Deliverable

D3.6 OSLO Extensions for the Digital Twin - current status

<table>
<tr><td>Project Acronym:</td><td colspan="2">DUET</td></tr>
<tr><td>Project title:</td><td colspan="2">Digital Urban European Twins</td></tr>
<tr><td>Grant Agreement No.</td><td colspan="2">870697</td></tr>
<tr><td>Website:</td><td colspan="2">www.digitalurbantwins.eu</td></tr>
<tr><td>Version:</td><td colspan="2">1.0</td></tr>
<tr><td>Date:</td><td colspan="2">09 June 2021</td></tr>
<tr><td>Responsible Partner:</td><td colspan="2">AIV</td></tr>
<tr><td>Contributing Partners:</td><td colspan="2">IMEC, TNO</td></tr>
<tr><td>Reviewers:</td><td colspan="2">Internal
Walter Lohman (TNO)
Philippe Michiels (IMEC)
External
Pieter Morlion
Yannis Charalabidis
Andrew Stott</td></tr>
<tr><td rowspan="2">Dissemination Level:</td><td>Public</td><td>X</td></tr>
<tr><td>Confidential – only consortium members and European Commission</td><td></td></tr>
</table>

This project has received financial support from the European Union's Horizon 2020 Programme under grant agreement no. 870697

Revision History

Revision	Date	Author	Organization	Description
0.1	18.03.2021	Lieven Raes	AIV	Initial structure
0.2	25.03.2021	Dwight Van Lancker	AIV	Start section 3: OSLO
0.3	01.04.2021	Bert Van Nuffelen	AIV	Extending section 3
0.4	30.04.2021	Dwight Van Lancker	AIV	Discussing air quality data + Linked Data Event Streams
0.5	30.04.2021	Bert Van Nuffelen	AIV	Extending the OSLO Toolchain
0.6	17.05.2021	Bert Van Nuffelen	AIV	DUET Principles
0.7	24.05.2021	Bert Van Nuffelen & Dwight Van Lancker	AIV	Creating connectors and artifacts (S6)
0.8	26.05.2021	Pieter Morlion	MORE LION	external review
0.9	28.05.2021	Dwight Van Lancker	AIV	Start section 1 + 2, and adopting feedback
1.0	09.06.2021	Bert Van Nuffelen & Dwight Van Lancker	AIV	Final version

Table of Contents

Figures

Code listings

1. Executive Summary

Data interoperability is a key challenge in a digital twin context. A digital twin combines data from different sources into a digital reflection of reality. To facilitate the aggregation of data sources, the semantics of the data has to be unambiguous. This deliverable presents an approach to reach a common semantic understanding tackling the challenges preventing achieving this goal.

The approach is based on the one used by the Flemish interoperability program, known as Open Standards for Linked Organizations (OSLO). The Flemish government wants to facilitate the flow of data by breaking down the vertical data silos of today. OSLO is the cornerstone of the Flemish implementation of the European Interoperability Framework, by providing governance, processes, and methods for standardization of semantic and technical data standards. The principles of the Semantic Web (Linked Data) form the technological foundation. These are elaborated and turned into a data publishing ecosystem to support both the user of the data standards as the editors. For the latter Open Source, editorial development and publication tooling have been created, called the OSLO Toolchain. Using the OSLO toolchain, editors can render the machine-readable specifications in a coherent way.

Until before, the OSLO Toolchain only supported the Dutch language. To be able to adopt the Toolchain on an international level, it had to be extended to support any language. This extension is demonstrated in this deliverable by creating an English data model for an air quality dataset of Pilsen. However, this data model was not aligned with existing data standards in the air quality domain. To increase interoperability, the air quality dataset of Pilsen should be mapped on an existing data standard. This could be the OSLO Air & Water data standard, a standard covering observations in the air and water domain, where DUET helped design it.

2. Introduction

2.1 Objective of this document

Combining multiple datasets to answer complex questions, allowing data-based policymaking is the true power of a Digital Twin. However, to be able to combine multiple datasets, agreements must be made on the definitions of properties, how datasets must be connected, ...

Flanders is committed to create and use unambiguous data standards to exchange information and OSLO is the way to go. With OSLO, Digital Flanders (before, the Flemish Information Agency), together with the academic sector, private partners and citizens focus on semantic interoperability. These semantic standards are very important for intergovernmental dialogue and the reuse of information by the private sector.

2.2 Previous work

The principles of OSLO were already mentioned in a previous deliverable where its concepts were discussed briefly (D3.8 Twin Data Broker Specifications and Tools). This deliverable discusses the process and methodology of OSLO more in detail and aims to be self-contained.

2.3 Structure of this document

In this document, we explain how Flanders tries to improve interoperability by looking at the process and methodology that is used. Furthermore, this document shows how Flanders extended their methodology to create standardized specifications to become multilingual so that international projects such as DUET can adopt it.

In section 3, we discuss the Flemish interoperability program, known as Open Standards, as Linked Organizations (OSLO), and its connection with the European Interoperability Framework (EIF). We conclude this section by explaining how data standards in Flanders are being developed.

Section 4 and 5 focusses more on how DUET can adopt the principles of OSLO, whereas section 6 provides insights on the layer of technical interoperability. In section 7, we map real-world datasets within the air quality domain, to existing data standards.

3. Open Standards for Linked Organisations (OSLO)

3.1. The European Interoperability Framework

Connecting digital systems with each other is often a non-trivial task hindered by small and larger differences resulting from the design decisions of the involved digital systems. To have smooth data exchange, not solely a technical bridge has to be built, but all data interoperability challenges have to be resolved. The European Commission has recognized that getting a grip on data interoperability is crucial for public governments as they offer digital services throughout many different systems. To aid in streamlining the processes supporting end-to-end digital services within and across member states the European Interoperability Framework (EIF) has been established. The EIF categorizes the data interoperability challenges into 4 main categories: legal, organizational, semantic, and technical interoperability. It also provides a set of principles that can be used to drive towards a more interoperable digital environment (European Commission, 2017).

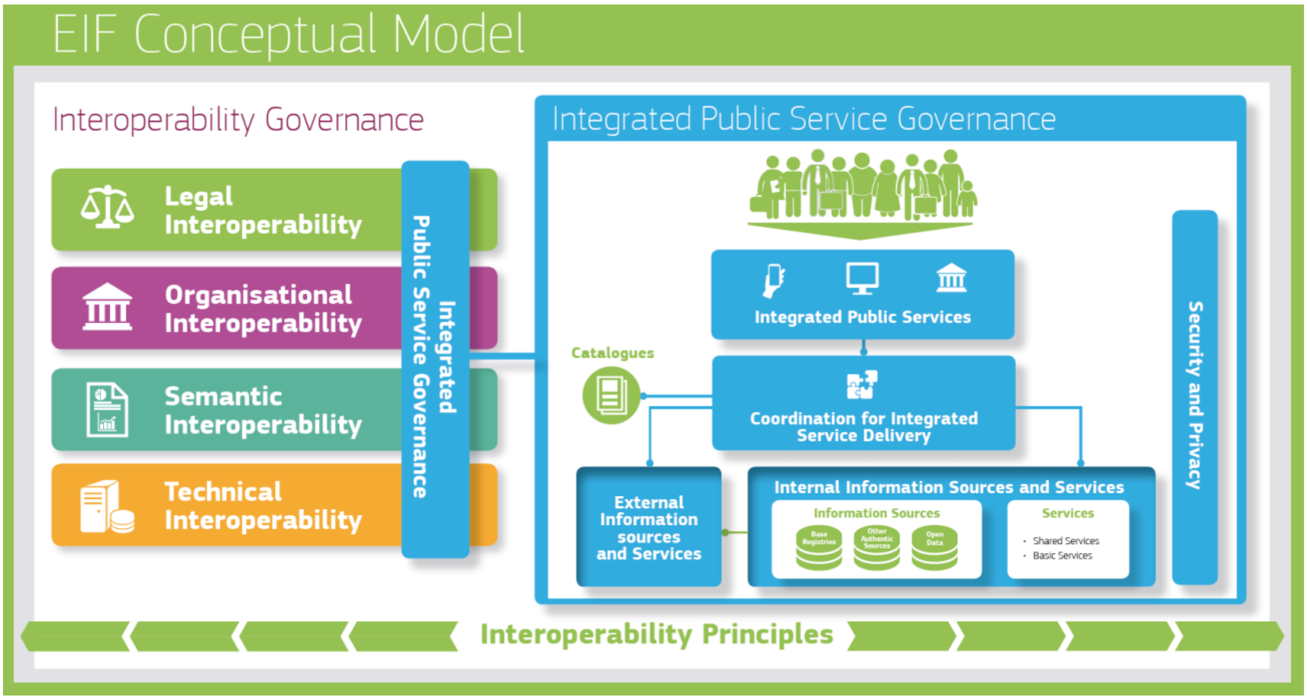

Figure 1: Model that categorizes interoperability in 4 groups: legal, organizational, semantic and technical (Source: https://ec.europa.eu/isa2/eif_en)

3.2. OSLO, the Flemish instantiation of the EIF

The Flemish government is committed to unambiguous standards for the exchange of information. The public services to citizens in Flanders are supported by various specialized applications from different software providers. With Open Standards for Linked Organizations (OSLO), the aim is to ensure greater coherence, better comprehensibility, and findability of data and services.

3.3. A brief history of OSLO

The first phase of OSLO was started in 2012 and ended in 2015 as a grassroots initiative. Local governments (i.e. cities) realized that their public services (more than 1000 services) were supported by over 250 different information systems. This initiated the first OSLO project as a public-private partnership by the Flemish Organization for ICT in Local Government (V-ICT-OR) and co-funded by Flemish ICT service providers and Flemish Government Administrations. The project was also supported by a wider community, including Local, Regional, and Federal administrations, non-profit organizations, academic partners, and the European Commission program Interoperability Solutions for European Public Administrations (ISA) (Buyle et al., 2016).

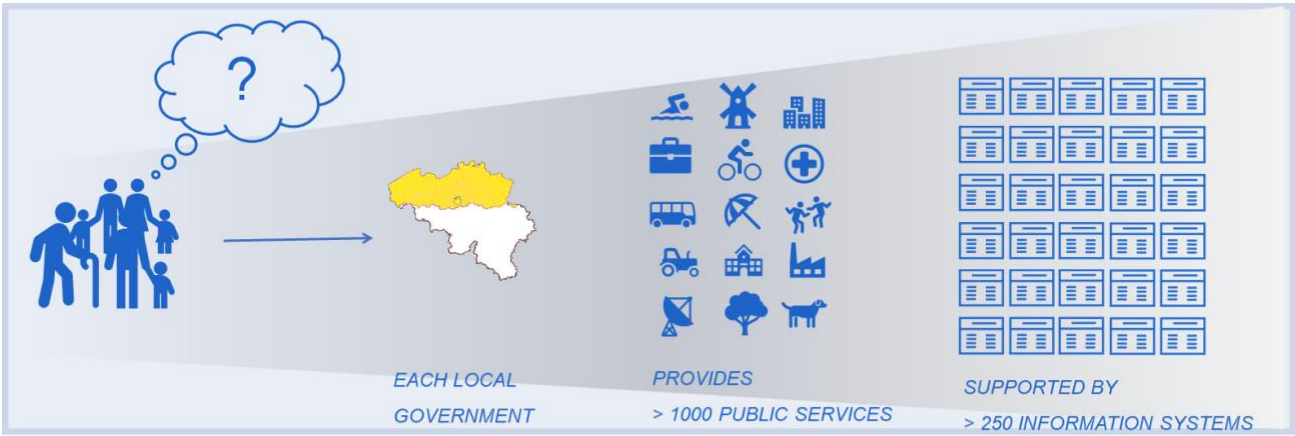

Figure 2: OSLO history

Later onwards the OSLO project became part of the strategic vision *Vlaanderen Radicaal Digitaal* of the Flemish government for digitizing its public administration. In that context, OSLO is integrated into the cross-government governance body *Stuurorgaan voor Vlaamse Informatie en ICT beleid* as the project that steers and guides the semantic data interoperability activities with the Flemish government.
The figure below describes the structure and activities of the OSLO project.

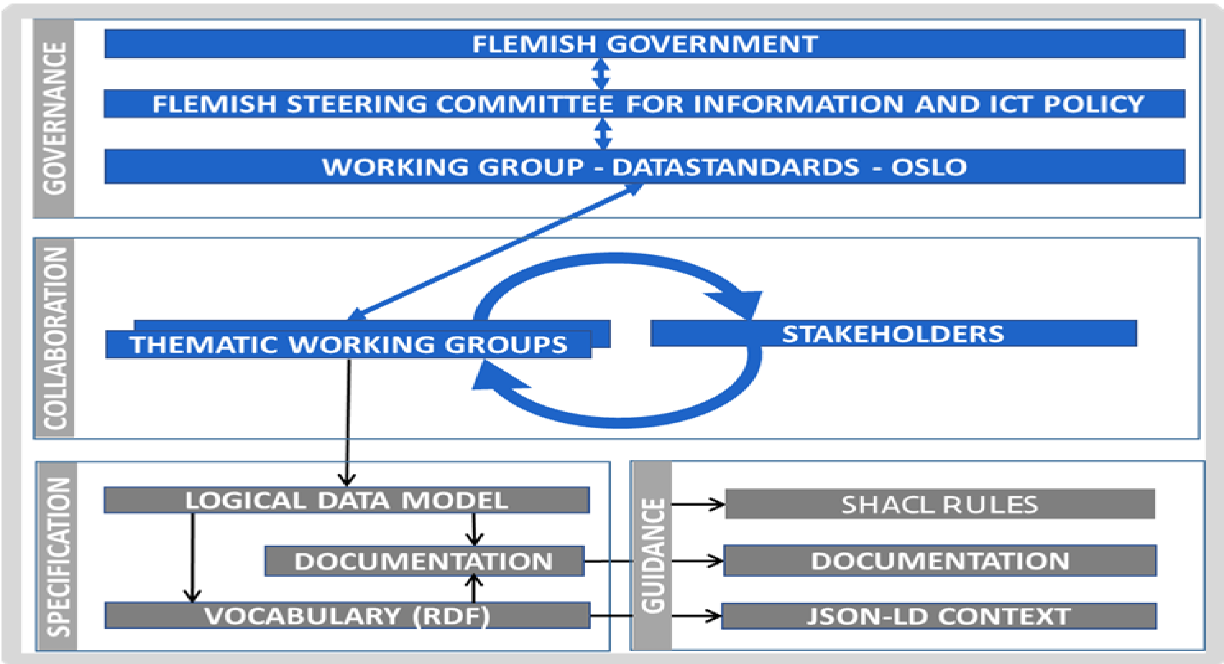

Figure 3: Structure and activities of the OSLO project

The Working Group (WG) Data Standards presented by OSLO is not the only interoperability WG that resides under the Steering Committee. There are also WGs concerning (data) security and authentic & authoritative sources. The latter is of special interest because they create interoperability from a different perspective than a semantical data standard: namely from the data perspective. Authentic & authoritative sources are supported by legislation expressing that all administrations must use these data in their business processes. To facilitate this transition, the agency responsible for the authentic & authoritative sources initiated a data standardization track under the supervision of the WG data standards (Buyle et al., 2016).

The above shows that semantic data interoperability via OSLO has become a well-integrated part of the digitalization of the Flemish Government. To our knowledge this is unique in the world: it does end-to-end standardization, not for a single domain but very distinct domains with strong governance based on open and public consensus building. Today is OSLO reaching out beyond its official legal framework: the Belgian government has adopted the approach for creating consensus at the inter-federal level (more info here).

Even more, the methodology is also used in international projects to facilitate the consensus-building around data standards, for example, the "Collaborative, Secure, and Replicable Open Source Data Lakes for Smart Cities" project (ODALA). ODALA is a strategic European project with the goal to improve data management in cities. An important aspect to accomplish this goal is to agree on the data model for the data that will be used. ODALA is using the OSLO methodology to create a standardized data model to exchange information about air and water quality. The creation of this standard is discussed in more detail in section 4.1.

3.4. OSLO concepts

3.4.1. Process & Methodology

As mentioned in the previous section, it is necessary to follow the OSLO process and methodology, to get acknowledged as an OSLO data standard. The process is based on best practices and guidelines used by ISA, W3C, and Open Stand. It is shown in the figure below. The most important item to note is that although the Flemish Government is initiating and endorsing the work, the work will be conducted in full public transparency. Anyone is welcome to join the working group: usually there are representatives from the government, industry, and academia in the working groups, but occasionally interested citizens join. This open approach working towards consensus enforces the adoption.

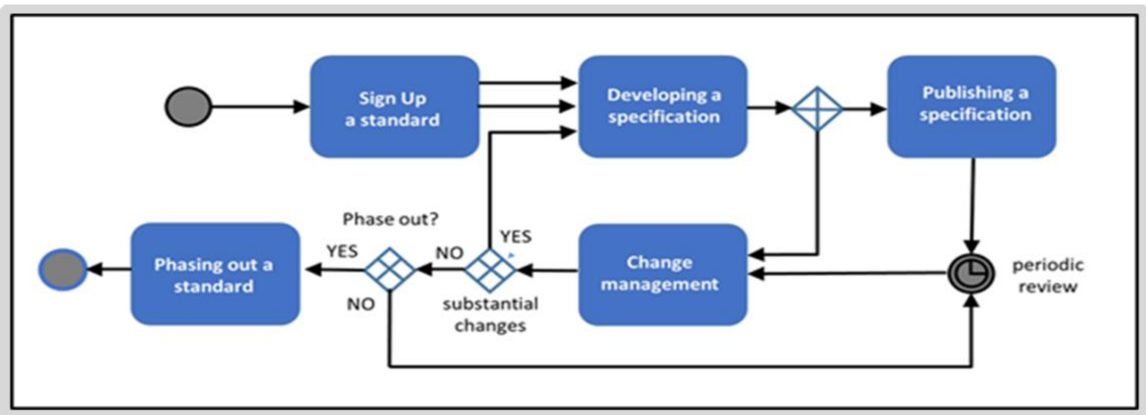

Figure 4: Formal process that needs to be followed in order to get acknowledged as an OSLO data standard

To ensure transparency, every OSLO data standard and the processes around, are published on one single website: https://data.vlaanderen.be/. In the OSLO standards registry, an overview is provided of all the completed data standards, but also the progress of standards in development can be found there. All presentations and reports of WGs are made available through detailed pages on the standards registry (one for every data standard). Via the detail pages, visitors can find the source of the data standard, which is a repository on GitHub. GitHub's issue management is used as the online forum for discussing topics.

The OSLO process & methodology impose guidelines and rules for the modelling of the data standards. The first cornerstone of these is the publication of data standards according to the principles of the Semantic Web. That means that each term that is defined is given a unique identifier: more precisely a persistent dereferenceable URI. The rules for publishing these are found in OSLO Modelleerregels (OSLO modelling rules). The second cornerstone is reuse. By reusing international standards throughout the semantic web principles, the OSLO standards become an integrated part of an international network of data. For those terms that do not have an existing URI that can be reused, a URI will be created. The combination of both is a strong mix that re-enforces each other. By directly referring to international agreements the local Flemish model is ready for sharing data across borders, but also international agreements can observe the extensions and adopt more easily the common practices shared across the usages.

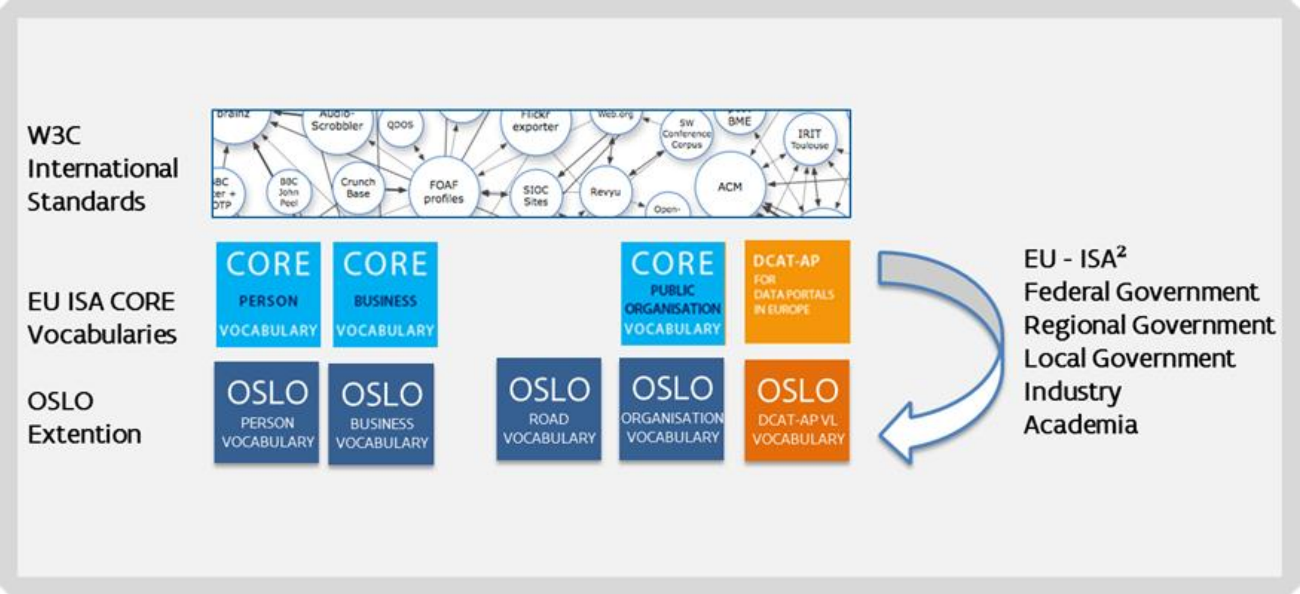

Figure 5: OSLO process & methodology

3.4.2. OSLO Toolchain

Digital Flanders has developed an ecosystem of tools, processes, and governance that allow the designing of data standards in a decentral approach, embedded in global governance. This ecosystem relies on Open-Source software and is executed by open continuous integration systems requiring minimal costs for the organization. The setup uses "a separation of concerns" approach: modelling, provenance, and the publication with a maximal of automation to facilitate expectations for humans and machines are decoupled.

The core process, usually referred to as OSLO Toolchain, works as follows. First, the semantic data modelling starts with an annotated UML document which is then converted to human machine-readable and machine-processable artifacts of which HTML pages, JSON-LD context files, and SHACL files are examples. This whole flow is described as a circle-ci configuration, which allows continuous integration. Initiating the process from an annotated UML document is a well-motivated design choice: it allows to have a uniform graphical representation with a coherent semantic web representation. The other way around is much more difficult to achieve.

In 2020, a new version of the Toolchain, Toolchain 2.0, was made available and supports *versioning*. The main driver for Toolchain 2.0 was the need for improved decentralization and versioning support without the loss of overall coherency. Today, each data standard has its own repository, which functions as the single source of truth for the data standard and also preserves the lifecycle of the data standard. By registering publication-ready commits of their repository in a central repository, a coherent view is created.

To facilitate the adoption of data standards, the core process is surrounded by tools and other guidelines:

- A URI strategy for persistent identifiers, together with an implementation example on the domain data.vlaanderen.be
- A compliance checking set up to check if data is compliant with an OSLO data standard. This setup uses SHACL and is based on the ISA testbed validator.
- Process provenance through a registry of data standards
- Tooling for the publication of code lists with persistent identifiers
- A playground to create your OSLO compliant payloads.

The methodology and setup have been applied in real-world projects where the data standards, vocabularies for broad reuse, and application profiles for usage in a generic application context, are turned into implementation models using the same Toolchain approach (Buyle et al., 2019).

***Note**:*
The objective of the OSLO Toolchain is not to become a toolchain for publishing the ultimate data standard, but a toolchain enabling that data standards are published using the same foundation, i.e. the Semantic Web so that differences and correspondences are made visible. It relies on elements in common to most data modeling languages: Classes, Properties, and Instances, identified with a URI.
Each of these entities is described with an interpretation in the real world (definitions) and preferred label. Only a few data modeling constructs are supported: subclass, subproperty, domain, and range. No additional "logical inference" statements like rules, OWL formal semantics, negation, ... are used.

Exploiting additional formal semantics to enable automated reasoning is possible when these do not violate the real-world interpretation and are not imposed as mandatory when using the data standard. This lightweight formal semantics approach creates a highly reusable base layer for sharing information between systems.

4. DUET principles

The data interoperability challenges, described in the previous sections in the generic context of the public sector, reappear in the smart city context in which DUET is situated. Therefore, the same solution principles can be applied to address these challenges. By establishing a stable semantical foundation for the data that is exchanged throughout the DUET platform, (re)use will be stimulated.

The contributions of DUET positioned in the EIF are mostly on the **technical** and **semantic** layers. The organizational and legal challenges are not in the scope of this technical work package. The impact on these is mostly indirect by stimulating and engaging responsible organizations in pilots, experiments, or demonstrations. Based on these experiences organizations might engage themself to invest in a more interoperable data ecosystem. Similarly, but even more challenging are legislative changes. DUET does not aim for this and therefore no direct legislative impact that is connected with DUET activities is to be expected.

OSLO tackles the semantic interoperability challenges by creating vocabularies and application profiles, the latter of which also help tackling the challenges at a technical level. However, this deliverable focuses on the **semantical** contributions of DUET, particularly the methods and means provided to facilitate semantical agreements. Technical agreements and contributions are documented in other project deliverables, such as D3.1 IoT stack and API specifications, D3.2 IoT stack and API specification v2, and D3.8 Digital Twin data broker specification and tools v1, which both focus on the technical architecture of Digital Twins. Other examples are D3.5 Cloud design for model calibration and simulation and D3.8. Nevertheless, the objective of data interoperability solutions is to reduce the impedance between the intentional semantics and the operational aspects. Therefore this deliverable also includes a novel technical approach, called Linked Data Event Streams, that allows sharing data closely bound to semantics.

4.1. Designing a commonly agreed data model

Rather than initiating a track itself to create a data model, DUET participated in the data modelling track for air and water quality that was initiated by the ODALA project. The data model to measure air and water quality has been designed according to the OSLO process and methodology, allowing contributions from the ODALA partners and external input from projects such as DUET.

To develop a data model, multiple public working group sessions have taken place or are yet to come:

- December 17th, 2020 → A business workshop determining the context, scope, and objectives of the data model
- February 9th, 2021 → A first thematic workshop to define the first draft model
- March 4th, 2021 → A second thematic workshop to discuss the proposed model
- Next activities

During these sessions, the air and water quality model has been elaborated step by step. Comments and contributions are made here: https://github.com/Informatievlaanderen/OSLOthema-airAndWater/. In this modelling different existing (meta)models were being considered and brought together into a coherent data model. It turned out that to reach an air and water quality model there must be agreement on the meta model to be used. So this ODALA track not only created a dedicated sensor-specific model but also a common basis aligning different IoT ecosystems. Initially, DUET started with the idea to create a common minimal data model. The ODALA Air & Water Quality track shows that there is an ability to create a common basis connecting different IoT ecosystems at a semantic level. It includes aspects from SSN, SOSA, OM, and FIWARE. Although there is no obligation by DUET to align on a common model, DUET users can use this integrated data model to start their publication of data in other domains than Air and Water quality. By doing so data will be anyhow more reusable.

The ODALA Air & Water Quality model contains multiple components currently published on the OSLO test environment. The first component is the Core model. As mentioned above, the model aligns several existing data standards. It is based on ISO Observations & Measurements[1], and extended with SSN/SOSA Sensors & Sampling[2], objects from SSN/SOSA Extensions[3] (e.g. ObservationCollection) and the Device object from FIWARE[4]. Metadata aspects not covered by previous standards are added according to the Dublin Core standard. As it is possible to describe many objects with this model, such as observations, sensors, devices, … it does not yet tell what is being observed and measured. Such an agnostic Core model design ensures that it can also be applied in other use cases than air and water. For example, the Dutch version of the Core model is being used in another OSLO trajectory, called 'Soil and Subsoil' where Observations are used to describe soil samples and the object SampleProcess for ground drillings. Another use case could be counting loops, as we could interpret counting loops as Sensors and the road segments they measure as SamplingFeatures. This way, we could use the model to measure road traffic.

The other components of the ODALA Air & Water Quality track are a model to observe and measure air quality and a model to observe and measure water quality. Both models use and extend the Core model in the same way. The Observation object becomes more detailed as air and water quality parameters are added. Also, the DomainObject from ISO Observations & Measurements was too abstract and meaningless for a certain domain, and thus a generic Air/Water object was added to the model. These objects can be typed according to an official code list or being substituted by classes such as OGC HY_WaterBody or OGC HY_Canal.

Note*: This model is still being developed. If it is being referred to in future deliverables, it is possible that changes have been applied to it and that things mentioned in this deliverable are no longer valid.*

4.2. Supporting the publishing of semantics

To address the data interoperability challenges one can also apply the OSLO concepts in a more practical setting. Namely, instead of focussing on consensus building, the OSLO concepts facilitate the documentation of the data in a common technical format & approach. Many data sources are accessible but not documented in the same way and at the same level of detail. Best practices such as in the REST API community to use OpenAPI are a step forward. But they focus on the technical structures; the business perspectives are often

[1] https://www.ogc.org/standards/om
[2] https://www.w3.org/TR/vocab-ssn/
[3] https://www.w3.org/TR/2020/WD-vocab-ssn-ext-20200116/
[4] https://fiware-datamodels.readthedocs.io/en/test_next_version/Device/Device/doc/spec/index.html

documented in additional documentation. These principles on how to manage Smart Data are discussed in D3.2 IoT stack and API specifications v2.

The OSLO methodology to publish the semantics according to the principles of the Semantic Web complement the OpenAPI guidelines. OSLO applied these principles to make them technically neutral, applicable therefore to any technical context, being a SOAP service to a REST JSON API.

To demonstrate this practical approach the Czech dataset of air quality sensors[5] has been modelled and published using a DUET specific setup of the OSLO toolchain. The setup and extensions that are made to make it applicable for DUET are discussed in the next chapter.

The Czech air quality sensor model[6], visible in the image below, is a direct translation of the data provided by the JSON feed. The first draft of the model has been designed independently from the publishers of the dataset. It is therefore rather sparse and requires improvements. The next step in improving this data model is reaching out to the publishers and capturing their feedback. Observe that the model is also rather flat and does not apply any data modelling principles to increase reuse, like common code lists, mapping on external generic vocabularies, etc. When there is room for adapting the structure of the feed, then even the alignment with/adoption of a common data model as the Air and Water quality data model can be considered.

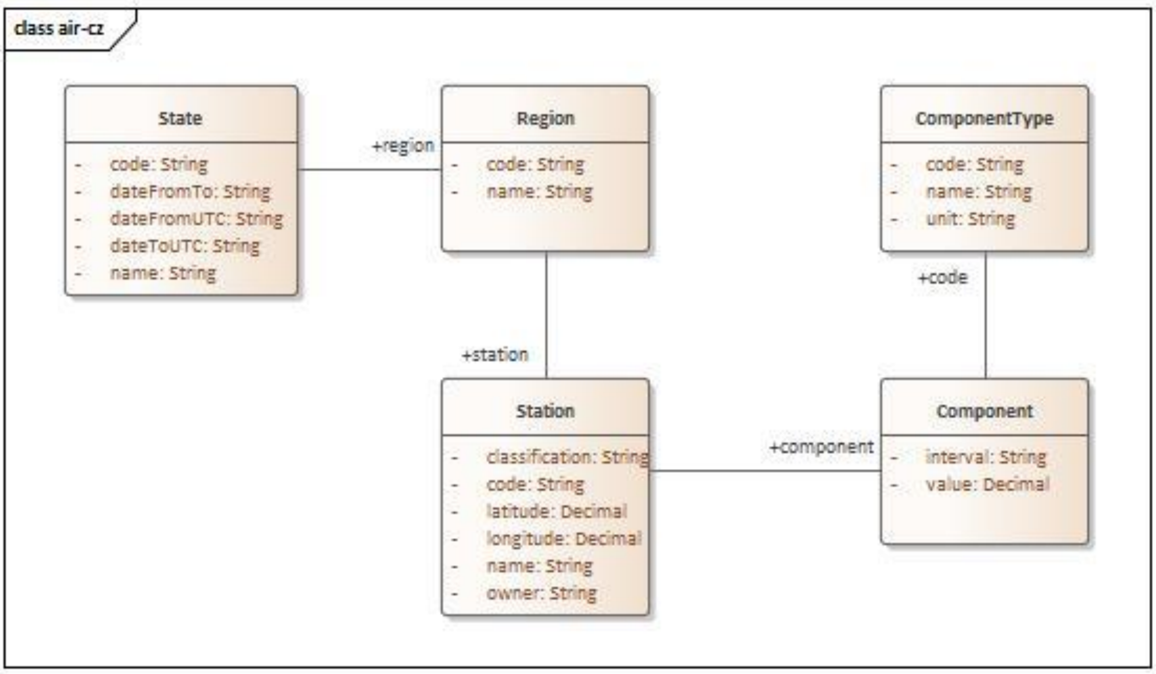

Figure 6: Czech air quality sensor data model that was used to render the specification available at https://duet.dev-vlaanderen.be/doc/applicationprofile/air-quality-czech/index.html

Publishing this model can inspire other air quality sensor publishers, e.g. the city of Athens, to also publish their air quality data. But more importantly, it will aid developers that access the data throughout the DUET platform to consume the data much easier. The semantic layer provides the means to understand how and in which context the data can be used. E.g.

[5] See: https://www.chmi.cz/files/portal/docs/uoco/web_generator/actual_hour_data_cze.json
[6] See: https://duet.dev-vlaanderen.be/doc/applicationprofile/air-quality-czech/index.html

What this experiment shows is that semantically describing a dataset using the OSLO toolchain can be started with minimal effort. And based on this initial version an alignment process can be started.
As already indicated, it is not necessary that the data is expressed in accordance with a universal data model. The OSLO approach does not require it but makes the variety of perspectives more controllable. Namely by aligning the technical language in which the models are expressed. The relationship between two data models can be studied and discussed in the same setting, in a technology-neutral approach. Tooling such as the mapping language RML (https://rml.io/) could be used for this. Deliverable D3.2 IoT Stack and API specification v2 discusses this in more detail.

The DUET data catalogue is a means to allow DUET users to find information about a dataset. The published model URI can be used in the DUET data catalogue to provide the semantics (schema) of the data. The OSLO toolchain approach allows for every DUET data source a data model in an implementation-neutral way. The DUET data catalogue can benefit from this approach.

5. The OSLO toolchain for DUET

5.1. Setup & configuration

The OSLO toolchain is a configuration of several GitHub repositories that are connected through continuous integration pipelines resulting in a static website. In addition to the OSLO Toolchain, the OSLO ecosystem also contains additional supporting tools, also Open Source, that can be deployed and take benefit from the outcomes of the OSLO toolchain. The deployed setup and how users can contribute are (briefly) explained in this section.

5.1.1 GitHub repository

As mentioned in the previous sections, the single source of truth for an OSLO data model is its GitHub repository. This repository contains all information that is needed by the OSLO Toolchain to create the various specifications (HTML page, RDF, JSON-LD context and SHACL template) and must be maintained by the editor of the data model throughout the process of creating a data standard.

The repository that is used for DUET can be found at: https://github.com/Informatievlaanderen/duet-vocabularies.

5.1.2 Starting the Toolchain

Once all information is present in the GitHub repository, the editor can start configuring the inputs that are needed by the OSLO Toolchain. The configuration happens in a specific configuration file called, publication.json and below is an example of the possible content of this file. Again, this requires another GitHub repository. For DUET, this repository is available at https://github.com/Informatievlaanderen/duet.

```
{
        "trigger" : 3,
        "urlref" : "/doc/vocabulary/air-cz",
        "branchtag" : "master",
        "name" : "air-cz-voc",
        "filename" "config/air-cz-voc.json",
        "examples" : false,
        "navigation" : {}
}
```

Code listing 1: Possible content of a publication.json configuration file

The properties worth mentioning are the **urlref** which specifies at which path this specification should be published. The property **name** refers to the name of the data model in the UML model, which is created by using the program Enterprise Architect. In this case, we are configuring the publication of a vocabulary. If one wants to publish another specification (a vocabulary or application profile), then another JSON object must be included in the configuration file.

Once the configuration is done and the changes have been pushed to the GitHub repository, a continuous integration process using circle-ci is activated. This process will start a series of scripts, which together are called the OSLO Toolchain. The configuration is found at https://github.com/Informatievlaanderen/duet/blob/master/.circleci/config.yml.

The process consists of a data aggregation phase in which all data about a specification is collected, and an artifact generation phase in which based on the aggregated data the desired artifacts are created. The generated artifacts are HTML pages, vocabularies in RDF format, JSON-LD contexts, SHACL templates, examples along with reports indicating possible errors. All the results are stored in a dedicated GitHub repository ready to be published as a static website. For DUET this repository is available at https://github.com/Informatievlaanderen/duet-generated.

5.1.3 Showing results on the website

The mentioned outcome in the previous section is being pushed on another GitHub repository which is used to display the results on the website. That means that the HTML specification will be available at the URL that was configured in the configuration file, machines will be able to read the vocabulary of the data model, etc…

The website for DUET is available at https://duet.dev-vlaanderen.be/. All generated specifications can be found on this website.

5.2. Extensions

As explained in a previous deliverable (D3.1 IoT stack and API specifications v1), collaborating across Europe poses challenges to the creation of semantic data standards. Language and cultural differences must be bridged but should not be avoided by defaulting to a single communication language. For the adoption and the quality of the data standard allowing access to the terminology in the user's language is a must. As DUET is creating a platform that aims to be applicable in Europe's variety of languages and cultures implemented by local development teams embedded in their local context, having a tool suite that enables them to create data standards in any language is an asset. International agreements get connected with local variants.

Semantic Web technology is by design very suited for the European multilingual challenge. The data format RDF has incorporated language support that does not require structural adaptations when switching from non-language awareness to language awareness. Other generic data serialization formats require structural adaptations. In JSON this is typically done by adding an additional attribute or turning a value into a complex object with the languages as an index. Even XML, which has extensive language support, is vulnerable. Because adding language support typically means that the cardinality constraints change instead of a single value for the attribute *name,* for example, there will be more values (often one per language). If multilingualism has not been foreseen, then the impact is often a structural change on the XSD.

Although the OSLO toolchain is built upon Semantic Web technology, adaptations to the toolchain are required to incorporate multilingualism. It included obvious changes like providing templates in alternative languages, but also adaptations to the process flow to facilitate the translation process. Multilingualism brings an additional dimension to the publication process: how to keep the translations in sync with the original data model? Out of the OSLO experience keeping the existing data models coherent and consistent is already a challenge. It requires substantial attention from the responsibility to react and adapt the models swiftly. Translations make this even more complex: terms appear or disappear, labels and definitions change, etc. For that reason, translations are maintained in the toolchain as first-class citizens. For each data model a unique translation object per language is defined which is created automatically. (See below an extract from the translation object for the Czech air quality sensor model with base language English and target language Dutch.) This translation object contains all translatable content (labels, definitions, usage notes) that is part of the data model. When a data model is reprocessed then the new translation object is compared with the previous version. Any change will be reported so that the ones responsible can take the appropriate action. This can be adding a new translation, but also revalidating existing translations when a change has been detected in the master text.

```
{
  "baseURI": "https://duet.dev-vlaanderen.be/ns/air-cz#",
  "classes": [
    {
      "EA-Guid": "{4C3E533D-0698-427f-9369-46C9D45C0EE0}",
      "name": "Component",
      "label": {
        "en": "Component",
        "nl": "Enter your translation here"
      },
      "definition": {
        "en": "A measurement for an air quality component",
        "nl": "Enter your translation here"
      }
    }, ...
}
```

Code listing 2: Translation object for the Czech air quality sensor model

In https://github.com/Informatievlaanderen/duet/tree/master/templates HTML rendering templates for the publication of DUET specifications in English and Dutch are stored. And the following extract from the SHACL template shows the result for the English language.

```
 <https://duet.dev-vlaanderen.be/shacl/air-cz-ap_en#ComponentShape> a
shacl:NodeShape;
  shacl:closed false;
  shacl:property [
    shacl:definition "the value measured in the interval"@en;
    shacl:maxCount 1;
    shacl:minCount 1;
    shacl:name "value"@en;
    shacl:path <https://duet.dev-vlaanderen.be/ns/air-cz#value>
  ], ...
```

*Code listing 3: **Extract from the SHACL template***

6. Creating connectors and artifacts

Agreeing to a common semantic perspective is important and necessary work, but it is a means to facilitate actual data exchange between systems. Therefore, the semantic layer cannot be considered isolated from the technical layer. The OSLO ecosystem provides and invests in tools and approaches to bridge these layers so that the semantic agreements are taken into account when implementing a data exchange. Also, the inverse direction is important: technical agreements may be lifted into the semantics agreements.

In this section, we describe two existing supports and one novel addition to the OSLO ecosystem, which are relevant to the DUET ecosystem.

6.1. Conformance checking

When implementing a data exchange, conformance with the semantics is expected. And by preference the checking is automated, removing any human interaction. Conformance is, however, a notion covering many aspects.

One aspect is *coverage:* the number of terms that are used in accordance with one or more data models. The '*in accordance with'* measure depends on the nature of the data model.

- A term used in implementation is in accordance with a term in a *vocabulary* if the semantics of the implemented term corresponds with the definition in the vocabulary and the additional basic constraints on the domain, range, and subclassing give rise to semantic incompatibilities.
- A term used in implementation is in accordance with a term in an *application profile* if it is in accordance with the vocabulary where the term is defined and if the additional constraints expressed as cardinalities, code lists, and usage notes are satisfied.

The coverage ratio can be viewed from two perspectives: a) from the implementation perspective to the data models or b) from the data models to the implementation. In the first case, one can discover terms in the implementation that have no semantics yet or are used in a way that is not covered by any data model under consideration. Looking from the other perspective, one can discover the absence of expected information. This is a typical assessment for mandatory properties in application profiles.

Another aspect is *level*: the strength of the syntactic and semantic connection with a data model. The first level is syntactic connection. That means that the names of the terms used in the implementation are corresponding to the labels in the data model. A direct usage, although often implicitly expected, is seldom possible because implementations tend to follow other naming conventions. Sometimes additional disambiguation is required.
A common hurdle that blocks syntactic correspondence is that the language expectation for the semantics is the language of the main target users (e.g. the official language of the country), while technical teams desire to provide the implementation in English to facilitate the development of applications by an international developer team.

The second level addresses the syntactic hurdles. Instead of binding the semantics via an implicit connection via the syntax, the connection is made explicit. The implementation provides a means by which each term in a data stream is connected with the appropriate semantic term identified by its URI. This connection should be available in a machine-processable way so that the mapping is unambiguous. Users of the data are thus able to retrieve the followed semantics without applying their own interpretation based on syntactic similarities.

The next and last level of conformance states that the data can be validated w.r.t. to the constraints expressed in the data model. Within the Semantic Web, the reference approach for this is SHACL. SHACL is a W3C recommendation defining a language for expressing constraints on an RDF graph. To support reaching this level the OSLO ecosystem has adopted the ISA testbed as the component for validating RDF data w.r.t. SHACL constraints. And for each data model, the OSLO toolchain also produces a SHACL specification.

This connects the levels with each other: at the first level the connection with the semantics is implicit relying on simple syntactic correspondences; the second level any data is convertible using an automated process to the semantic web and in the third level the outcome of the conversion process is validatable. To conclude, the discussed conformance measures are defined as implementation technology agnostic - as no technology choice has been made -, but measure the conformance as the strength of the connection with the Semantic Web because in that formalism the data models are expressed.

6.2. Sticky semantics

The conformance statement is on-purpose data publishing technology agnostic. This avoids the need for technology harmonization and opens up for adoption of the semantics in the data publishers' technology. The drawback of this is that designing generic building blocks and generic business processes which can be combined like Lego blocks is much harder. Ecosystems like INSPIRE (geospatial) or FHIR (health) combine semantic and technological harmonization, restricting much more the freedom of the participants of these ecosystems to the benefit of more generic components. The OSLO ecosystem is more open. Nevertheless, minimal investments by data publishers have to be made to obtain conformance: namely expressing how the data is connected with the semantics. This expression should be best done embedded in the data: sticking the semantics to the data. When the data from one party goes to another party and then to a third party then the semantics are carried with.

To obtain sticky semantics it must be supported by the data formats in which the data is shared. In the Semantic Web, the data format RDF does this by design: every term is represented by a URI. And if these URIs are dereferenceable: i.e., linked to the semantics, then any data expressed as RDF is stuck to the semantics. Other data formats do not have this embedding. For one popular data format, JSON, there is however JSON-LD. It is a specific form of JSON that expresses in a JSON-friendly way a mapping to the RDF data format. JSON-LD is a W3C Recommendation. Therefore, to have sticky semantics JSON data should be expressed as JSON-LD. Due to the limits of the JSON-LD mapping language this activity may affect the JSON structures used to obtain an RDF structure in accordance with the data model.

For other data formats, e.g., XML, such a standardized mapping language does not exist. In those cases, the advice is to create a mapping implemented as a program, translating the data into the RDF graph. Providing this mapping as an executable service and sharing the program code with the data consumers provide the data consumers with all the means to understand the data. This service then becomes the canonical interpretation w.r.t. the semantics of the data model. Whereas switching from JSON to JSON-LD represents limited effort (often already done by the data publisher when investigating a data model) providing an executable interpretation for other languages represents more work. But since this is explaining to the consumers what the data actually means, this is work a data publisher has to do.

The OSLO toolchain supports sticky semantics to create JSON-LD mappings (called contexts) for the data models it publishes. These reduce the adoption of turning a JSON data feed into a JSON-LD data feed with sticky semantics.

6.3. Semantics for changing data

Datasets are not static entities. Even though data models might suggest this by only modelling a desired state of affairs, in practice data changes. Where-as the above contributions mostly suffice to address data with a more or less static publication approach (dumps, REST APIs, SPARQL endpoints, ...), dealing with change requires more.

The Linked Data Event Streams (https://w3id.org/ldes/specification) specification, is a novel approach that incorporates both technical standardization challenges as the connected semantic challenges in the context of publishing changes to data. It is a contribution to the Semantic Web ecosystem enabling a semantic perspective on the publication of data changes. So that from a semantic perspective data challenges as data synchronization can be addressed.

6.3.1 Linked Data Event Streams

In Flanders, the re-use of public sector information is embedded in the legislation and as a result, Flanders has an Open Data portal containing over 4000 datasets. However, due to a lack of interoperability, it ensures that for interconnecting and interpreting these sources of information, a great effort is still required from public administrations, businesses, and citizens. As already mentioned in chapter 3, to break down these data silos, the Flemish Government applied the principles of Linked Data in a public sector context, resulting in the Flemish Interoperability initiative, OSLO.

As the lack of interoperability was (is) a problem, the way how data is published is too. If we take a closer look at the different strategies to publish data, then we can divide them into two categories: querying APIs and data dumps. In the case of a data dump, users always have to fully download the latest version to stay up to date with the dataset, and once they fail to download the latest version, working with an outdated view of the dataset. With a querying API on the other hand, users can query the dataset without first having to download the entire dataset. However, trying to meet the needs of their re-users, data publishers will have to provide and maintain an increasing amount of such querying APIs as specific end-user features are solved by creating feature-specific APIs. So, to avoid synchronization problems with data dumps, and maintenance problems of an always increasing amount of querying APIs, trade-offs need to be made.

Data publishers must accept that they will not be able to implement any querying API on their own, but that there are other organizations with other interests that can take up parts of the processing. This resulted in Linked Data Event Streams (LDES), which is a strategy that allows re-users to synchronize with the history of the dataset but also retrieve the latest updates of the data collection. An LDES extends the principles of a regular event stream, but it publishes interoperable data re-using existing machine-readable data standards.

The definition of an LDES is a collection of immutable objects, such as an observation made by a sensor or the version of an address.

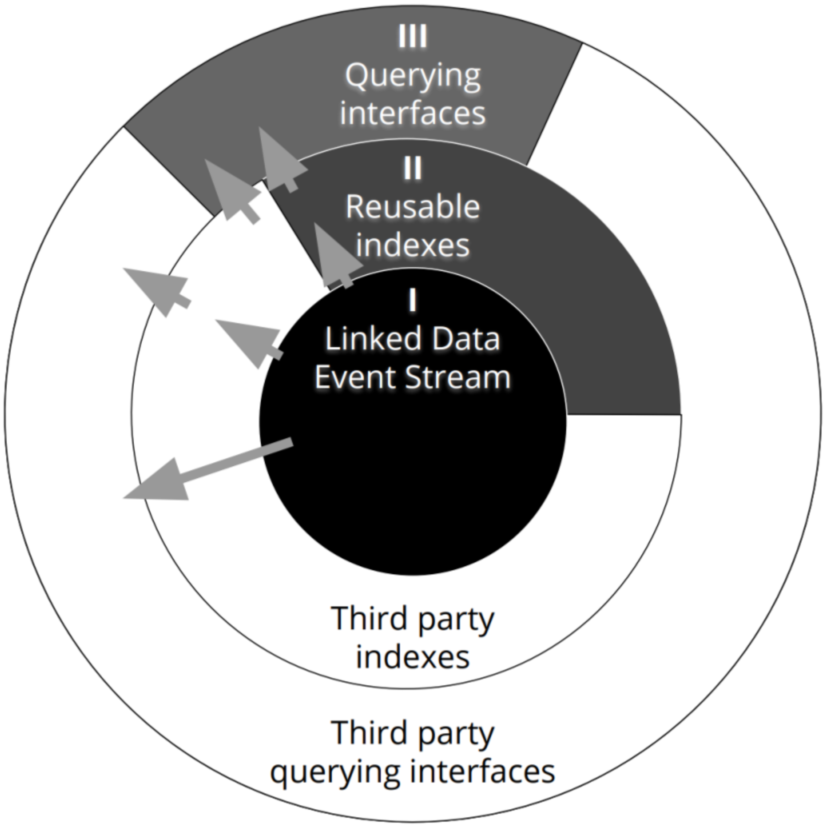

Figure 7: Example of a three-layered ecosystem on how to publish and process data

The figure above represents a three-layered shield illustrating the entire ecosystem of how data should be published and processed. In the core (I), there is the Linked Data Event Stream, the part that is mandatory for the organizations that are the authoritative source of the dataset. However, querying over the LDES can result in slow queries. Therefore, the second layer (II) consists of the creation of reusable indexes. For example, the creation of a geospatial index of addresses so that data consumers only have to download the part that they are interested in, instead of the original LDES. The third layer (III) in this ecosystem consists of querying APIs which are built on top of the indexes or if none exist, due to lack of financial resources or other priorities, on top of the LDES. Both the second and third layers can be done by the data controller but can equally be done by third parties (Van Lancker et al., 2021).

7. Data examples

This deliverable has focused on publishing air quality data from different regions as a knowledge graph. In the first example, we publish air quality data from Flanders as a Linked Data Event Stream. In the second data example, air quality from Pilsen is published in a machine-readable format, JSON-LD.

7.1 Flemish Region Air quality data

To improve air quality in cities, it is important to have an objective image of the air quality in cities. In Flanders, this is done by measuring air quality through very accurate sensors of the Flemish Environment Agency (VMM). However, over a distance of 20 meters, the air quality may have already changed. Therefore, mobile sensors are needed to create a more fine-grained network of air quality sensors. In Antwerp, a project was done where they fitted highly sensitive air quality sensors to the roofs of Bpost delivery vans who drive through the city all day long[7].

The sensors observe and publish the air quality at a high frequency, and each observation of air quality can be considered an *event*. Looking at the sensors only, some of them are mobile, which means a sensor can have different locations throughout the day. So, each sensor has different *versions* over time. Taking these concepts in mind, it is possible to publish this fast-updating dataset as a Linked Data Event Stream. At the moment of publishing this dataset as an LDES, no OSLO data model for this domain existed yet. As this has already been mentioned multiple times in this deliverable, ODALA has initiated a data modeling track for air and water quality. Once this track has been finished, it is the intention to modify this dataset, so it complies with the OSLO data model for air quality.

Because of the lack of an OSLO data model, the Semantic Sensor Network ontology[8] (SSN) was used to model sensors and observations. Listing 1 (in JSON-LD format, without context to save space) shows an observation object in JSON-LD format described with the SSN ontology as part of the LDES:

- **id** → Each observation has its unique identifier so that others can refer to it
- **memberOf** → This property is a property of the LDES specification and not SSN, and it indicates to what collection this observation belongs.
- **madeBySensor** → This property describes the relation between the observation and the sensor that made the observation
- **resultTime** → Indicates at what time the observation was made
- **hasFeatureOfInterest** → According to SSN, this property should be used to describe the entity whose quality was observed (e.g. PM10). However, here it was used to indicate at which location the observation was made.
- **observedProperty** → Relation with the property that was observed in this observation
- **hasResult** → Holds the value of the observed property and its unit.

[7] See: https://www.imeccityofthings.be/en/projecten/bel-air

[8] See: https://www.w3.org/TR/vocab-ssn/

```
{

"id":"https://streams.datapiloten.be/observations#6d8cdef14e8a0e41ee92bf37b1fe54a9"
,
  "memberOf":"https://streams.datapiloten.be/observations",
  "madeBySensor":"{IDENTIFIER OF THE SENSOR WHO MADE THE OBSERVATION}",
  "resultTime":"2020-07-06T17:29:13.185Z",
  "hasFeatureOfInterest":{
     "asWKT":"POINT (4.41591365262866 51.2375716958195)"
  },
  "observedProperty":"cot:PM25",
  "hasResult":{
     "numericValue":3.154094934463501,
     "unit":"m3:MicrogramPerCubicMetre"
  }
}
```


Code listing 4: Example of an air quality observation object

Listing 2 shows how sensors are being published as an LDES. However, when implementing the LDES for sensors, the model had to be extended, because compared to observations, a sensor does not know the concept of time. To solve this problem, the model was extended with the property **dcterms:isVersionOf**, to indicate which sensor this is a version of.

- **id** → Uniquely identifying each version of a sensor, so that one can refer to it
- **isVersionOf** → This property is needed to be able to speak about sensors in time
- **prov:generatedAtTime** → This timestamp corresponds to the time at which the sensor published its information
- **hasGeometry** → Holds the coordinates of the location of the sensor
- **observes** → Indicates which properties are being observed by that sensor


```
{ "id":"https://streams.datapiloten.be/sensors#lora.3432333857376518@2020-06-
30T14:32:57.784Z",
   "isVersionOf":"https://streams.datapiloten.be/sensors#lora.3432333857376518",
   "prov:generatedAtTime":"2020-06-30T14:32:57.784Z",
   "memberOf":"https://streams.datapiloten.be/sensors",
   "hasGeometry":{
      "asWKT":"POINT (4.8376199416816235 50.977539075538516)"
   },
   "observes":[
      "cot:PM1",
      "cot:O3",
      "cot:PM25",
      "cot:NO2",
      "cot:PM10"
   ]
}
```


Code listing 5: Example of a version object of a sensor

7.2 Czech Republic air quality data

In Pilsen, air quality from 6 official and certified meteorological stations are available as Open Data in JSON and XML format (and HTML). This dataset is updated each hour and contains only the latest state of each measuring station.

As this dataset is not yet available in a machine-readable format (RDF), OSLO created a data model that maps to the data model that is used in the original dataset. No alignments were made with existing standards (OSLO Air & Water, SSN/SOSA, …). The data model can be found on the DUET environment within OSLO: https://duet.dev-vlaanderen.be/doc/applicationprofile/air-quality-czech/index.html. This data model was used to transform the air quality data in Pilsen from JSON to JSON-LD, visible in Listing 3. Compared to the original dataset, each property and object are now uniquely identified and available in a machine-readable format, JSON-LD.


```
{
"@context":"https://gist.githubusercontent.com/ddvlanck/027248d8c0b9f52021875d65b1e4463b/raw
/d29496f26f550d3739cbb0969dd9ae8261fc3438/czech-air-quality-context.jsonld",
   "@graph":[
      {
         "@id":"http://example.org/id/state/Czech_Republic",
         "@type":"State",
         "State.code":"CZ",
         "State.name":"Czech Republic",
         "State.dateFrom (utc)":"2021-05-19 11:00:00.0 UTC",
         "State.dateTo (utc)":"2021-05-19 12:00:00.0 UTC",
         "State.dateFromTo":"2021-05-19 13:00 - 14:00 CEST",
         "State.region":[
            {
               "@id":"http://example.org/id/region/Prague",
               "@type":"Region",
               "Region.name":"Prague",
               "Region.code":"A",
               "Region.station":[
                  {
                     "@id":"http://example.org/id/station/ABREA",
                     "@type":"Station",
                     "Station.code":"ABREA",
                     "Station.name":"Praha 6-Brevnov",
                     "Station.owner":"CHMI",
                     "Station.classification":"B/U/RN",
                     "Station.latitude":"50.084385",
                     "Station.longitude":"14.380116",
                     "Station.component":[
                        {
                           "@id":"http://example.org/id/component/ABREA/SO2",
                           "@type":"Component",
                           "Component.code":"http://example.org/id/componentType/SO2",
                           "Component.interval":"1h"
                        },
                        {
                           "@id":"http://example.org/id/component/ABREA/NO2",
                           "@type":"Component",
                           "Component.code":"http://example.org/id/componentType/NO2",
                           "Component.interval":"1h",
```

```
                          "Component.value":"12.5"
                        }
                      ]
                    }
                  ]
                }
              ]
            },
            {
              "@id":"http://example.org/id/componentType/SO2",
              "@type":"ComponentType",
              "ComponentType.code":"SO2",
              "ComponentType.unit":"sulphur dioxide",
              "ComponentType.name":"µg/m³"
            },
            {
              "@id":"http://example.org/id/componentType/NO2",
              "@type":"ComponentType",
              "ComponentType.code":"NO2",
              "ComponentType.unit":"nitrogen dioxide",
              "ComponentType.name":"µg/m³"
            },
            {
              "@id":"http://example.org/id/componentType/CO",
              "@type":"ComponentType",
              "ComponentType.code":"CO",
              "ComponentType.unit":"carbon monoxide",
              "ComponentType.name":"µg/m³"
            },
            {
              "@id":"http://example.org/id/componentType/O3",
              "@type":"ComponentType",
              "ComponentType.code":"O3",
              "ComponentType.unit":"ozone",
              "ComponentType.name":"µg/m³"
            },
            {
              "@id":"http://example.org/id/componentType/PM10",
              "@type":"ComponentType",
              "ComponentType.code":"PM10",
              "ComponentType.unit":"particles PM10",
              "ComponentType.name":"µg/m³"
            },
            {
              "@id":"http://example.org/id/componentType/PM2_5",
              "@type":"ComponentType",
              "ComponentType.code":"PM2_5",
              "ComponentType.unit":"particles PM2.5",
              "ComponentType.name":"µg/m³"
            }
          ]
        }
```


Code listing 6: Air quality data from Czech Republic in JSON-LD format (only part of the dataset was transformed)

8. Conclusions

Interoperability is a key factor when developing a Digital Twin because interoperability ensures that different data sources can be combined. To be interoperable, it is necessary to have a commonly agreed data model(s) aligned with existing data standards. In Flanders, the Flemish Interoperability program provides a structure and governance body, called the OSLO process, to co-create, develop and promote data standards. OSLO is an embodiment of the European Interoperability Framework (EIF) at the Flemish Region level. The collected experience, processes, and methods in OSLO are shared within the local community and offered to the broader community to build a more sustainable ecosystem of interconnected data standards. For instance, OSLO facilitated the initiation of a similar program, called ICEG, at the Belgian Interfederal level. This deliverable has positioned DUET in the context of the EIF because the many key stakeholders of DUET are public organisations. OSLO shows that their investment in addressing (semantic data) interoperability is also a facilitator for creating a Digital Twin.

Despite being based on the Semantic Web, the adoption of the OSLO Toolchain is hindered by the used natural language, namely Dutch. Enabling the slumbering internationalisation of the OSLO Toolchain is a key step towards the adoption by other parties across Europe. Within DUET, the necessary updates have been made to achieve a Toolchain that supports English as a prime language. The new Toolchain is demonstrated by extracting a semantically annotated data model for an air quality dataset of Pilsen. It also has been used in collaboration with another project, ODALA. The technical internationalisation improvements by DUET facilitated applying the OSLO process and method in the ODALA project to create a data standard on Air and Water Quality. This data standard can now be applied in DUET. This is one of the following steps to be taken: the air quality data to which DUET has access and is building its pilots can be used to assess the applicability and appropriateness of the new data standard. This is a top-down exploration.
Alternatively, we can explore how a bottom-up approach, as applied to the Pilsen dataset by semantically lifting the data, can contribute to an easier integration into a Digital Twin. A better understanding of these dynamics, including a cross-border component such as exploring a common understanding of Air and Water quality, will improve the processes and methods to make data more interoperable. In this exploration, we aim to reduce any hurdles for application in DUET. One of the possible approaches is to apply the methodology to other data sources in the context of the DUET pilot cases.

An improved understanding of the data is a foundation for the understanding of the data models. Data models in DUET are mathematical abstractions of the world that can provide insights into today's situation and predict the future. Understanding how these data models relate to (terms in) data standards and how data standards can facilitate the application or reuse of existing data models are explored. Today this is an open area of work. Data standards focus on defining a precise as a possible term but with a wide application area. However, they ignore the essential contextual variations important to data models: the same air quality mathematical model can be applied to a city with personal car traffic or only public transport traffic. The outcome is still levels of air pollution, but the interpretation is based on the parameterisation of the mathematical model.

The last topic that has been addressed in this deliverable are improvements to the technical uptake of the data standards. We showed a novel data publishing strategy called Linked Data Event Streams: a lightweight approach to connect evolving datasets and even context data. Although the primary use case for a Linked Data Event Stream is a time-based fragmentation, any dataset can be published as a Linked Data Event Stream, including useful indexes. Work in this area is essential as the more semantically disclosed data can be provided, the reuse of that data is eased because the effort for understanding is reduced. Building and exploring such bridges and possibilities are important because it is a step towards understanding what adequate building blocks could be to inject data into a Digital Twin. Whereas data standards are highly technically agnostic, this track of work explores the technical key principles to make in a technical semantical way the data described in a data standard available. Other future work in this area will look for answers to questions such as the best (serialisation) format for data is and how the format connects to the data semantics.

9. References

1. Buyle, R., De Vocht, L., Van Compernolle, M., De Paepe, D., Verborgh, R., Vanlishout, Z., ... & Mannens, E. (2016, November). OSLO: Open standards for linked organizations. In Proceedings of the international conference on electronic governance and open society: Challenges in Eurasia (pp. 126-134).
2. European Commission (2017). New European Interoperability Framework. https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf
3. Colpaert, P. (2021). Linked Data Event Streams. https://w3id.org/ldes/specification
4. Van Lancker, D., Colpaert, P., Delva, H., Van de Vyvere, B., Rojas Meléndez, J., Dedecker, R., Michiels, P., Buyle, R., De Craene, A., & Verborgh, R. (2021, May). Publishing base registries as Linked Data Event Streams. Proceedings of the 21th International Conference on Web Engineering. Accepted for publication.
5. Buyle, R., Van Nuffelen, B., Mechant, P., & Mannens, E. (2019). The Rise of Linked Organizations in Flanders. https://2020-eu.semantics.cc/sites/2020-eu.semantics.cc/files/SEMANTICS_CC_2019_Linked_Organisations_Raf_Buyle_Bert_Vannuffelen_v_final%20-%20Kopie.pdf