Deliverable 1.3: Roadmap for integrating vocabularies across partner data infrastructures

Authors: Ewan O'Connor

Work package no	WP1
Deliverable no.	D1.3
Lead beneficiary	FMI
Deliverable type	X R (Document, report) ☐ DEC (Websites, patent filings, videos, etc.) ☐ OTHER: please specify .....
Dissemination level	X PU (public) ☐ CO (confidential, only for members of the Consortium, incl. Commission)
Estimated delivery date	M24
Actual delivery date	24/03/2026
Version	Final
Reviewed by	Ann Mari Fjaeraa
Accepted by	Ewan O'Connor
Comments	

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

Contents

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

1. The CARGO-ACT project

Aerosol, clouds and trace gases are short-lived atmospheric constituents, with residence times in the atmosphere ranging from seconds or minutes to a few weeks. Some of the short-lived atmospheric constituents are also climate forcers. From the 6th IPCC report (IPCC, 2021) on the physical basis of climate change, the climate feedbacks induced by short-lived climate forcers (SLCFs) are assessed as having a significant cooling effect but with low confidence. The same species are responsible for approximately 8 million premature deaths globally (Health Effects Institute, 2024). Mechanisms by which SLCFs affect climate are more complex than for greenhouse gases and require the definition of more than 20 essential climate variables (GCOS-240).

The pan-European Aerosol, Clouds and Trace Gases Research Infrastructure (ACTRIS) provides long-term, high-quality time series of observed properties of these short-lived atmospheric constituents. This is achieved through ground-based in situ and remote sensing measurements at surface stations. The Research Infrastructure and its measurement processes involves all steps from the observation and data collection at the station, quality assurance and quality control (QA/QC) by Central Facilities (CFs), to data curation and management by a central Data Centre, where QA/QC and data management are organised in units by topic and observation type. In CARGO-ACT, ACTRIS and its Central Facilities (CFs) are collaborating with their U.S. American counterparts at the Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) Facility, the National Aeronautics and Space Administration (NASA), and the National Oceanic and Atmospheric Administration (NOAA).

The overarching goal of CARGO-ACT is to deliver a clear roadmap for sustainable global cooperation between key atmospheric monitoring organisations in Europe and in the United States. This cooperation aims to provide all users, in the scientific community and beyond, with the best possible services for accessing and using information from monitoring climate- and air quality-relevant properties of aerosol, cloud and trace gases in the atmosphere.

2. Scope for vocabulary convergence

The goal of CARGO-ACT is to begin the work towards achieving a level of interoperability between CARGO-ACT partner networks sufficient to enable the creation of global atmospheric datasets combined from multiple atmospheric data repositories with a consistent description of the data and products, their sources, uncertainties and processing.

This interoperability can be achieved through the use of an ontology, a framework providing a formal naming and definition of the categories, properties, and relations between the concepts and data in a particular subject field. An ontology can be a controlled list of vocabulary terms for metadata or can also provide a semantic structure to the variable descriptions. The I-ADOPT Framework ontology is

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

such an example and it provides machine-readable variable descriptions in which the components are mapped to FAIR vocabulary terms (see Appendix A).

"Deliverable 1.2: Roadmap for convergence of FAIR Enabling Resources" identified vocabulary interoperability as one of the 5 topics where improvements in implementation and approach among CARGO-ACT partners would have the largest impact on improving user experience. The deliverable also considered the use cases for interoperability and outlined possible approaches for achieving these.

3. Existing controlled vocabularies and glossaries

The locations of controlled vocabulary lists for partner networks in CARGO-ACT are shown in Table 1. Most of these are currently available as documentation intended for human use and, apart from the ACTRIS vocabulary, these are not yet machine-readable. NFAN provides extensive README files for their datasets.

Table 1: CARGO-ACT network level controlled vocabularies and glossaries

Network	Title	Type	Link
ACTRIS	ACTRIS Vocabulary	Controlled vocabulary	https://vocabulary.actris.nilu.no/skosmos/actris_vocab/en/
ARM	ARM Data File Standards Version: 1.3	Controlled vocabulary	https://armgov.svcs.arm.gov/publications/programdocs/doe-sc-arm-15-004.pdf
ARM	Glossary	Glossary	https://armgov.svcs.arm.gov/resources/glossary
MPLNET	MPLNET Product Information	Controlled vocabulary	https://mplnet.gsfc.nasa.gov/product-info/

ACTRIS and ARM provide landing pages for all data objects in their respective data portals; for ACTRIS, clicking on the variable name on the landing page (Fig. 1) will take you to the definition on the ACTRIS vocabulary server, for ARM, hovering over or clicking on a variable name on the landing page will retrieve the variable definition and display it in a pop-up box (Fig. 2);

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

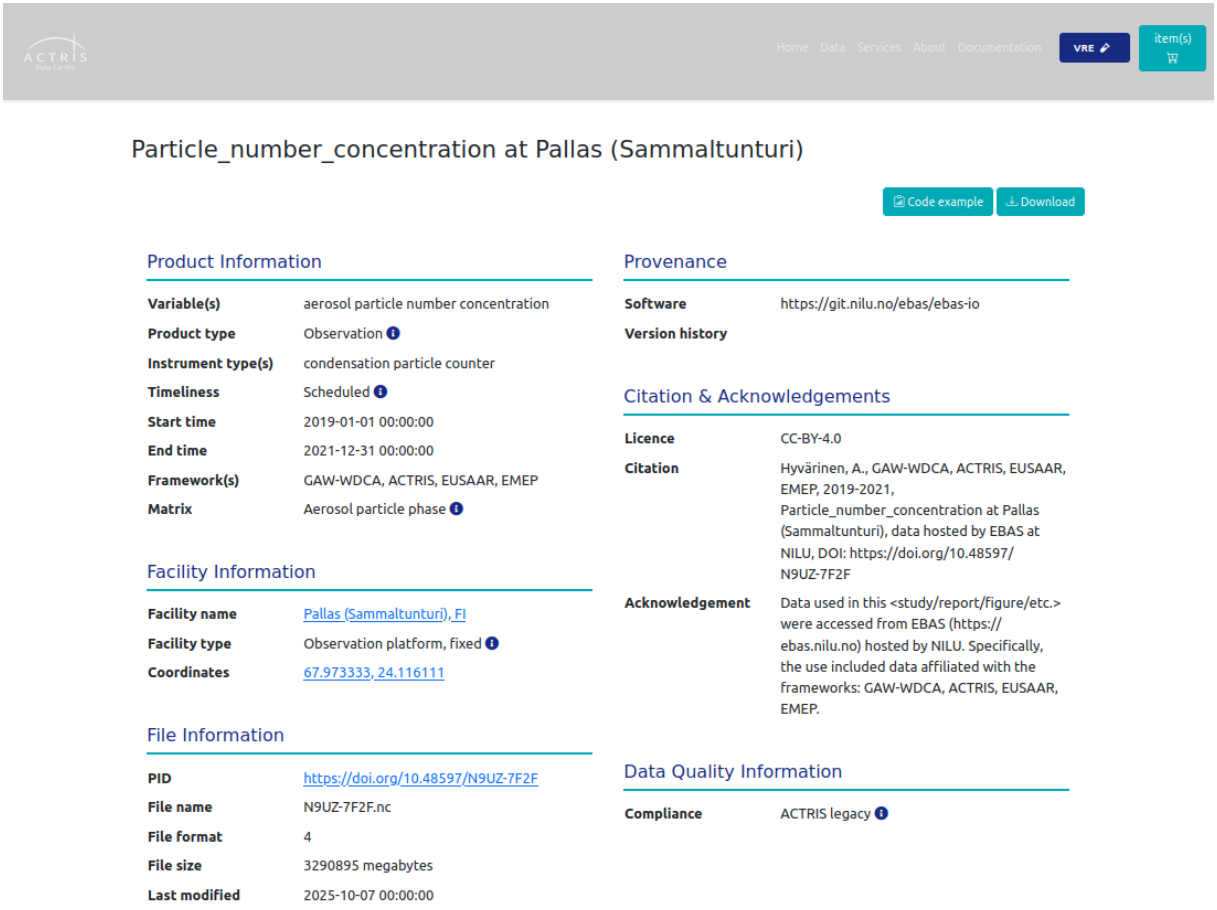

Figure 1: ACTRIS data portal landing page for variable "*Aerosol particle number concentration*". Clicking on the blue i symbol next to product type will take you to the ACTRIS vocabulary server.

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

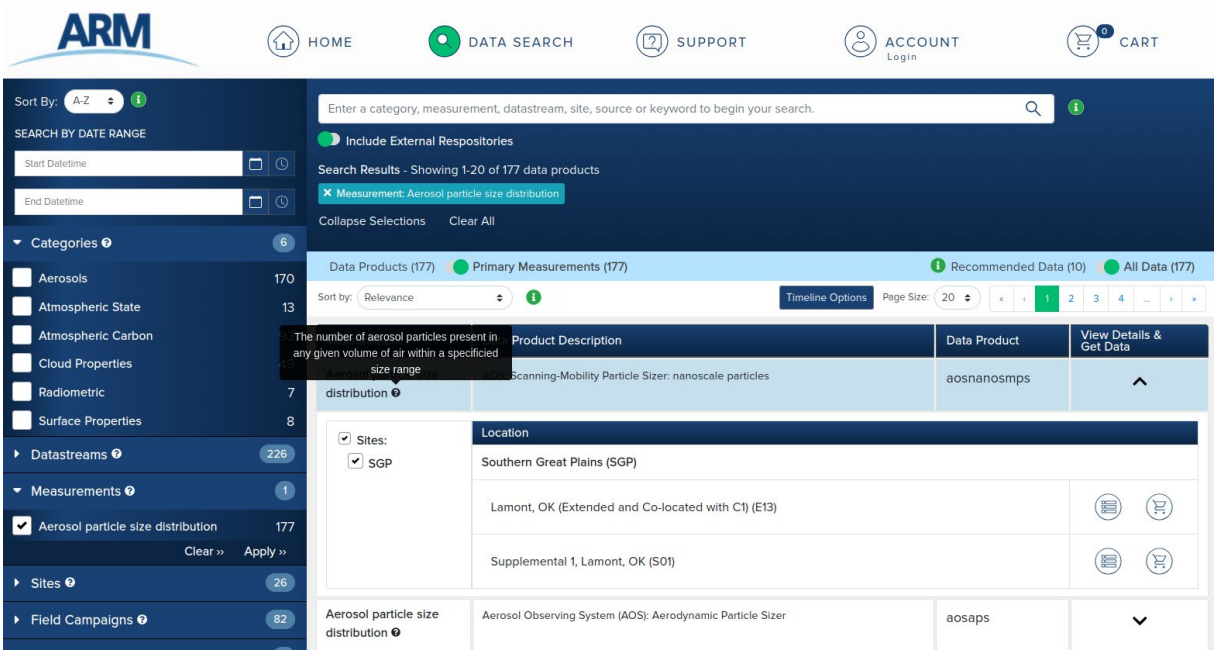

Figure 2: ARM data portal landing page for variable "*Aerosol particle size distribution*" with the variable definition in a pop-up box (white text on black background).

Interoperability via metadata exchange between CARGO-ACT partner data portals has already been shown in "Deliverable 1.2: Roadmap for convergence of FAIR Enabling Resources". This made use of the ACTRIS vocabulary API to identify the appropriate metadata name and then search for these in the ACTRIS data portal using the search API.

The goal, however, is not just interoperability between the partner networks, but interoperability with other global portals and stakeholders who may maintain their own ontologies and controlled vocabularies. Relevant examples for the atmospheric domain include those given in Table 2, which include controlled vocabularies and glossaries for satellite agencies (or committees representing coordination between satellite agencies), numerical weather prediction and climate modelling centres, and globally coordinated surface measurement networks.

CARGO-ACT vocabulary should also converge with, or be understood within, these frameworks too where relevant. As discussed in "Deliverable 1.2: Roadmap for convergence of FAIR Enabling Resources", this can be achieved by linking to existing items from these vocabularies when relevant rather than creating new ones, but note that direct correspondence may not always be possible, whether due to a lack of suitable descriptors for a different viewing geometry (with respect to satellite) or instrument type and operation, or that observed and modelled parameters are fundamentally different (with respect to model variables).

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

Table 2: CARGO-ACT stakeholder controlled vocabularies and glossaries

Stakeholder	Title	Type	Link
NASA	GMCD Keyword Viewer	Controlled vocabulary	https://www.earthdata.nasa.gov/data/tools/gcmd-keyword-viewer
WMO	wmds: WIGOS Metadata Standard: Semantic standard and code tables	Controlled vocabulary	https://github.com/wmo-im/wmds
KCEO	KCEO Glossary	Glossary	https://ec-jrc.github.io/KCEO-Glossary/glossary_topology/
CF Metadata	CF Metadata conventions	Controlled vocabulary	https://cfconventions.org/
CEOS	EO Glossary	Glossary	https://github.com/ceos-org/eo-glossary/
WMO-GAW	WMO/GAW Glossary of QA/QC-Related Terminology	Glossary	https://www.empa.ch/web/s503/gaw_glossary
CMIP	CVs and MIP Tables	Controlled vocabulary	https://wcrp-cmip.org/cvs-and-mip-tables/
OGC	OGC API - Records	vocabulary	https://github.com/opengeospatial/ogcapi-records

Issues such as different communities using different standard names in their controlled vocabulary for the same variable can be solved by mapping, where the standard names are linked together so that, for example, searching for one standard name for a variable will return results matching all standard names for that variable. Exact matches with other controlled vocabularies can also be recorded in machine-readable vocabulary lists. Both examples are shown Fig.3, where for the variable "*liquid droplet total column mass content*", there are two **entry terms** (*liquid water path* and *lwp*) which are used within the ground-based cloud profiling community (but not the satellite community) and an **exactly matching concept** in another vocabulary.

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

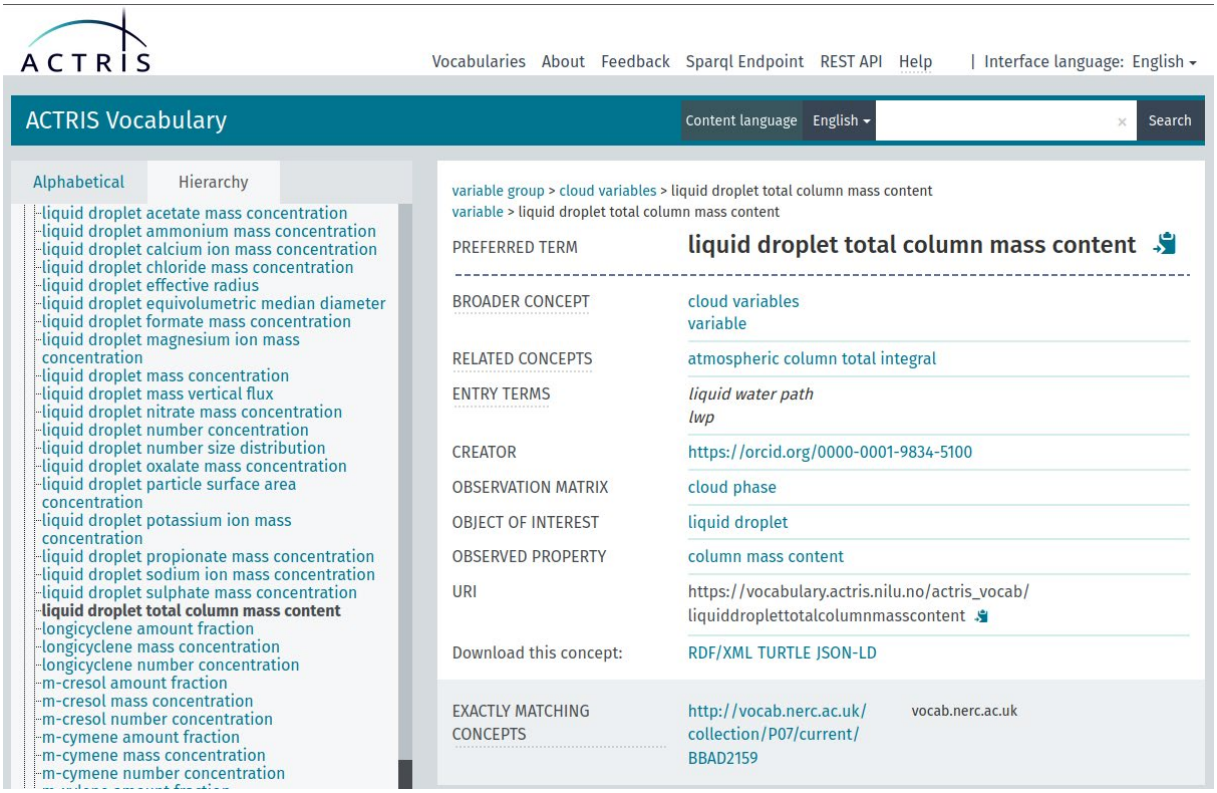

Figure 3. An example of mapping both to multiple community terms and to matching concepts in another vocabulary

4. Current status

A common structure of top-level concepts for vocabulary describing QA and QC (quality assurance and quality control) for the observation of aerosol particle, cloud particle, and trace gas was agreed upon between the CARGO-ACT partner networks: "Deliverable 2.2: Agreement on common vocabulary for describing instrument traceability and calibration, quality assurance and quality control". This vocabulary formally describes the procedures described in "Deliverable 2.3: Recommendations for common calibration and operation procedures".

An interoperable ontology should also contain strict definitions describing measurement uncertainties and how they are calculated for each instrument and parameter, as described in "Deliverable 2.4: Recommendation for a common approach in measurement uncertainty estimation". This is particularly relevant, for example, when a global dataset comprises products generated from a heterogeneous mix of instruments, because each instrument has its own calibration procedure, quality control methods, and ways of estimating uncertainty.

Note that interoperability has also been applied to data access vocabulary to ensure that common descriptors and metrics are used: see "Milestone 2: Common standard for measuring data access".

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

The CARGO-ACT QA/QC vocabulary is hosted as part of the ACTRIS vocabulary, and uses the same maintenance mechanism (issue tracker on GitHub). Open collaboration continues on the CARGO-ACT QA/QC vocabulary in response to new methodologies being developed by the research community and requests from users.

As seen in Fig. 3, the ACTRIS vocabulary server is built on Skosmos, an open source web-based SKOS browser and publishing tool https://skosmos.org/, where SKOS stands for Simple Knowledge Organisation System https://www.w3.org/TR/skos-reference/ developed by The World Wide Web Consortium (W3C) https://www.w3.org/.

The vocabulary server is not static and has regular updates as new developments in processing, quality assurance and quality control, and other aspects evolve, and the vocabulary evolves with it. The responsibility for maintaining the server lies with the ACTRIS data centre and will continue to be supported after the CARGO-ACT project finishes.

With netcdf and hdf5 file formats, comprehensive metadata can be stored in both the files and in a database. Storing comprehensive metadata within the data file provides an excellent reference for local use, but it is very inefficient for performing metadata searches in large file collections. One practical advantage though is that multiple names (from multiple or updated vocabularies) can be encoded in the variable/global attributes and be available until fully deprecated. This also helps with backwards compatibility for long-running operational systems making use of the data.

5. Roadmap and recommendations

Figure 4 displays an actionable roadmap for converging vocabularies.

First perform a landscape analysis across contributing networks and assess multiple aspects, not just product/variable names and metadata. Items to assess include:

- Measurement metadata
    - Geophysical product
    - Instrument
- Operational procedures
    - Calibration
    - QA/QC methods
    - Processing applied
        - Software version used
- Uncertainty estimation
- Attribution
    - Licence
    - Acknowledgements
    - Citation
- Data metrics

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

Adopt an agreed upon controlled vocabulary (and optionally semantic attribute description) suitable for the domain for each of the items assessed.

Map this controlled vocabulary with the users (stakeholders) needs; note that the vocabulary items can be mapped rather than match names precisely.

Create a vocabulary server or use an existing vocabulary server for creating links on landing pages to enable machine-to-machine interoperability.

Exchange metadata with other portals (to increase data discoverability).

Figure 4: Schematic outlining steps required for converging vocabularies across multiple network data repositories.

Recommendations

It is recommended to undertake the task of converging vocabularies in small steps rather than tackling all items at once. Start with a subset of variables or measurements first before expanding to the full list. This helps in understanding what is required to come to agreement for each variable and means that individual variable vocabulary items can be made interoperable immediately rather than waiting to converge the entire vocabulary.

Generating suitable quality control and other vocabulary helps clarify exactly what is being done to the data, not only to the users but also to the community producing the data.

When attempting to establish converging vocabularies, find common agreement with what a 'global user' such as a space agency or a global modelling centre would require. This gives a common focus rather than just imposing one networks' implementation over all others. Finding common agreement can be challenging for established networks, especially as there may be issues of backwards compatibility with ongoing operations involving invested stakeholders.

Certain file formats such as netcdf or HDF5 are recommended since they permit multiple naming/mapping in their metadata attributes.

Make regular updates to the ontology or vocabulary to ensure that it keeps pace with developments in the data acquisition. Responsibility for this lies with both the data centres and the wider research community.

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093

Appendix

I-ADOPT framework

The [I-ADOPT Framework Ontology] is a lightweight ontology primarily designed to facilitate interoperability between existing variable description models (including ontologies, taxonomy, and structured controlled vocabularies). The [ontology working group] is a member of the Research Data Alliance (RDA).

WORKING GROUP

InteroperAble Descriptions of Observable Property Terminology WG (I-ADOPT WG)

One aim of I-ADOPT is to enable the consistent generation of new variable names through the use of the schema shown in Fig. A1.

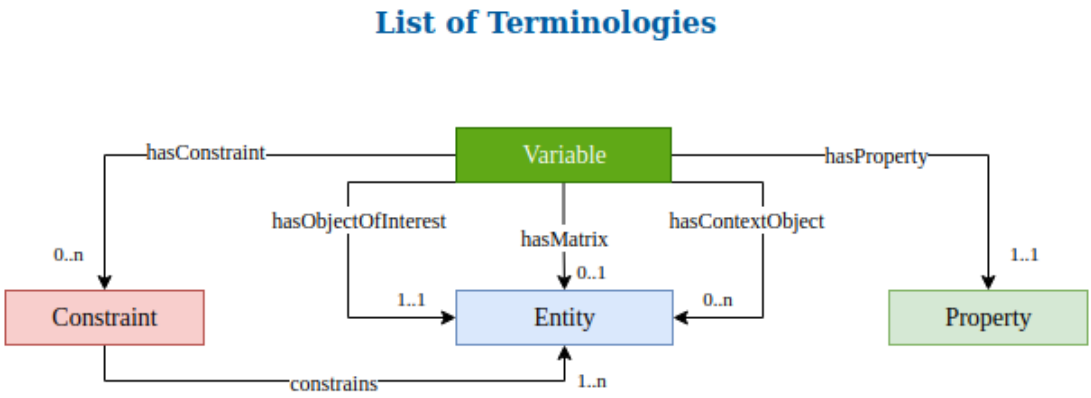

Figure A1: Visualisation of the lightweight version of the I-ADOPT ontology, with the 4 classes Variable Constraint, Entity, and Property, as well as the relations between them (source: **I-ADOPT Framework Ontology**).

Creating new variable descriptions in the I-ADOPT Framework means:

1. decompose the variable in atomic components
2. assign a description role for each component
3. annotate each component with concepts from terminologies

Then, compose the variable name by applying the I-ADOPT grammar consistently in the form:

Property of Constraint Object of Interest in Constraint Matrix of Constraint Context Object

CARGO-ACT (www.cargo-act.eu) is supported by the European Commission under the Horizon Europe (HORIZON) – Strengthening the international dimension of ESFRI and/or ERIC research infrastructures, HORIZON-INFRA-2023-DEV-01-07, Grant Agreement #: 101132093