

# Semantics in data spaces

Semantic Treehouse as Vocabulary Hub

SEMANTiCS 2024 Tutorial

Jelte Bootsma

Linda Oosterheert

Cornelis Bouter



# Agenda



- 10:30 Welcome & get-to-know
  - Introduction to data spaces
  - Semantics in data spaces
  - Vocabulary Hub: Semantic Treehouse + demo
  - Challenge 1: Semantics in communities
- 12:00 Break
- 13:00 Challenge 2: Message vs Linked Data based interoperability
  - Do it yourself!
  - Future challenges
- 14:30 End

SEMANTICS-2024

# Your tutorial leaders from **TNO**



Linda Oosterheert

Consultant  
Data sharing and AI



Jelte Bootsma

Consultant  
Data Ecosystems



Cornelis Bouter

Scientist  
Semantic Interoperability

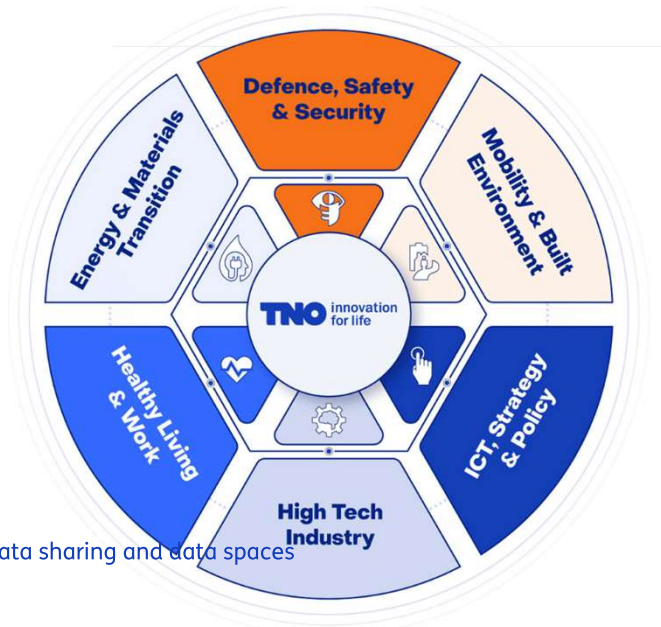
# TNO is the Dutch Research & Development Institute

As a statutory organisation, established by Dutch law, TNO has an **independent position** that allows to give objective, scientifically founded judgments.

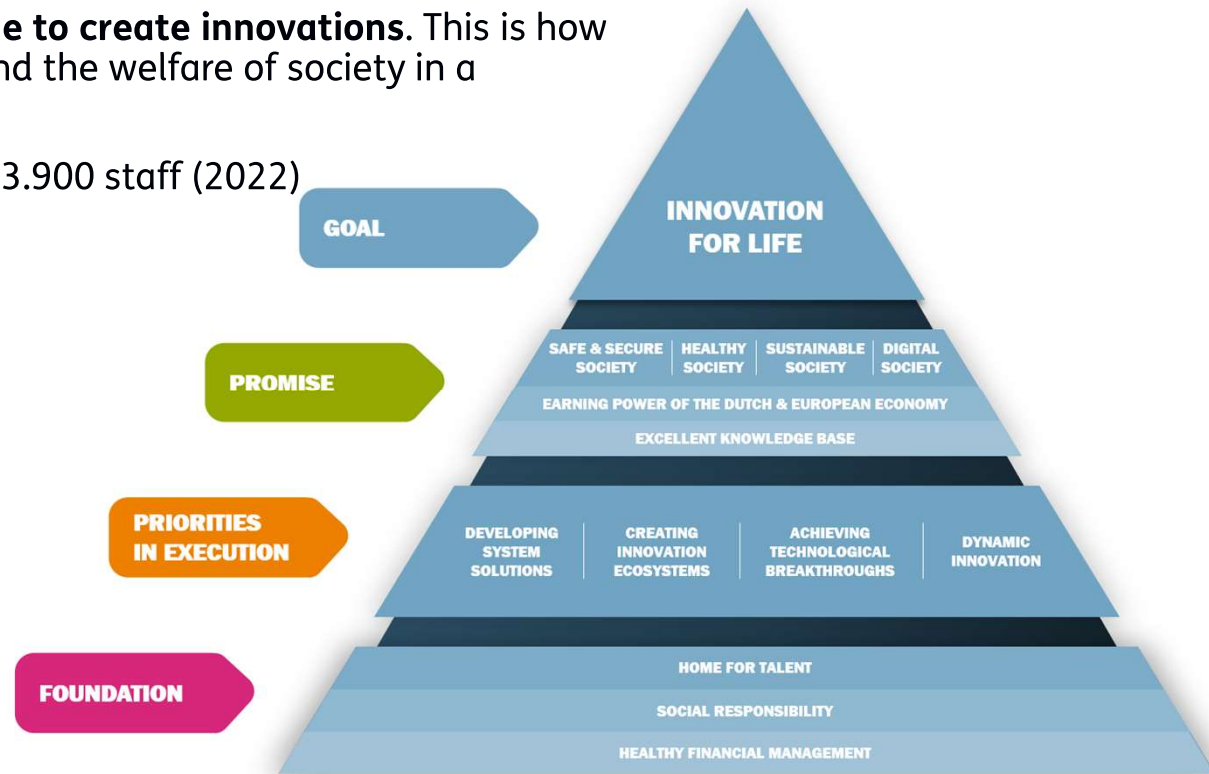
TNO's mission: **TNO connects people and knowledge to create innovations.** This is how we strengthen the competitiveness of companies and the welfare of society in a sustainable way.

**TNO is not-for-profit**, with € 590 mln revenues and 3.900 staff (2022)

TNO's activities are organized in 6 units



Guest lecture on data sharing and data spaces



# TNO ICT, Strategy & Policy

TNO ICT, Strategy & Policy (ISP) works on breakthroughs to help solve societal challenges. We do so in a multidisciplinary way by connecting stakeholders in ecosystems across and within a wide range of sectors.

## Research & Innovation portfolio clusters:

- Digital Infrastructure (incl. 5G/6G/Verticals/Cloud/Immersive/XR)
- Data sharing & AI
- Trusted ICT & Quantum
- Embedded Systems Innovation
- Transitions & Transformations (Innovation, Industrial & Urban)

## 400+ experts in 9 research groups:

- Advanced Computing Engineering
- Cyber Security Technologies
- Data Science
- Applied Crypto & Quantum Algorithms
- Networks
- Embedded Systems Innovation
- Data Ecosystems
- Strategic Business Analysis
- Strategy & Policy



## All major societal innovation challenges require some form of data sharing

1



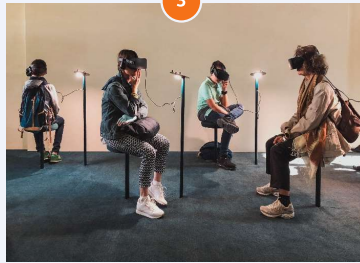
Smart Energy

2



Smart Cities

3



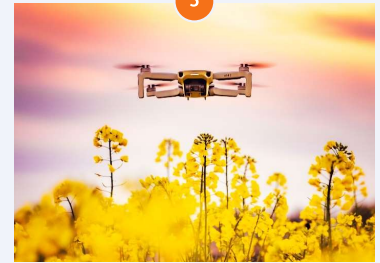
Smart Society

4



Smart Climate Solutions

5



Smart Agriculture

6



Smart Production

7



Smart Resources

8



Smart Health

9



Smart Security

10

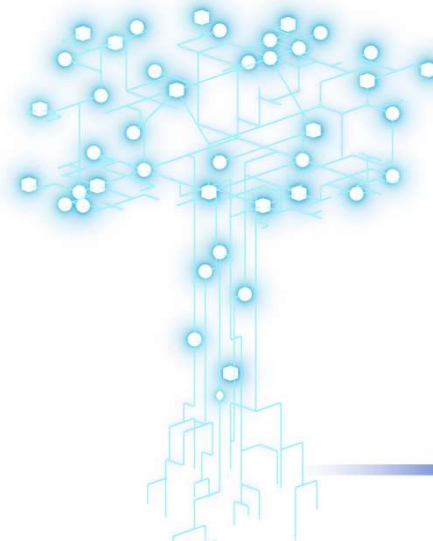


Smart Mobility

# The challenges addressed in the tutorial

**How to establish semantic interoperability in practice in data sharing communities?**

**How to bridge the gap between message based and Linked Data based interoperability?**



# Challenge 1 of this tutorial

**How to establish semantic interoperability in practice in data sharing communities?**

- Data-sharing initiatives need to come together to establish clear agreements on how data will be shared.
- Different stakeholders bring varied interests and viewpoints.
- It's important to ensure that all perspectives are considered and incorporated into the agreements.





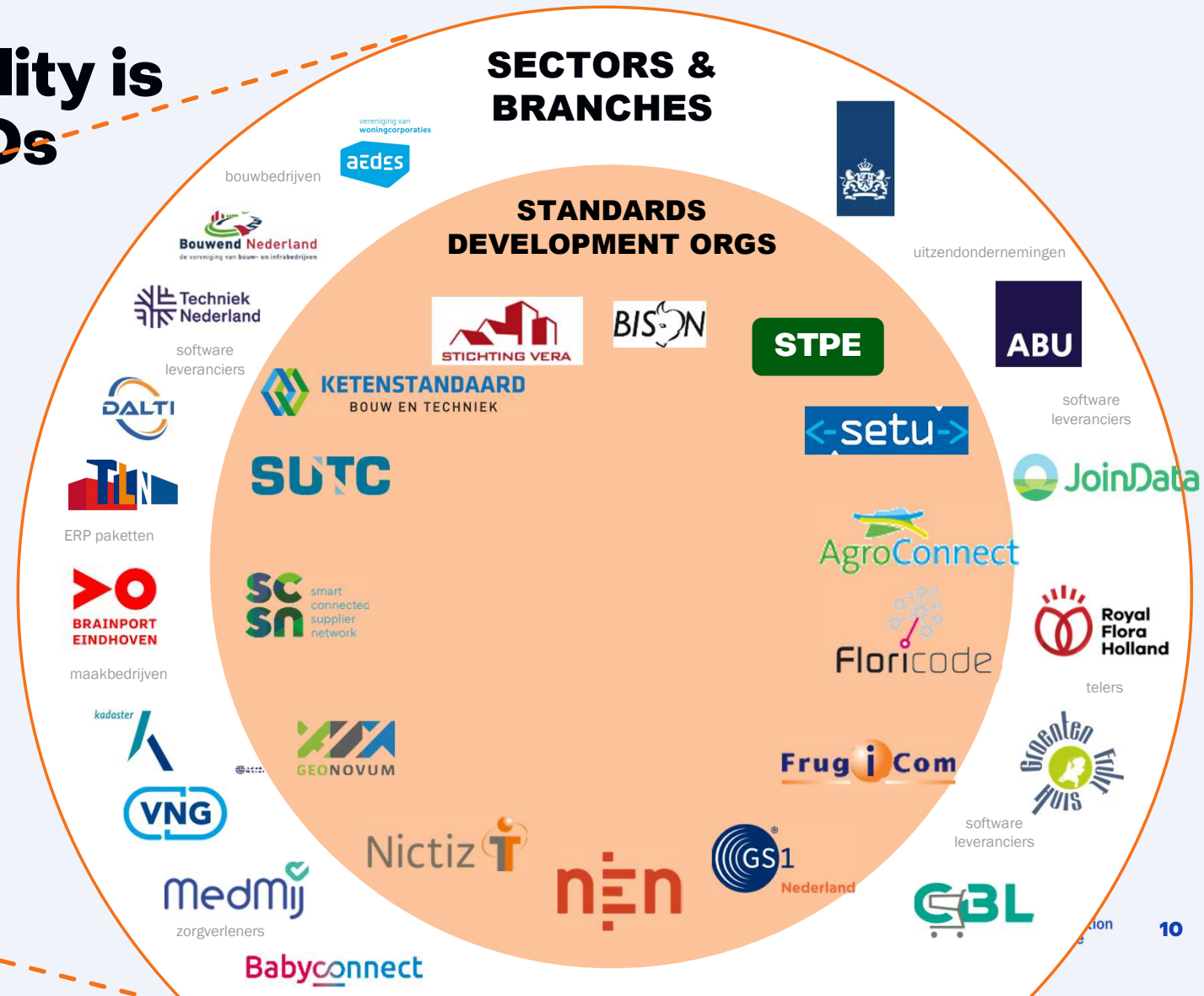
# Digitalization agenda of NL is ambitious



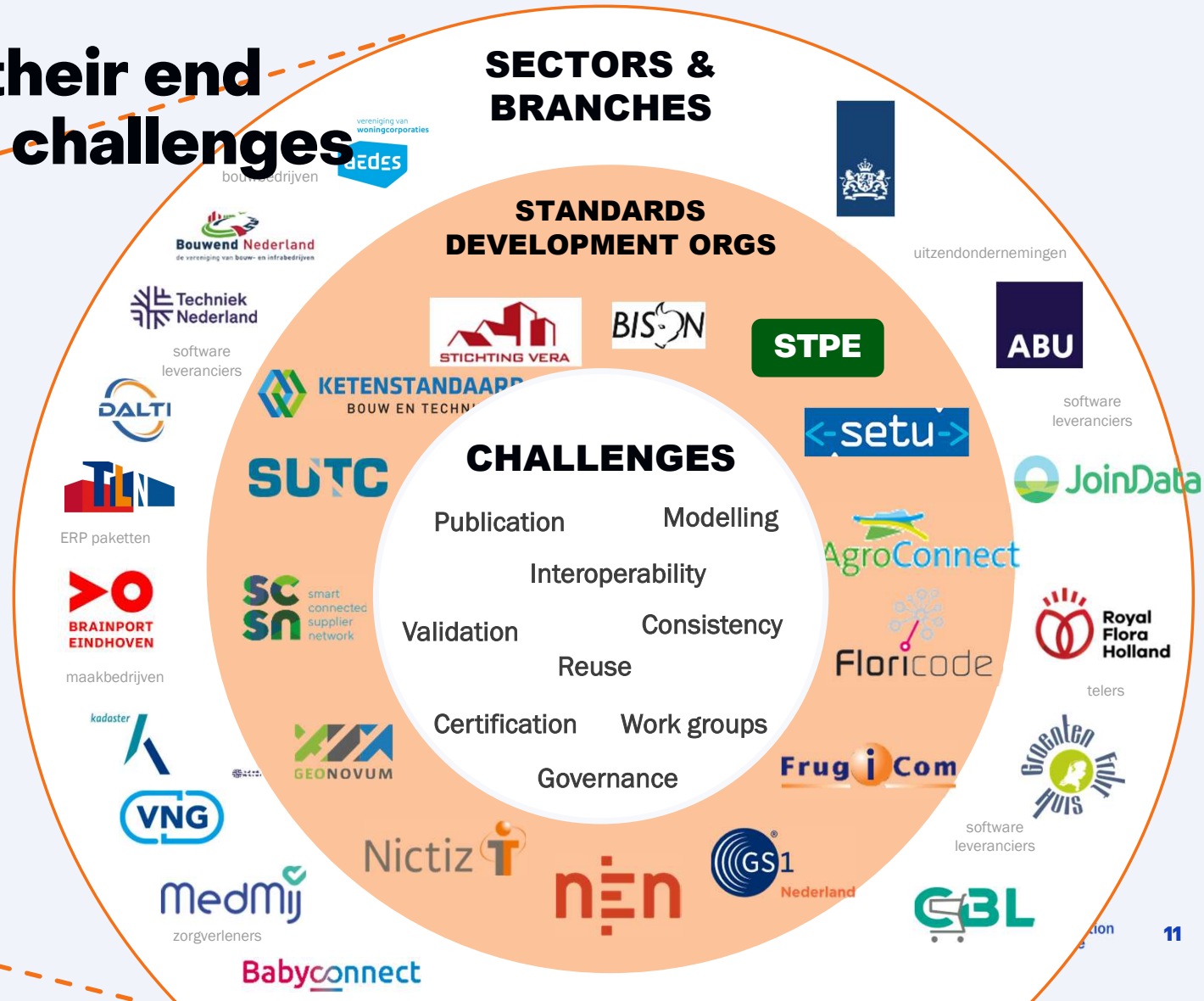
## SECTORS & BRANCHES



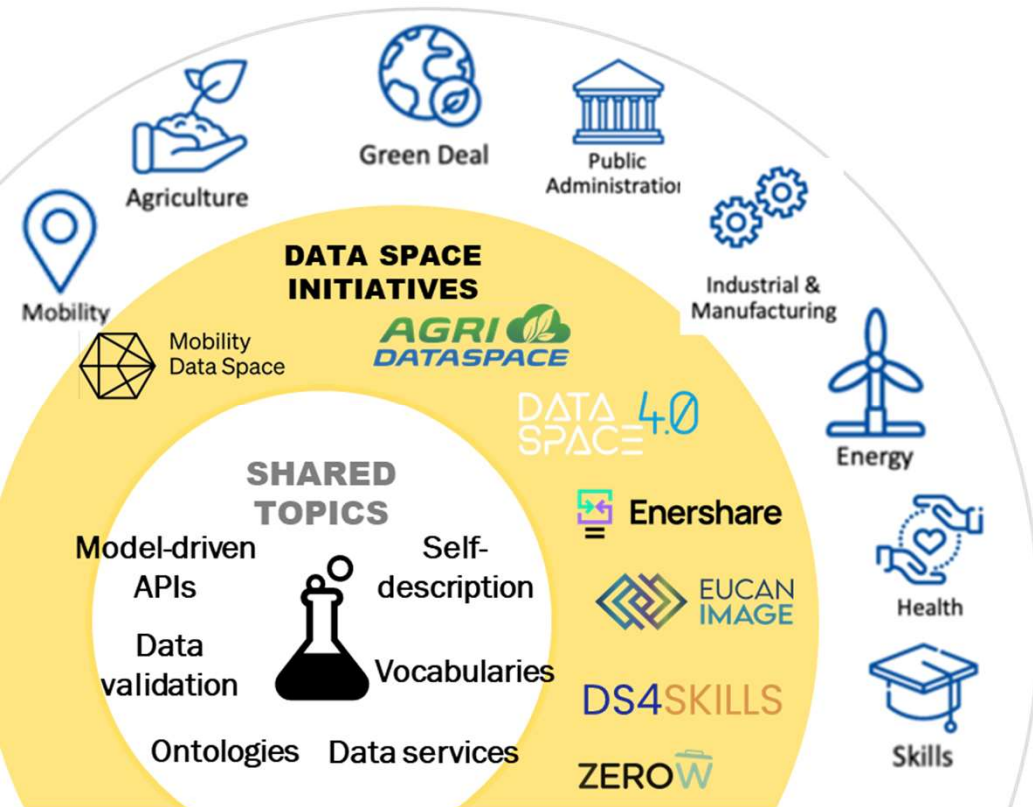
# Data interoperability is facilitated by SDOs



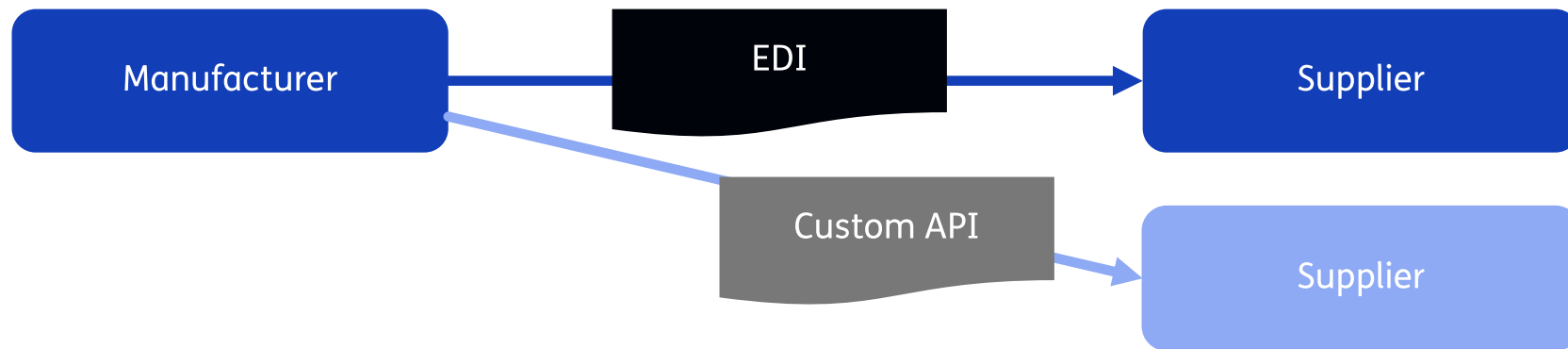
# These SDOs and their end users face similar challenges



# EU data space initiatives face the same challenges for semantic interoperability

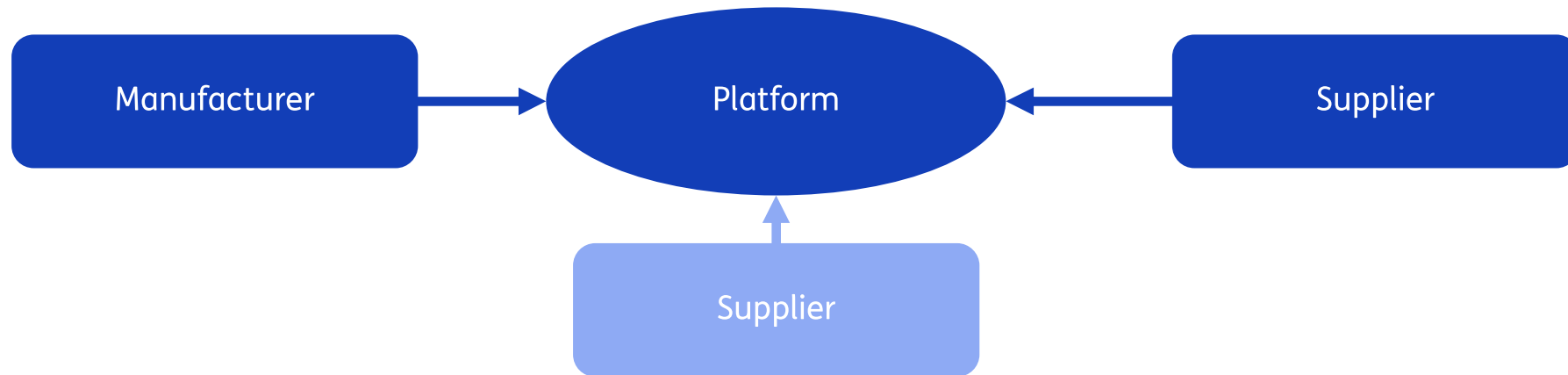


# Sharing data: Peer-to-peer data exchange



- Companies create digital connections with their business relations.
- The connections are tailored to the needs of the involved companies.
- ICT integrators are often involved.
  
- Limitations
  - Hard to scale
  - Difficult to organize for SMEs

# Sharing data: platforms

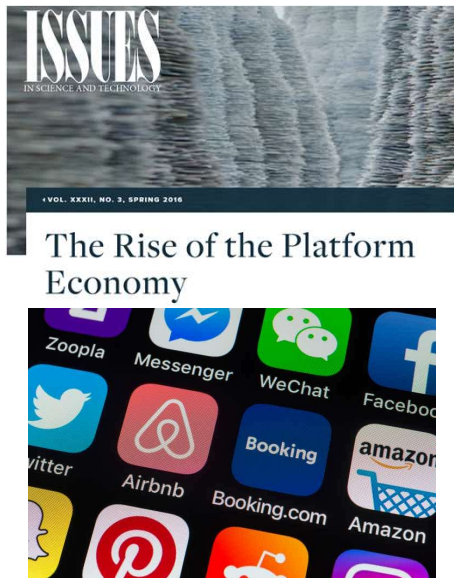


- Data is collected in a centralized platform.
- All companies adopt the same way of working (or a mapping is made).
- Advantage: economy of scale.

# Data spaces enables control of your own data

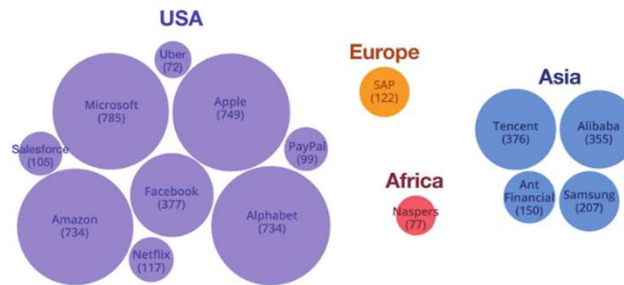
1

Platforms are great! Easy to use for end-users, large customer base, and standardized way of working for everyone.



2

Europe lost the race for platforms to the USA and China.



3

Disadvantages of platforms

- Winner-takes-all-model: Low profits for majority of users**
- Lack of data sovereignty: control of own data**
- Lack of strategic autonomy: strong dependency on USA and China**

# Data sharing challenges



**Data sovereignty**  
Being in control of your own data



**Scalability**  
Quickly connect many companies





## Data spaces are part of a European strategy that supports data sharing

EC WANTS TO REGAIN INITIATIVE BY REGULATIONS&INVESTMENTS

A real **data economy**, on the other hand, would be a powerful engine for innovation and new jobs. And this is why we need to secure this data for Europe and make it widely accessible. **We need common data spaces** – for example, in the energy or healthcare sectors. This will support innovation ecosystems in which universities, companies and researchers **can access and collaborate on data**.

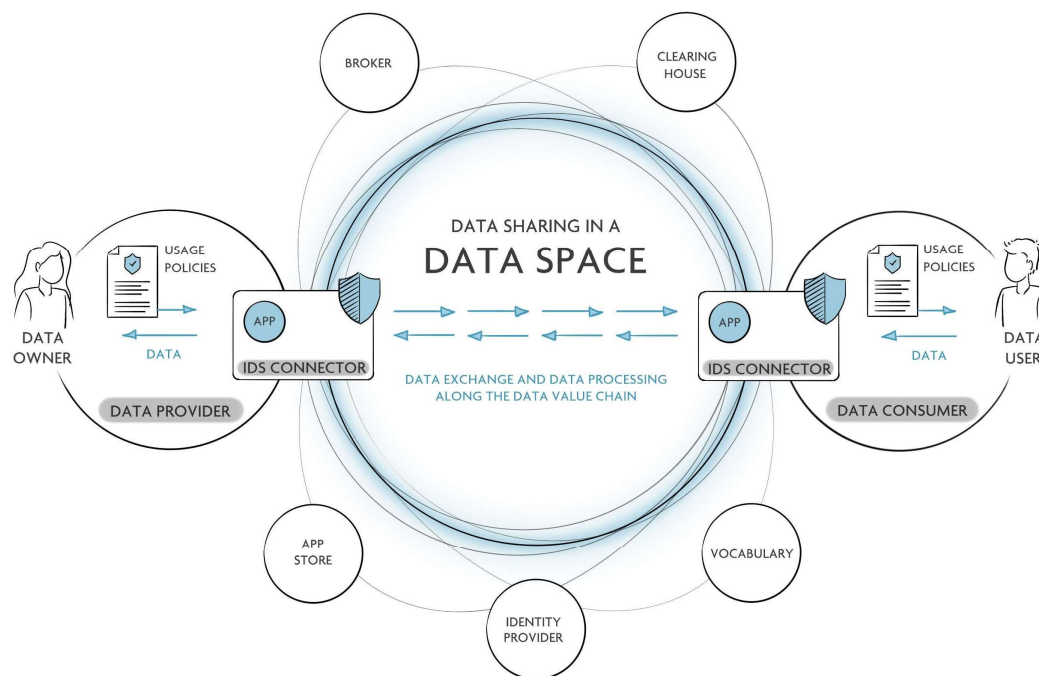
(State of the Union Address on 16 September 2020)



# Data spaces are European answer to platforms



Decentral digital collaboration between organizations in a sector conform European values: level-playing field & data sovereignty.

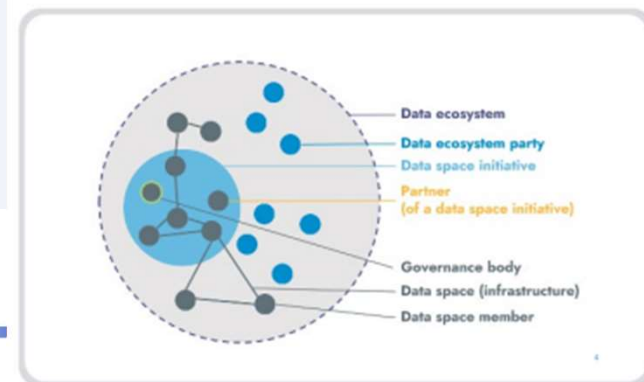
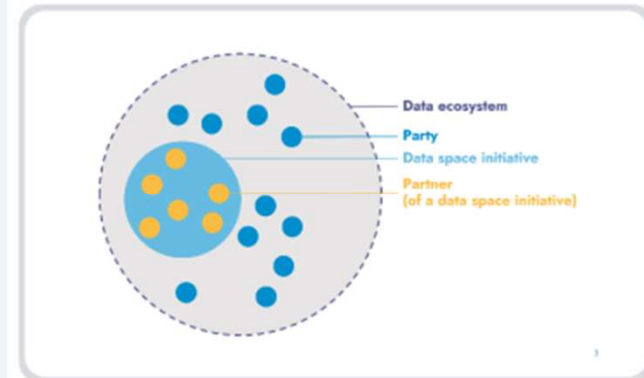
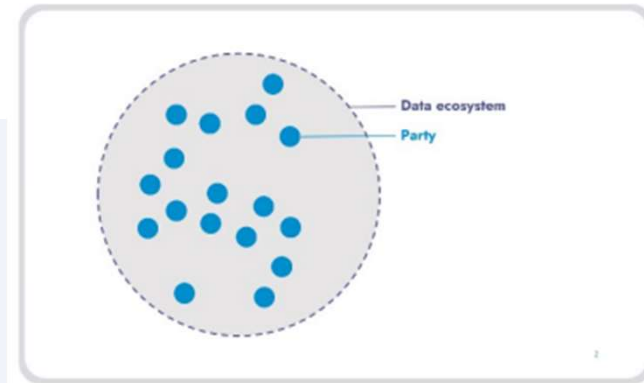


# Defining data spaces

“A **business ecosystem** is a purposeful business arrangement between two or more entities to create and share in **collective value** for a common set of customers (...)” [1]

“**Data ecosystem** can perceive as a ‘universe’ where the data ecosystem parties (legal and natural persons) engage in data sharing.” [2]

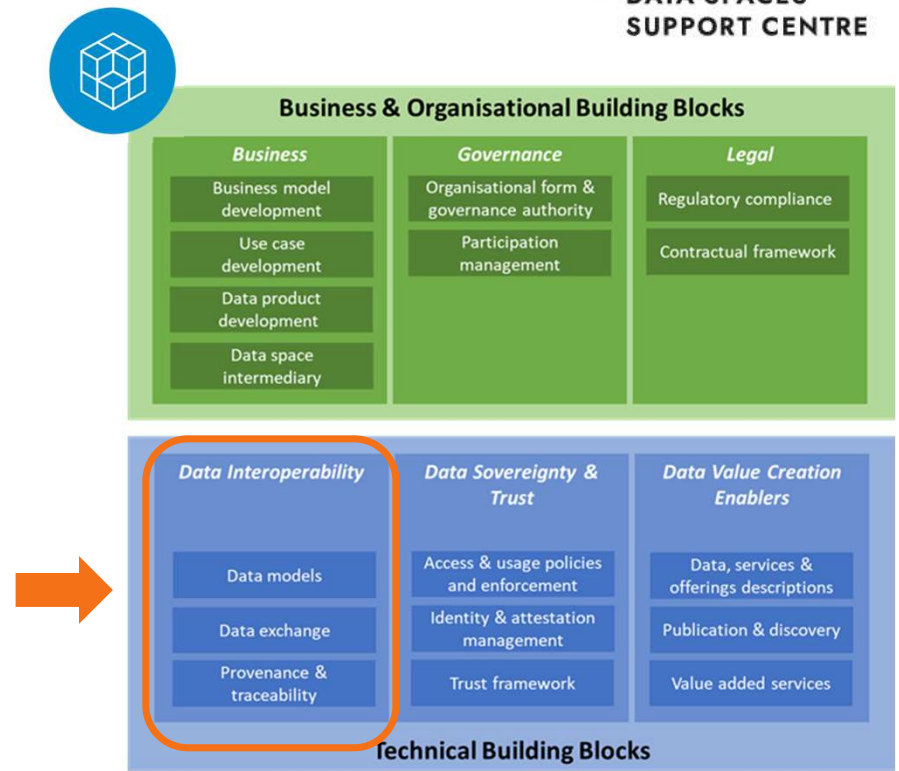
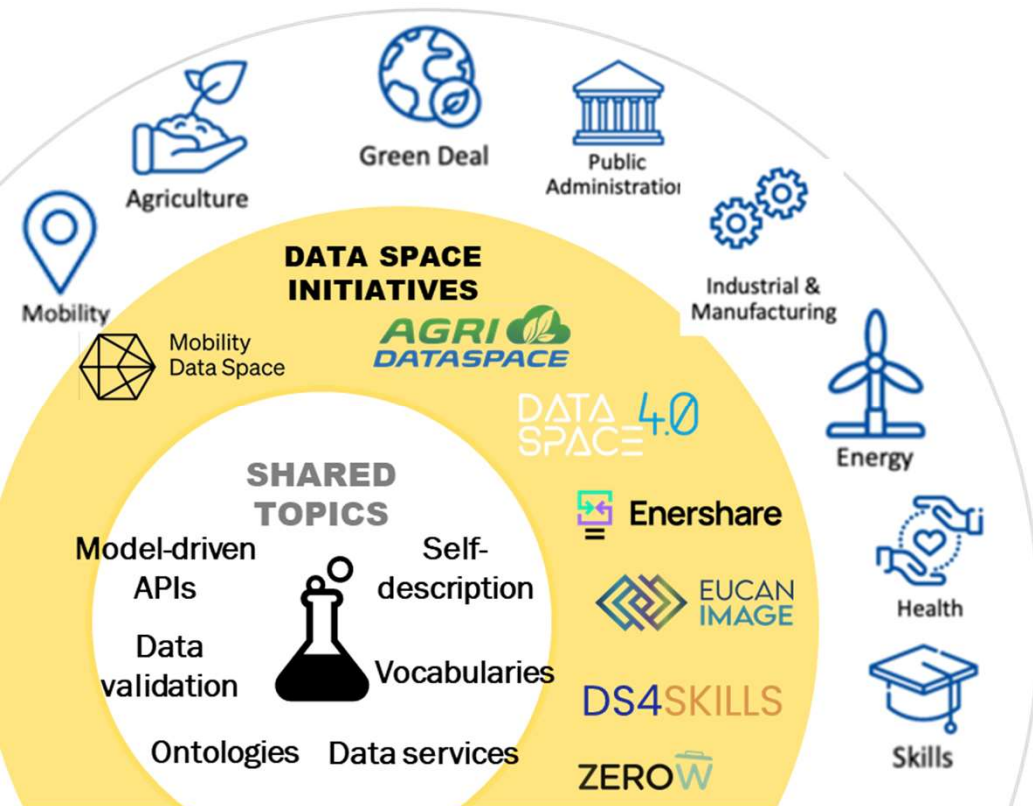
“**Data space** is an **infrastructure** that enables data transactions between different data ecosystem parties based on the **governance framework** of that data space.” [2]



[1]: [What business ecosystem means and why it matters | EY - Netherlands](#)

[2]: [DSSC Glossary | Version 1.0 | March 2023 - Glossary - Data Spaces Support Centre](#)

# Data spaces require solutions for data interoperability



# Building blocks

- Semantic interoperability is found primarily in the DSSC building block Data Models & formats
- But, also important for metadata in other building blocks!



Domain Language



Connectors, Validation and Transformations

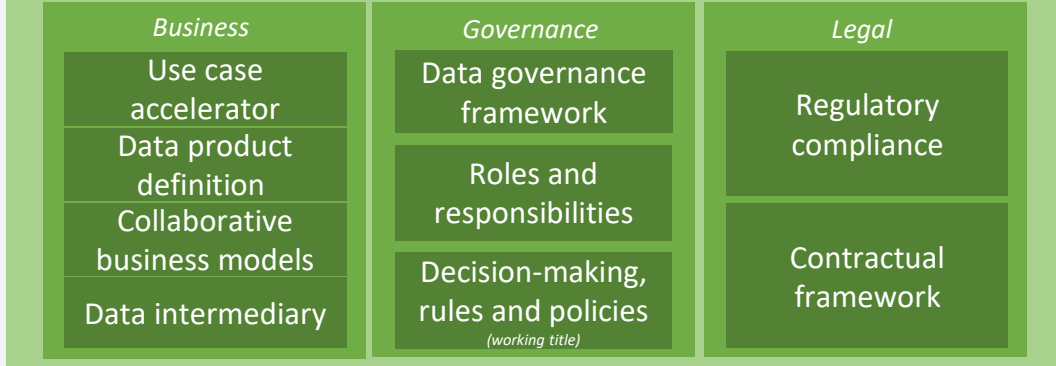


Vocabulary hubs & (Federated) catalogues

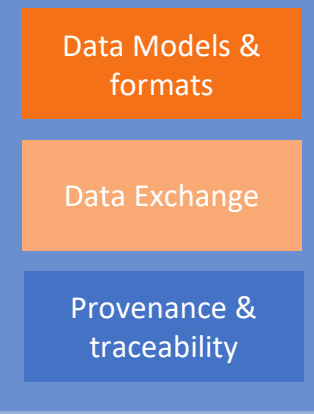


**DATA SPACES  
SUPPORT CENTRE**

## Organisational & Business Building Blocks



## *Data Interoperability*



## *Data Sovereignty & Trust*



## *Data Value Creation*



## Technical Building Blocks

# We introduce...



**Semantic  
Treehouse**

Online community platform  
for **business & IT**  
to work collaboratively  
on **data models**

GitHub



Online community  
platform for **developers**  
to collaboratively work  
on **software**

SwaggerHub



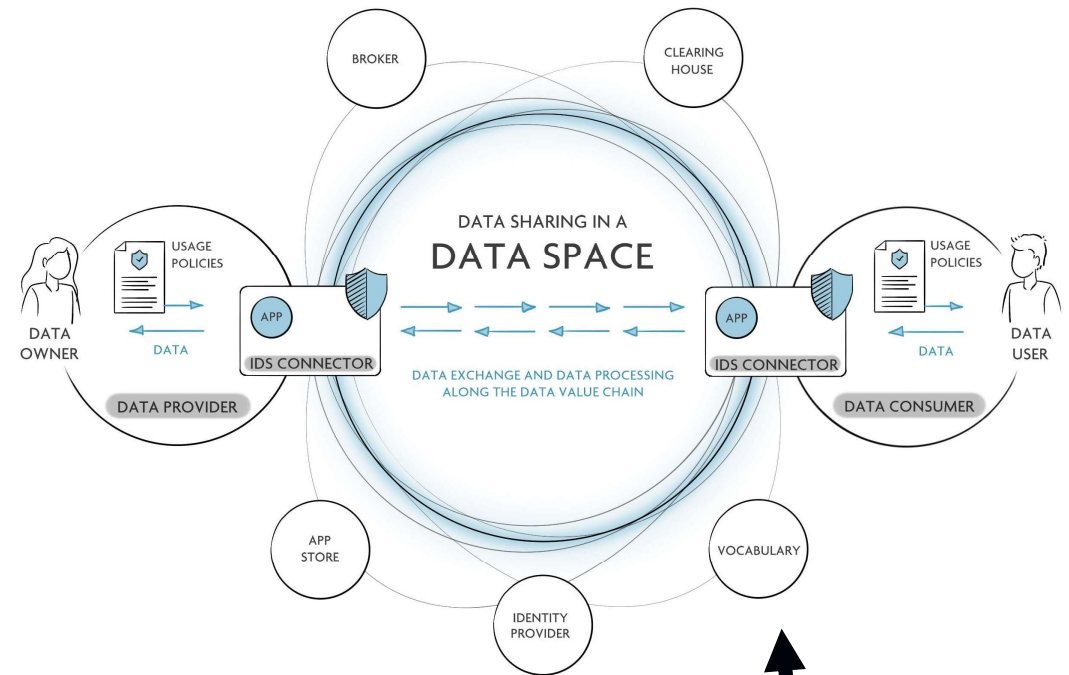
Online platform  
for **developers**  
to collaboratively  
on **APIs**

# We introduce...



## Semantic Treehouse

Online community platform  
for **business & IT**  
to work collaboratively  
on **data models**



In data space terms:  
a **Vocabulary Hub**

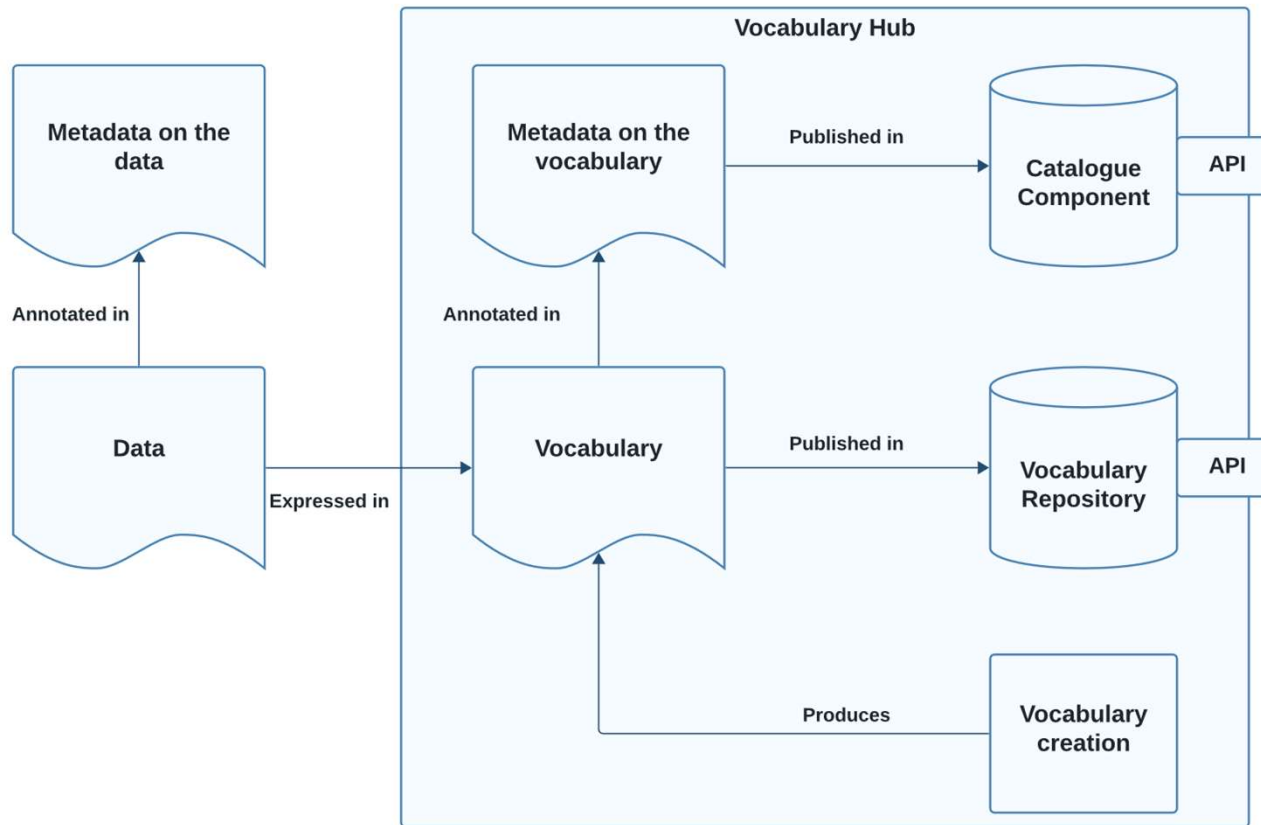


## Key components of a vocabulary hub

- **Vocabulary creation and editing**; offering flexibility for starting from scratch or integration of existing vocabularies.
- **Vocabulary repository**; providing storage of all the distributions of vocabularies: any specification that can be used to consistently describe data.
- **Catalogue**; enabling metadata descriptions and easy access to vocabularies within the vocabulary repository, facilitating their discoverability and reuse.

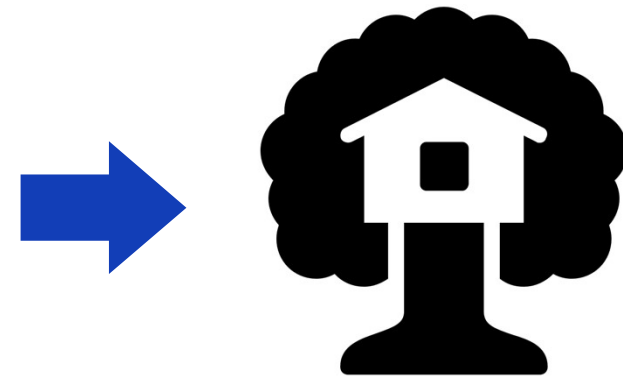


# Key components of a vocabulary hub





# Integrated solution for community driven data model management



Synoniemen:

- Berichtenbibliotheek
- Gegevens definities catalogus
- Modellen platform
- Canoniek Data Model (CDM)

# Integrated solution for community driven data model management

**Documentation**

**Codelists**

**Contact list**

**Maintenance**

**Examples**

**E-mail**

**Workgroups**

**DEMO**



Synoniemen:

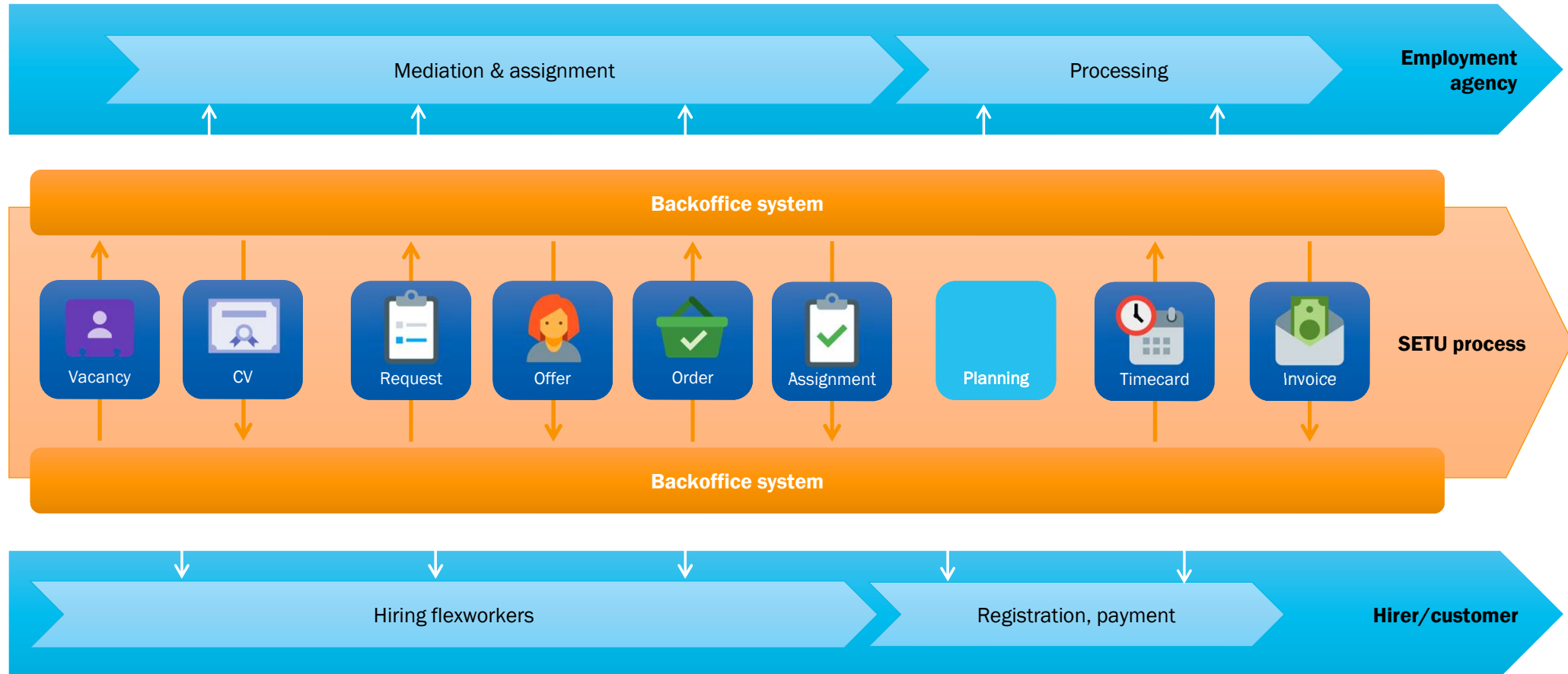
- Berichtenbibliotheek
- Gegevens definities catalogus
- Modellen platform
- Canoniek Data Model (CDM)

# Use Case: Semantic Standards in the Staffing Industry

**The SETU HR standards enable employment agencies, hirers and HR software suppliers to solve digital integration challenges in a modern, flexible and cost-efficient way.**



# Use Case: SETU HR standards



# The community: different stakeholders

- For this case, multiple stakeholders are involved in the development of the standard:

## 1. Staffing Customer

Alice values exchanging data that enables a good match, including skill sets as well as availability at the right time.



## 2. Staffing Supplier

Mark has the best interests of his human resources in mind and considers matching availability and efficiency the most important; maybe they could carpool together?



## 3. Software Provider

Sarah values having a consistent standard and, where possible, basing it on existing standards; this way, she only needs to implement it once and can save costs.



## 4. Planner

David values having a clear overview in human resource planning. Each planning needs to be unique, and any changes must be easy to implement.

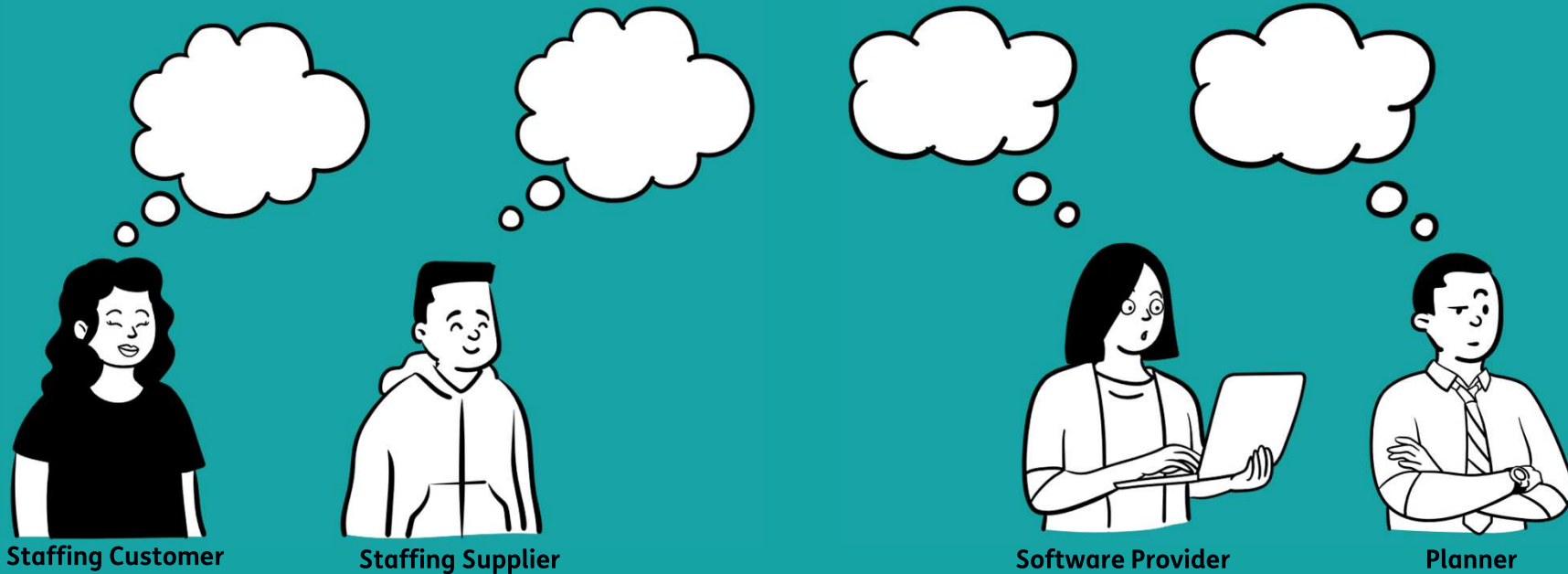


# Agreeing on semantics in a community

A standard for planning & scheduling?



[LINK TO MENTIMETER](#)





# But, are we doing the right thing here?

HOW STANDARDS PROLIFERATE:  
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



# But, are we doing the right thing here?



HR OPEN  
STANDARDS

HOW STANDARDS PROLIFERATE:  
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



BREAK

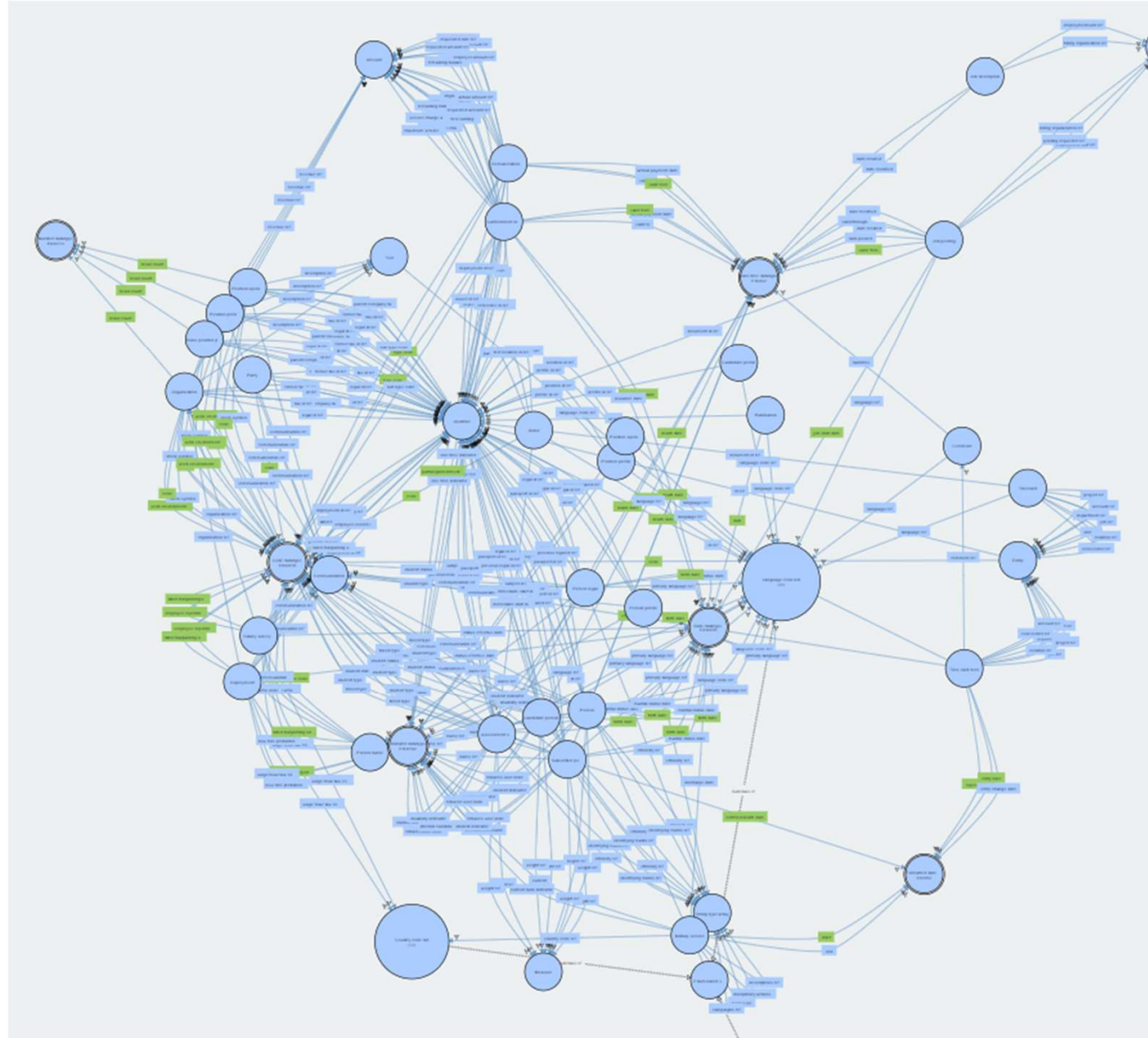


SEMANTICS-2024

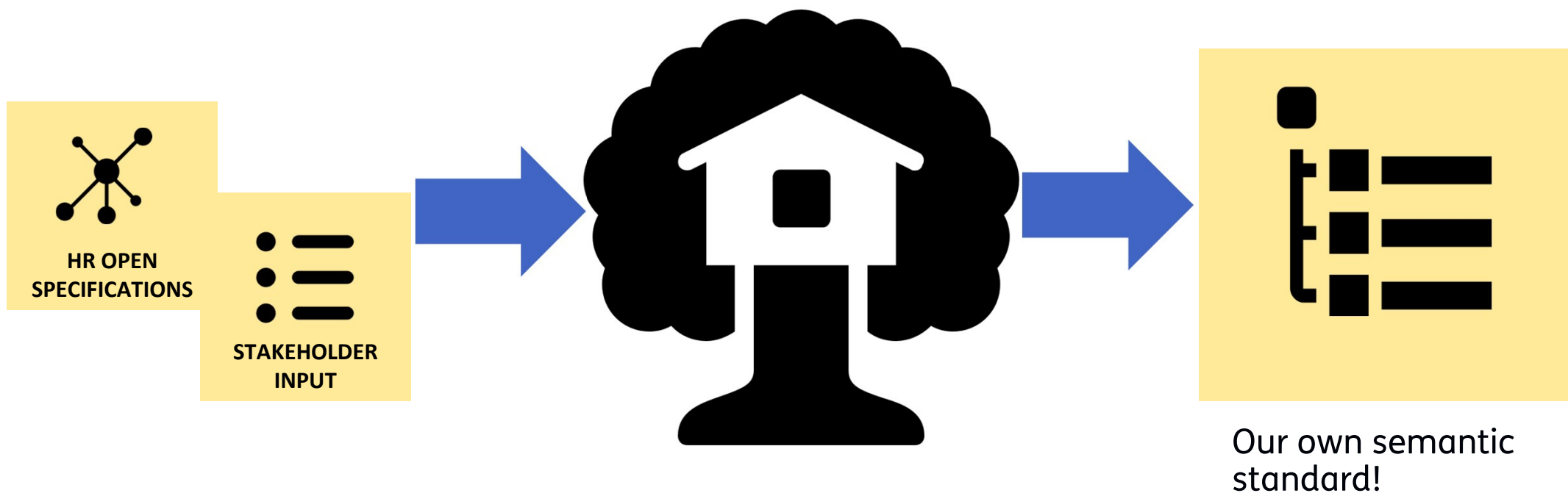
# Building upon an existing standard



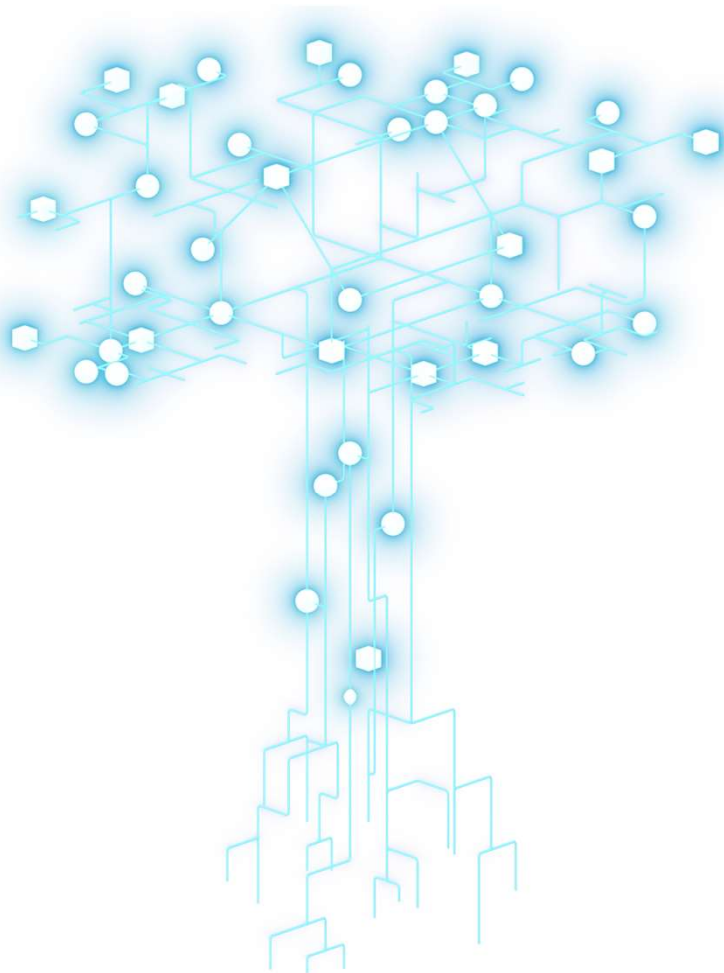
**HR OPEN**  
STANDARDS



# DEMO TIME!



## Challenge 2 of this tutorial



**How to bridge the gap  
between message based  
and Linked Data based  
interoperability?**

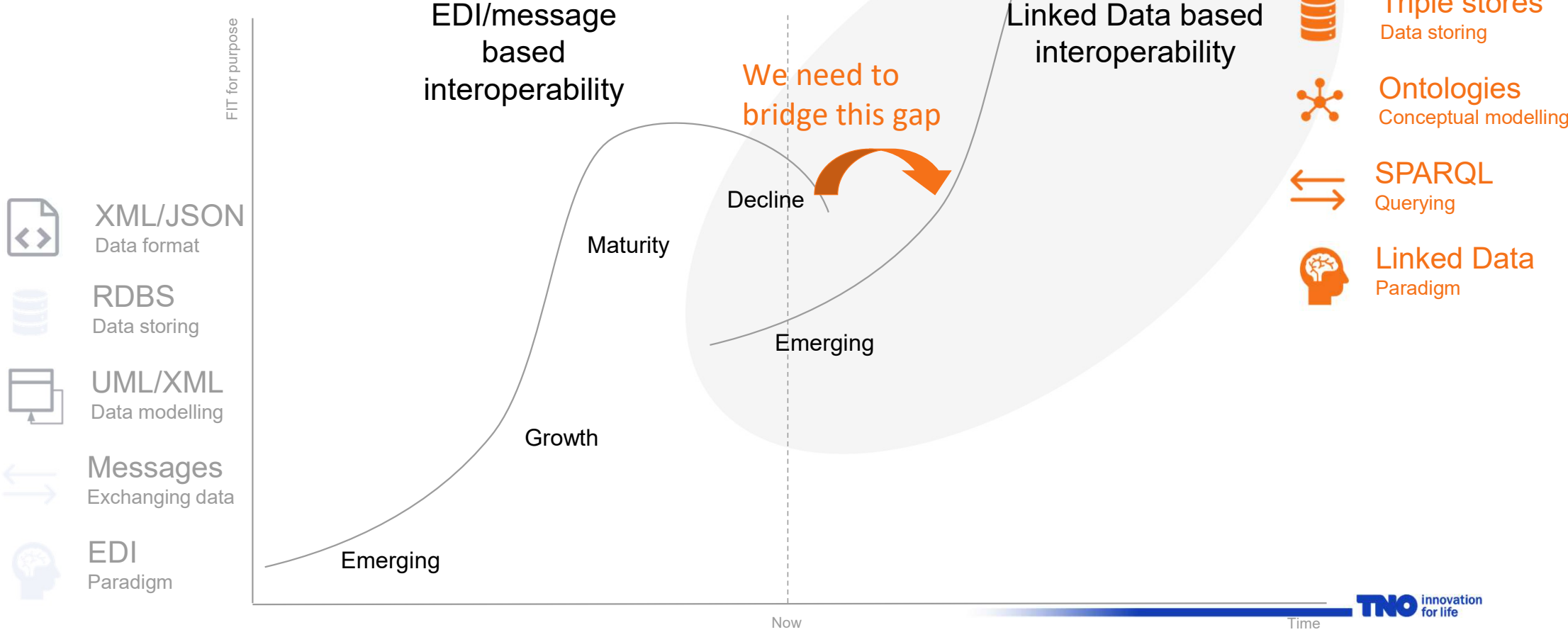
- No big-bang, but facilitate the incremental adoption of semantic (web) technologies like rdfs/owl, shacl and rml.
- We recognize that message based data sharing paradigm is still relevant for majority of use cases. We bridge these worlds with our wizard technology.

# The semantics community introduces semantic web technologies....

- We see the benefits in terms of:
  - explicit formalized semantics for FAIR data,
  - keeping data at the source and linking data sources
  - and the flexibility offered in models, queries and APIs.
- But those technologies are hard to work with due to many reasons, incl. mismatch with current IT stack, lack of knowledge, immature TRL, concerns about scalability and performance.
  - Is this still true?! Or a persistent perception?



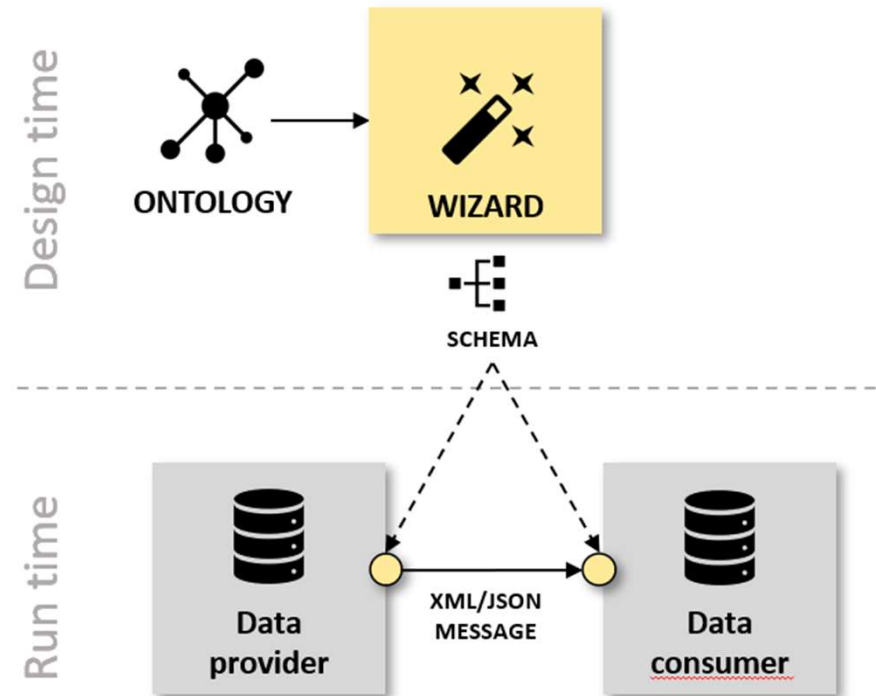
# Enabling the transition towards linked data based interoperability



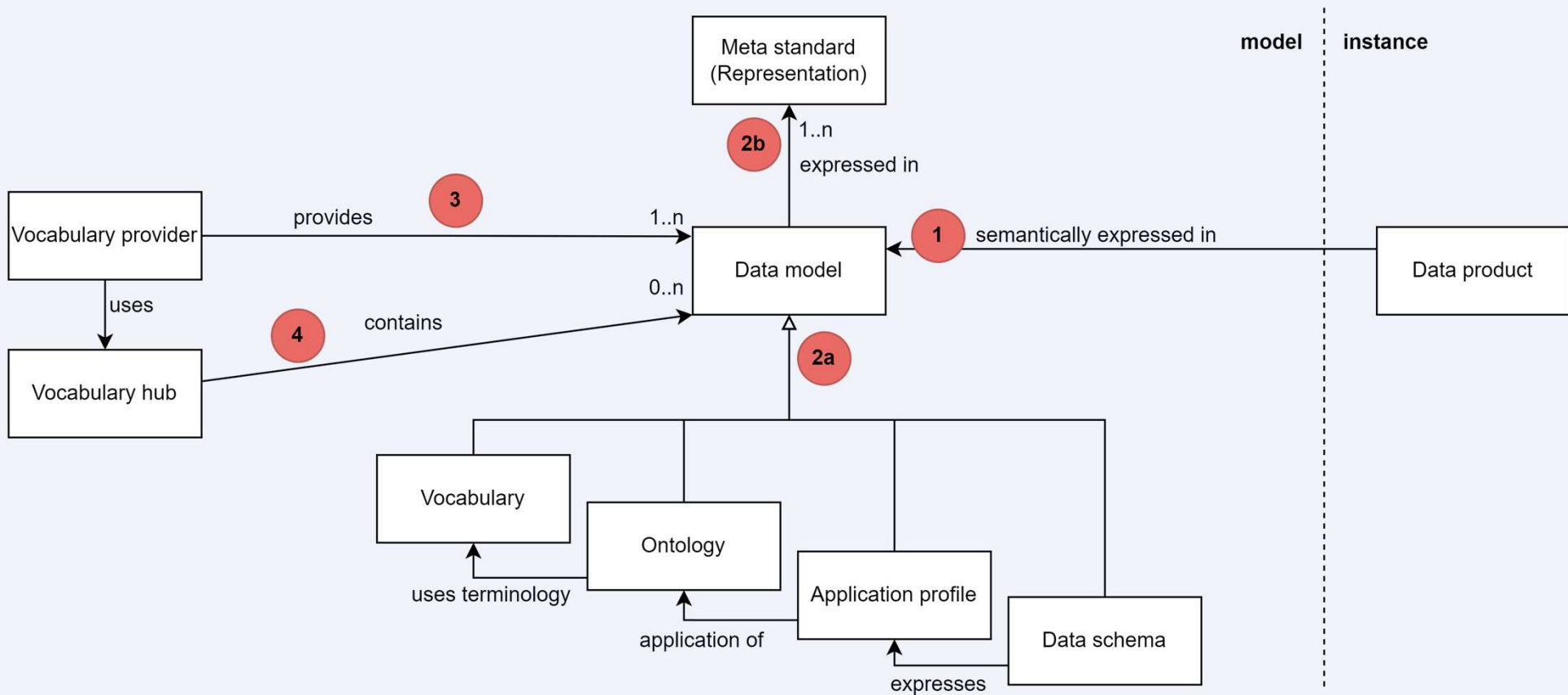


# Wizard approach

1. Prepare the message model specification by defining the **ontologies to import** (owl, rdfs, shacl, json schema), set the namespaces, define a root element and select the class that serves as the entry point to the ontology.
2. Model the **information requirements for the use case** by selecting from the ontologies. This shapes an abstract message tree (AMT). The wizard component allows users to 'cherry pick' the relevant classes and properties from the ontology and keeps the link
3. Let the wizard **generates a technology-specific syntax** binding for the AMT. The resulting schema can be used for the data plane.



# Data model?

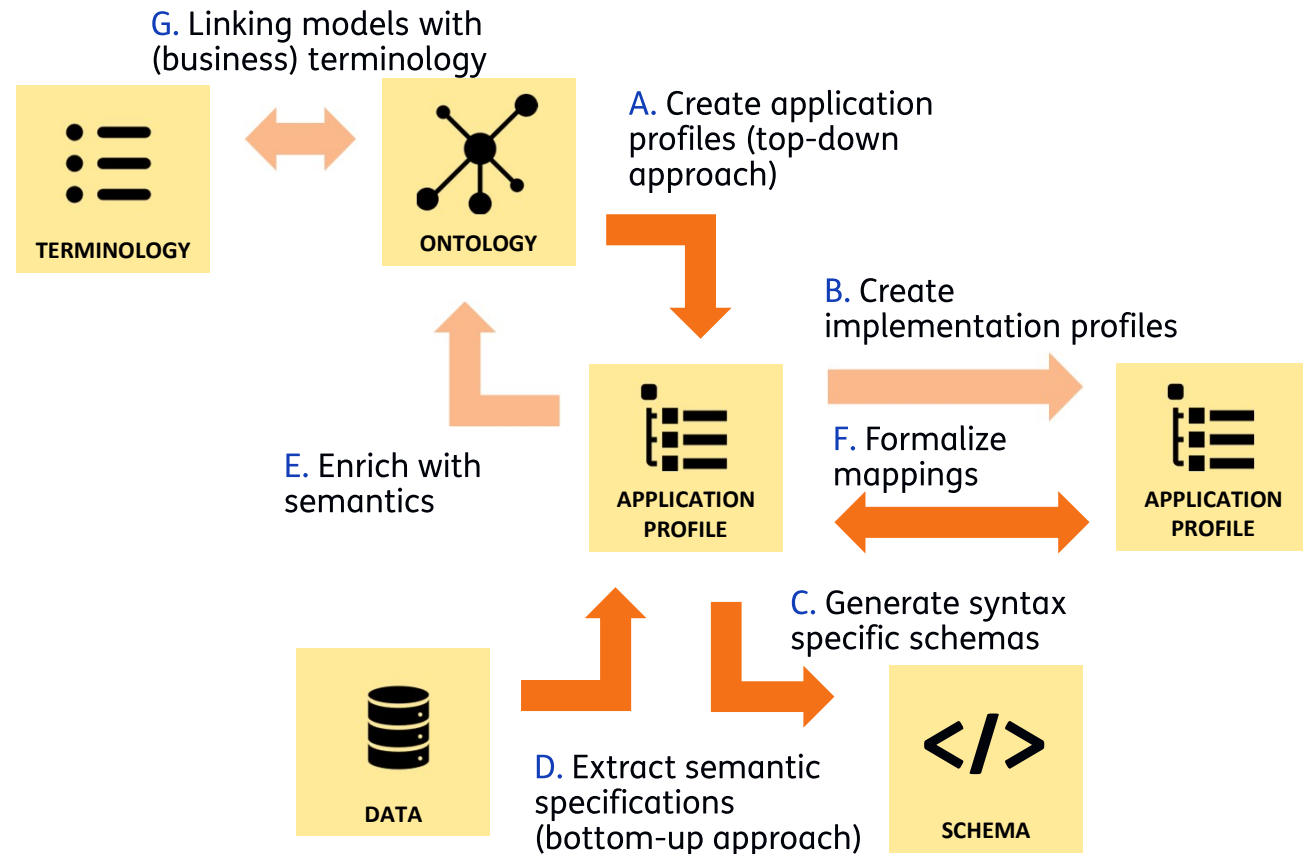


# STH connects models at different levels

We recognize different users with different needs and different views regarding to semantic interoperability.

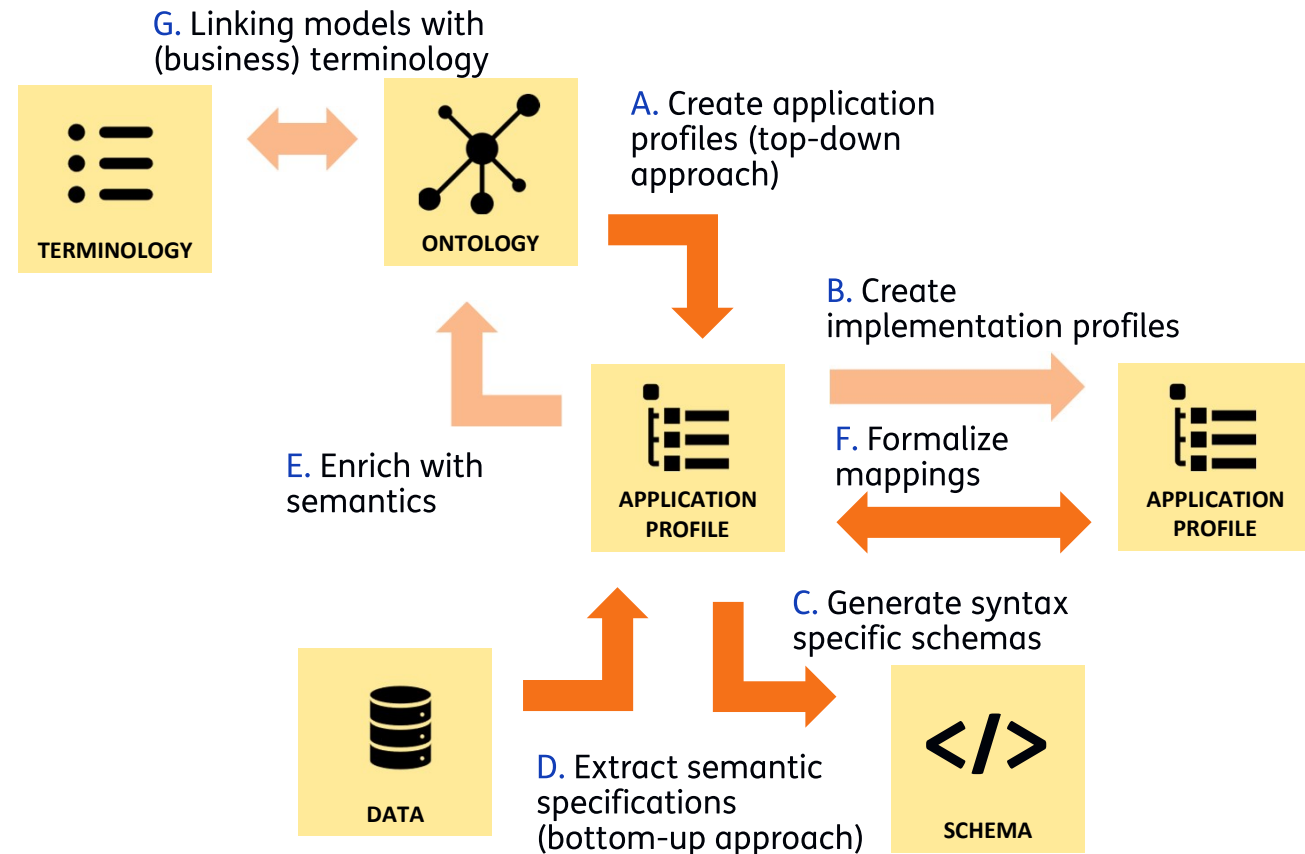
Therefore, Semantic Treehouse provides several entry points to make explicit and formalize the data models used to enable semantic interoperability:

- A. Create application profiles from existing ontologies (RDFS/OWL/SHACL or json schema sets). This results in message definitions for specific data sharing use cases.
- B. Create implementation profiles based on existing message specifications. E.g. to model company or use case specific restrictions and sub sets.
- C. Generate syntax / technology specific schema's from existing message specifications. E.g. to use in API specifications.



## STH connects models at different levels (2)

- D. Extract message model specifications from existing data sets or data samples (e.g. CSV documents). This allows to kickstart formalizing the data models used.
- E. Enriching message model specification by linking to existing ontologies.
- F. Formalize mappings between existing message model specifications
- G. Linking models with (business) terminology





DIY

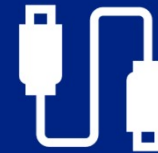


# Future challenges – work in progress

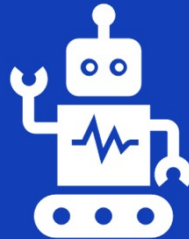
**How to support the reuse and diversity of existing data models and formats across data spaces?**



**Integrating semantic technologies into data space architectures**



**How to reduce the amount of manual labour in developing and mapping semantic standards?**

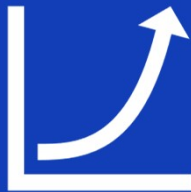


**How to improve data quality and the quality of data models?**



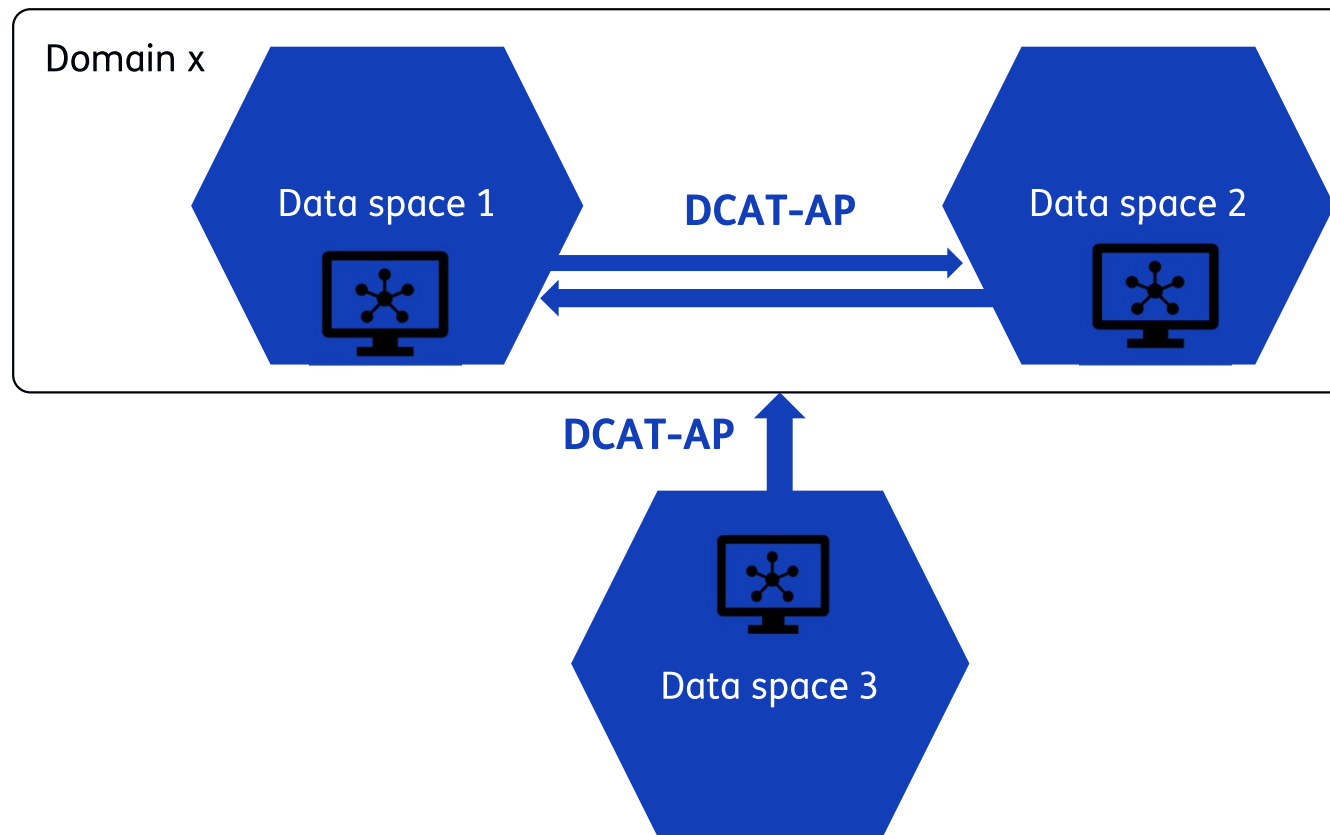
# Future challenge – work in progress 1

**How to support the reuse and diversity of existing data models and formats across data spaces?**



- Publishing and sharing data models using open standards and adopting FAIR principles.
- We recognize the importance of separation of concerns on different levels of information modelling (MIM) and the role of DCAT-AP to enable Vocabulary Hub federation.

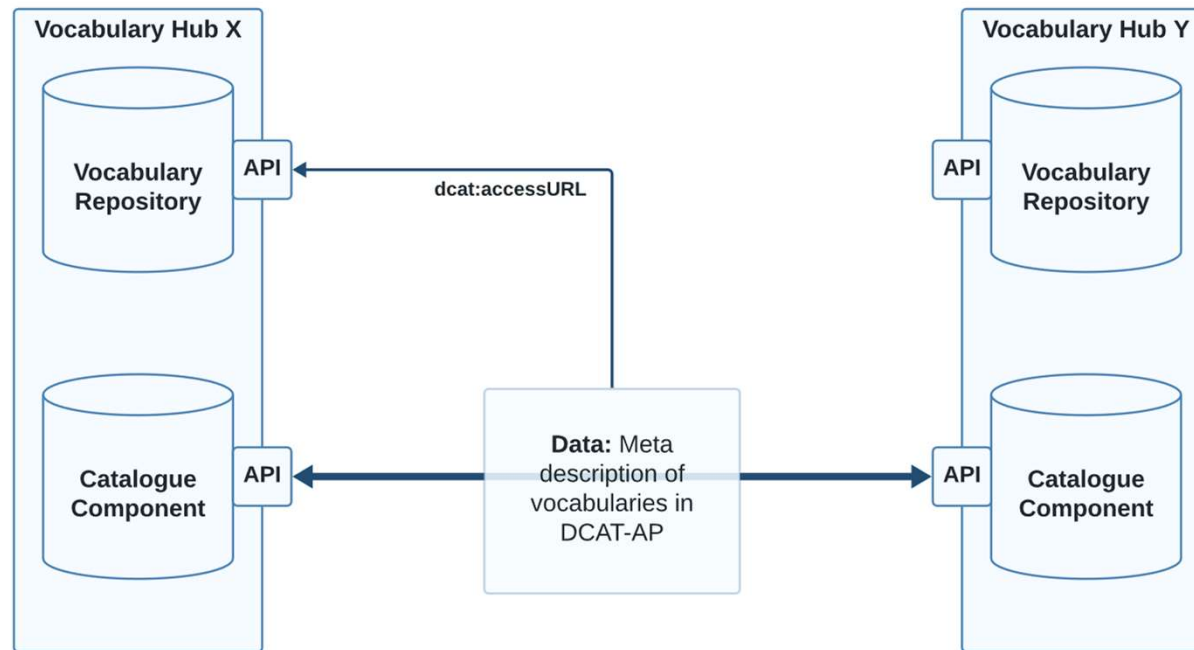
# Establishing semantic interoperability across data spaces: federation





# Exchanging vocabularies using DCAT-AP

DCAT-AP can be used to facilitate the exchange of vocabularies between different vocabulary hubs by standardizing the metadata of vocabularies.

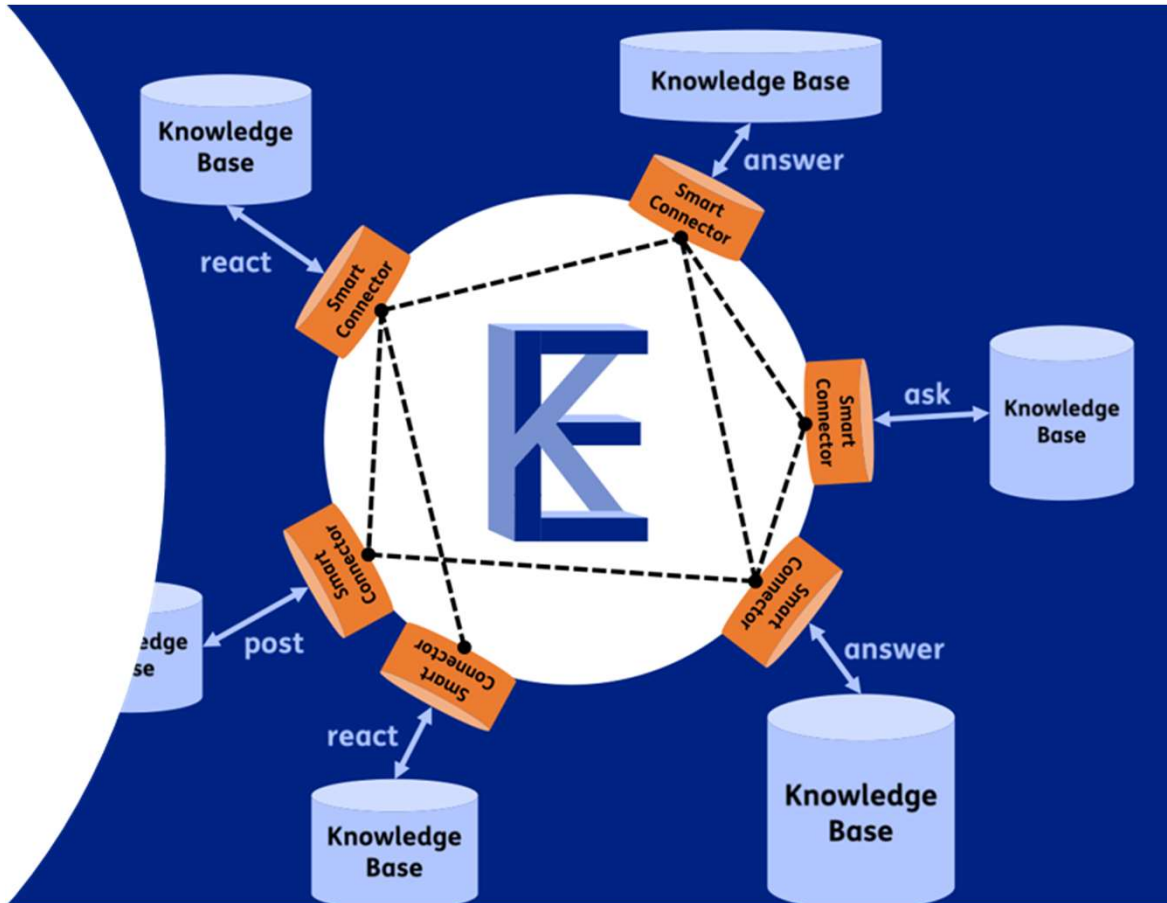


# Interested in more? Read our [position paper](#) and [try it out!](#)

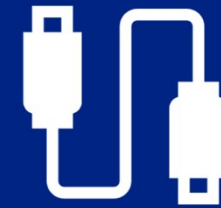
The screenshot shows the Semantic Treehouse interface. On the left is a teal sidebar with the Semantic Treehouse logo and navigation links: Specifications, Codelists, Validator, Issues, Groups, ENERSHARE project, and BD4NRG project. The main content area is titled 'Specifications' and shows a list of 'ENERSHARE vocabularies'. A 'public' filter is active. Two ontology cards are visible: 'Platoon Ontology' (v1.2 - 2024-03-14) and 'Enershare Maintenance Ontology' (v - 2023-12-21). The 'Export to DCAT' button on the second card is circled in blue.

<https://www.semantic-treehouse.nl/> or join us on [Discord](#)

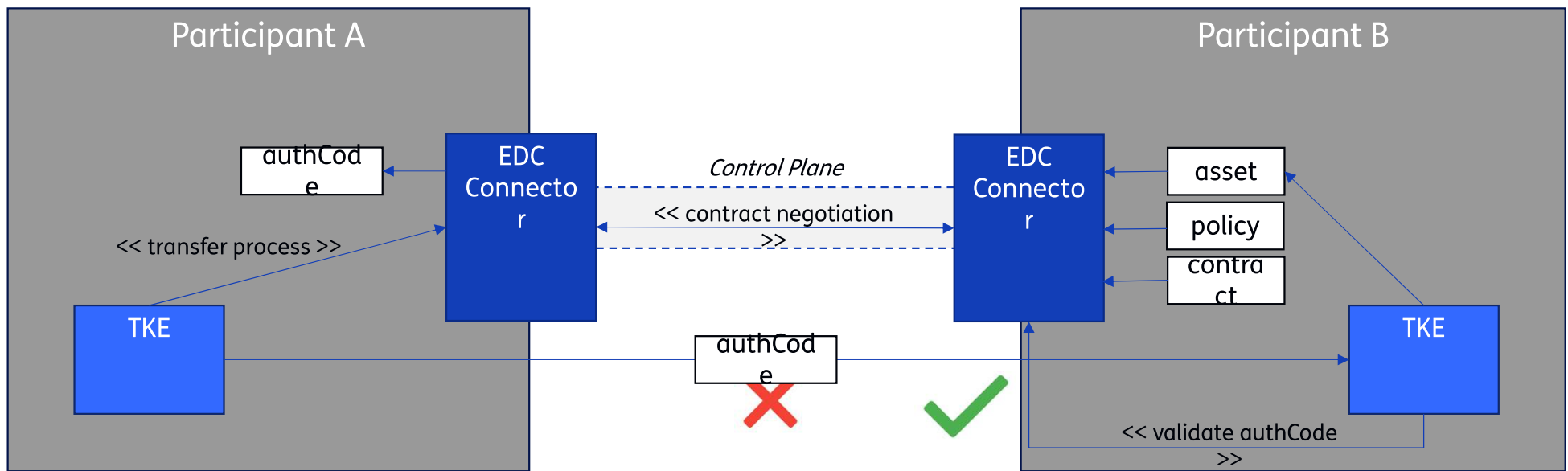
# Future challenge – work in progress 2



**Integrating  
semantic technologies  
into data space architectures**



# TKE - EDC integration: 2 participants



# Future challenge – just started 1

- *“Metadata, you see, is really a love note – it might be to yourself, but in fact it’s a love note to the person after you, or the machine after you, where you’ve saved someone that amount of time to find something by telling them what this thing is.”*
- It is important for the vocabulary provider to support in creating and maintaining high quality data models.
- We recognize the 5-star model for open data and vocabulary use

**How to improve data quality  
and quality of data models?**



# 5 star open data by Tim Berners-Lee

**5 ★ OPEN DATA**

BY EXAMPLE COSTS & BENEFITS SEE ALSO

## 5 ★ OPEN DATA

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.

Diagram illustrating the 5-star Open Data scheme, showing a staircase with five steps, each representing a star level. The steps are labeled with icons: PDF, XLS, CSV, RDF, and LOD. A yellow arrow points upwards from the bottom step to the top step.

**5 ★ OPEN DATA**

BY EXAMPLE COSTS & BENEFITS SEE ALSO

## BY EXAMPLE ...

Below, we provide examples for each level of Tim's 5-star Open Data plan. The example data used throughout is 'the temperature forecast for Galway, Ireland for the next 3 days':

- ★ make your stuff available on the Web (whatever format) under an open license<sup>1</sup> [example ...](#)
- ★★ make it available as structured data (e.g., Excel instead of image scan of a table)<sup>2</sup> [example ...](#)
- ★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)<sup>3</sup> [example ...](#)
- ★★★★ use URIs to denote things, so that people can point at your stuff<sup>4</sup> [example ...](#)
- ★★★★★ link your data to other data to provide context<sup>5</sup> [example ...](#)

# 5-star model for vocabulary use

source: [Five Stars of Linked Data Vocabulary Use](#)



Data without any vocabulary. e.g. 'LA temp 37.'



There is dereferenceable **human readable** information about the used vocabulary.



The information is available as **machine-readable** explicit axiomatization of the vocabulary.



The vocabulary is linked to other vocabularies



Metadata about the vocabulary is available

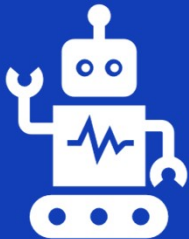


The vocabulary is linked to by other vocabularies.

# Future challenge – Data to vocabulary mapping

```
{ "data_battery_level":0.34,  
  "data_building":"urn:Building:SmartThings:d563d271-4855-4728-8367-  
  cfb205cf3da",  
  "data_room":"urn:Room:SmartThings:38f802d2-bb3c-4bf4-ba28-23c5ae3a64d0",  
  "data_status":0,  
  ...}
```

**How to reduce the amount of manual labor in developing and mapping semantic standards?**



In: "data\_battery\_level"  
Out: saref:hasValue

In: "data\_battery\_level"  
Out: <meas\_uri> rdf:type saref:Measurement .  
<meas\_uri> saref:hasValue <input> .  
<device> saref:isMeasuredIn <uom> .



**Sensor Device JSON data**

search

- 1...1 device\_link\_type\_description
- 1...1 data\_building
- 1...1 data\_room
- 1...1 data\_battery\_level**
- 1...1 device\_manufacturer\_description
- 1...1 device\_model\_description
- 1...1 device\_parent\_serial\_number
- 1...1 device\_product\_id
- 1...1 device\_serial\_number
- 1...1 device\_type\_description
- 1...1 message\_timestamp
- 1...1 message\_type\_description
- 1...1 message\_type\_id

AI mapping suggestion

**saref:hasValue**

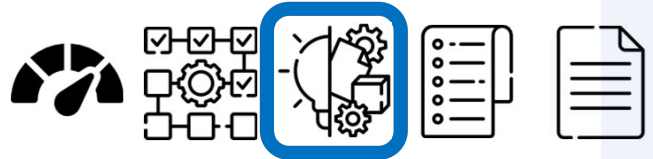
*A relationship defining the value of a certain property, e.g., energy or power. Note that, even if numeric values are expected to enable reasoning, measurement values could use other datatypes.*

**Justification:** saref:hasValue is used for data\_battery\_level because it defines the value of a certain property, such as the battery level.

**ACCEPT SUGGESTION**

**SAREF** v 1.3

- saref:hasModel
- s4bldg:isContainedIn
- saref:hasTemperature
- saref:hasState**
- saref:hasModel
- saref:hasValue
- saref:hasTemperature
- s4bldg:isContainedIn
- saref:hasModel
- saref:hasValue
- saref:hasTemperature
- s4bldg:isContainedIn**
- saref:hasModel
- saref:hasValue
- saref:hasTemperature
- s4bldg:isContainedIn
- saref:hasModel
- saref:hasValue
- saref:hasTemperature
- s4bldg:isContainedIn
- saref:hasModel**
- saref:hasValue
- saref:hasTemperature
- s4bldg:isContainedIn
- saref:hasValue**
- saref:hasModel
- saref:hasTimestamp





SEMANTIC  
TREEHOUSE

# Thank you

Want to know more or collaborate  
on the challenges presented?

Reach out to Linda Oosterheert:

[linda.oosterheert@tno.nl](mailto:linda.oosterheert@tno.nl)



**TNO** innovation  
for life