SMART DICTIONARY MVP

Project Summary Report

EDM Council
December 2, 2022
Dr. Paweł Garbacz, Paul Worrall, and Dr. Jan Kroh

EXECUTIVE SUMMARY

This document describes the work done over the course of 9 weeks to deliver a "**Proof of Concept" MVP prototype,** to demonstrate and verify the capabilities and applicability of a **"SMART DICTIONARY"** approach to the harmonization of ESG and sustainability data.

KEY OBSERVATIONS:

1. The concept of a Smart Dictionary, to define terms and conditions, identify concepts and concept relationships, will enable a harmonization of disparate taxonomies that is **practical, feasible and realistic**

2. To build a successful Smart Dictionary, relevant Subject Matter Experts will be needed to work alongside Data Management professionals to address the complexities of the many different and disparate disclosures and emissions standards that exist in the industry today.

3. The results of this MVP will demonstrate that executing a standard, open collaboration approach with SMEs and data professionals will serve as the basis for a "scaffolding" for this and all future projects and Use Cases in building out the shared Smart Dictionary, reducing costs and time to market.

INSPIRATION

Governments across the world are implementing regulations to require companies to disclose information about their impact on the climate and their approach to risk management. There are many variants of global, regional and industry specific standards that guide how and what should be reported. This makes preparing reports resource intensive for preparers and difficult to understand, compare and contrast - for those making decisions based on their contents. These standards also define the type of information to disclose but do not provide any clear guidance on the digital form it should take. This makes automation for both publishers and consumers of the information difficult and limits the interoperability needed when combining it with other information to make informed decisions.

At DSD Lab we have successfully managed to bring together all of the fundamental "components" required for an effective and efficient development and delivery of complex SMART DICTIONARY solutions in decentralized working environments. One can think of this as a foundational "scaffolding" that is and will be used for the current and future Use Cases and projects:

1. The DSD Lab can successfully gather subject matter experts (SMEs), conceptual data modeling experts, project and program managers and the wider "communities of interest" to engage in a collaborative development process

2. The collaborative development process is a proven methodology, which include input gathering, management of working groups, SMART DICTIONARY modelling, review and acceptance processes, and controlled release management

3. The environment for collaborative SMART DICTIONARY development is based on a proven and battle-tested INFRASTRUCTURE of hardware, software (Jira, Confluence, Github, Jenkins, Ontology Publisher and OntoViewer) and project-related processes.

Our vision is to enable the different standards, and the information disclosed against them, to be expressed digitally. This will make it easier to combine, compare and contrast with greater interoperability and with increased automation. We are proposing this can be achieved using a SMART DICTIONARY.

TABLE OF CONTENTS

THE DSD-Lab SMART DICTIONARY MVP

PROBLEM STATEMENT

1. Mandatory GHG disclosures are here, international disclosure standards are converging, but there are differences in the thousands of – for example - emissions factors underlying this information.

2. There are many sets of GHG emissions factors applied to corporate transactions to estimate GHG emissions.

3. Emission factors are extremely helpful and reflect the best information on the carbon implications of business activities. In some cases, emission factors can overlap and have estimates, making them difficult to compare and contrast disclosures.

SMART DICTIONARY Project: BACKGROUND

The project has built on the EDM Council's experience of facilitating collaboration between users, domain experts and technologists to produce knowledge systems (e.g. other SMART DICTIONARIES). A small core team of knowledge experts, software engineers and domain experts have worked in an iterative, open and inclusive exercise to build a Minimum Viable Product (MVP). The team has identified use cases that demonstrate how **digital disclosures will improve interoperability and automation** for both publishers and consumers.

The MVP includes the means for a broader audience to be engaged in the future to extend and evolve the SMART DICTIONARY to address other standards and information requirements in the evolving Environmental, Sustainability and Governance space.

In order to coordinate continuous integration of new and revised material, facilitate collaboration across topic teams, and ensure continuous quality improvement, EDM Council supports the development of the DevOps framework created to automate aspects of "robust testing" to guarantee a minimum level of quality.

The individual tests are not necessarily novel. What is new, however, is a **portable, open-source infrastructure that automatically runs these tests as an integral part of the SMART DICTIONARY integration and publication process**. This framework is designed for either a single development environment or a cross organizational activity.

COP27 MVP USE CASE

There are DIFFERENCES between standards & related emission factors that make it IMPOSSIBLE TO COMPARE. Therefore, for the purpose of this MVP we have adopted a simple use case - **Shipping Goods from the US to the UK** - which can be illustrated as follows:
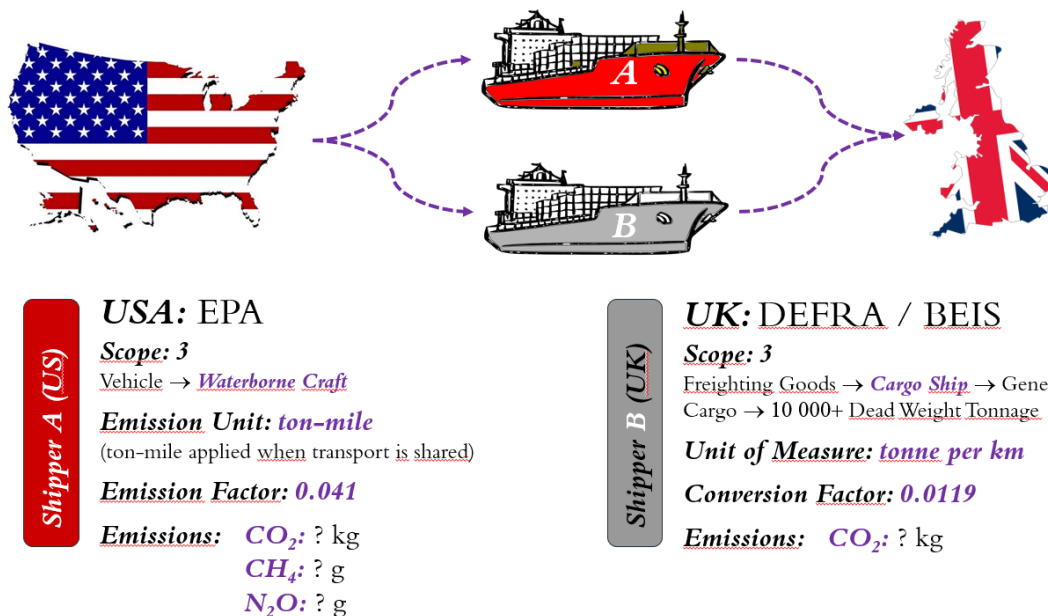

Figure 1: Smart Dictionary comparison of standards concepts for ocean going transport

Shipper A (US)

USA: EPA
Scope: 3
Vehicle → *Waterborne Craft*

Emission Unit: ton-mile
(ton-mile applied when transport is shared)

Emission Factor: 0.041

Emissions: CO_2: ? kg
CH_4: ? g
N_2O: ? g

Shipper B (UK)

UK: DEFRA / BEIS
Scope: 3
Freighting Goods → *Cargo Ship* → General Cargo → 10 000+ Dead Weight Tonnage

Unit of Measure: tonne per km

Conversion Factor: 0.0119

Emissions: CO_2: ? kg

COP-27 MVP SCOPE

The objective of the MVP Use Case was to demonstrate the value of the Smart Dictionary in addressing the challenges of disparate standards by demonstrating its ability to show the cross-cutting concerns over standards, and the interdependencies and relationships between the concepts in them. The scope of the MVP focused on addressing selected issues as described in the Carbon Call Paper (https://carboncall.org). The paper explains the issues around allocating (or attributing) the carbon costs to products and services and integrating it with the financial reporting systems.

Climate change is a global issue where companies tackle the data exchange for carbon accounting within a complex supply chain. Like financial accounting, carbon accounting quantifies the impact of an organization's business activities – though instead of financial impact, it measures climate impact and then tries to understand and apply the financial impacts. Financial and carbon accounting should quantify and understand how a company contributes to climate change. Wherever they are in the world, regulated entities have been subject to an increasing amount of fragmented regulatory regime changes, legislation, reviews, and queries. Carbon is another reporting regime

impacting companies that are already challenged with a broader view often associated with material compliance risks and costs.

Currently, the lack of consistency in applying different regulatory regimes and guidance on applying standards, defining, and accessing GreenHouse Gasses (GHG) scope data, including CO_2 and carbon, has major business implications. For example, in a supply chain with multiple participating companies, when the recipient of the supply chain wants to calculate the total carbon footprint of a product, it is virtually impossible since each supply chain participant has stored data using its own standards, data formats, and defined units. In addition, the upstream suppliers generally disclose data at the estimated overall corporate footprint level instead of at the individual product level. Product level disclosures would help interoperability significantly.

The SPECIFIC OBJECTIVE of the MVP has been to demonstrate how a SMART DICTIONARY supports automatic identification of emission factors within a supply chain across multiple emission factor standards.

MVP Timeframe

This has been a short project (MVP) with limited time and resources, so a number of precautions have been taken to manage any impact on the best achievable outcome. The team itself has significant experience in producing knowledge systems of this kind and has the skills to build a fully working system.

The project has been working in short and frequent cycles delivering usable functionality to solicit and reflect feedback from stakeholders. Therefore it was important that the stakeholders with the right domain knowledge are engaged and active during the project. It was difficult to know all the initiatives in this space and their scope, so the MVP sort to address this by being open and inclusive in the approach by demonstrating the capacity for reuse and integration with other solutions and existing investments.

COP-27 SMART DICTIONARY MVP DELIVERABLES

The DELIVERABLES of the SMART DICTIONARY MVP include…

1. **THE SMART DICTIONARY PROTOTYPE**

   The SMART DICTIONARY MVP used the *USA Environmental Protection Agency's (EPA) Emission Factors* for organizational greenhouse gas reporting and the *UK's Department for Business, Energy & Industrial Strategy conversion factors* for company reporting of greenhouse gas emissions (BEIS, aka DEFRA) to demonstrate the benefits of a SMART DICTIONARY. The information provided by both organizations was in the form of spread sheets, unsuitable for machine processing.

   The initial Smart Dictionary definition framework was created by looking at the concepts, terms and categories used in each standard. These enable users to determine the emission factors to use in CO2e calculations for a particular type of transport.
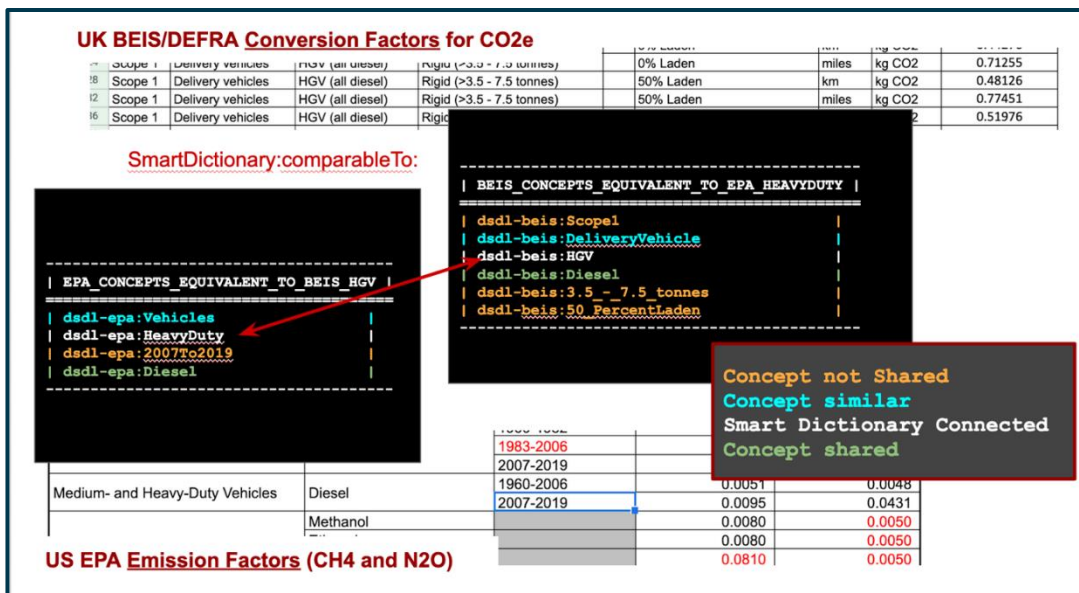

Figure 2: Smart Dictionary comparison between standard definitions for Heavy Duty Vehicles

   The Smart Dictionary definitions were limited in scope to one specific type of Passenger Vehicle, Small and Heavy Goods Vehicle and Ocean Transport. Just these two standards and this limited set of transport types provided Smart Dictionary definitions that demonstrated how vastly different just two standards can be.

2. **A VISUALISATION of the definitions to show their connections, relationships and interdependencies (using EDM Council's Onto-Viewer).**
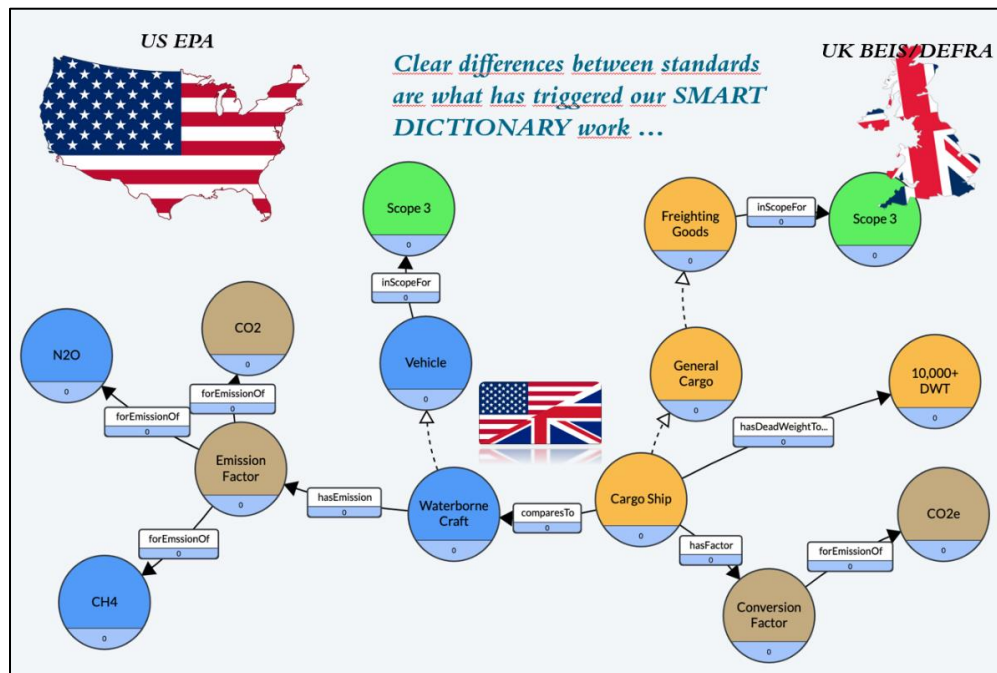

Figure 3: Visualization, note differences in gases reported left and right

The Smart Dictionary showed clearly **what concepts were shared** (although defined independently), those that appeared similar and those not shared (Figure 1: Smart Dictionary comparison of standards concepts for ocean going transport. The Smart Dictionary also showed that both standards have a **material difference** in how they break down the constituent types of gases in emissions and how they are translated to CO2e. (**Error! Reference source not found.**).

The differences that the Smart Dictionary prototype is showing between how the USA and UK guidance for emission factors makes it clearer what is contributing to emission reporting using each standard. The Smart Dictionary is showing what could be done to have standards converge and interoperate.

The Smart Dictionary prototype has also shown how the terms, concepts, attributes, and their definitions can be made clearer to facilitate comparing and contrasting between different requirements. It established a pattern for how questions of standards can be asked and how they might be tagged and mapped between each other. There is now a basis for a reusable 'scaffolding' creating the linkages between standards to share and repurpose knowledge across digital disclosures in a systematic, saleable and automated way.

The work has also identified a number of architectural and philosophical decisions that need to be made in subsequent phases. These will help framing requirements and priorities for sponsors, users, Subject Matter Experts and the DSD Lab team working on the project.

3. A PROVEL COLLABORATIVE DEVELOPMENT ENVIRONMENT

The EDM Council's standard Ontology-Publisher methodology was used at the basic **INFRASTRUCTURE** for **Collaborative Development Environment** on which to grow and extend the scope of this **SMART DICTIONARY.**
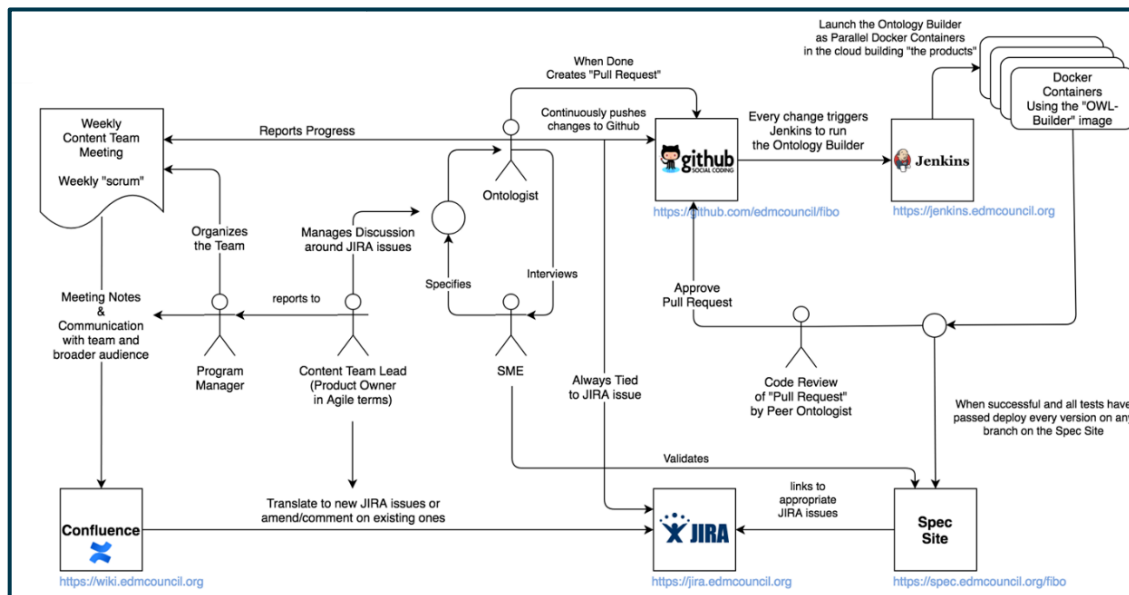
The Council's infrastructure is depicted in the diagram above. It is a combination of proven processes and technology components – all working together.

This Collaborative Ontology Development Environment framework was initially designed for the continuous development of **FIBO** (Financial Industry Business Ontology), is battle tested and has been proven over the past 10 years. It is now being used in our **IDMP-O** (Identification of Medicinal Products ontology) and **IOF/NIST** (National Institute of Standards and Technology) projects.

4. Ability to PUBLISH DATA using SMART DICTIONARY definitions (through an integration of various semantic web tools).

This initial Smart Dictionary work had three types of data. First, is the meta-data that describes the concepts, terms and categories that factual data should be described or tagged as. Second, are the emission factors numerical values, for use in calculating emissions, published by both the EPA and BEIS/Defra organizations. And third, the CO2e numbers and the data required to calculate them such as units for distance travelled or load carried.

The Smart Dictionary prototype has created a process and pattern for how meta-data for machine readability, needed to get the scale and interoperability, can be published and controlled. Each specialist domain, e.g. SME from EPA and BEIS/Defra, can use open standard data publishing tools to build on the Smart Dictionary concepts to develop and publish their own Smart Dictionary "sub dictionary". This can be achieved from a spread sheet, a specialist tool like Protégé or a custom developed software. Publishing meta-data in this way is a World Wide Web Consortium (W3C) open standard with a lot of support and choice.

The emission factors, and other public data in the future, can be tagged using the meta-data defined in the Smart Dictionary and published on any website. Again, this can be done from a spread sheet, a web publishing tool or one of many other techniques onto a website. This makes it machine readable from the website in an open and interoperable way.

In this project the scope was on only a few data points, so they were manually created and tagged in a text editor for use in the many test queries that were created, e.g. for calculating emissions for various transport types (

Figure 4. Calculation of CO2e for a gasoline passenger car using EPA emission factors).


```
CO2e for CH4 from Gasoline Passenger Car made in 2000

select ?emission ?factor ?unit ?value ?ch4GwpFactor ?distanceTravelled ?CO2e where {
    BIND( 135 AS ?distanceTravelled )

    dsdl:CH4 dsdl:hasGlobalWarmingPotential ?ch4GwpFactor .

    ?emission dsdl:appliesTo (  dsdl:PassengerCars dsdl:2000 dsdl:Gasoline )  .

    ?emission dsdl:hasFactor ?factor .

    ?factor a dsdl:CH4 ;
        dsdl:hasUnitOfMeasure ?unit ;
        dsdl:hasValue ?value .

    BIND( (?distanceTravelled * ?value * ?ch4GwpFactor ) AS ?CO2e )

}
```

emission	factor	unit	value	ch4GwpFactor	distanceTravelled	CO2e
:EmissionExamplePassengerCar	:CH4FactorExample1	"g / miles"	0.0175	25	135	59.0625

Figure 4. Calculation of CO2e for a gasoline passenger car using EPA emission factors

TECHNICAL DEVELOPMENTS

Smart Dictionary OWL Artefacts

The **SMART DICTIONARY** has been developed and expressed in the *OWL DL* (Ontology Web Language) - the language of the World Wide Web using Description Logic for machine readable inference processing. By taking advantage of the OWL DL Web standard, the SMART DICTIONARY will use **machine intelligence and network graph capabilities** to express, classify and link data in ways that were not previously possible.

In the context of our MVP use case this approach (and solution) can be illustrated in the following way:

Collaborative Ontology Infrastructure
We set up a (temporarily private) GitHub repository that stores all artefacts developed within the scope of this MVP:
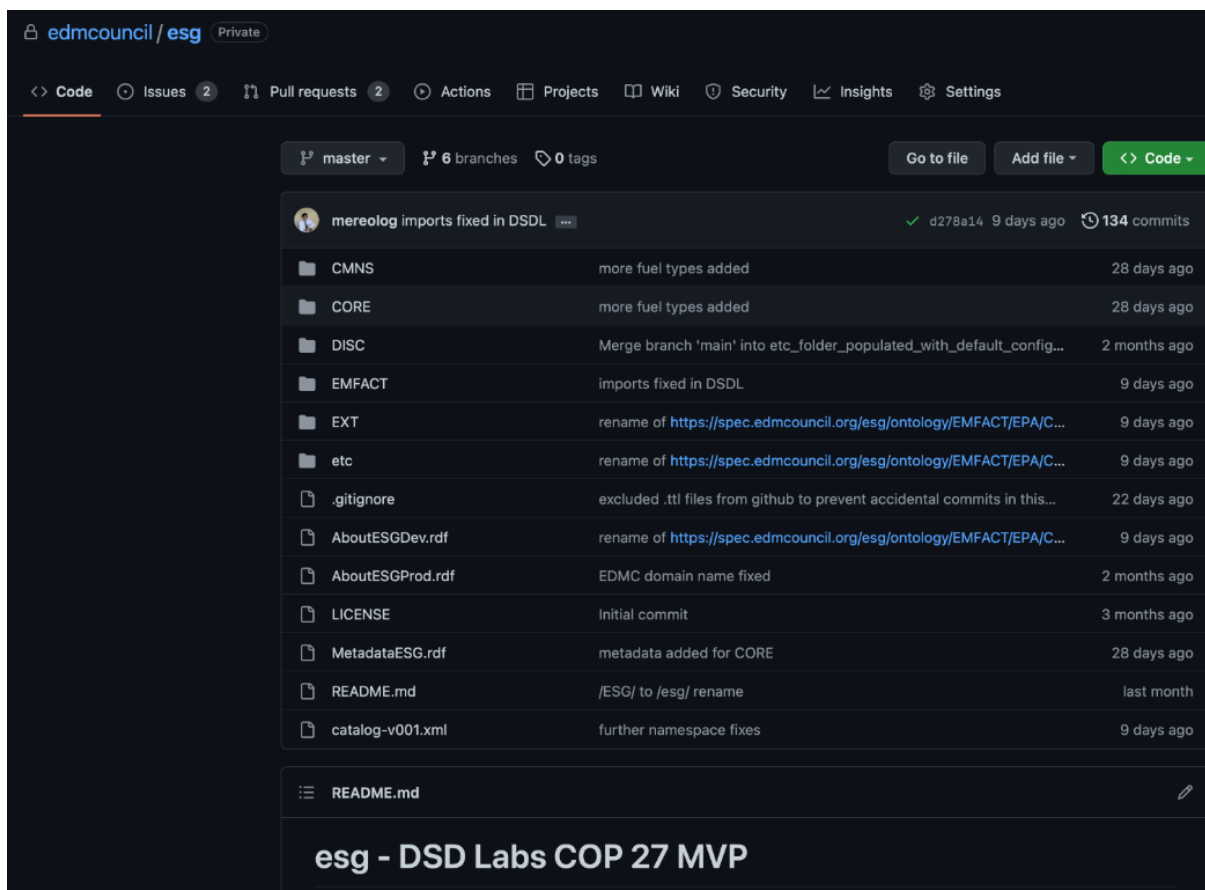
The Smart Dictionary files, which are versioned controlled in this repository, are processed by means of the EDMC Jenkins infrastructure. Every change in the Smart Dictionary is run through a testing pipeline (developed and used in the development of FIBO, IDMP-O and IOF/NIST):

https://jenkins.edmcouncil.org/blue/organizations/jenkins/esg/detail/master/19/pipeline
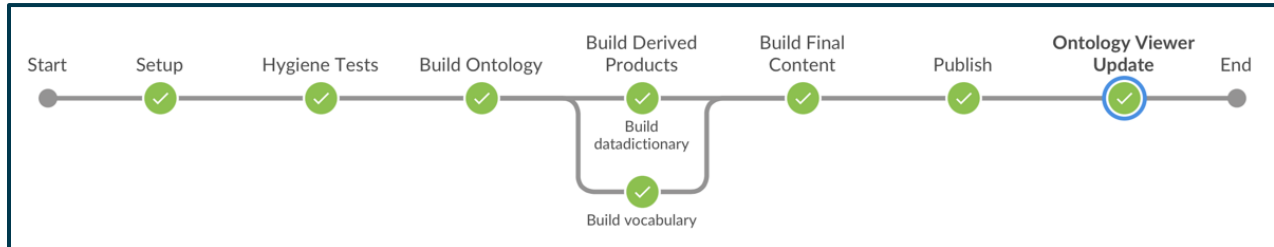
Onto-Viewer

The final outcome of the Jenkins orchestrator process described in the previous section can be accessed through the Onto-Viewer web application. The app provides both the textual and visual explanations of the terms being defined in the Smart Dictionary.
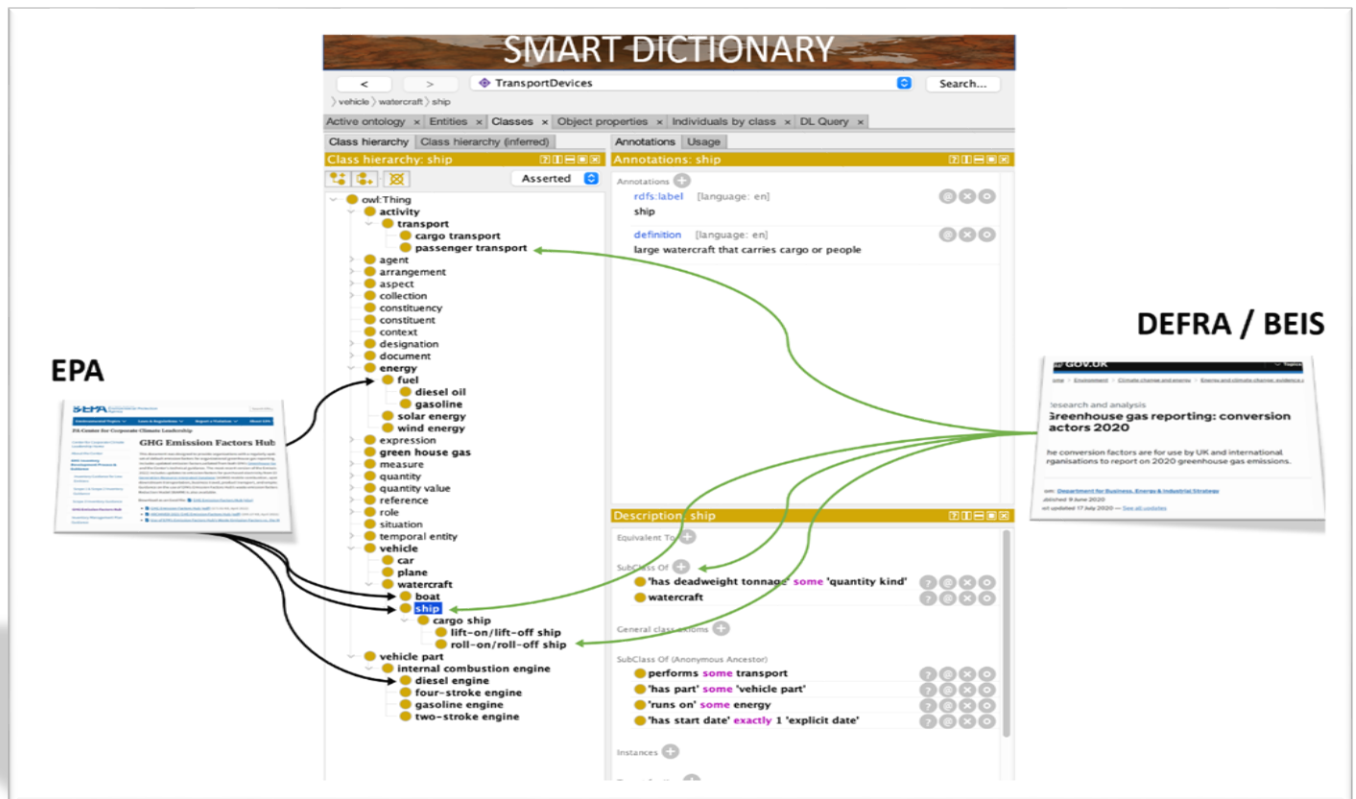
Benefits of the SMART DICTIONARY

By decomposing the existing standards down to their granular attributes, a SMART DICTIONARY will enable the discovery of where standards differ and build a 'roadmap' for harmonization.

In this MVP, the SMART DICTIONARY demonstrated how the 'end user' tasked with comparing the carbon emissions of two different vessels using different emission standards, can be mapped and harmonized such that a true comparison of the vessel's emissions can be made.

This concept of mapping disparate standards via the SMART DICTIONARY **is not limited to the carbon Use Case**. Standards for all ESG and sustainability data can be similarly mapped, thus creating a 'cross-walk' function across multiple standards as they exist in the industry today.

This will take the combined efforts of the Standards bodies, Subject Matter Experts and data management professionals to work collaboratively to address each of the various ESG standards and Use Cases that challenge the industry today. Once developed, mapping real data to the SMART DICTIONARY will enable standardized, digital, machine-readable data that can be aggregated, analyzed and trusted.

Without good data, you can't make good decisions.

The SMART DICTIONARY is a pivotal tool in making this possible!

OUTLOOK: SMART DICTIONARY Project, PHASE 1

The "**Proof of Concept" MVP prototype** presented in this document represents **"Phase 1"** of the complete buildout of the DSD-Lab Smart Dictionary. The prototype shows how the terms, concepts, attributes, and definitions (for this example, emission factors described in different standards - EPA, DEFRA/BEIS, TCFD) can be made clearer to **facilitate comparing and contrasting different requirements**.

The prototype achieved its objectives:

☑ The concept of a Smart Dictionary, to define terms and conditions, identify concepts and concept relationships, will enable a harmonization of disparate taxonomies that is **practical, feasible and realistic**

☑ To build a successful Smart Dictionary, relevant Subject Matter Experts will be needed to work alongside Data Management professionals to address the complexities of the many different and disparate disclosures and emissions standards that exist in the industry today.

☑ The results of this MVP will demonstrate that executing a standard, open collaboration approach with SMEs and data professionals will serve as the basis for a "scaffolding" for this and all future projects and Use Cases in building out the shared Smart Dictionary, reducing costs and time to market.

The prototype demonstrated how a Smart Dictionary can enable trusted reporting against different standards -- with less effort, reprocessing, or rewriting.

There is now an understanding of how the linkages between standards like these could be created and how a reusable "scaffolding" can be implemented to share and repurpose knowledge across digital disclosures – thus systematically improving interoperability and automation.

It is from this prototype that the DSD Lab plans to deliver the attributes, definitions, and relationships of the common and shared SMART DICTIONARY for the first proposed Use Case (Carbon Disclosure data), and subsequent Use Cases to follow.