TC DATA
A Framework for Open, Trusted Distributed Data Infrastructures

Presented by: Diego López

3rd April 2025

The (Increasingly Distributed) Data Landscape

- Data exchange is the essential functionality of any *data* network
  - The requirements for such data exchanges evolve rapidly as their applications evolve

- Consolidation of agent applications
  - Acting autonomously according to intent
  - Able to collaborate and select appropriate data
  - We often use the term AI to refer to them

- Distributed and multi-domain
  - Very much like network themselves
  - Requiring solutions in such spaces

- A central item in the ETSI Technology Radar

The notion that "Data are the New Oil" has been around for quite some time. Generally credited to mathematician Clive Humby: [2006]
Data, like oil, are not useful in its raw state. They need to be refined, processed and turned into something useful.
Their value lies in their potential

The TC DATA Initiative

A call to action on data technologies was presented to the ETSI Board

- Requesting a decisive ETSI positioning on data solutions
  - Leveraging the results of existing fruitful initiatives
  - Consolidating open data solutions
  - Fostering innovation in base technologies
  - Exploring different application domains
  - Enhancing privacy and security
  - Solidifying ETSI's presence in the data sphere

TC DATA was approved by the end of January

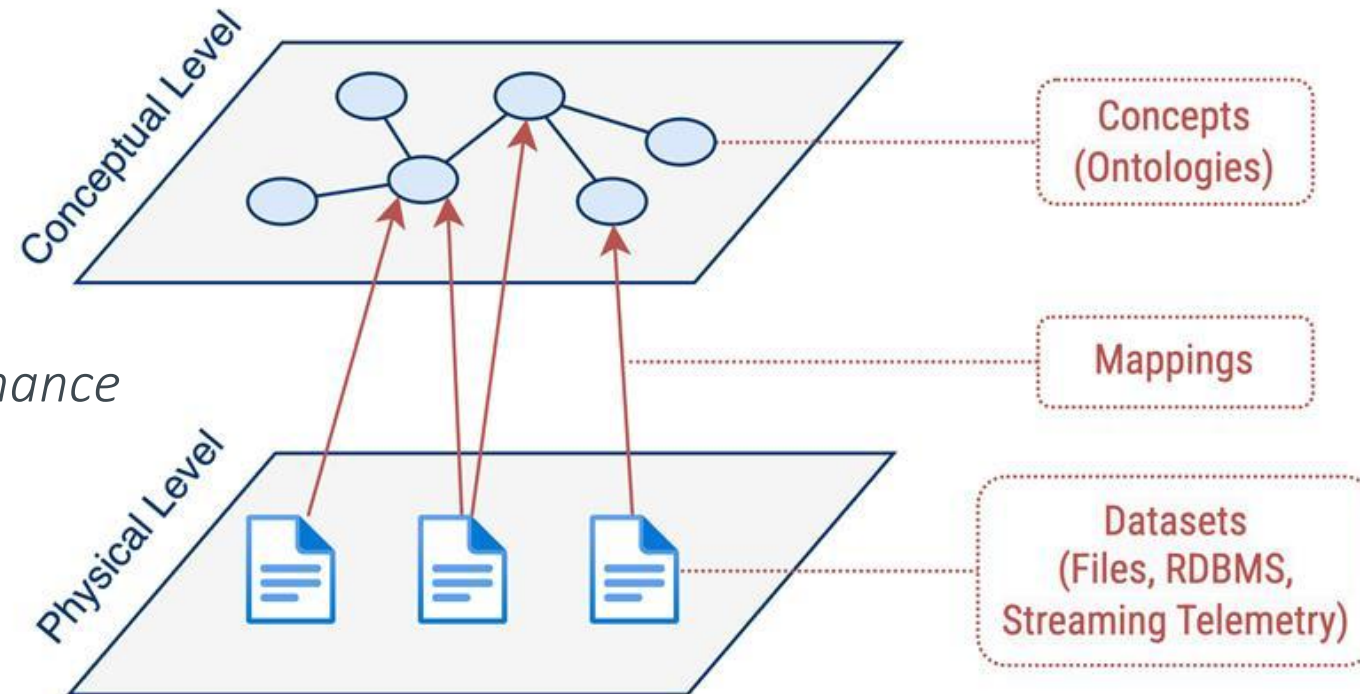- And we are currently working in making it take off

The Scope

- Develop deliverables to support the deployment and operation of distributed data solutions
  - Connectivity: data in transit
  - Storage: data at rest
  - Compute: data in process

- Address European policy and regulatory requirements and engage with other regulatory bodies
  - Ensuring relevant global, regional, and national requirements

- Provide input on technical aspects of the ETSI responses to governmental requests on data solutions
  - Special emphasis on the European Data Act and to the data-related aspects of the European AI Act

- Collaborate with open-source initiatives relevant for the data domain standardization
  - Reference implementation and interoperability testing

- Cooperate with other European and international standards organizations
  - Avoid duplication of work and promote harmonization

The Proposed Activities

- Providing a centre of expertise in the area of data infrastructures, services and applications
  - In coordination with other ETSI activities
- Developing technical standards to support data interoperability and semantic interoperability
- Maintaining and evolving specifications related to data solutions and published by other ETSI TGs
  - A list of initial agreements and interests
    - TC SmartM2M and TC ESI
    - ISG CDM, ISF CIM, and ISG PDL
- Supporting the development and the maintenance of semantic and data models
  - SAREF (including the SAREF open portal)
  - NGSI-LD
- Supporting the transposition in ETSI of the outputs of oneM2M
- Supporting the maintenance and evolution of relevant industry data standards

Open Data Infrastructures

- Support the autonomous use of data by agent applications
  - And by any other data-driven technology

- Aligned with FAIR principles
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable

- According to an appropriate *data governance*
  - Data access control
  - Data consistency
  - Data privacy preservation

- Leveraging domain dataspaces
  - Facilitating their use and integration

Conceptual Level

Concepts (Ontologies)

Mappings

Physical Level

Datasets (Files, RDBMS, Streaming Telemetry)

# Distributed Data Infrastructures – Data Products

- FAIR principles for data productization

- Decentralized data management paradigms
  - Data fabric, data mesh

- Knowledge graph as an essential enabler
  - Graph representation of data with the knowledge about them

- Data plus semantic metadata
  - Grounds on formal, shared representations of knowledge
  - Thesaurus, ontology, taxonomy…

- Data integration
  - Heterogenous data
  - Distributed data silos
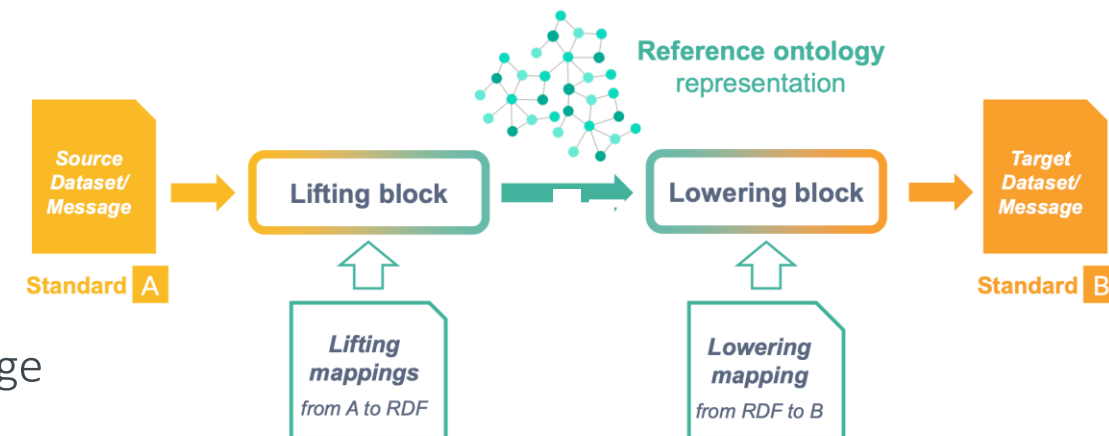
Distributed Data Infrastructures – Ontologies

- Challenges in ontology development
  - Skills on semantic modelling
  - Time-consuming and mostly manual task
  - Requires domain expertise
  - Tooling
- Linked Open Terms (LOT)
  - Standard methodology rather than standard ontologies
  - Facilitate development of ontologies for use cases
  - Reuse other (standard) ontologies for interoperability
  - Embraces agile software development practices
  - Open-source tooling
- ETSI SAREF as reference
  - IoT landscape
  - Generalization

LOT Methodology http://lot.linkeddata.es
@Ontology Engineering Group
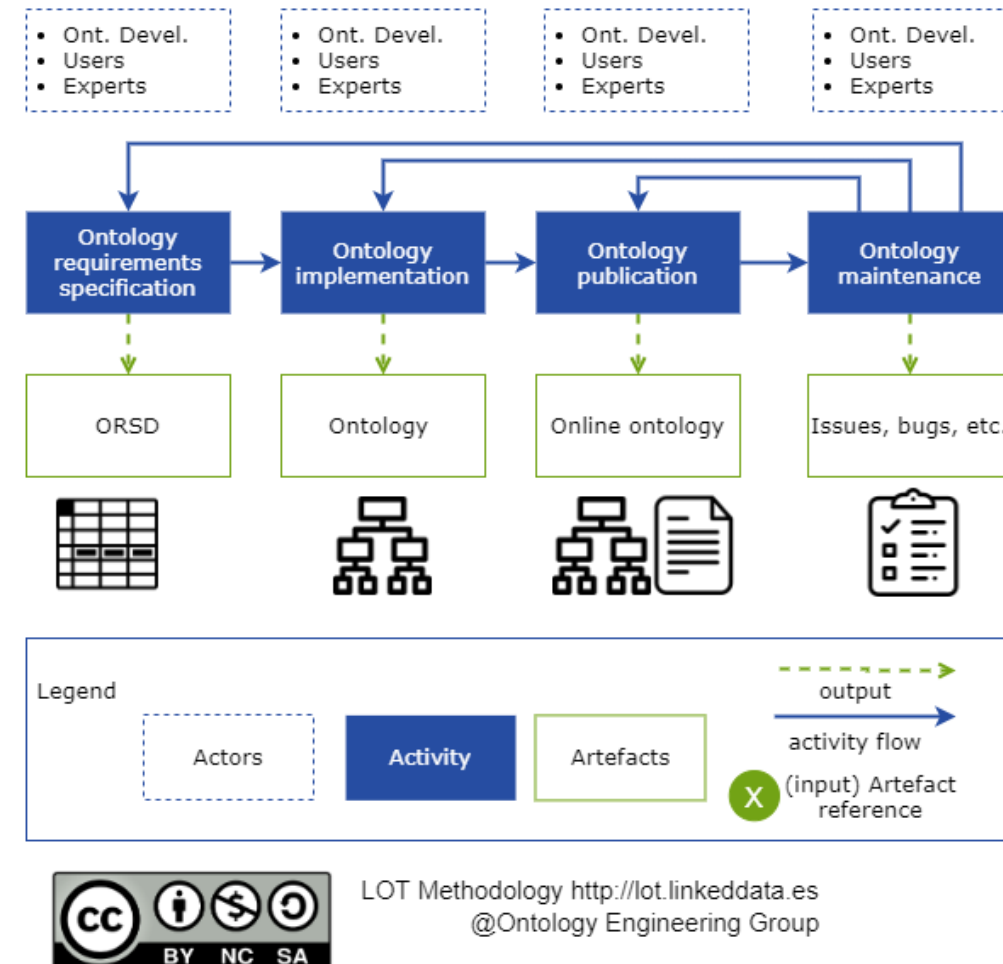
Trusted Data Infrastructures - Identities

- Agent applications require proper identification of workloads
  - And their connection to human identities
  - Not only an essential enabler
  - Suitable to become a network service
- Current issues
  - Identity management is human-centric
  - Simple policies
  - Accounting and auditability
  - Support identity traceability
  - Security and privacy implications
- Exploring first applications in data governance
  - Based on OAuth tokens and policy agents
  - Recursion becomes essential
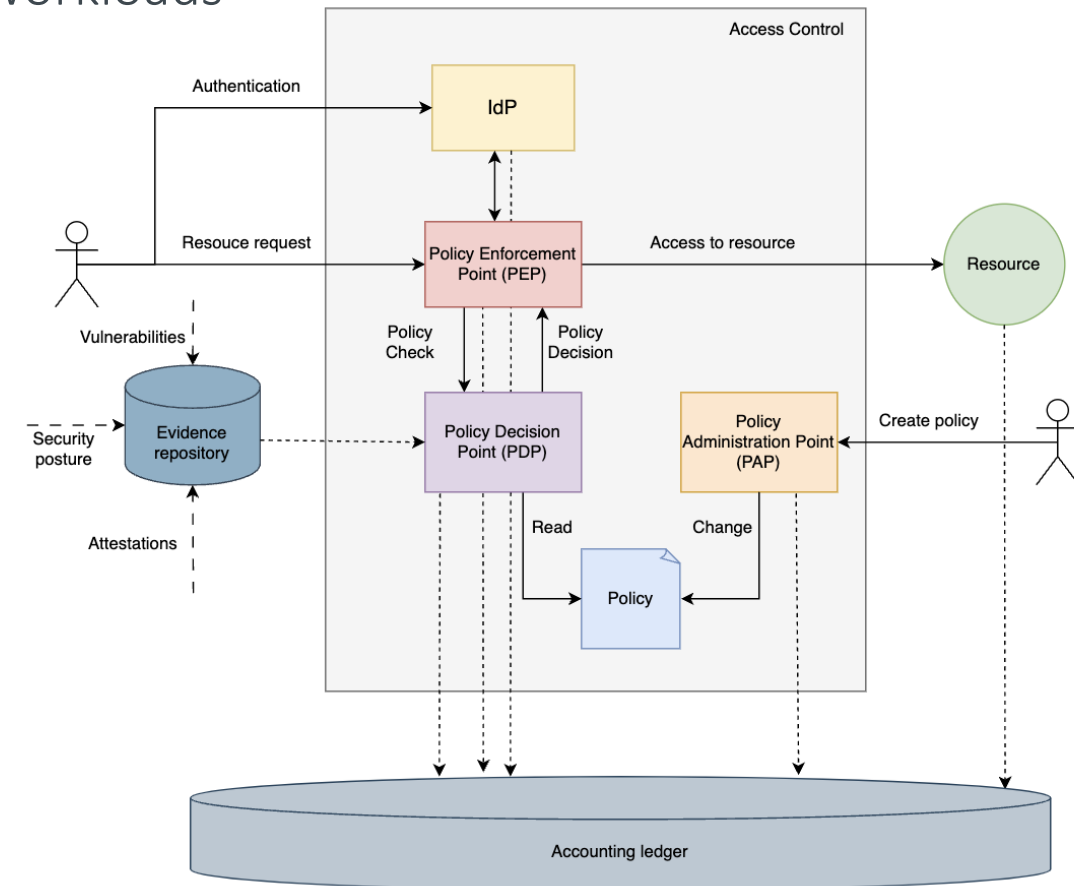
Trusted Data Infrastructures – Provenance and Evidence

- *Provenance*: A documented trail accounting for the origin of a piece of data and where it has moved from to where it is presently
 - Assurance of data origin and integrity
  - Whenever the dataset is used beyond an original online flow
  - Use of data intermediaries, such as data lakes
  - AI/ML training and validation
  - Audit trails, including forensics evidence

- *Evidence*: Supported by a transparent notary service
  - As a common, reliable source
  - Verifiable, auditable, distributed
  - Source for policy decisions

- Open issues
  - Recursion in compose datasets
  - Granularity
  - First approaches based on concise signing and supply chain assurance
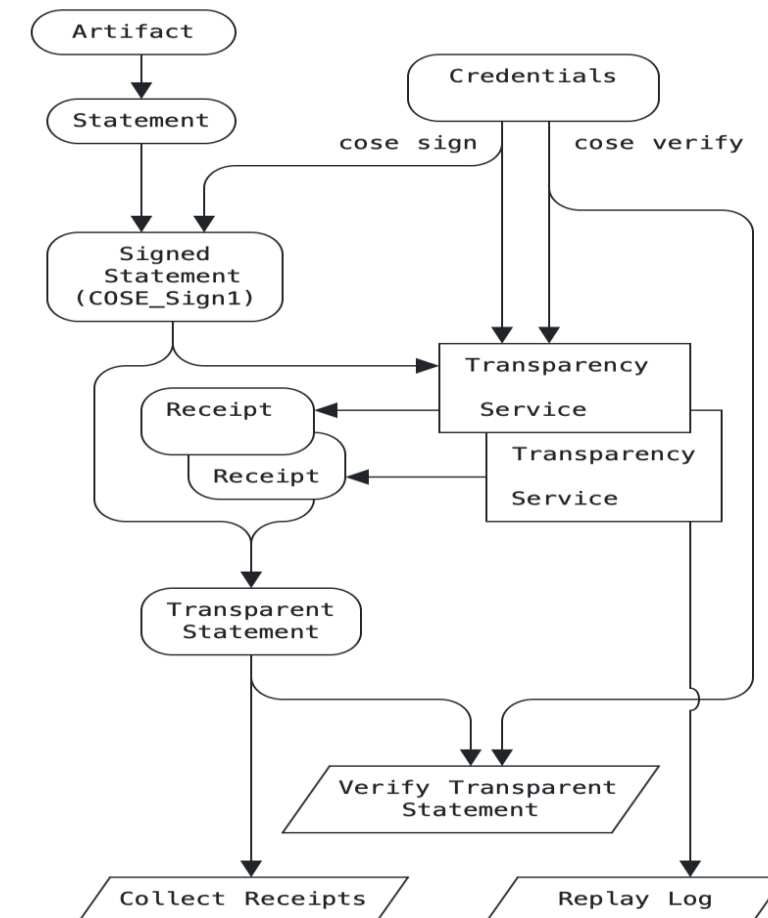
Application Scenario: A MEE Graph for Distributed Services

- Structured around three main functions

- Modeling
  - Identify the guidelines to be applied, expressed as intents
  - Evaluate their feasibility and assign service levels
  - Set the autonomic mechanisms in support of the service levels

- Enactment
  - Apply the guidelines to a given situation, as a policy decision
  - Identity, so every entity involved in a particular decision is uniquely identified
  - Evidence, allowing to make the most complete evaluation possible of the status of such entity

- Explainability
  - Support the auditing processes, guaranteeing the evaluation of the decisions and related actions
  - In the light of the applicable models and evidence
  - Involved identities and the relevant evidence for each decision