21] https://pubchem.ncbi.nlm.nih.gov/compound/2244#section=Chemical-and-Physical-Properties

16	**Data type in more detail: keywords**	image from sample after corrosion experiment; aggregated data evaluate the amount of corrosion	data from spectroscopic experiment e.g. chemical composition of a surface	key/value pair	MSDS sheets for chemicals/materials
17	**Is the structure documented?**	No	No	NA	NA
18	**Link to documentation**				
19	**How is the metadata given?**	column header for aggregated data	column header	NA	
20	**Does it follow a schema?**				
21	**Name/link of schema followed**				
22	**Where is the data now? (Data source)**	local	local	public service / database	local
23	**On a platform? (yes/no)**	No	No	Yes	No
24	**Platform name and acronym**			PubChem	
25	**Platform home URL**			[URL][22]	
26	**Data provider (name or type)**	IFAM	IFAM	Pubchem	depends
27	**Data owner (name or type)**	IFAM	IFAM	public	depends
28	**Data access rights (license, registration or fee needed)**	confidential	confidential	Free	
29	**Interface (connector, etc)**			REST API	
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	No	would be possible	No
31	**Name+acronyms of the ontology(ies)**			PubChemRDF	
32	**Documentation on the ontologies (paper URL)**			[URL][23]	
33	**URLs of the ontologies**			[URL][24]	

[22] https://pubchem.ncbi.nlm.nih.gov/
[23] https://jcheminf.biomedcentral.com/articles/10.1186/s13321-015-0084-4
[24] https://pubchemdocs.ncbi.nlm.nih.gov/rdf

34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**				
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	few meta data (experiment, material)	few meta data (experiment, material)	chemical identifier, physico-chemical property(s)	chemical identifier

5. Showcase #4. Structural adhesives: Fatigue behaviour

Showcase #4 is developed by Fraunhofer (FHG) IFAM and SIEMENS (SISW), and concerns structural properties of joints, in a combination of experiments and ML modelling. Below, in Table 5 we show the data that are produced and consumed in Showcase #4, grouped into three main categories by "data type": Adhesive and substrate properties, Joint properties, and Fatigue data. As described earlier, the data concerning the ML modelling will be presented for Showcase #9 (see Section 11) to avoid repetition. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 5: Data involved in Showcase #4

1	Data type (high level)	Adhesive and substrate properties	Joint properties	Fatigue data
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I/O	I/O	I/O
3	Probability of usage in the showcases (sure, high, low)	Sure	sure	sure
4	In which showcase(s)?	4	4	4
5	We have access already	Yes	Yes	Yes
6	Structure (File format(s) etc)	.XLSX (Excel)	.XLSX (Excel)	.XLSX (Excel)
7	Intermediate structure within the use case (File formats we convert to etc)	CSV	CSV	CSV
8	Folder structure (yes/no + notes)	No	No	No
9	File size (order of magnitude, including all types)	100 kB	100 kB	100 kB
10	Open-source format?	No	No	No
11	File identifier			
12	Name of file (e.g., it contains relevant information/metadata)	Not completely consistent, but related to materials and sample	Not completely consistent, but related to materials and sample	Not completely consistent, but related to materials and sample
13	Changes in time (NA=not applicable)	NA	NA	NA
14	Binary, human readable? (B, H)	H	H	H
15	Concrete example files (paths)	Stored internally in DOME 4.0		

16	**Data type in more detail: keywords**	Commercial Name, Data Sheet Information (e.g. tensile strength, tensile modulus, glass Transition temperature, lap shear strength)	Geometry (type - e.g., lap shear, butt- and parameters) and materials (substrate + adhesive), properties	Sample lifetime (i.e., number of cycles till a crack forms or other criteria is satisfied)
17	**Is the structure documented?**	No	No	No
18	**Link to documentation**	N/A	N/A	N/A
19	**How is the metadata given?**	column headers	column headers	column headers
20	**Does it follow a schema?**	Not, but common variables in the field	Not, but common variables in the field	Not, but common variables in the field
21	**Name/link of schema followed**			
22	**Where is the data now? (Data source)**	Commercial data Sheets	Experimental, from publications and reports (external and internal to Fraunhofer)	Experimental data (publications in .pdf, or reports from projects → Mainly human extraction, from images or tables). Output from ML too.
23	**On a platform? (yes/no)**	No	no	no
26	**Data provider (name or type)**	Adhesive supplier or Fraunhofer	Fraunhofer	Fraunhofer
27	**Data owner (name or type)**	Fraunhofer	Fraunhofer	Fraunhofer
28	**Data access rights (license, registration or fee needed)**	So far not open, fee needed in the future (for non-publicly funded projects)	So far not open, fee needed in the future (for non-publicly funded projects)	So far not open, fee needed in the future (for non-publicly funded projects)
29	**Interface (connector, etc)**	none	none	none
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	No	No
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**			
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**			

36	**Notes**	Note: .pdfs are on-line, need to be accessed and processed)	Note: Some samples do follow standards, e.g. ASTM D1002 (it depends on the application, e.g. railways, aircraft). NOT in our data. Note: .pdfs are on-line (so can be platforms), but need to be accessed and processed	Excel tables from experiments

6. Showcase #5. Production equipment tools and service catalogues (metals, plastics, high-tech)

Showcase #5 is developed by INTRA and concerns the connection to and data sharing aspects with MARKET4.0 and other digital marketplaces, in the context of equipment and services for manufacturing. Below, in Table 6 we show the data that are produced and consumed in Showcase #5, grouped into four main categories by "data type": Equipment/Service information from MARKET4.0 catalogue, Information on DOME 4.0 platform tools, Product catalogue of material marketplace and International Data Spaces (IDS) equipment data. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 6: Data involved in Showcase #5

1	Data type (high level)	Equipment/service information from MARKET4.0 catalogue	Information on DOME 4.0 platform tools	Product catalogue of material marketplace	IDS Equipment data
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I/O	I/O	I/O	O
3	Probability of usage in the showcases (sure, high, low)	sure	high	high	High
4	In which showcase(s)?	5	5, maybe 6	5, maybe 2	5
5	We have access already	Yes	No	No	No
6	Structure (File format(s) etc)	JSON	JSON, preferably	JSON, preferably	JSON
7	Intermediate structure within the use case (File formats we convert to etc)	N/A			No
8	Folder structure (yes/no + notes)	Data is organized based on API methods (i.e. each method retrieves a subset of the data).			No
9	File size (order of magnitude, including all types)	In total around 500KB. Comes in pages so page size will limit the data size produced.	N/A	N/A	A few MBs at most.
10	Open-source format?	Yes	Yes, preferably	Yes, preferably	Yes

11	**File identifier**	Each catalogue entry has a unique id. The id is integer.	N/A	N/A	Each machine has a unique id.
12	**Name of file (e.g., it contains relevant information/metadata)**	N/A			N/A
13	**Changes in time (NA=not applicable)**	Depends on the suppliers.			Depends on the suppliers.
14	**Binary, human readable? (B, H)**	H	N/A	N/A	H
15	**Concrete example files (paths)**	[URL][25] and [URL][26]			Requires IDS infrastructure.
16	**Data type in more detail: keywords**	Production equipment/service descriptions for MARKET4.0 catalogue items coming from different suppliers.	Description of tools MARKET4.0 users can access over DOME 4.0. Endpoint information to forward users.	Catalogue description of material marketplace.	Machine properties e.g. material names processed, processing capabilities, etc.
17	**Is the structure documented?**	No			Yes
18	**Link to documentation**	N/A			Documentation available internally for DOME4.0 development purposes
19	**How is the metadata given?**	JSON property names			JSON property names
20	**Does it follow a schema?**	Yes, MARKET4.0 internal schema			Yes, MARKET4.0 internal schema
21	**Name/link of schema followed**	N/A			N/A

[25] http://platform.market40.eu/search/offerings?key=&page=0&size=12
[26] http://platform.market40.eu/search/offerings?key=prima&page=0&size=12

22	**Where is the data now? (Data source)**	MARKET4.0 portal	DOME 4.0 infrastructure	DOME 4.0 infrastructure or material marketplace infrastructure	Data is available over an IDS connection. The data is stored on data provider infrastructure.
23	**On a platform? (yes/no)**	yes			yes
24	**Platform name and acronym**	MARKET4.0			IDS
25	**Platform home URL**	[URL][27]		N/A	
26	**Data provider (name or type)**	Various MARKET4.0 partners.			MARKET4.0 suppliers
27	**Data owner (name or type)**	MARKET4.0			MARKET4.0 suppliers
28	**Data access rights (license, registration or fee needed)**	Free			Free for those that have a valid IDS certificate
29	**Interface (connector, etc)**	GUI, API	API, preferably	API, preferably	IDS API
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	Not yet - Ontologies are being developed (in WP3)		No
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**	not used			not used
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	Basic metadata e.g. machine name, manufacturer name, materials processed if present.			Basic metadata e.g. machine name, manufacturer name, materials processed if present.

[27] http://platform.market40.eu/index.html

7. Showcase #6. Turnkey services & custom workflows integrating simulations and data

Showcase #6 is developed by EPFL and addresses molecular simulations. Below, in Table 7 we show the data that are produced and consumed in Showcase #6, grouped into three main categories by "data type": Atomic structure, Simulation Settings and AiiDA Export data. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 7: Data involved in Showcase #6

1	Data type (high level)	Atomic structure	Simulation Settings	AiiDA export
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I/O	I	I/O
3	Probability of usage in the showcases (sure, high, low)	Sure	Sure	high
4	In which showcase(s)?	6	6	6
5	We have access already	Yes	Yes	Yes
6	Structure (File format(s) etc)	.CIF, .XYZ, .aiida	JSON, .aiida	.aiida
7	Intermediate structure within the use case (File formats we convert to etc)	.tar.gz	.tar.gz	.tar.gz
8	Folder structure (yes/no + notes)	Only the .aiida files have an internal structure, details here: [URL][28]	Only the .aiida files have an internal structure, details here	Only the .aiida files have an internal structure, details here
9	File size (order of magnitude, including all types)	kB-MB	kB-MB	MB-GBs
10	Open-source format?	Yes	Yes	Yes
11	File identifier	Yes (UUID)	Yes (UUID)	Yes (UUID)
12	Name of file (e.g., it contains relevant information/metadata)	Can be the UUID but no strict convention		
13	Changes in time (NA=not applicable)	No	No	No
14	Binary, human readable? (B, H)	B, H	B, H	B

[28] https://aiida.readthedocs.io/projects/aiida-core/en/latest/internals/data_storage.html

15	**Concrete example files (paths)**	Stored internally in DOME 4.0	Stored internally in DOME 4.0	Stored internally in DOME 4.0
16	**Data type in more detail: keywords**	atomic coordinates and species	typically, a dictionary of key-value pairs containing settings for a calculation	export file containing an AiiDA provenance graph (including all results and calculations)
17	**Is the structure documented?**	Yes	Yes	Yes
18	**Link to documentation**	[URL][29]	[URL][30]	[URL][31]
19	**How is the metadata given?**	None for XYZ, column headers for CIF	key-value pairs	Included in the export as provenance graph attributes
20	**Does it follow a schema?**	No	Some do	No
21	**Name/link of schema followed**		e.g., [URL][32]	
22	**Where is the data now? (Data source)**	Can come from various sources: clients HDD, OPTIMADE API, MaterialsCloud, etc	Comes from the client	Can be: supplied by client or downloaded from materialscloud.org
23	**On a platform? (yes/no)**	yes	No	Yes
24	**Platform name and acronym**	MaterialsCloud		MaterialsCloud
25	**Platform home URL**	www.materialscloud.org		www.materialscloud.org
26	**Data provider (name or type)**	Various (including EPFL)	The client	The client or materialscloud.org (various authors)
27	**Data owner (name or type)**	Various (typically the person who created it)	The client	The author

[29] https://www.iucr.org/resources/cif/documentation/cifguide https://en.wikipedia.org/wiki/XYZ_file_format

[30] https://aiida.readthedocs.io/projects/aiida-core/en/v1.6.3/topics/data_types.html#topics-data-types-core-base-iterable

[31] https://aiida.readthedocs.io/projects/aiida-core/en/latest/howto/share_data.html?highlight=create%20archive

[32] https://aiida.readthedocs.io/projects/aiida-core/en/v1.6.3/topics/data_types.html#topics-data-types-core-base-iterable

28	**Data access rights (license, registration or fee needed)**	Various (many free - Creative Commons, many commercial, e.g. ICSD[33])		Free and open (typically)
29	**Interface (connector, etc)**	OPTIMADE RestAPI	Interactively supplied	Export file or RESTAPI
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	No	No
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**	OPTIMADE		
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	Metadata (e.g. the chemical composition of the atomic structure, the volume, perhaps symmetry information, but not atomic coordinates)	Basic metadata (creation date, modification date, creator, etc)	Basic metadata (creation date, modification date, creator, etc)
37	**Specific libraries and/or tools required to read data**	Many python tools, e.g. AiiDA, ASE		AiiDA
38	**What we use to read it**	AiiDA python library		

[33] https://www.fiz-karlsruhe.de/en/produkte-und-dienstleistungen/inorganic-crystal-structure-database-icsd

8. Showcase #7. Formulated consumer products

Showcase #7 is developed by UKRI and combines experiments and modelling to address formulated consumer goods, typically in the fluid phase. Below, in Table 8 we show the data that are produced and consumed in Showcase #7, grouped into two main categories by "data type": Chemical and Physical properties and Molecular simulation (input) data. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 8: Data involved in Showcase #7

1	Data type (high level)	Chemical and Physical properties	Molecular simulation (input) data
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I	I/O
3	Probability of usage in the showcases (sure, high, low)	sure	high
4	In which showcase(s)?	7	7
5	We have access already	Yes	Yes
6	Structure (File format(s) etc)	GUI: Various download options available, including .XLSX and .PDF. JSON from the REST API.	Software specific. E.g., for DL_MESO_DPD[34], input files are: CONTROL, FIELD, CONFIG (optional)
7	Intermediate structure within the use case (File formats we convert to etc)		
8	Folder structure (yes/no + notes)		no (Typically, each simulation is run in a separate folder)
9	File size (order of magnitude, including all types)	10 kB	<kB (CONTROL and FIELD); CONFIG depends on simulation size (about 100kB per 1000 particles)
10	Open-source format?	No	Yes
11	File identifier		N/A
12	Name of file (e.g., it contains relevant information/metadata, is a unique identifier)	Compound name (for .PDF), Chemeo-generated code (for .XLSX)	File names need to be as expected by software
13	Changes in time (NA=not applicable)	Generally, no. But it could, if source changes (see Notes)	NA
14	Binary, human readable? (B, H)	H	H

[34] https://www.scd.stfc.ac.uk/Pages/DL_MESO.aspx

15	**Concrete example files (paths)**	Stored internally in DOME 4.0	
16	**Data type in more detail: keywords**	Compound chemical formula, SMILES, CAS number, IUPAC INCHI, alternative names, properties (e.g., critical temperature)	Simulation parameters for dissipative particle dynamics, DPD (both model and numerical settings), e.g.: population of species, force field parameters, boundary conditions, initial simulation state.
17	**Is the structure documented?**		Yes
18	**Link to documentation**		DL_MESO User Manual
19	**How is the metadata given?**	Row header (for .XLSX)	Keywords, own format
20	**Does it follow a schema?**	No, but includes widely used notations such as SMILES, CAS number, IUPAC INCHI[35]	No
21	**Name/link of schema followed**	N/A	N/A
22	**Where is the data now? (Data source)**	On Chemeo platform	local machines at UKRI
23	**On a platform? (yes/no)**	Yes	no
24	**Platform name and acronym**	Chemeo	N/A
25	**Platform home URL**	https://www.chemeo.com/	N/A
26	**Data provider (name or type)**	Ceondo	UKRI or modeller
27	**Data owner (name or type)**	Ceondo GmBH and original sources (NIST, etc)	UKRI or modeller
28	**Data access rights (license, registration or fee needed)**	Free	depends on system
29	**Interface (connector, etc)**	GUI, REST API	N/A
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	Not yet, but could be used
31	**Name+acronyms of the ontology(ies)**		VIMMP Ontologies
32	**Documentation on the ontologies (paper URL)**		[URL][36]
33	**URLs of the ontologies**		[URL][37]

[35] https://www.inchi-trust.org/
[36] https://doi.org/10.1007/978-3-030-68597-3
[37] https://gitlab.com/vimmp-semantics/vimmp-ontologies

34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**		
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	Metadata to identify the compound (name and synonyms, identifiers), data source	
36	**Notes**	"Note: Cheméo is only indexing the data, follow the source links to retrieve the latest data." Note 2: amount of information varies with compounds	Metadata to be attached to identify the material to which the model applies (could be given as text within the input files, not regulated)
37	**Specific libraries and/or tools required to read data**		DL_MESO_DPD and associated utilities
38	**What we use to read it**		DL_MESO_DPD and associated utilities

9. Showcase #8. Semantic Analytics of Manufacturing Assets

Showcase #8 is developed by BOSCH and addresses manufacturing. Below, in Table 9 we show the data that are produced and consumed in Showcase #8, grouped into four main categories by "data type": Welding raw data , Welding Knowledge Graph (KG) data , ML Model data , ML Model prediction results. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 9: Data involved in Showcase #8

1	Data type (high level)	Welding raw data	Welding KG data	ML Model data	ML Model prediction results
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I	I/O	O	O
3	Probability of usage in the showcases (sure, high, low)	Sure	Sure	Sure	Sure
4	In which showcase(s)?	8	8	8	8
5	We have access already	Yes	Yes	Yes	Yes
6	Structure (File format(s) etc)	csv	ttl	py json	Csv, json
7	Intermediate structure within the use case (File formats we convert to etc)				
8	Folder structure (yes/no + notes)	Single feature csv table Folder of time series csv table	No	No	No
9	File size (order of magnitude, including all types)	Few MB	Few MB	NA	Few MB
10	Open-source format?	Yes	Yes	Yes	Yes
11	File identifier	The primary key column of the single feature table contains all identifiers,	Project name as filename	Project name as filename	Project name as filename

		which will be used for the filenames of time series tables			
12	**Name of file (e.g., it contains relevant info/metadata)**	user determined	user determined	user determined	user determined
3	**Changes in time (NA=not applicable)**	Depends on availability of additional training data	Depends on availability of additional training data	Depends on availability of additional training data	Depends on availability of additional training data
14	**Binary, human readable? (B, H)**	H	H	B	H
15	**Concrete example files (paths)**	Stored internally in DOME 4.0	Stored internally in DOME 4.0	Stored internally in DOME 4.0	Stored internally in DOME 4.0
16	**Data type in more detail: keywords**	Welding raw datasets that contain welding control setting, parameters, sensor measurements and quality indicators	Structured welding data constructed from welding raw data with domain ontologies as the knowledge base and KG schema	Machine learning model information	Predicted quality indicators, edge or node prediction and classification
17	**Is the structure documented?**	Yes	Yes	Yes	Yes
18	**Link to documentation**	NA	NA	NA	NA
19	**How is the metadata given?**	Column headers	KG schema	ReadMe file	Column headers and filenames
20	**Does it follow a schema?**	Relational data schema	KG schema	No	No
21	**Name/link of schema followed**	NA	NA	NA	NA
22	**Where is the data now? (Data source)**	Bosch internal device	Bosch internal device	Bosch internal device	Bosch internal device
23	**On a platform? (yes/no)**	No	No	No	No
26	**Data provider (name or type)**	Bosch internal	Bosch internal	Bosch internal	Bosch internal
27	**Data owner (name or type)**	Bosch internal	Bosch internal	Bosch internal	Bosch internal

28	**Data access rights (license, registration or fee needed)**	Closed	Closed	Closed	Closed
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	Yes	Yes	Yes
31	**Name+acronyms of the ontology(ies)**	NA	RSW ontology Welding core ontology	ML ontology	ML ontology
32	**Documentation on the ontologies (paper URL)**	NA	[URL][38]	URL	URL
33	**URLs of the ontologies**	NA	NA	NA	NA
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**	ISO 14327:2004(E)	ISO 14327:2004(E)	ISO 14327:2004(E)	ISO 14327:2004(E)
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	Sample welding machine dataset	Sample KG	NA	Sample results
37	**Specific libraries and/or tools required to read data**	NA	NA	NA	NA
38	**What we use to read it**	Excel	Txt, Protege	Txt, json reader	Txt, image reader

[38] https://www.sciencedirect.com/science/article/pii/S1570826821000391

10. Showcase #9. Virtual development of composite materials

Showcase #9 is developed by SIEMENS (SISW) and addresses virtual development of composite materials which, for example, is relevant for the automotive industry. Below, in Table 10 we show the data that are produced and consumed in Showcase #9, grouped into four main categories by "data type". Please note that, as already mentioned, these data types will also be relevant to showcases #3 and #4. For clarifications on the row headers, cf. Annex 1 – Elucidations.

Table 10: Data involved in Showcase #9 (and also in showcases #3 and #4)

1	Data type (high level)	ML model - Training Data	ML Model - Definition and Evaluation	ML Model - Design Space	ML Model - Prediction and results
2	Within the showcases, is it Input, Output or both? (I, O, I/O)	I	I/O	I	I/O
3	Probability of usage in the showcases (sure, high, low)	sure	sure	sure	sure
4	In which showcase(s)?	(3),4,9	(3),4,9	(3),4,9	(3),4,9
5	We have access already	Yes	Yes	Yes	Yes
6	Structure (File format(s) etc)	CSV	.py	.py, CSV	.py, CSV, JSON
7	Intermediate structure within the use case (File formats we convert to etc)				
8	Folder structure (yes/no + notes)	No	No	No	No
9	File size (order of magnitude, including all types)	few 100 kB	NA	NA	few MB
10	Open-source format?	Yes	Yes	Yes	Yes
11	File identifier	Typically, one CSV file condensing all labelled input data			Export per project in CSV file with projectname as filename
12	Name of file (e.g., it contains relevant information/metadata)	user determined	user determined	user determined	user determined
13	Changes in time (NA=not applicable)	Depends on availability of additional training data	Depends on availability of additional training data	Depends on availability of additional training data	Depends on availability of additional training data

14	**Binary, human readable? (B, H)**	H	H	H	H
15	**Concrete example files (paths)**	Stored internally in DOME 4.0		Stored internally in DOME 4.0	Stored internally in DOME 4.0
16	**Data type in more detail: keywords**	Data which is needed to train the ML-model; it contains information on material property, composition, testing conditions,...	Assign the available training data to either input or output for the ML. Define the structure of the ML-model.	Define the material design space of interest	Predict material candidates based on the combination of the ML-model, design space and desirable target properties
17	**Is the structure documented?**	Yes	Yes	Yes	Yes
18	**Link to documentation**	Raw_data[39]	Predictor[40]	Design_Space [41]	Prediction[42]
19	**How is the metadata given?**	column headers	UI - interactive graph, graphical models	UI - input fields	column headers
20	**Does it follow a schema?**	Citrine	Citrine	Citrine	Citrine
21	**Name/link of schema followed**	Raw_data	Predictor	Design_Space	Prediction
22	**Where is the data now? (Data source)**	Raw data comes from the End-user and is uploaded (via API or GUI) to Citrine informatics, cloud tool	Direct input by the user - evaluation report is on the Citrine platform	Direct input by the user	Direct input by the user and on the Citrine platform (results)
23	**On a platform? (yes/no)**	yes	yes	yes	yes
24	**Platform name and acronym**	Citrine Informatics (tool provider, data sink)	Citrine Informatics (tool provider, data sink)	Citrine Informatics (tool provider, data sink)	Citrine Informatics (tool provider, data sink)
25	**Platform home URL**	Citrine Platform (citrine-platform.com)	Citrine Platform (citrine-platform.com)	Citrine Platform (citrine-platform.com)	Citrine Platform (citrine-platform.com)

[39] https://citrineinformatics.github.io/citrine-python/formulations_example.html#example-raw-data
[40] https://citrineinformatics.github.io/citrine-python/formulations_example.html#training-a-predictor
[41] https://citrineinformatics.github.io/citrine-python/formulations_example.html#defining-a-design-space
[42] https://citrineinformatics.github.io/citrine-python/formulations_example.html#proposing-new-formulation-candidates

26	**Data provider (name or type)**	Fraunhofer; SABIC	SISW	SISW	SISW
27	**Data owner (name or type)**	Fraunhofer; SABIC	SISW, Fraunhofer; SABIC	SISW; Fraunhofer; SABIC	Fraunhofer; SABIC
28	**Data access rights (license, registration or fee needed)**	Registration; fee needed	Registration; fee needed	Registration; fee needed	Registration; fee needed
29	**Interface (connector, etc)**	GUI, API	GUI, API	GUI, API	GUI, API
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	No	No	No	No
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**	NA	NA	NA	NA
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	All input data			All output data

11. Platforms entering the showcases

In Table 11 and Table 12 below, we summarise the properties of the platforms that enter the showcases. Please note that two of the columns (OPTIMADE and IDS ecosystem), rather than to individual platforms, refer to emerging standards/initiatives that aim to support interoperability.

In particular, we distinguish data providers in three categories (cf. Row 11), based on the different workflow they will require for integration within DOME 4.0:

- **Data (D)**: i.e., data that is there/available, one sends a request and gets the data
- **Data On Demand (DOD)**: i.e., data to be computed (one sends a request, gets a receipt, then the data - possibly after some time)
- **Interactive App (IA)**: i.e., one logs in and "does" something on the platform

Table 11: Platforms involved in the showcases (first set)

1	Platform name and acronym	Vessel Finder	AIS Hub	Elevation data	Weather	PubChem
2	Platform home URL	www.vesselfinder.com		http://www.viewfinderpanoramas.org/Coverage%20map%20viewfinderpanoramas_org3.htm	api.openweathermap.org	https://pubchem.ncbi.nlm.nih.gov/
3	Platform query URL (i.e., query end-point)	Need purchase to know	[query URL][43]	N/A	[query URL][44]	[query URL][45]
4	Platform documentation (via URL or similar)	https://www.vesselfinder.com/realtime-ais-data	https://www.aishub.net/api	http://viewfinderpanoramas.org/dem3.html	https://openweathermap.org/api	https://pubchemdocs.ncbi.nlm.nih.gov/about
5	Scope	Ship location	Ship location	Topography	Weather	open chemistry database at the National Institutes of Health (NIH)
6	Interface	REST	REST	GUI	REST	Web or REST API

[43] https://data.aishub.net/ws.php?username=A&format=B&output=C&compress=D&latmin=E&latmax=F&lonmin=G&lonmax=H&mmsi=I&imo=J&interval=K

[44] https://api.openweathermap.org/data/2.5/weather?lat={lat}&lon={lon}&appid={API key}

[45] https://pubchem.ncbi.nlm.nih.gov/rest/pug_view/data/compound/1234/JSON

7	**Does the API specification comply with a "standard" (yes/no + name + URL)**		Yes	No API	Yes	Yes
8	**How to access (registration or fee needed)**	Fee	Fee	Free	Free for now	Free
9	**Enters in Showcase(s) #**	1	1	1	1	3
10	**Type (P=provider, C=consumer, P/C=both)**	P	P	P	P	P
11	**Provider of: data (D), Data on Demand (DOD), Interactive App (IA)**	DOD	DOD	IA	DOD	D

Table 12: Platforms involved in the showcases (second set)

1	Platform name and acronym	Materials Cloud	The Open Databases Integration for Materials Design (OPTIMADE)	Citrine Informatics	MARKET 4.0	Chemeo (Cheméo)	IDS ecosystem
2	Platform home URL	www.materialscloud.org		Note: You get a specific URL AFTER registering	http://platform.market40.eu/index.html	https://www.chemeo.com/	Peer to peer ecosystem no centralized URL is available
3	Platform query URL (i.e., query end-point)	Multiple, e.g.: [query URL][46]	e.g., [query URL][47] and [query URL][48]		[query URL][49] Other API methods are available as well (e.g., querying the results by category id)	[query URL][50]	Requires IDS infrastructure.
4	Platform documentation (via URL or similar)	https://www.materialscloud.org/explore/connect	https://github.com/Materials-Consortia/OPTIMADE/blob/develop/optimade.rst	Citrine Python client documentation	http://platform.market40.eu/index.html#!/help	https://www.chemeo.com/notes/about/	https://internationaldataspaces.org/
5	Scope	Computational materials data	Atomic structure data	Tools for Material informatics (e.g., ML models)	Production equipment and service marketplace	Chemical and physical data for the process industry	Private supplier data

[46] https://www.materialscloud.org/mcloud/api/v2/discover/mc3d/compounds
[47] https://oqmd.org/optimade/
[48] https://www.crystallography.net/cod/optimade/
[49] http://platform.market40.eu/search/offerings?key=&page=0&size=12
[50] https://www.chemeo.com/api/v1/

6	**Interface**	REST API	REST API	GUI or API	GUI, Web API	GUI and REST API	IDS API
7	**Does the API specification comply with a "standard" (yes/no + name + URL)**	Yes - OPTIMADE	Yes - OPTIMADE		No	No	No
8	**How to access (registration or fee needed)**	Open: no registration or fee	Typically free and no registration	At a cost, registration needed	Access to the catalogue is open	Free	Valid IDS certificate required
9	**Enters in Showcase(s) #**	6	6	(3),4,9	5	7	5
10	**Type (P=provider, C=consumer, P/C=both)**	P/C	P	P/C	P/C	P	P
11	**Provider of: data (D), data on demand (DOD), interactive app (IA)**	D, IA	D	DOD, IA	D	D	D
12	**Note**		OPTIMADE is the standard, but there are many OPTIMADE data providers, a list can be found at [URL][51]	Secure platform ISO 27001			

[51] https://www.optimade.org/providers-dashboard/

12. File formats entering the showcases

In Table 13 below, we summarize the file formats entering the DOME 4.0 showcases, together with basic information (references, domain) and relevance for our project. Whenever available, we indicate the DOI assigned to the format within FAIRsharing[52], a community-driven project to support FAIR principles. Note that file formats that are specifically connected to a given software (e.g., .cdb for Ansys) are not included in the table.

Table 13: File formats involved in the showcases

File format name	Extension	Domain (N/A if generic)	Format documentation (via URL)	Format FAIRsharing DOI	Level of understanding DOME4.0 needs to have (0=know,1=load, 2=parse)	Priority (high, medium, low)	Enters in showcase(s) #
AiiDA export file	.aiida	workflows	[URL][53]		0	low	6
Crystallographic Information Framework (CIF) - Model and format	.CIF	atomistic	[URL][54]	[DOI][55]	0	medium	6
Comma-Separated Values (CSV)	.CSV	N/A	[URL][56]	Awaiting DOI assignment	0	high	3, 4, 9
Geographic Tagged Image File Format (GeoTIFF)	.geotiff	spatial	[URL][57]	[DOI][58]	0	low	1
Height	.HGT	topography	[URL][59]		0	low	1

[52] https://fairsharing.org/; https://www.nature.com/articles/s41587-019-0080-8
[53] https://aiida.readthedocs.io/projects/aiida-core/en/v2.0.0b1/internals/storage/sqlite_zip.html

[54] https://www.iucr.org/resources/cif/spec/version1.1
[55] https://doi.org/10.25504/FAIRsharing.zr52g5

[56] https://datatracker.ietf.org/doc/html/rfc4180
[57] https://docs.ogc.org/is/19-008r4/19-008r4.html
[58] https://doi.org/10.25504/FAIRsharing.cdd9bf
[59] https://gdal.org/drivers/raster/srtmhgt.html

Joint Photographic Experts Group Format (JPEG)	.jpg	N/A	[URL][60]	[DOI][61]			3
JavaScript Object Notation	.JSON	N/A	[URL][62]	[DOI][63]	0,1,2	medium	5
Portable document format (PDF)	.PDF	N/A	[URL][64]	See [URL][65]			4
Resource Description Framework (RDF)	.RDF	N/A	[URL][66]	[DOI][67]			1
Standard for the Exchange of Product (STEP) Data	.STEP	CAD design	[URL][68]				2
Gzip-compressed archive	.tar.gz or .tgz	N/A	[URL][69]				6
Terse RDF Triple Language (Turtle, TTL)	.ttl	N/A	[URL][70]	Awaiting DOI assignment			8
Text	.txt	N/A	[URL][71]		1		1
Microsoft Excel format	.XLSX	N/A	[URL][72]				2, 3, 4
XYZ atomistic data	.XYZ	atomistic	[URL][73]		0	medium	6

[60] https://www.fileformat.info/info/mimetype/image/jpeg/index.htm
[61] https://doi.org/10.25504/FAIRsharing.nggj0j
[62] https://www.json.org/json-en.html

[63] http://dx.doi.org/10.17487/RFC8259

[64] https://www.iana.org/assignments/media-types/application/pdf
[65] https://doi.org/10.25504/FAIRsharing.0ade3e
[66] https://www.iana.org/assignments/media-types/application/rdf+xml
[67] https://doi.org/10.25504/FAIRsharing.p77ph9

[68] https://en.wikipedia.org/wiki/ISO_10303-21
[69] https://www.gnu.org/software/gzip/
[70] http://www.w3.org/TR/turtle/
[71] https://developer.mozilla.org/en-US/docs/Web/HTTP/Basics_of_HTTP/MIME_types/Common_types
[72] https://www.iana.org/assignments/media-types/application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
[73] https://en.wikipedia.org/wiki/XYZ_file_format

As a general note on text-based files and data, we underline that the chosen *character set* (*charset*) is a delicate issue: we recommend using US-ASCII as a default or its superset UTF-8 (often, the character encoding can be chosen in software settings).

13. Conclusions / Next steps

In this document we have collected information on the data to be handled within the nine DOME 4.0 showcases, providing a rich overview of the specific domains, file formats and platforms involved. This information has already and will continue to support the DOME 4.0 platform development. The project is still in its initial phase (M18 out of 48 months), so there might be some changes to aforementioned aspects of the showcases in the future, but we are confident that the bulk of the information provided in this document will remain relevant.

Content-wise, we have collected diverse data types (about 30), encompassing experiments, observations and simulations (e.g., creep experimental data, weather data, simulation settings, ML model design space, etc). Concerning file formats: about 20 have been identified, with both general ones (.XLSX/Excel, JSON, CSV) and domain-specific ones (e.g., .XYZ for molecular simulations and .HGT for topography). Most formats are human readable, and metadata tend to be given as key/value pairs, row/column headers, or in the file name/ID tag. We note that few schemas are used, however concepts often belong to domain-specific "standards". We have identified topics and domain-specific keywords and pointed out those that will be most relevant for DOME 4.0. Whenever the use cases are already using ontologies or have identified relevant ones, references to those have been collected too. About 10 platforms (intended as web-based services) will enter the showcases, including both data and service providers; usually, they come with both GUI and API interfaces.

While this document concludes the activities of Task 4.1, DOME 4.0 WP4 will continue working on the execution of the showcases, in close connection with the core technical WPs as well as the business model-focused WP7.

14. Deviations from Annex 1

There are no deviations from Annex 1.

15. Acknowledgements

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 – Universita 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.

16. Table of Abbreviations

Abbreviation	Explanation
APDL	Ansys Parametric Design Language
API	Application Programming Interface
ASE	Atomic Simulation Environment
CAD	Computer-Assisted Design
CRS	Coordinate Reference System
DOI	Digital Object Identifier
DPD	Dissipative Particle Dynamics
FE	Finite Elements
GF	Glass Fiber
GHS	Globally Harmonized System
GUI	Graphical User Interface
HDD	Hard Disk Drive
HSP	Hansen Solubility Parameter
ICSD	Inorganic Crystal Structure Database
IDS	International Data Spaces
KG	Knowledge Graph
ML	Machine Learning
MSDS	Material Safety Sata Sheet
NSST	Neutral Salt Spray Test
REST	REpresentational State Transfer
SW	Software
URL	Uniform Resource Locator
UUID	Universally Unique Identifier
XPS	X-ray Photoelectron Spectroscopy

Annex 1 – Elucidations

In Table 14 below, we briefly explain what is expected for each row in the tables used for the individual showcases. Most categories are self-explanatory.

Table 14: Elucidations on the entries of the data tables used in Sections 2-10

#	Row name	Explanation, additional information, examples
1	**Data type (high level)**	Data by content at a mid/high level. It can involve multiple file formats
2	**Within the showcases, is it Input, Output or both? (I, O, I/O)**	Input/Output with respect to the showcase workflow
3	**Probability of usage in the showcases (sure, high, low)**	Gives an estimate of usage probability
4	**In which showcase(s)?**	Points to the DOME 4.0 showcase(s) number
5	**We have access already**	Yes/No. “We” is intended as “the DOME 4.0 consortium”
6	**Structure (File format(s) etc)**	
7	**Intermediate structure within the use case (File formats we convert to etc)**	
8	**Folder structure (yes/no + notes)**	E.g., specifies if files are organized in a folder with a certain structure that we need to preserve to be able to process the data with a specific tool.
9	**File size (order of magnitude, including all types)**	
10	**Open-source format?**	Yes/No. Specifies if the format can be freely used and implemented by anyone.
11	**File identifier**	Gives information on the eventual file identifier (e.g., used within the data source)
12	**Name of file (e.g., it contains relevant information/metadata)**	Specifies if the file name contains relevant information, metadata, identifiers.
3	**Changes in time (NA=not applicable)**	Yes/No/NA. Allows to flag data that is expected to vary in time.
14	**Binary, human readable? (B, H)**	B/H.
15	**Concrete example files (paths)**	Points to example files available publicly (URL) or internally
16	**Data type in more detail: keywords**	Specifies the data content in further detail, expanding Row #1
17	**Is the structure documented?**	Yes/No - Specifies if the data structure is documented
18	**Link to documentation**	
19	**How is the metadata given?**	For example: as column/row headers, key/value pairs.
20	**Does it follow a schema?**	Yes/No - Specifies if data follows a data schema

21	**Name/link of schema followed**	Points to the schema followed
22	**Where is the data now? (Data source)**	Points to the current location of the data, which may be different once DOME 4.0 is finalized (e.g., data could be currently on client local machines)
23	**On a platform? (yes/no)**	Yes/No. Specifies if data is already available via a web-based platform.
24	**Platform name and acronym**	(EXTRA – only shown if not empty)
25	**Platform home URL**	(EXTRA – only shown if not empty)
26	**Data provider (name or type)**	
27	**Data owner (name or type)**	
28	**Data access rights (license, registration or fee needed)**	
29	**Interface (connector, etc)**	(EXTRA – only shown if not empty) Interface of the platform
30	**Are ontologies already used in the showcase in relation to these data? (Yes/No)**	Yes/No
31	**Name+acronyms of the ontology(ies)**	(EXTRA – only shown if not empty) Points to the ontology(ies) name and acronym
32	**Documentation on the ontologies (paper URL)**	(EXTRA – only shown if not empty) Points to the ontology(ies) documentation
33	**URLs of the ontologies**	(EXTRA – only shown if not empty) Points to the ontology(ies) source files
34	**Standards (ISO-like, or "de facto") used in showcase in relation to these data**	Points to the “standards” (in a broad sense) used
35	**Subset/aspect of the data DOME needs to be able to 'understand' (text)**	Points out the data aspects (group at mid/high level) that DOME 4.0 needs to understand, as opposed to those that can be opaque for DOME and are left to the user to deal with
36	**Notes**	(EXTRA – only shown if not empty)
37	**Specific libraries and/or tools required to read data**	(EXTRA – only shown if not empty)
38	**What we use to read it**	(EXTRA – only shown if not empty)

Annex 2 – Example of queries and additional requirements

In Table 15 below, we show some examples of queries (as text or machine formula), while in Table 16 we list additional requirements from the data consumers.

Table 15: Example queries

Show case #	Data topic	User type	Query text	Query (machine formula)	Trigger (e.g., triggers a calculation)
1	Air quality	City council	Create sensor at latitude-longitude	Usually an HTTP POST with JSON data. For geospatial queries [URL][74]	Yes
2	lightweight construction	DOME client (e.g. company looking for material properties for specific application)	Find materials with creep rates lower than xxx/s and elastic modulus exceeding YYY MPa and a specific weight below ZZZ kg/m3		
3	Corrosion protection additives	DOME client	Give me a predicted performance of a set of possible structures		
4	Structural adhesive	DOME client	Show me adhesives with fatigue strength higher than X MPa		
5	MARKET4.0 catalogue	DOME client	Return all equipment in the MARKET4.0 catalogue in pages. The parameter key is used to search by name of equipment, supplier, etc. When the key values is empty	http://platform.market40.eu/search/offerings?key=&page=0&size=12	No

[74] https://newcastle.urbanobservatory.ac.uk/api_docs/doc/sensors-json/

			all equipment are returned.		
6	Atomic structure	DOME client (e.g. company developing materials)	Return all materials with Cobalt plus Lithium plus O2	http://www.crystallography.net/cod/result?formula=Co+Li+O2	Both
6	CO2 adsorption energy	DOME client (e.g. company developing materials)	Show me the CO2 adsorption energy from AiiDA calculation 4fc8801e-1000-429c-bc89-f8ece8d99c00	https://www.materialscloud.org/explore/curated-cofs/details/4fc8801e-1000-429c-bc89-f8ece8d99c00?nodeType=NODE	No
6	Electronic structure		Show me the band structure for AiiDA node 87e21964-8186-43f3-9efc-ea0d0086f3ba	https://www.materialscloud.org/explore/2dstructures/details/87e21964-8186-43f3-9efc-ea0d0086f3ba?nodeType=NODE	No
7	Chemical compounds	Dome client (e.g., user looking for standard identifier for compounds)	Look for CTAB (GUI); In the REST API, one can look selectively for the SMILES formula and IUPAC INCHI of CTAB (Cetrimonium Bromide)	https://www.chemeo.com/search?q=ctab	No
8	Manufacturing machine/control	Welding experts	Return the number of welding programs and welding operations for all welding machines	See [example query][75]	No
9	Composite materials	DOME client (e.g. material consumer)	How can I select the best material for the Melt Flow Index and with		

[75] SELECT ?wm (COUNT(distinct ?prog) AS ?CountProgram)
COUNT(distinct ?op) AS ?CountOperation)
WHERE {
?wm rdf:type rsw-kg:WeldingMachine .
?wm rsw-kg:hasWeldingProgram ?prog .
?op rsw-kg:performedByMachine ?wm . }

			high mechanical properties? i.e. Which material ingredient combination allows me to hit target material properties?		

Table 16: Examples of additional requirements from data consumers

Data consumer (type or name, if it is a concrete data sink/platform)	Show case #	Requirements from consumer on data structures (text)	Requirements from consumer on metadata (text)	Any other requirements from consumer on data/metadata (text)
Data sink outside DOME	5	The consumer understands the JSON format of the data.	The consumer understands the JSON format of the data.	The consumer has to have an IDS connector and a valid certificate to access the data peer to peer.
AiiDALab	6	AiiDALab expects to get a URL pointing to an OPTIMADE endpoint specifying a particular atomic structure	None	Identification of user via an access token
Welding experts	8	The consumers understand the data schemata; or there exists some mechanism that ease the data access, e.g., query selection, form-based query generation	None	The consumer has knowledge on the domain of the data.
Data provided by the material supplier	9, (3,4)	.py script used to ingest the data on the Citrine platform	The consumer understands the provided data	