Data Management Plan in CONNECT-NM

Michał Pecelerowicz (NCBJ) – RL1/WP4 leader

10.06.2025

Table of contents

# Overview of the Data Management Plan

- **Purpose**
    - Fundational document for data management,
    - Defines a standardized methodology for managing, storing, and preserving research data across the CONNECT-NM partnership,
- **Scope**
    - Applies to all partners generating data; covers the full data lifecycle and diverse research lines,
- **Key components**
    - Tools, data classification, metadata standards, licensing, access policies, responsibilities,
- **FAIR principles**
    - Aligned with FAIR principles - the foundation for data integration, metadata design, and access policies across the project,
- **NM-KOS alignment**
    - Closely aligned with NM-KOS.

NM-KOS - overview

- **Nuclear Materials Knowledge Organization System (NM-KOS)**, a specialized semantic-based tool, that facilitates effective data cataloguing, storage, organization, retrieval, and search.

- NM-KOS adopts W3C standards to support interoperability and enable the creation of a federated data model. In this model, NM-KB acts as a central access point where users can query across multiple, distributed databases using SPARQL.

- Two-layer architecture:
  - Data Layer – Data Management System (DMS)
  - Semantic Layer – Knowledge Base (KB)

- **DMP is mainly about the definition of the best practices for the setup of a local DMS and provides methodological guidance for DMS**
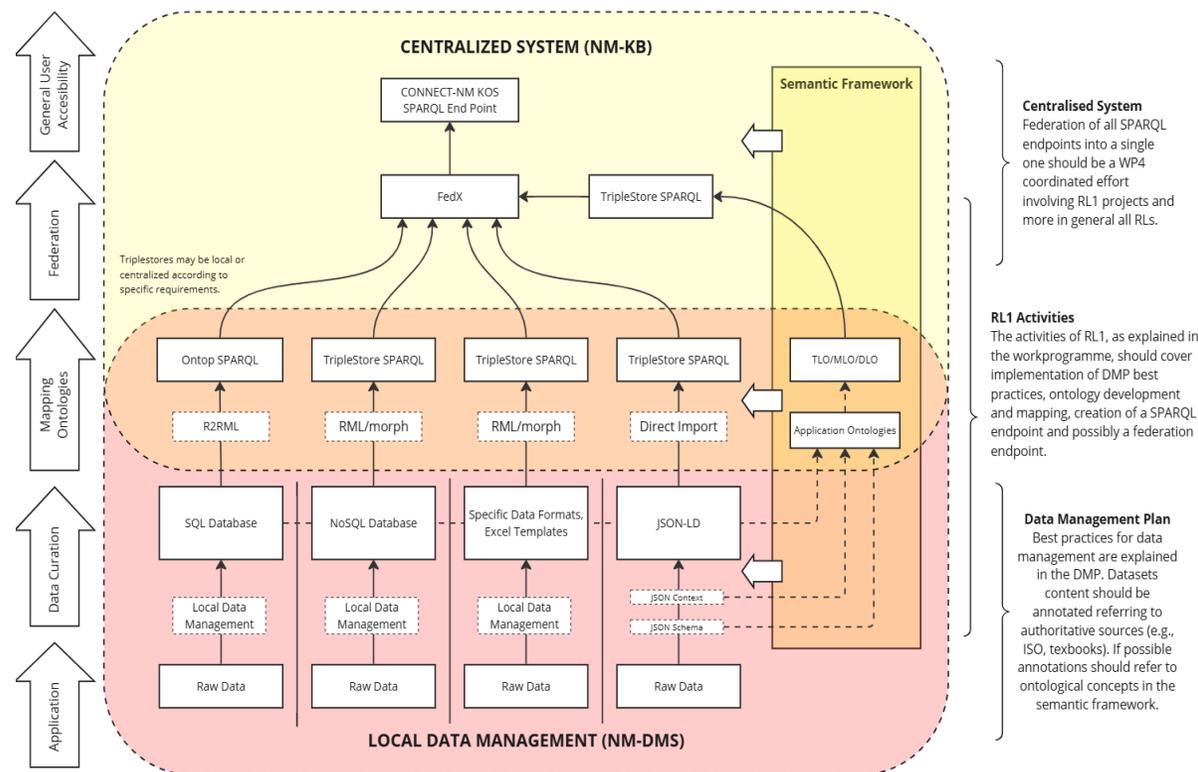
NM-KOS – data integration

- Different integration levels:

    - **Partial Integration** (data cataloguing) – involves metadata-level integration using DCAT metadata in RDF format. Datasets metadata will be provided by each data generating partner and uploaded in the NM-KB, to ensure datasets findability within the NM-KOS. Data interoperability will be limited at the catalogue level.

    - **Full Integration** – requires the mapping and virtualisation of the datasets/databases to fully enable data federation. Full integration requires the development (or re-use) of an application ontology and mapping. Interoperability is thus specifically addressed – this approach provide a shared and understandable language across different domains.

Data Summary, Policies and Classification

- **Origins of data at CONNECT-NM:**
    - Experimental, observational, images, texts, numerical data. Originate mostly from Research Lines 2-5.
- **Storage policy:**
    - **Local** storage (institutional premises) or **trusted repository** storage,
    - NM-KOS design, based on **distributed databases**, supports partners specific approaches to data management, and removes the need for a centralised data repository. Advantages: full ownership and security
    - **Each partner is responsible for the storage of the data generated during the project activities. No centralised data storage/repository system is foreseen for the CONNECT-NM project.**
    - Access to data **does not mean always making data open by storing them in a third-party public repository**. Instead, data owners may **opt to store data locally in their institution/company providing public access only to the metadata**, clearly specifying who can access the data under what conditions or whom to ask for permission.
- **Data Classification:**
    - **Raw (internal) data:** typically large, application-specific, and stored in non-standard or inconsistent formats that limit their reuse and compatibility with RDF tools.
    - **Mappable data:** reusable, domain-independent, and stored in consistent formats compatible with RDF tools, enabling their integration into the NM-KOS.
    - **Partners are required to classify data as RAW or MAPPABLE to enable appropriate NM-KOS integration.**

Data Formats and Standards

- NM-KOS allows partners to keep their existing file formats and databases

- Proprietary formats should be avoided or kept only in internal repositories

- Preferred: tabular data, standard formats (CSV, XML, JSON, RDF, etc.) and structured databases (SQL/NoSQL)

- Widely recognized technologies will ensure interoperability and NM-KOS integration

During the periodic interproject meetings, the CONNECT-NM tasks in charge of RLs guidance (T5.2, T6.2, T7.2 and T8.2) should support the projects in the definition of the most suitable database technologies and file formats.

This can be supported by WP4.

Making Data Findable

- **Persistent identifiers:**
    - To ensure that CONNECT-NM data are accessible and easy to locate, **each data generating partner should assign a Persistent Identifier (PID) to their data and/or metadata**. This can be achieved by creating a record in a trusted or recommended repository.
    - Prominent types of persistent identifiers include the **Digital Object Identifier (DOI)** and the **Handle System;**
    - A list of general purpose repositories for the creation of PIDs for data and/or metadata can be found here: https://erc.europa.eu/manage-your-project/open-science.

Making Data Findable

DCAT metadata

- DCAT is an RDF vocabulary that facilitates interoperability between web-published data catalogues. It enables **standardised descriptions** of datasets and services, improving **discoverability, aggregation, and decentralised publishing**. This supports **federated search** across multiple catalogues using a **common query structure**.

- **Metadata** for each CONNECT-NM generated datasets must be **provided in RDF compatible format using DCAT vocabulary to WP4**. Such metadata will be **uploaded into the NM-KB** and will provide a way to collect in a shared knowledge space, information about all datasets generated by CONNECT-NM activities.

- Minimum metadata requirements mandatory for all datasets: dcat:Dataset, dcat:Distribution, dcterms:title, dcterms:creator, dcterms:issued, dcat:version, dcterms:license, dcterms:accessRights, dcterms:description, dcat:landingPage, ...
(full list of ~20 fields available in the DMP).
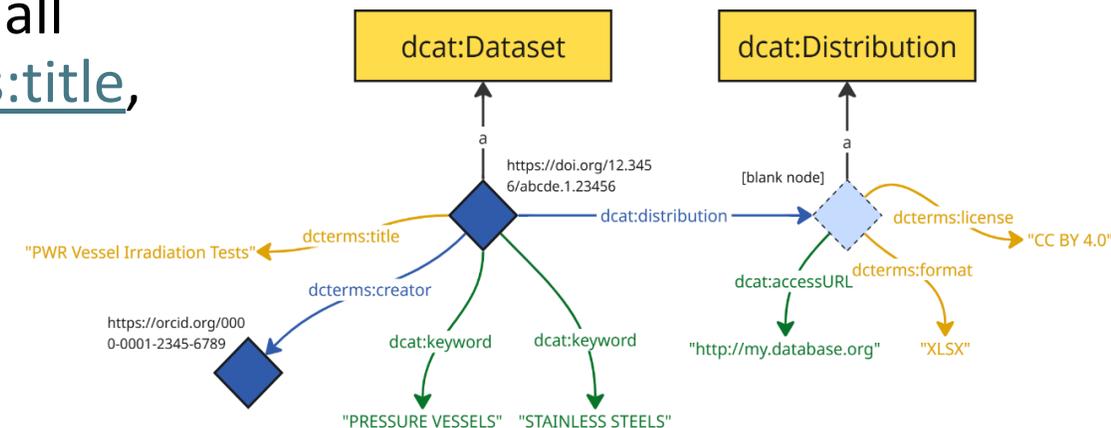
Making Data Accessible

- Data in NM-KOS becomes **accessible** only after **integration**. There are two levels:
  - **Partial** (DCAT-based cataloguing)
  - **Full** (ontology + mapping)

- The choice depends on **data type** and **research line**
  - **Raw data** – no integration or **partial** integration, **mappable data** – **partial** integration **or full** integration
  - **RL1** requires **full** integration. **Other RLs** require at least **partial** (DCAT)
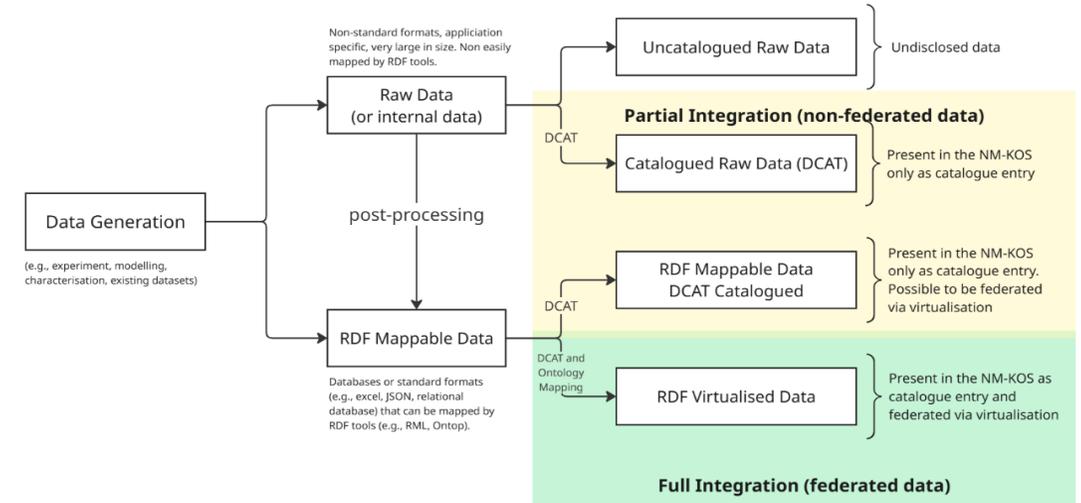
- Institutions must **identify types of datasets** through internal **data inventory process**.
  - **Data is generated through processes like experiments or simulations.** This results in either **raw** or **mappable** data.
  - **Raw data** may be large or in non-standard formats. **Mappable data** fits standard formats (e.g., JSON) and can be mapped into RDF.
  - Post-processing can turn raw into mappable.
  - **Raw** data may remain private or be **catalogued.** **Mappable** data can be **catalogued** or **fully integrated**.

Making Data Accessible

- **Data Integration ≠ Data Disclosure**
    - NM-KOS integration improves visibility through metadata and/or semantic mapping, however it **does *not* automatically imply disclosing the data itself**.
    - Each data provider defines whether their data is openly accessible, available under conditions (such as an NDA), or limited to metadata exposure only.

- **Data**
    - Each data owner must **ensure open access to their dataset via their repository of choice**, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', **unless exceptions apply** [CONNECT-NM GrA]

- **Metadata**
    - Metadata of deposited data **must be open** under a Creative Common Public Domain Dedication (CC 0) or equivalent (**to the extent legitimate interests or constraints are safeguarded**), in line with the FAIR principles [CONNECT-NM GA]

- **Data supporting publications**
    - For data supporting publications, open access is required in a trusted repository, **unless there are valid, documented reasons not to** [Horizon Europe-funded publications]

Making Data Interoperable and Reusable

- **Semantic Interoperability:** achieved Using Semantic Web Standards and ontologies to enable cross-domain data integration and machine-readability.

- **Data reusability**: achieved through persistent identifiers, rich metadata, and decentralized access.

Resource allocation

- Each CONNECT-NM partner is expected to **allocate resources to support basic data management tasks** (metadata preparation, and alignment with the FAIR data practices of the project). Additional support may be provided through Work Package 4 and the RL1 projects, particularly for tasks involving semantic mapping or integration with the NM-KOS infrastructure.

- **RL1 work programme** requires that projects funded on this research line to mainly focus on data management activities and to the development and deployment of the semantic layers of the NM-KOS. For these projects, **the full integration of their datasets is mandatory**, and projects resources should support the achievement of this objective.

- To support this process, each CONNECT-NM funded project should designate **a dedicated data steward** responsible for coordination. The data steward will be responsible for: coordinating the provision of domain-specific metadata, populating or reviewing metadata templates, ensuring timely and accurate documentation of datasets, communicating data updates or structural changes to the NM-KOS Data Management Team

Thank you!

Michal.Pecelerowicz@ncbj.gov.pl

Introduction to the CONNECT-NM Data Management Plan and Nuclear Materials Knowledge Organisation System Architecture

- **ID4.5.1 NM-KOS Architecture**

Emanuele Ghedini – University of Bologna, WP4/RL1 Advisor, T4.5 (PSP) Leader

CONNECT-NM Online Webinar, 10/06/2025
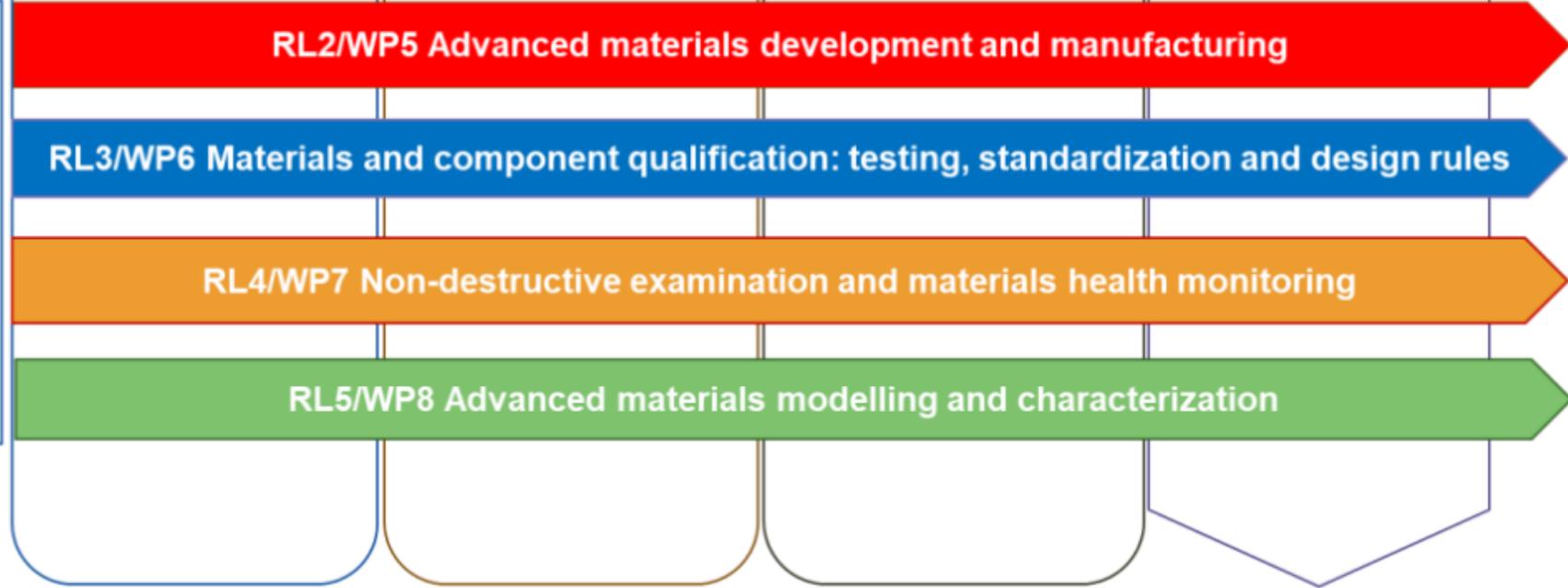

Co-funded by the European Union

CONNECT-NM (the slide of doom)

Internal Deliverable ID4.5.1

This report provides high-level documentation for the CONNECT-NM **Nuclear Materials Knowledge Organisation System (NM-KOS)** architecture and its default functionality. The NM-KOS architecture is based on a two-layer approach, splitting the responsibilities into Data Management (DM) and Knowledge Management (KM). In addition to presenting the architecture, this report outlines the design choices regarding available semantic and software solutions and provide a justification for their selection

GHEDINI, E., Paponetti, I. M., Arenas-Guerrero, J., Ferrucci, B., & Pecelerowicz, M. (2025)
CONNECT-NM Internal Deliverable ID4.5.1: Documentation of the NM-KB Architecture and Default Functionality
EURATOM2027 CONNECT-NM Project.

https://doi.org/10.5281/zenodo.15504436

CONNECT- NM
Coordination of the European Research Community on Nuclear Materials for Energy Innovation

A EURATOM Cofund Action
Co-funded by the European Union

Start date of project	01/10/2024
Duration	60 months
Reporting period	RP1

Work Package 4 – Nuclear Materials Knowledge & Data Management (RL1)

**Internal Deliverable ID4.5.1: Documentation of the NM-KB Architecture and Default Functionality**

Author(s) name and affiliation	(1) Emanuele Ghedini	UNIBO
	(2) Michal Pecelerowicz	NCBJ
	(3) Ilaria Maria Paponetti	UNIBO
	(4) Julian Arenas Guerrero	UPM
	(5) Barbara Ferrucci	ENEA
Date of issue	04/04/2025	
Date of final approval	04/04/2025	
Dissemination Level		
PU	Public	X
CO	Confidential, only for partners of CONNECT-NM and the EC	
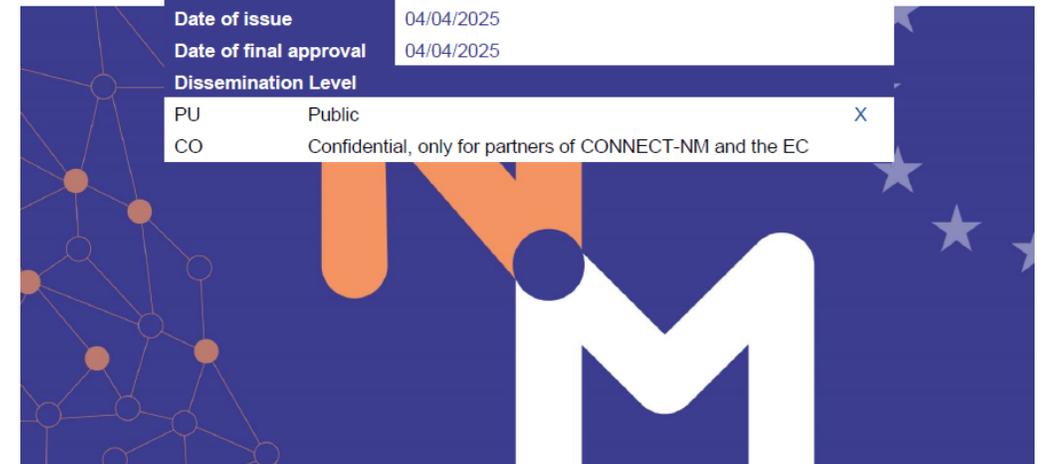
IAEA and Knowledge Management

Managing Nuclear Knowledge with Semantic Technologies

Puja Daya, IAEA Office of Public Information and Communication

MAY 27 2021

The nuclear sector is knowledge, information and data intensive. These require sophisticated and well-planned management. The IAEA is exploring the use of innovative technologies to help overcome this problem. (Photo: IAEA)

https://www.iaea.org/newscenter/news/managing-nuclear-knowledge-with-semantic-technologies

IAEA Nuclear Energy Series

No. NG-T-6.15

Exploring Semantic Technologies and Their Application to Nuclear Knowledge Management

Basic Principles

Objectives

Guides

Technical Reports

https://www.iaea.org/publications/13469/exploring-semantic-technologies-and-their-application-to-nuclear-knowledge-management

IAEA and Knowledge Management

Managing Nuclear Knowledge with Semantic Technologies

"With the past decade having seen vast technological improvements in **collecting**, **managing**, **understanding**, **presenting**, **sharing** and **using knowledge gained from data and information** all over the world, access to the latest updates have become just a web search away."

"**Semantic technologies** use a **wide range of tools** in a similar way the human brain does. It helps **process data and information and links them together** to create understanding and meaning."

"Semantic technology, can be used in the nuclear field to help experts and stakeholders **maintain, preserve, link and share nuclear knowledge**. "Making more use of semantic technology could address the challenge of **co-ordinating and compiling information from various institutions** by enhancing access to knowledge resources throughout the field on a larger scale."

"Semantic technology, especially when **combined with artificial intelligence, machine learning, modern taxonomies and ontologies**, represents a **powerful tool for managing the vast amount of nuclear data**, information and knowledge."

https://www.iaea.org/newscenter/news/managing-nuclear-knowledge-with-semantic-technologies

IAEA and Knowledge Management

Managing Nuclear Knowledge with Semantic Technologies

"Semantic technology improves the organisation of data and information – by **linking various sources so they can be shared and reused across sectors, organisations and scientific communities**. Through this improved knowledge, information and data organisation, **nuclear safety standards**, **recommendations**, **experiences**, **best practices** and **previous research** can become more widely available."
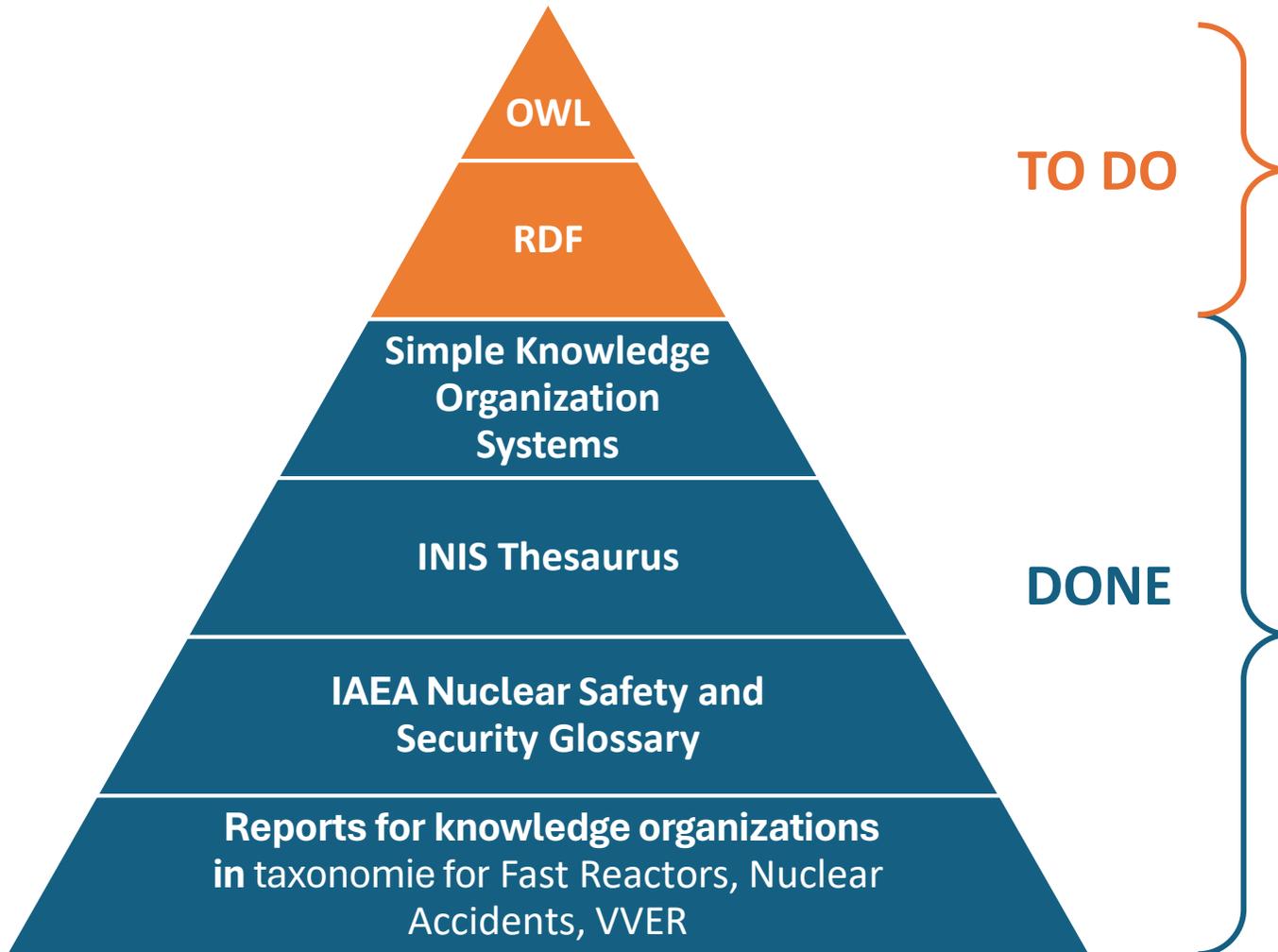
"If already existing data and information can be connected to other sources using semantic technology**, research and development in the field becomes easier and faster and could lead to more breakthroughs**."

"It is a **major challenge** to **adapt a systematic approach to nuclear knowledge** and make it **accessible to all relevant parties and organisations on a local, as well as on an international scale**.."

"Currently, many nuclear organisations host their knowledge on portals which act as central repositories for **thousands of documents with little or no metadata. Without metadata, it is difficult to search and gain access to the information required**. Even greater difficulty would be to establish links between diverse information resources."

"Semantic technology includes a wide range of tools, standards and methodologies, allowing for the processing of information based on its context and meaning. To encode semantics – the meaning of a word, phrase, or text – technologies such as **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)** could be successfully used to **codify the metadata**."

IAEA and Knowledge Management

OWL

RDF

Simple Knowledge Organization Systems

INIS Thesaurus

IAEA Nuclear Safety and Security Glossary

Reports for knowledge organizations **in** taxonomie for Fast Reactors, Nuclear Accidents, VVER

TO DO

DONE

IAEA Interest publishing Exploring Semantic Technologies for NKM (2021)

IAEA joint the development of a RICH,EXPANDABLE nuclear knowledge bases and portals.

Linked Open Data Cloud

Linked data is structured data which is interlinked with other data so it becomes more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers.
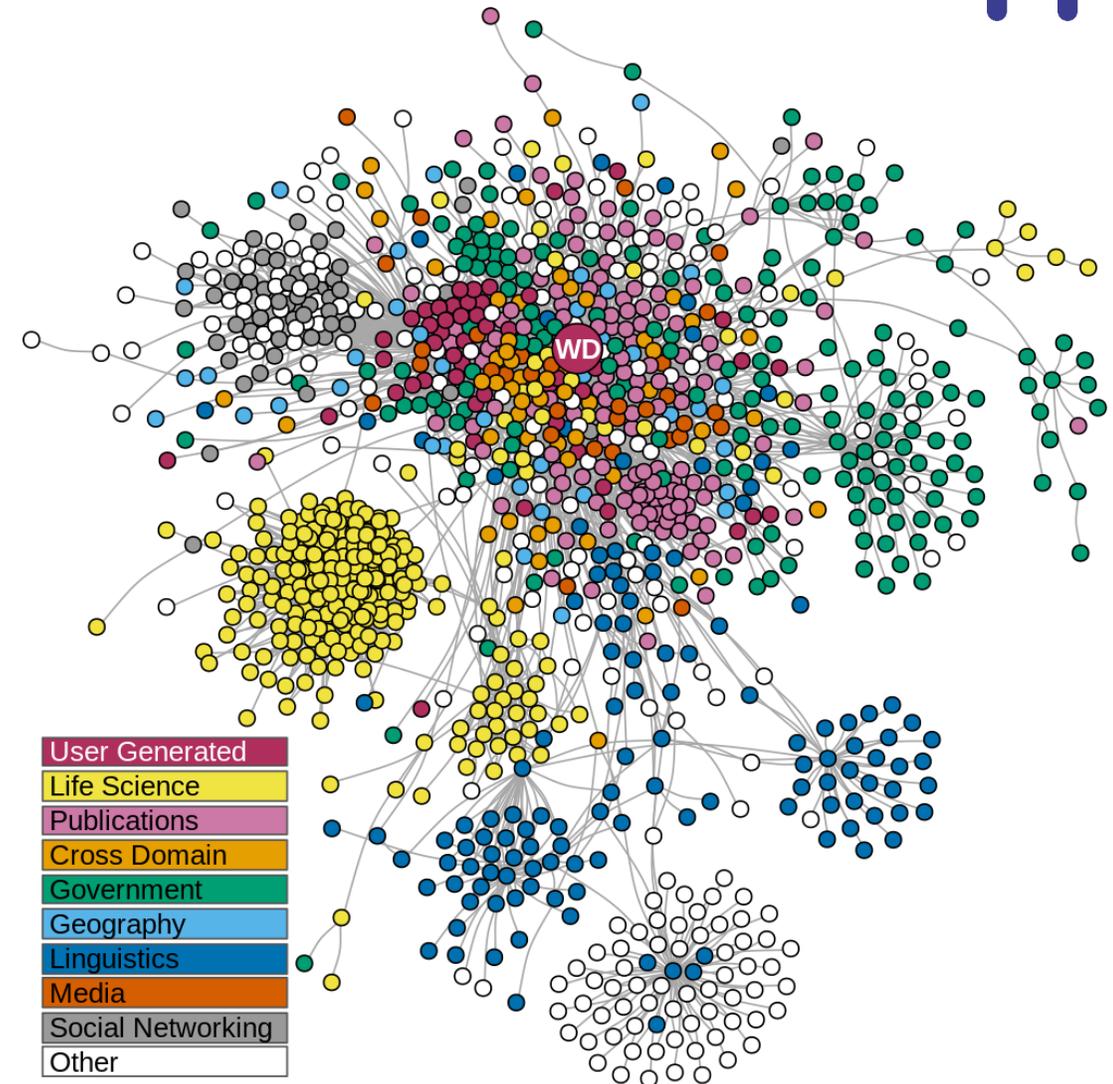
Part of the vision of linked data is for the Internet to become a global database.[1]

Tim Berners-Lee, director of the World Wide Web Consortium (W3C), coined the term in a 2006 design note about the Semantic Web project.[2]

https://en.wikipedia.org/wiki/Linked_data

The dataset currently contains **1314** datasets with **16308** links

https://www.lod-cloud.net/

Legend:
- User Generated
- Life Science
- Publications
- Cross Domain
- Government
- Geography
- Linguistics
- Media
- Social Networking
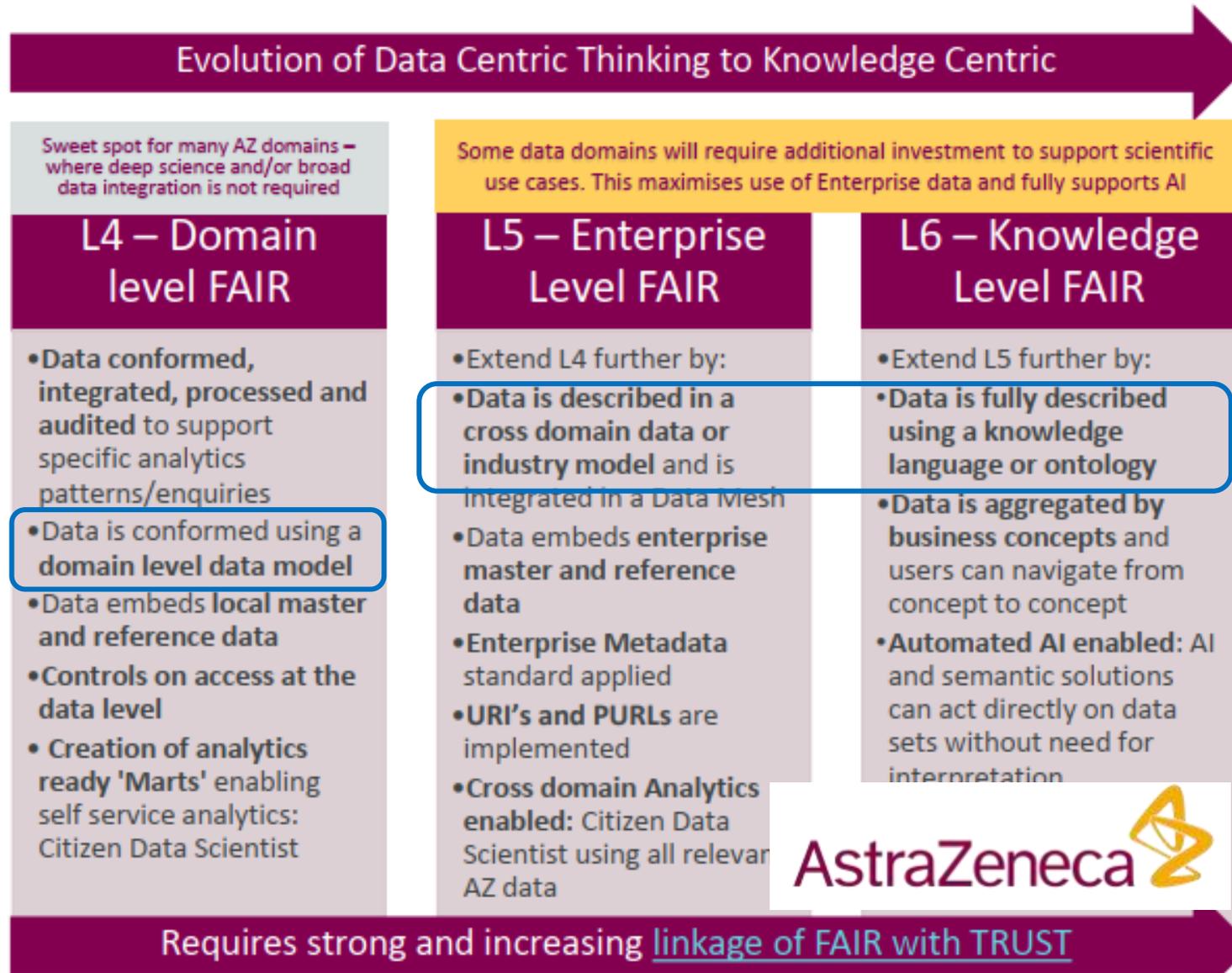- Other

From Data to Knowledge

AstraZeneca

FAIR Data-centric Information Architecture:

It's all fun and games until someone uses an "I"

Ben Gardner

Data Standards & Interoperability, Data Office, Data Science & Artificial Intelligence, R&D, AstraZeneca, Cambridge, United Kingdom
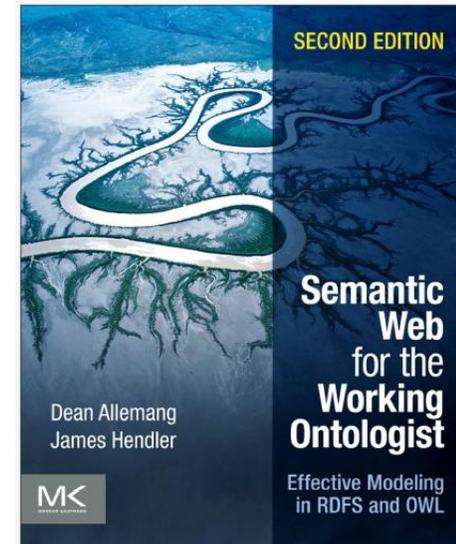
September 2023

https://www.slideshare.net/slideshows/fair-datacentric-information-architecturepptx/265566803

Evolution of Data Centric Thinking to Knowledge Centric

Sweet spot for many AZ domains – where deep science and/or broad data integration is not required	Some data domains will require additional investment to support scientific use cases. This maximises use of Enterprise data and fully supports AI	
L4 – Domain level FAIR	**L5 – Enterprise Level FAIR**	**L6 – Knowledge Level FAIR**
• **Data conformed, integrated, processed and audited** to support specific analytics patterns/enquiries	• Extend L4 further by:	• Extend L5 further by:
• Data is conformed using a **domain level data model**	• **Data is described in a cross domain data or industry model** and is integrated in a Data Mesh	• **Data is fully described using a knowledge language or ontology**
• Data embeds **local master and reference data**	• Data embeds **enterprise master and reference data**	• **Data is aggregated by business concepts** and users can navigate from concept to concept
• **Controls on access at the data level**	• **Enterprise Metadata** standard applied	• **Automated AI enabled:** AI and semantic solutions can act directly on data sets without need for interpretation
• Creation of analytics ready 'Marts' enabling self service analytics: Citizen Data Scientist	• **URI's and PURLs** are implemented	
	• **Cross domain Analytics enabled:** Citizen Data Scientist using all relevant AZ data	

### Requires strong and increasing linkage of FAIR with TRUST

Semantic Web Approach

The **W3C Semantic Web** provides a machine-readable format for the representation of data and their meaning, through RDF subject-predicate-object triplet structure, and incorporating logical elements with RDFS and OWL 2 languages.

RDF is providing a **graph structure** for distributed data on a cell-based approach, overcoming the limitation of tabular representation and enabling the AAA Slogan: "Anyone can say Anything about Any topic".

RDFS and OWL 2 provides the semantics through:
- **classes** defining the types of the entities
- **relations** between entities
- **axioms** imposing constraints
- **individuals** representing instances
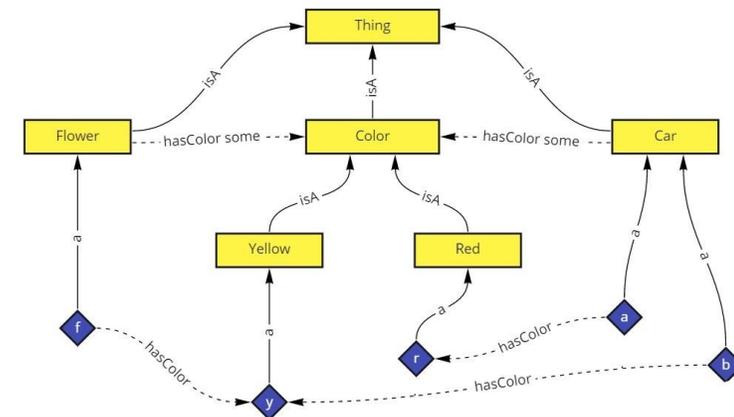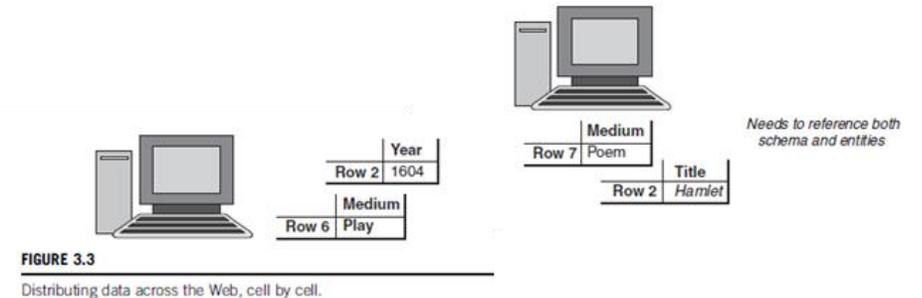
Distributed Data
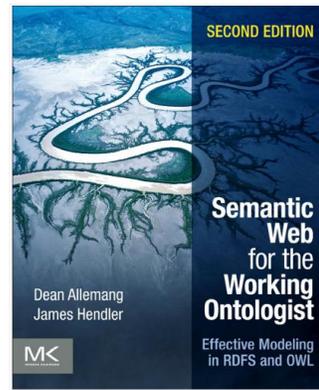
Data are most often represented in **tabular form**, in which each row represents some item we are describing, and each column represents some property of those items. The cells in the table are the values for those properties.

This approach is compact (*what you see is what you get* approach), provide fast access, but is **extremely rigid** if you want e.g. add a new column, or distribute part of the dataset in different places (like **dense matrix storage**).

Table 3.1 Tabular Data about Elizabethan Literature and Music

ID	Title	Author	Medium	Year
1	As You Like It	Shakespeare	Play	1599
2	Hamlet	Shakespeare	Play	1604
3	Othello	Shakespeare	Play	1603
4	"Sonnet 78"	Shakespeare	Poem	1609
5	Astrophil and Stella	Sir Phillip Sidney	Poem	1590
6	Edward II	Christopher Marlowe	Play	1592
7	Hero and Leander	Christopher Marlowe	Poem	1593
8	Greensleeves	Henry VIII Rex	Song	1525

Distributed Data
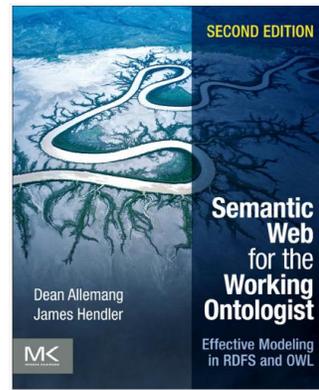
Distributed data systems are possible but have to provide **common data schema** (row based) or a **common id reference system** (column oriented), and **synchronisation** between servers to ensure data consistency.
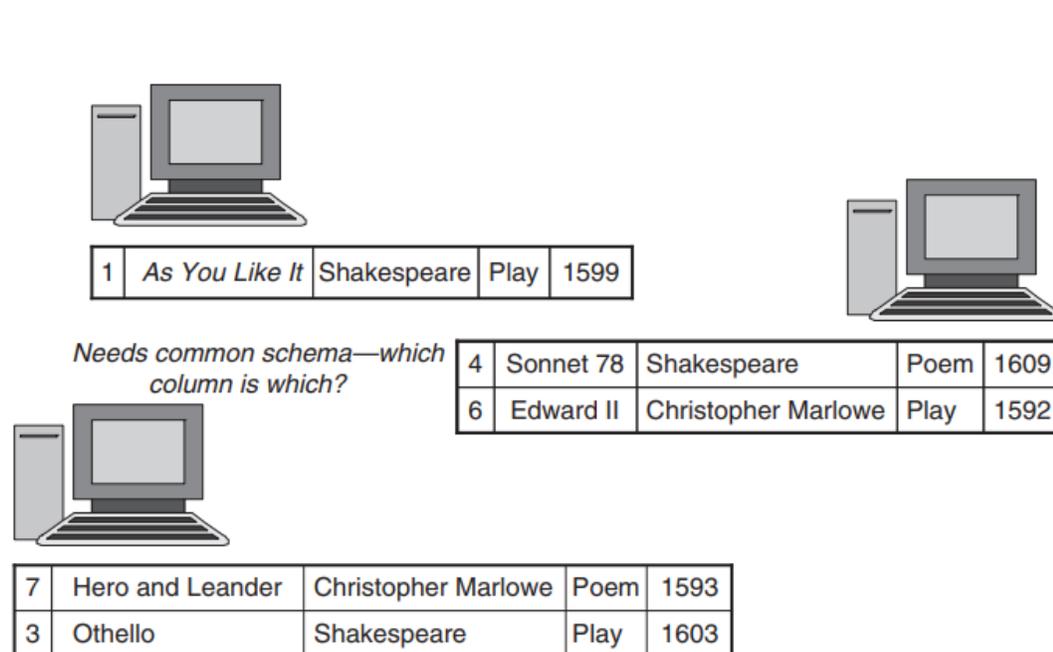
FIGURE 3.1

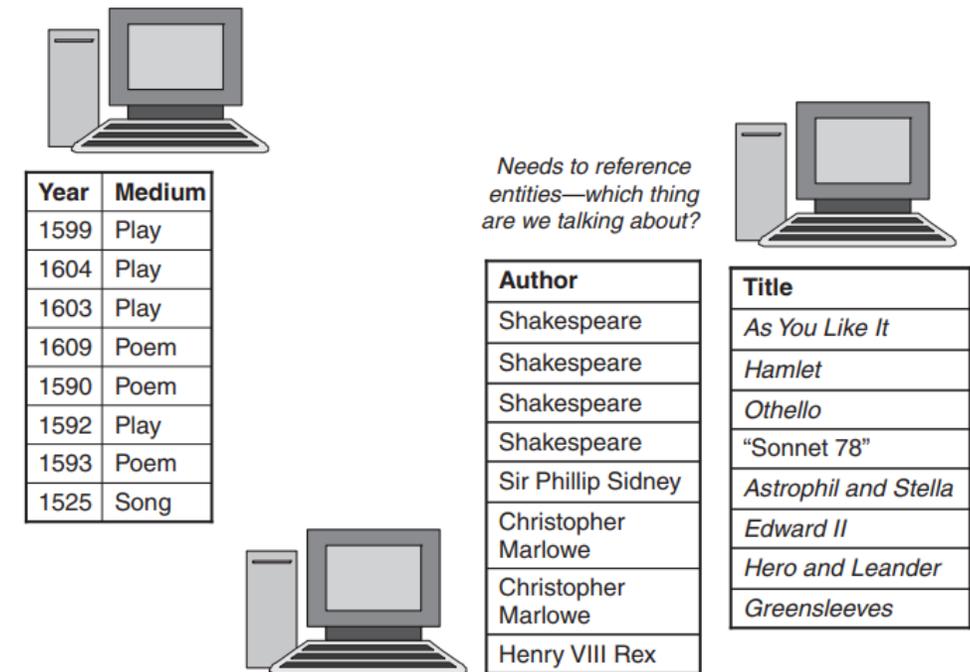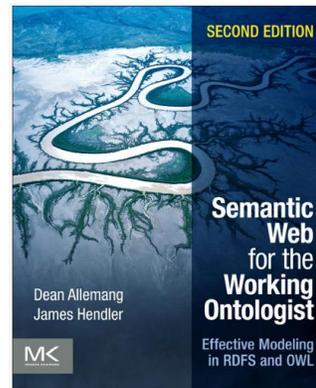Distributing data across the Web, row by row.

FIGURE 3.2

Distributing data across the Web, column by column.

Distributed Data

Information **distributed cell by cell**. Each machine is responsible for some number of cells in the table (similar to **sparse matrix storage**).

Pros: This system combines the **flexibility of both of the previous strategies**. Two servers can share the description of a single entity, and they can share the use of a particular property.

Cons: we need a **global reference for the column headings and a global reference for the rows**.
Each cell is represented with three values: a global reference for the row, a global reference for the column, and the value in the cell itself.

Table 3.2 Sample Triples

Subject	Predicate	Object
Row 7	Medium	Poem
Row 2	Title	Hamlet
Row 2	Year	1604
Row 4	Author	Shakespeare
Row 6	Medium	Play
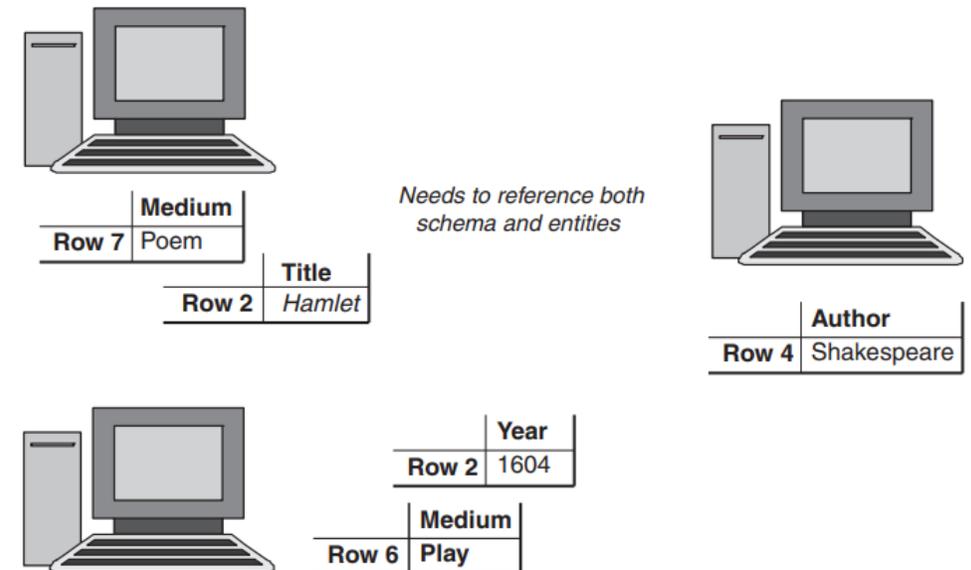
FIGURE 3.3

Distributing data across the Web, cell by cell.

Distributed Data
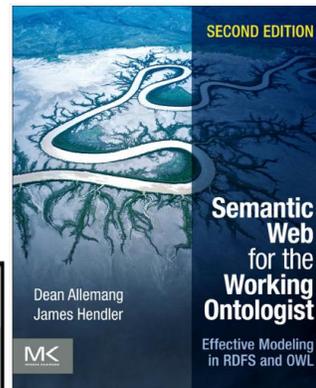
This third strategy is the strategy taken by **W3C RDF** (**R**esource **D**escription **F**ramework).

Since a cell is represented with three values, the basic building block for RDF is called **the triple**.

The identifier for the row is called the **subject**

The identifier for the column is called the **predicate**

The value in the cell is called the **object**.

Table 3.3 Sample Triples

Subject	Predicate	Object
Shakespeare	wrote	King Lear
Shakespeare	wrote	Macbeth
Anne Hathaway	married	Shakespeare
Shakespeare	livedIn	Stratford
Stratford	isIn	England
Macbeth	setIn	Scotland
England	partOf	UK
Scotland	partOf	UK
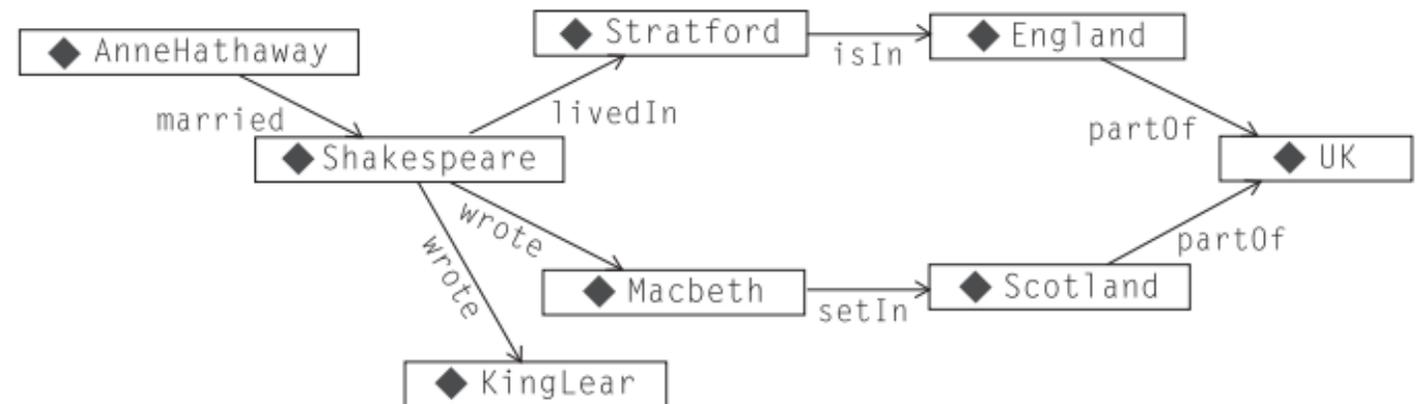
FIGURE 3.4

Graph display of triples from Table 3.3. Eight triples appear as eight labeled edges.

Distributed Data

Given this cell-based representation, it is easy to **connect together information about different domains** through dedicated relations.

In the example, *livedIn* and *setIn* connect information about Shakespeare and UK.
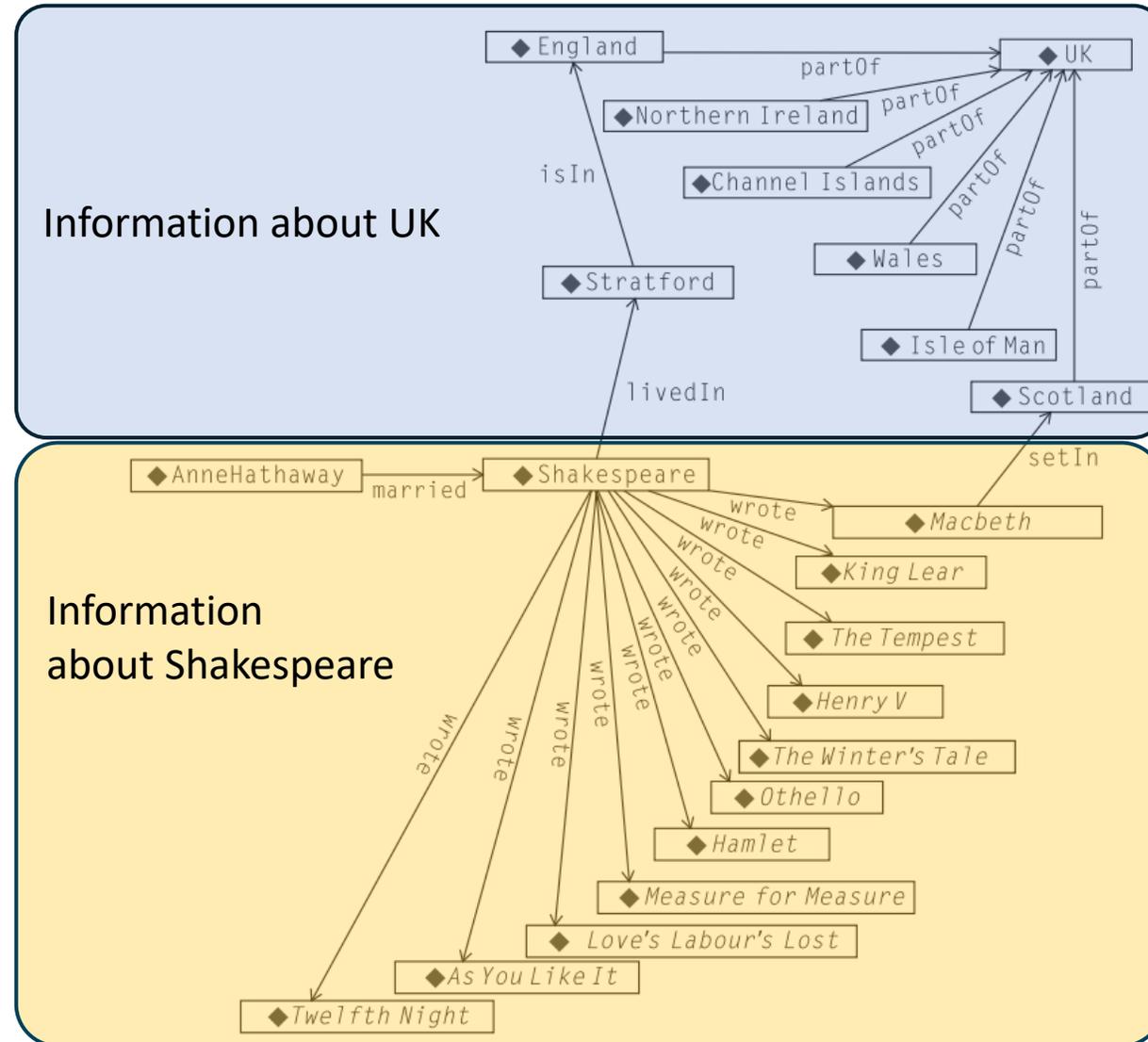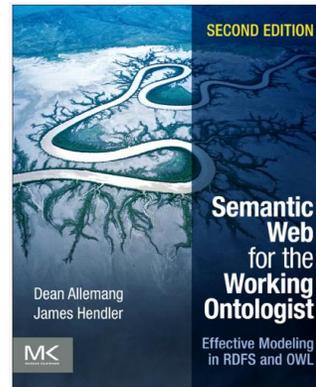
These are the so-called **Knowledge Graphs**

FIGURE 3.6

Combined graph of all triples about Shakespeare and the United Kingdom.

Triplestore

The term **Triplestore** typically refers to any system that has facilities for persistent storage of **Resource Description Framework (RDF) data triples** (i.e., subject-predicate-object).

Triplestores:

- are **frameworks** that are more than just RDF databases

- typically provide a so-called **SPARQL endpoint** to query the triplestore via SPARQL queries (pretty similar to SQL queries, but for RDF)

- typically, also provide **Application Programming Interfaces (APIs)** that allow **handling and manipulating RDF data** – that goes beyond simple storage and retrieval
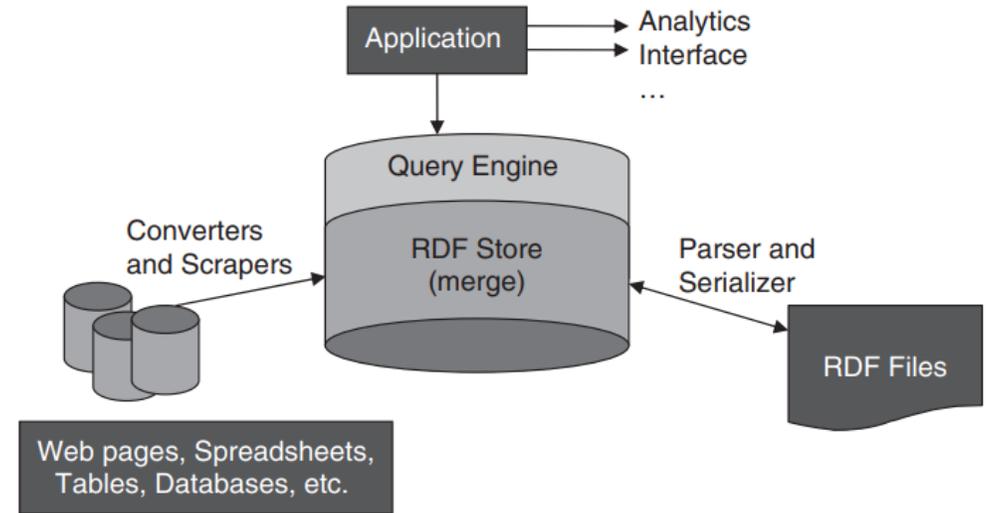

FIGURE 4.2

Application architecture for an RDF application.

CONNECT-NM KOS

The **NM-KOS architecture** is designed as a **two-layered system**, whose components relies upon well-established semantic technologies and industry ready software solutions:
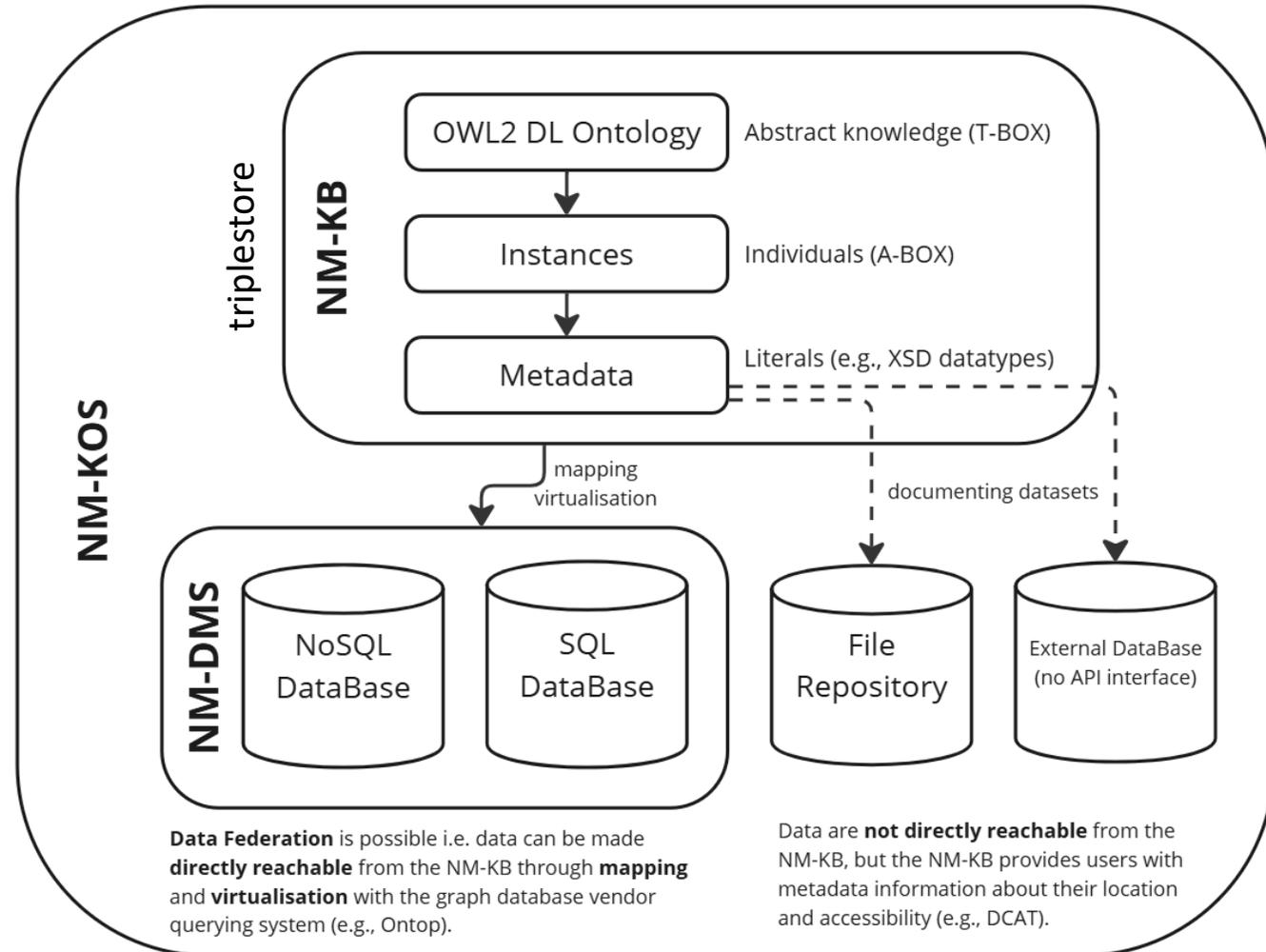
1. **Nuclear Materials Data Management System (NM-DMS)**, a set of distributed databases of datasets collections, storing the actual data (e.g., mechanical properties, microstructures), generated by the project activities or imported from external pre-existing sources;
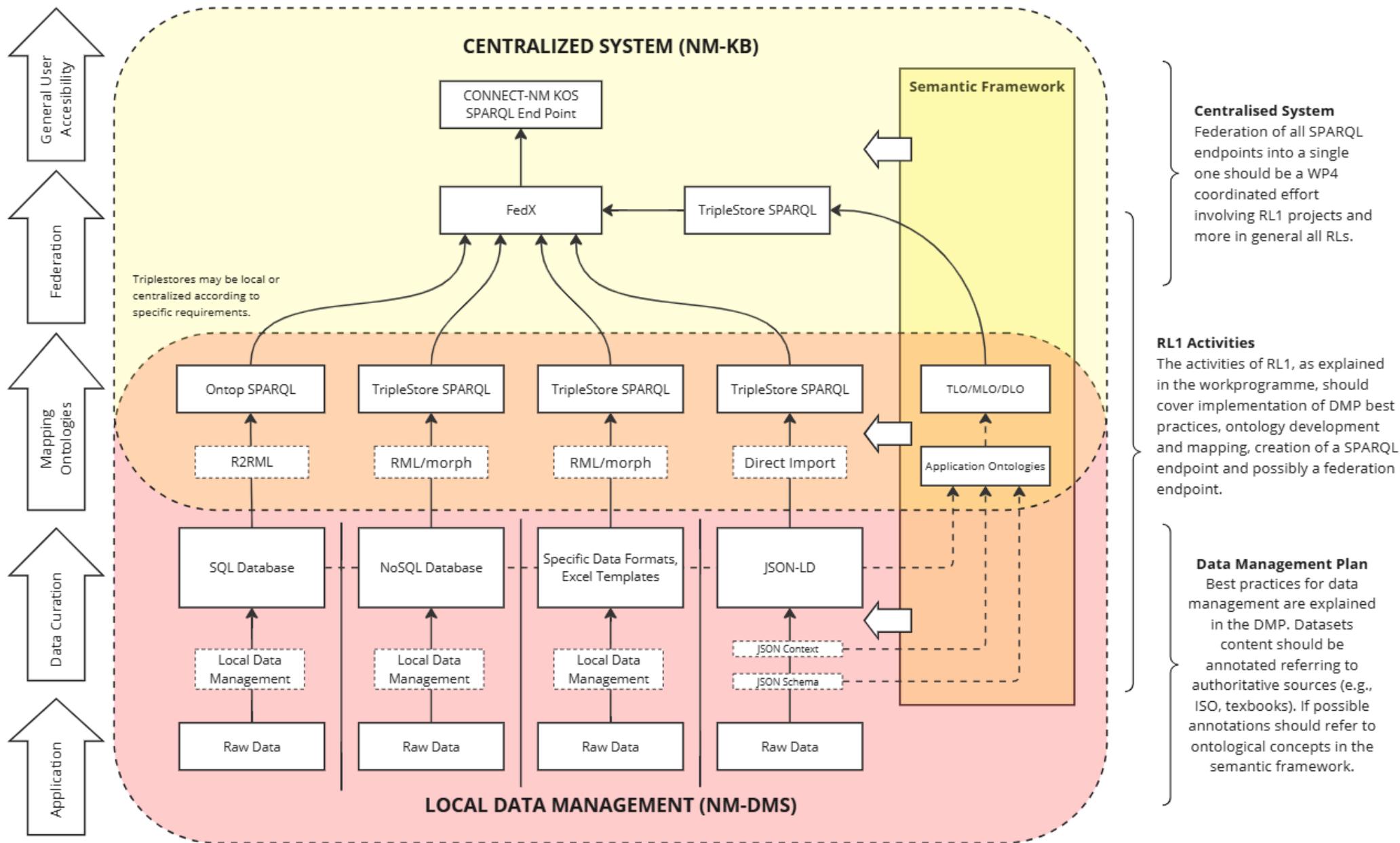
2. **Nuclear Materials Knowledge Base (NM-KB)**, a graph database (a.k.a. triplestore) that will host:
   a. the **abstract knowledge** (i.e., ontology framework);
   b. the information about **domain entities** (e.g., device, material sample, testing procedure, modelling workflow);
   c. the **metadata** needed to describe them (e.g., name, titles, unique identifiers);
   d. the **mapping** between KB entities and actual DMS data or other external data sources.

NM-KOS

triplestore

NM-KB

OWL2 DL Ontology	Abstract knowledge (T-BOX)
Instances	Individuals (A-BOX)
Metadata	Literals (e.g., XSD datatypes)

mapping virtualisation

documenting datasets

NM-DMS

NoSQL DataBase SQL DataBase File Repository External DataBase (no API interface)

Data Federation is possible i.e. data can be made **directly reachable** from the NM-KB through **mapping** and **virtualisation** with the graph database vendor querying system (e.g., Ontop).

Data are **not directly reachable** from the NM-KB, but the NM-KB provides users with metadata information about their location and accessibility (e.g., DCAT).

NM-KOS Architectural Schema

CENTRALIZED SYSTEM (NM-KB)

CONNECT-NM KOS SPARQL End Point

Semantic Framework

FedX ← TripleStore SPARQL

Triplestores may be local or centralized according to specific requirements.

Ontop SPARQL | TripleStore SPARQL | TripleStore SPARQL | TripleStore SPARQL | TLO/MLO/DLO

R2RML | RML/morph | RML/morph | Direct Import | Application Ontologies

SQL Database | NoSQL Database | Specific Data Formats, Excel Templates | JSON-LD

Local Data Management | Local Data Management | Local Data Management | JSON Context | JSON Schema

Raw Data | Raw Data | Raw Data | Raw Data

LOCAL DATA MANAGEMENT (NM-DMS)

General User Accesibility

Federation

Mapping Ontologies

Data Curation

Application

Centralised System
Federation of all SPARQL endpoints into a single one should be a WP4 coordinated effort involving RL1 projects and more in general all RLs.

RL1 Activities
The activities of RL1, as explained in the workprogramme, should cover implementation of DMP best practices, ontology development and mapping, creation of a SPARQL endpoint and possibly a federation endpoint.

Data Management Plan
Best practices for data management are explained in the DMP. Datasets content should be annotated referring to authoritative sources (e.g., ISO, texbooks). If possible annotations should refer to ontological concepts in the semantic framework.

NM-KOS Principles and Benefits

Federation

Data federation refers to the process of **aggregating data from different databases**, formats, and locations **without physically merging them into a single repository**.

Its main advantages include:

- **Unified Data Access** – Provides a single virtual view of data across different systems, eliminating silos.
- **Real-Time Data Integration** – Enables dynamic access to up-to-date data without requiring batch processing.
- **Cost Efficiency** – Reduces storage and maintenance costs by avoiding data duplication and centralization.
- **Scalability** – Easily integrates new data sources without major infrastructure changes.
- **Improved Decision-Making** – Allows organizations to analyse diverse data in real time for better insights.
- **Data Security & Compliance** – Maintains data in its original location, helping meet regulatory requirements.

The NM-KOS is designed as a **collection of distributed databases**, without the need for a **centralised data repository**. The Semantic Web enables **data federation** by facilitating seamless **integration**, **retrieval**, and **exchange of data** from diverse and heterogeneous sources.

This enables all players in the field of nuclear materials to **act independently** in terms of local data management, without the need for a unique, large, and costly infrastructure, very difficult to manage and sustain in the long run.

NM-KOS Principles and Benefits

Locality

Data locality means storing and processing data close to its source, enhancing efficiency and reducing costs.

Its key benefits include:

- **Lower Latency** – Reduces delays by minimizing data movement, crucial for real-time applications like IoT and autonomous systems.

- **Improved Performance** – Speeds up processing by keeping data near computing resources, optimizing frameworks like Hadoop and Spark.

- **Reduced Network Costs** – Saves bandwidth and cloud transfer fees by limiting unnecessary data transfers.

- **Better Security & Compliance** – Helps meet regulations (e.g., GDPR) by keeping sensitive data within legal boundaries.

- **Energy Efficiency** – Reduces power consumption in data centres by minimizing data movement.

- **Scalability for Edge Computing** – Supports faster decision-making in IoT, smart cities, and 5G networks.

By keeping data near where it's needed, **data locality enhances speed, security, and cost-effectiveness** in modern computing. The NM-KOS federated architecture enables the achievement of data locality principle, and all its benefits.

NM-KOS Principles and Benefits

Respect

The NM-KOS federated architecture **respects** the maintaining of **each data source's original storage methodology**, enhancing integration and efficiency.

Key benefits include:

- **Data Integrity & Consistency** – Preserves original data structures and data acquisition methods, ensuring accuracy and reducing transformation errors.

- **Compliance & Security** – Maintains data within its native system, helping meet **regulatory** and **security** requirements.

- **Flexibility for Innovation** – Enables the adoption of new storage technologies as needed without overhauling the entire system.

- **Optimized Performance** – Each system continues using its best-suited storage model (e.g., relational, NoSQL, object storage), avoiding performance bottlenecks.

- **Reduced Data Migration Costs** – Eliminates the need for costly and time-consuming data transformation or duplication.

The NM-KOS has no specific no requirements on local data storage technologies except for a) following the best practices required by the **CONNECT-NM DMP**, and b) using standard data storage technologies that can be **mapped to an RDF system and virtualised** (i.e., almost all main databases solutions, including Excel sheets).

NM-KOS Principles and Benefits

Security

The NM-KOS federated architecture enhances **security** by enabling locally controlled data access, compliance, and privacy while maintaining interoperability.

Key security benefits include:

- **Data Sovereignty & Compliance** – Data remains in its original location, helping organizations comply with regulations like **GDPR** and **HIPAA** without centralizing sensitive information.

- **Granular Access Control** – Semantic Web technologies like **OWL** and **RDF** enable **fine-grained access policies**, ensuring users access only relevant data.

- **Reduced Attack Surface** – Since data is not centralized, the risk of a single point of failure is minimized, making it harder for cyberattacks to compromise the entire system.

- **Improved Data Integrity** – Federated queries using **SPARQL** retrieve real-time data without modifying or duplicating it, reducing risks of tampering or inconsistency.

- **Enhanced Encryption & Secure Data Linking** – **Linked Data principles** ensure secure connections between datasets using encrypted URIs, preventing unauthorized modifications.

- **Decentralized Authentication** – **WebID and OAuth** enable secure identity management across federated sources without exposing login credentials.

- **Resilience Against Data Breaches** – Even if one data source is compromised, other sources in the federation remain protected, ensuring **better containment of security threats**.

Ownership I

The NM-KOS federated architecture allows organizations and data providers to maintain **organizations retain full ownership and governance** over their data while enabling seamless integration.

Key ownership benefits include:

- **Data Control & Sovereignty** – Organizations retain full ownership of their data, avoiding the risks of centralization while complying with **data sovereignty laws** (e.g., GDPR).

- **No Data Duplication** – Since data is accessed **on-demand** via **SPARQL queries** rather than being copied, ownership remains with the original source.

- **Decentralized Governance** – Each data provider can define and enforce its **own policies, security rules, and access controls**, ensuring autonomy.

- **Trust & Transparency** – **Linked Data principles** enable verifiable and traceable data sources, ensuring data integrity and authenticity.

- …

Ownership II

- **…**

- **Scalability Without Ownership Loss** – New data sources can be integrated **without requiring transfers of control**, preserving independence.

- **Custom Monetization Models** – Data providers can regulate **who accesses their data and under what conditions**, enabling **controlled sharing and monetization**.

- **Flexible Data Management** – Organizations can adopt **different storage methodologies** without compromising their control, ensuring long-term adaptability.

In the NM-KOS data owners **can map their entire database** and then **select which part can be shared in the KOS and with whom**, using a shell layer approach (e.g., share metadata about the existence datasets but not the actual data).

NM-KOS Principles and Benefits

Semantics I

The NM-KOS semantic framework should be based on the **EMMO ontology**, and follow the guidelines provided by EMMC (European Materials Modelling Council) and EMCC (European Materials Characterization Council). A **common ontology framework** provides a standardized, shared understanding of data across various domains and systems.

Key benefits include:

- **Unified Data Representation** – A common ontology ensures that data from different sources is represented using consistent concepts, reducing ambiguity, and fostering better understanding across systems.

- **Improved Interoperability** – Ontologies enable seamless integration and exchange of data between heterogeneous systems, supporting data sharing and collaboration across platforms.

- **Enhanced Data Discovery** – Semantic annotations and relationships defined in the ontology improve the discoverability and relevance of data, allowing users and systems to easily find and access information.

- **Automated Reasoning & Inference** – A shared ontology enables systems to automatically infer new knowledge or relationships, enhancing decision-making capabilities and operational efficiency.

- **Data Consistency & Accuracy** – Standardized terms and relationships ensure data consistency across diverse systems, reducing errors and conflicts in data interpretation.

- …

Semantics II

- …

- **Scalability & Flexibility** – New concepts, terms, and data sources can be easily incorporated into the ontology framework, allowing systems to scale and adapt as requirements change.

- **Enhanced Data Integration** – By providing a common vocabulary, ontologies facilitate the integration of disparate data sources, making it easier to merge, analyse, and derive insights from diverse datasets.

- **Support for Complex Queries** – Ontologies enable more sophisticated queries based on relationships and classifications, improving the precision and effectiveness of data retrieval.

- **Regulatory Compliance** – A shared ontology can help ensure consistent application of industry standards, regulations, and best practices across systems, aiding compliance efforts.

- **Collaboration & Knowledge Sharing** – A common ontology fosters collaboration across organizations, domains, and industries by providing a mutual understanding of key concepts and relationships.

By utilizing a **common ontology framework**, the NM-KOS can enhance **data integration, collaboration, and knowledge sharing**, while ensuring **consistency, scalability, and automation** across their systems.

https://github.com/emmo-repo/EMMO

NM-KOS Principles and Benefits

Scalability

The abovementioned principles enable the implementation in the NM-KOS of two types of **scalability** principles:

a) **Vertical Scalability:** since data owners face a series of **sequential improvements** of their databases/datasets incrementing step by step the semantic integration, from simple DMP compliancy up to a fully federated knowledge graph, according to their resources and capabilities.

b) **Horizontal scalability**: since the federated architecture greatly facilitates the introduction of **new knowledge sources**, as new **plug-ins** that can be added to the existing ones and limits the impact of potential disappearance of existing resources.

NM-KOS Principles and Benefits

Sustainability

The **sustainability** of the NM-KOS after the end of CONNECT-NM is **greatly facilitated by:**

- **the lack of a centralised framework**, that need resources to be maintained. Instead, a federated system allows for distributed data governance and ownership, which prevents the need for large, centralized infrastructures, **reducing the maintenance costs.**

- the usage of **existing well-established standards** (e.g., RDF) and **widely available commercial and free software tools** (e.g., triplestores) thus eliminating the need for software development and ensuring the availability of technical personnel that is already on the market.

Objective of RL1 calls is to **demonstrate the multi-level benefits of such semantic approach to data interoperability**, that should motivate users to spend some resources in data curation and semantic enhancement that will be beneficial to the overall NM-KOS.

Local Data Management
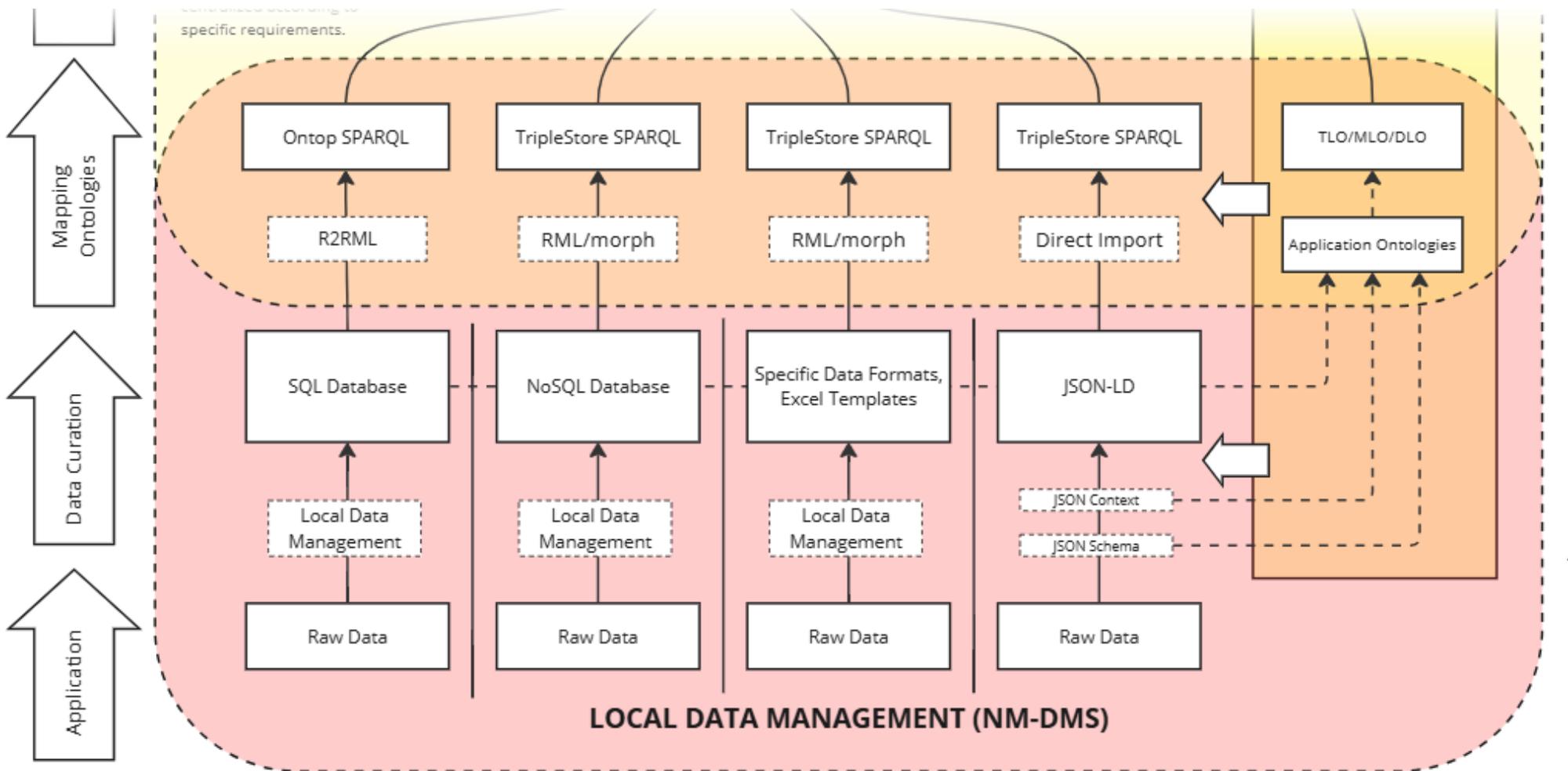
RL1 Activities
The activities of RL1, as explained in the workprogramme, should cover implementation of DMP best practices, ontology development and mapping, creation of a SPARQL endpoint and possibly a federation endpoint.

Data Management Plan
Best practices for data management are explained in the DMP. Datasets content should be annotated referring to authoritative sources (e.g., ISO, texbooks). If possible annotations should refer to ontological concepts in the semantic framework.

Full Integration

The **full integration with the NM-KOS**, giving the possibility to end users to gather specific data, up to the selection of a **single specific datum within a dataset**, can be achieved only in these two cases:

a) Local data are stored directly in a **local triple store** in the form of **linked data**, exposing a SPARQL endpoint directly connect to it. This requires the usage of **linked data formats** (e.g., JSON-LD) at the level of local data management, or **mapping of other tabular formats** (e.g., excel, CVS) and the usage of EMMO based application ontologies for their documentation.

b) Local data are stored in a **local database solution** that can be **virtualised** and connected to a SPARQL endpoint. This requires that the local database technology is **compatible with at least one RDF virtualisation platform**. This is usually not an issue since most of the existing database technologies have been virtualised by several free and commercial software providers.

As example, the Ontop 5.1.0 platform for RDF virtualisation of databases supports the following database systems: PostgreSQL, MySQL, MariaDB, SQL Server, Oracle, DB2, Snowflake, Databricks, Google BigQuery, AWS Redshift and DuckDB; and database federators such as Denodo, Dremio, Teiid, Apache Spark and Trino/PrestoDB/AWS Athena.

Partial Integration

If virtualisation is not possible, by choice or by technological limitations, there will be **a partial integration with NM-KOS**, limited to simple dataset metadata documentation.

This can be done using **DCAT (Data Catalog Vocabulary),** which is a **W3C standard** designed to facilitate the **discovery, sharing, and interoperability** of **datasets** across organizations and platforms, especially in **open data portals and data federations**. However, it does not provide the access to the specific data entry that a fully virtualised solution would grant.

https://www.w3.org/TR/vocab-dcat-3/

Full vs Partial Integration
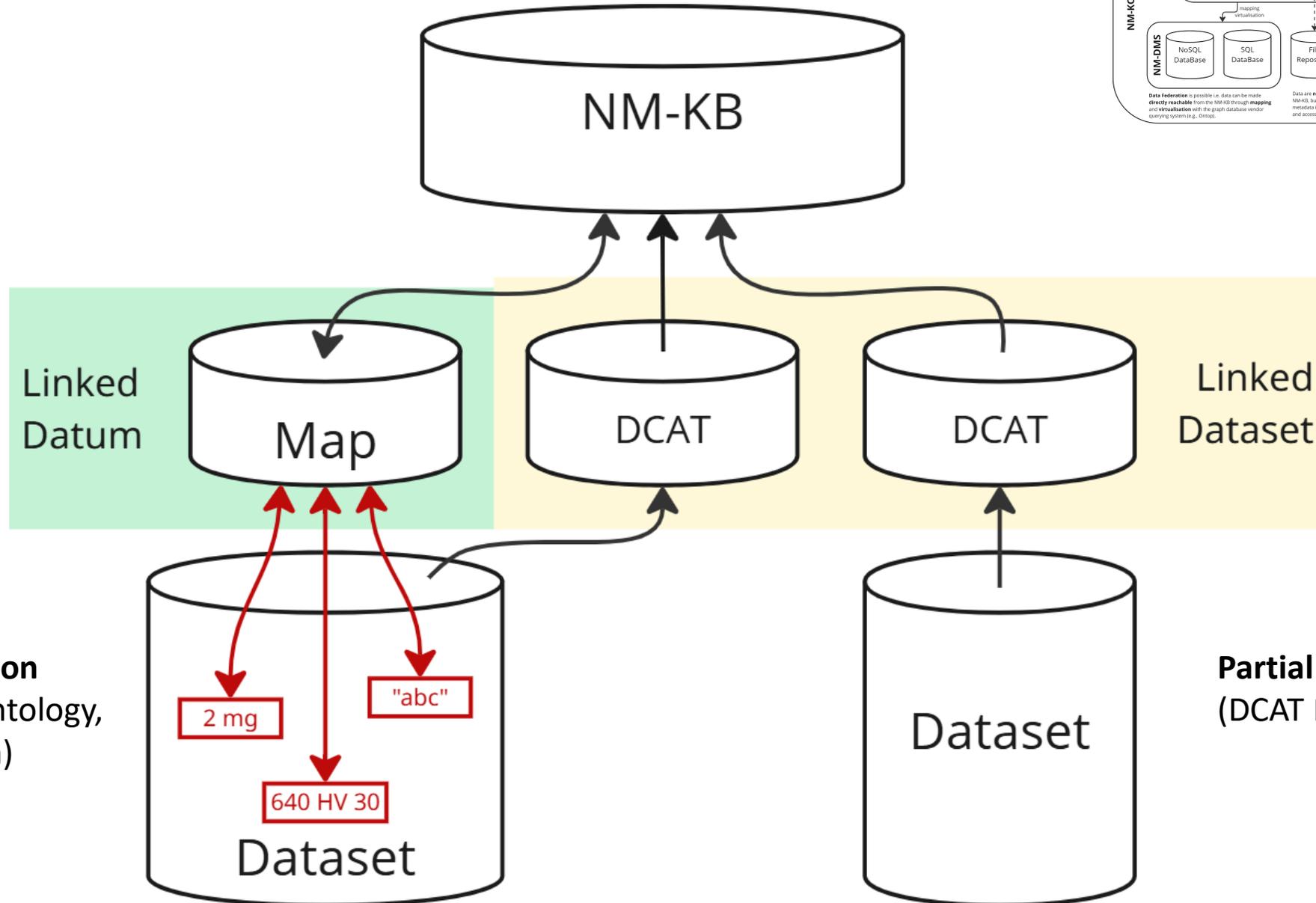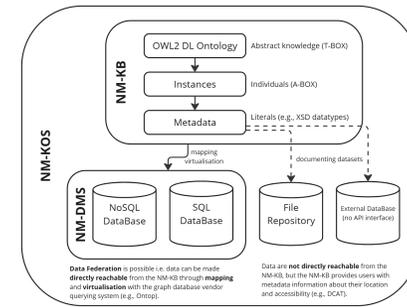
Full Integration
(mapping, ontology, virtualisation)

Partial Integration
(DCAT Metadata)

see
data.europa.eu

Metadata Example
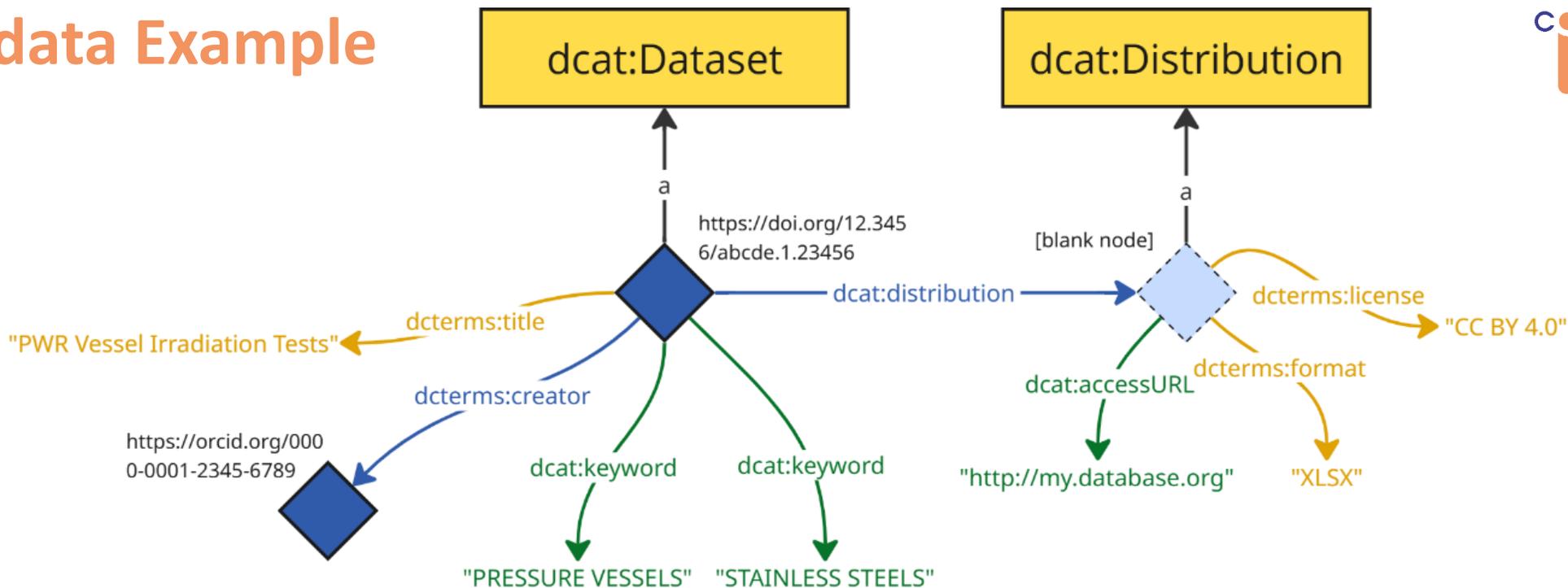

```json
{
    "@context": {
        "dcat": "http://www.w3.org/ns/dcat#",
        "dcterms": "http://purl.org/dc/terms/"
    },

    "@id": "https://doi.org/12.3456/abcde.1.23456",
    "@type": "dcat:Dataset",
    "dcterms:creator": {
        "@id": "https://orcid.org/0000-0001-2345-6789"
    },
    "dcat:keywords": ["PRESSURE VESSELS", "STAINLES STEELS"],
    "dcterms:title": "PWR Vessel Irradiation Tests",
    "dcat:distribution": {
        "@type": "dcat:Distribution",
        "dcat:accessURL": "http://my.database.org",
        "dcterms:license": "CC BY 4.0",
        "dcterms:format": "XLSX"
    }
}
```

Subject	Predicate	Object
_:b0	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/ns/dcat#Distribution
_:b0	http://purl.org/dc/terms/format	XLSX
_:b0	http://purl.org/dc/terms/license	CC BY 4.0
_:b0	http://www.w3.org/ns/dcat#accessURL	http://my.database.org
https://doi.org/12.3456/abcde.1.23456	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/ns/dcat#Dataset
https://doi.org/12.3456/abcde.1.23456	http://purl.org/dc/terms/creator	https://orcid.org/0000-0001-2345-6789
https://doi.org/12.3456/abcde.1.23456	http://purl.org/dc/terms/title	PWR Vessel Irradiation Tests
https://doi.org/12.3456/abcde.1.23456	http://www.w3.org/ns/dcat#distribution	_:b0
https://doi.org/12.3456/abcde.1.23456	http://www.w3.org/ns/dcat#keywords	PRESSURE VESSELS
https://doi.org/12.3456/abcde.1.23456	http://www.w3.org/ns/dcat#keywords	STAINLES STEELS

Software

There exists several triplestore implementations that are readily available for deployment.

The most relevant **commercial triplestores** are:
- **GraphDB**
- **Stardog**
- **AllegroGraph**
- **Oracle RDF Graph**
- **MarkLogic**

The most relevant **Open-Source triplestores** are:
- **Apache Jena TDB/Fuseki**
- **Virtuoso**
- **Blazegraph**
- **RDF4J**
- **GraphDB Free Edition**

Due to the easy-to-use web interface, and the fully functional free version (that may be converted to commercial) **GraphDB Free** may be the preferred entry-level choice for demonstration purposes.

Mapping

Declarative mapping between relational databases and RDF is a method that enables the transformation of structured data stored in relational databases (RDBs), or more in general as tabular data, into RDF (Resource Description Framework) format without physically moving or duplicating the data. It allows seamless integration of relational data into the Semantic Web while maintaining interoperability with existing database systems.

This approach enables **non-intrusive transformation** of data, since the original database structure remains unchanged, and data is **queried on demand** rather than being migrated, and **semantic Interoperability**, since data is represented in **RDF triples**, making it machine-readable and enabling linkage with external knowledge graphs.
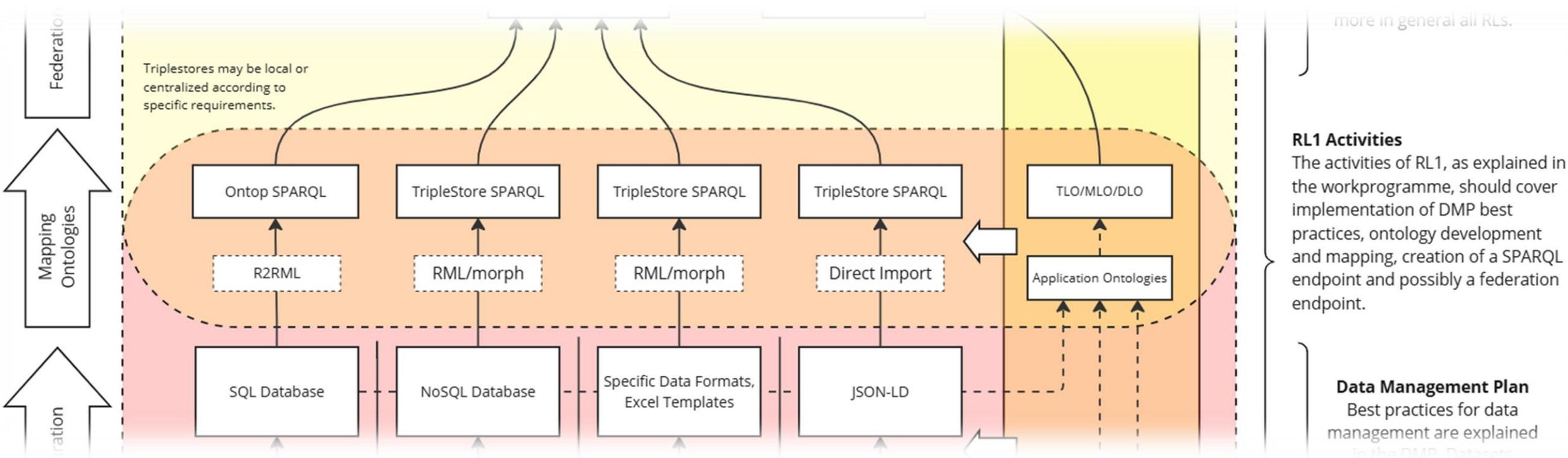
Mapping

Mapping can be implemented through different methodologies, such as:

- **R2RML (RDB to RDF Mapping Language)** – A W3C standard that defines how relational data should be mapped to RDF using **SQL-based mappings**.

- **Direct Mapping** – An automatic approach that converts relational schemas into RDF triples **without requiring custom mappings**.

- **OBDA (Ontology-Based Data Access)** – Uses ontologies to provide a **semantic layer** over relational databases, often implemented with **R2RML or SPARQL-to-SQL translation**.

- **OntoRefine** – the Ontotext GraphDB tool based on OpenRefine that allows fast cleaning, mapping, and transformation of any structured data (including CSV and Excel files) to RDF.

The final goal of mapping is to deploy and **expose a local SPARQL endpoint** to provide an RDF version for the local data, thus enabling federation with other local SPARQL endpoints.
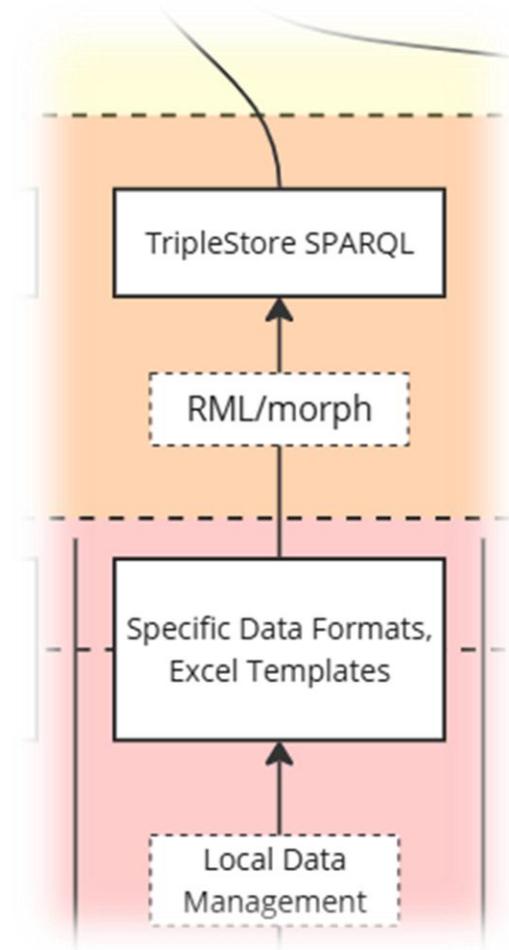
Mapping

Such activity requires **dedicated competencies in semantic technologies**, and a good understanding of **ontologies**, since the transformation of data into RDF format requires an underlying data model that is usually provided by the **development of a dedicated application ontology**.

Mapping is one of **the key activities** in **CONNECT-NM RL1 projects**, since it requires a direct investment of resources towards data semantical enhancement, that is usually one of the barriers to overcome in this field.

It is then expected that most of these mapping activities of the population of NM-KOS will be covered by RL1 projects.
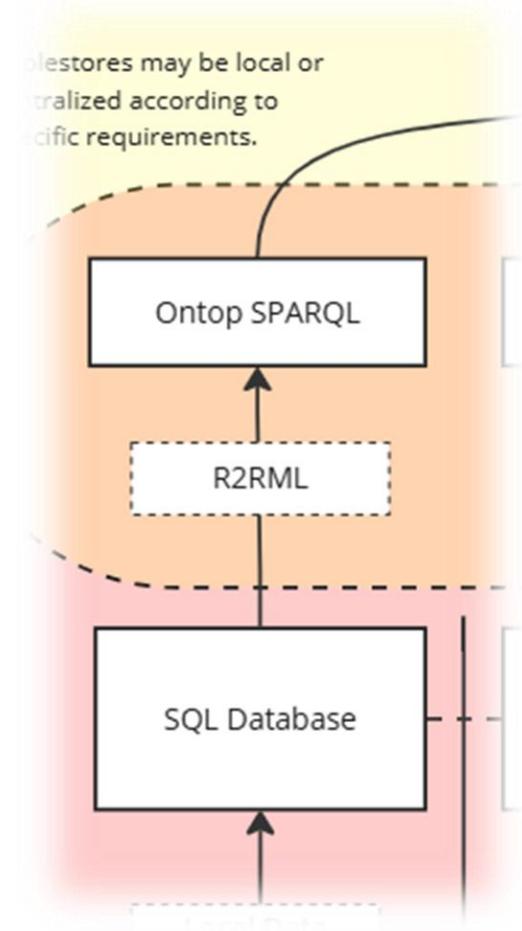
Federation

SPARQL Federation is a mechanism in the **Semantic Web** that enables querying **distributed RDF data sources** across multiple SPARQL endpoints. It allows users to retrieve and integrate data from different locations **without centralizing** it, making it a key technique for **federated knowledge bases**
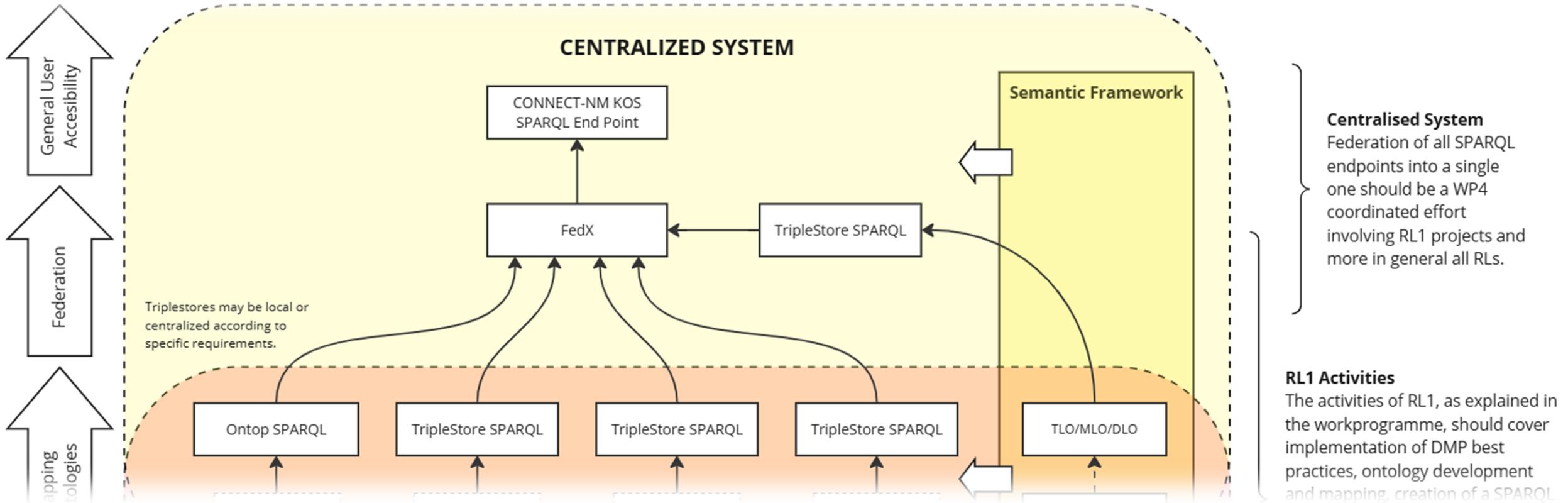
NM-KOS Architectural Schema

The NM-KOS architecture is **vertically scalable** for users.

It enables users to go through a **progressively enhanced documentation** and interoperability of their data.

Generic CONNECT-NM users may decide which level of detail and investment on data they want to achieve, with **minimum requirement defined by the DMP**, and **RDF semantic mapping achievable by RL1 funded project**

NM-KOS Architectural Schema

The NM-KOS semantic framework used for mapping and potentially for the documentation of datasets act at different levels of the architecture:

- at **data level** it provides references for the documentation of the datasets, where users may refer to specific Internationalized Reference Identifiers (IRI) of ontological entities (e.g., materials, processes, datatypes) to semantically enhance a data entry;

- at **RL1 level** for mapping databases to RDF, since RDF virtualisation and interoperability must refer to a specific set of types (i.e., classes) and relations (i.e., RDF properties) that are usually provided by an ontology;

- at **global level** as data/conceptual model for the construction of SPARQL queries to retrieve data-based knowledge through a common federated unique endpoint.

Semantic Framework

TLO/MLO/DLO

Application Ontologies

Centralised System
Federation of all SPARQL endpoints into a single one should be a WP4 coordinated effort involving RL1 projects and more in general all RLs.

RL1 Activities
The activities of RL1, as explained in the workprogramme, should cover implementation of DMP best practices, ontology development and mapping, creation of a SPARQL endpoint and possibly a federation endpoint.

Data Management Plan
Best practices for data management are explained in the DMP. Datasets content should be annotated referring to authoritative sources (e.g., ISO, texbooks). If possible annotations should refer to ontological concepts in the semantic framework.

<u>Semantic</u> Technology

From Ancient Greek *σημαντικός* (sēmantikós), meaning "relating to signs", derived from *σῆμᾰ* (sêma), meaning "mark".

Related to *σημαίνω* (sēmaínō), meaning "to indicate" and *σημεῖον* (semeion), the noun for "sign".

DATA → INFORMATION → INTERPRETATION → THINGS
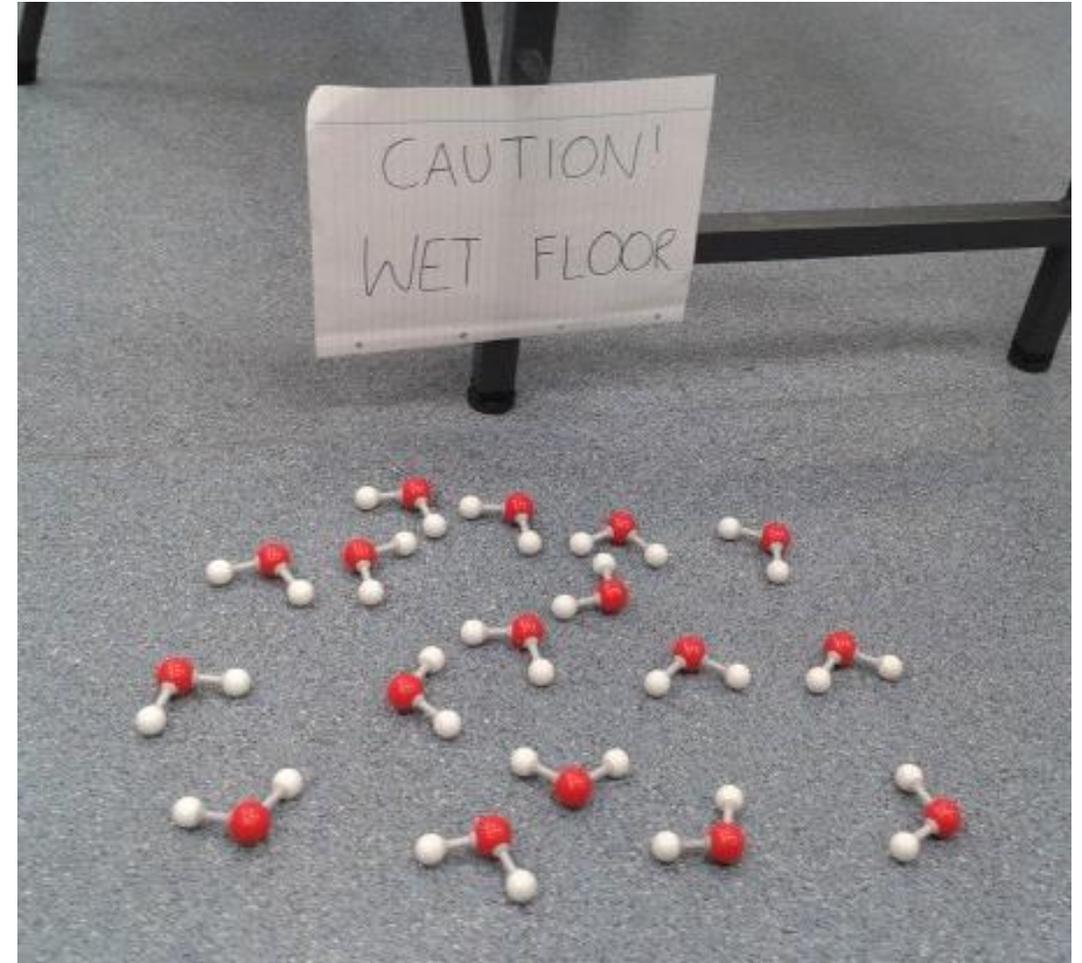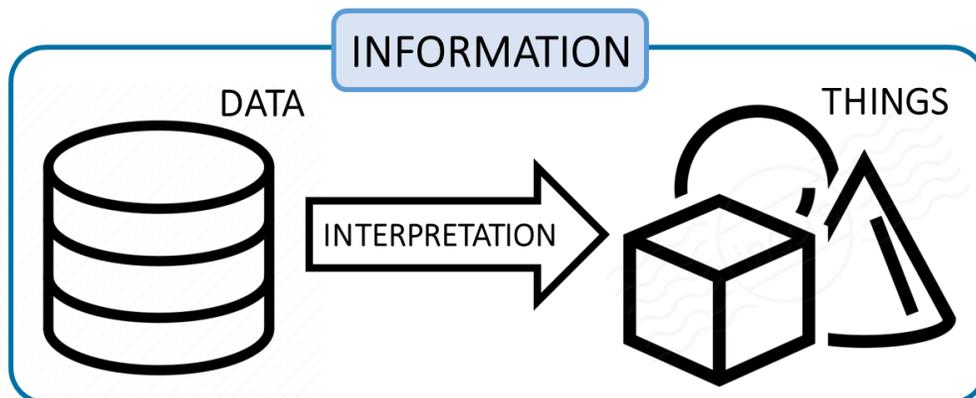
Be sure to clearly distinguish between the sign and the thing for which it stands...

Conceptualisation and Subjectivity
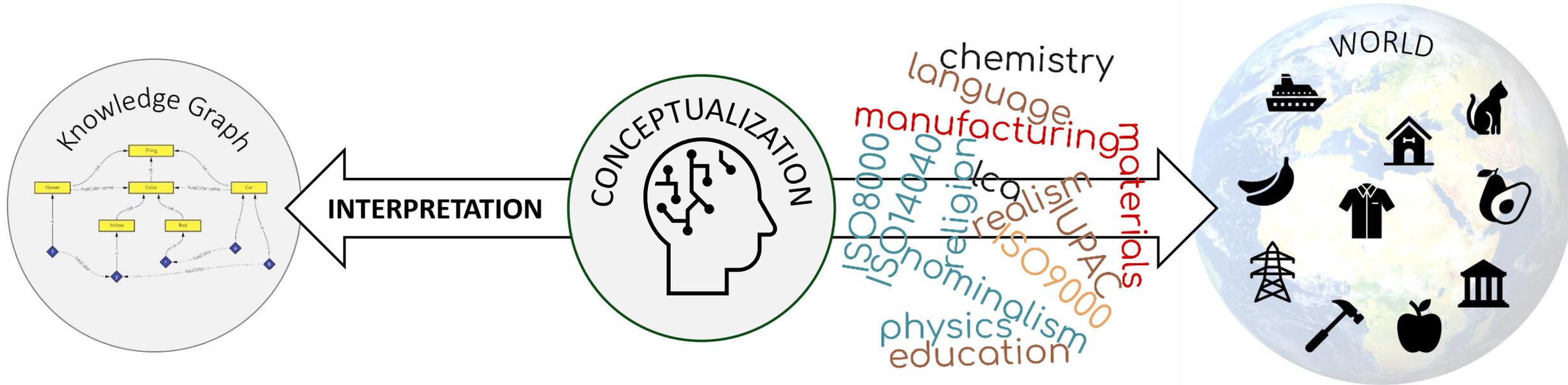
Several conceptualizations exist for **the same things**, so that almost each human being is going to provide a different interpretation for a single term.

Non-Univocal Scientific Conceptualisations

IUPAC Goldbook Definitions (Chemistry)
An **Atom** is a nucleus of Z positive charge and Z electrons
A **Molecule** is an electrically neutral entity consisting of more than one atom

Chemical-Physics Definitions
A **Molecule** is an electrically neutral entity consisting of nuclei and electrons.
An **Atom** is a singular type of molecule with one nucleus only.
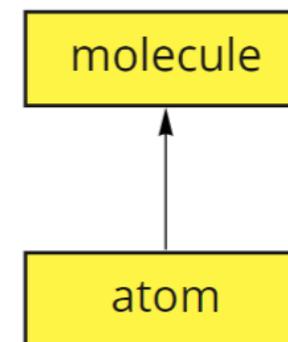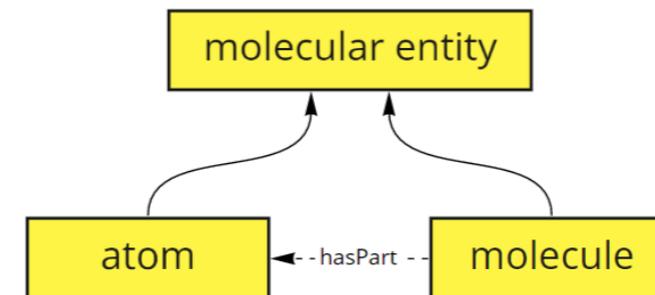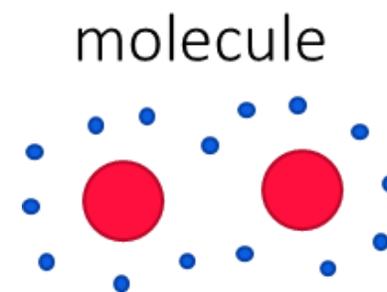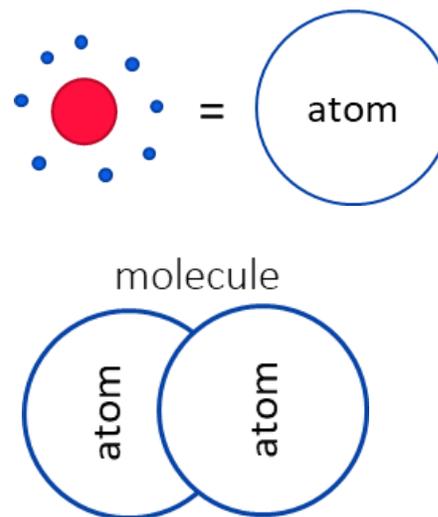
Non-Univocal Scientific Conceptualisations

ISO 9000
Quality management systems
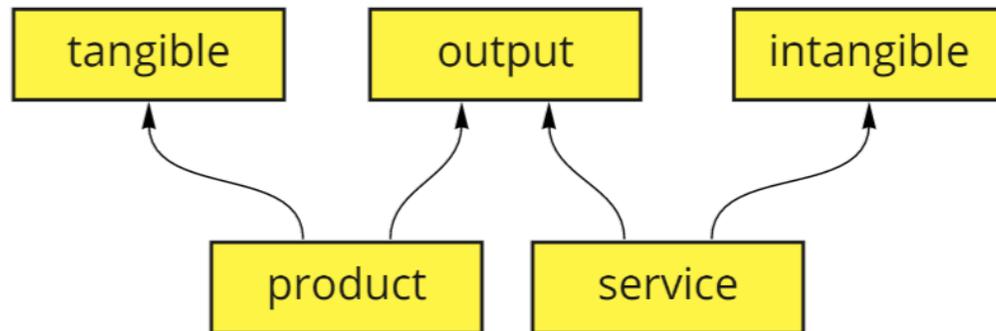
ISO 14040
Environmental management — Life cycle assessment
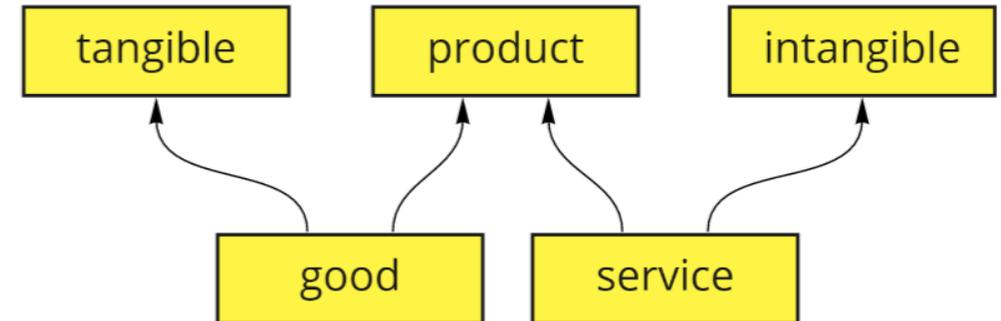
It seems that the definitions of product in ISO 9000 and ISO 14040 are somewhat incompatible, and for an ontology embracing one ISO means that the other ISO will be no more representable.

OWL 2 DL Ontology Example

Ontology Entities

class   relation   individual

Example of axioms:
```
ClassAssertion( :a :Car)
ClassAssertion( :r :Color)
ObjectPropertyAssertion( :hasColor :a :r )
```

Real-world objects
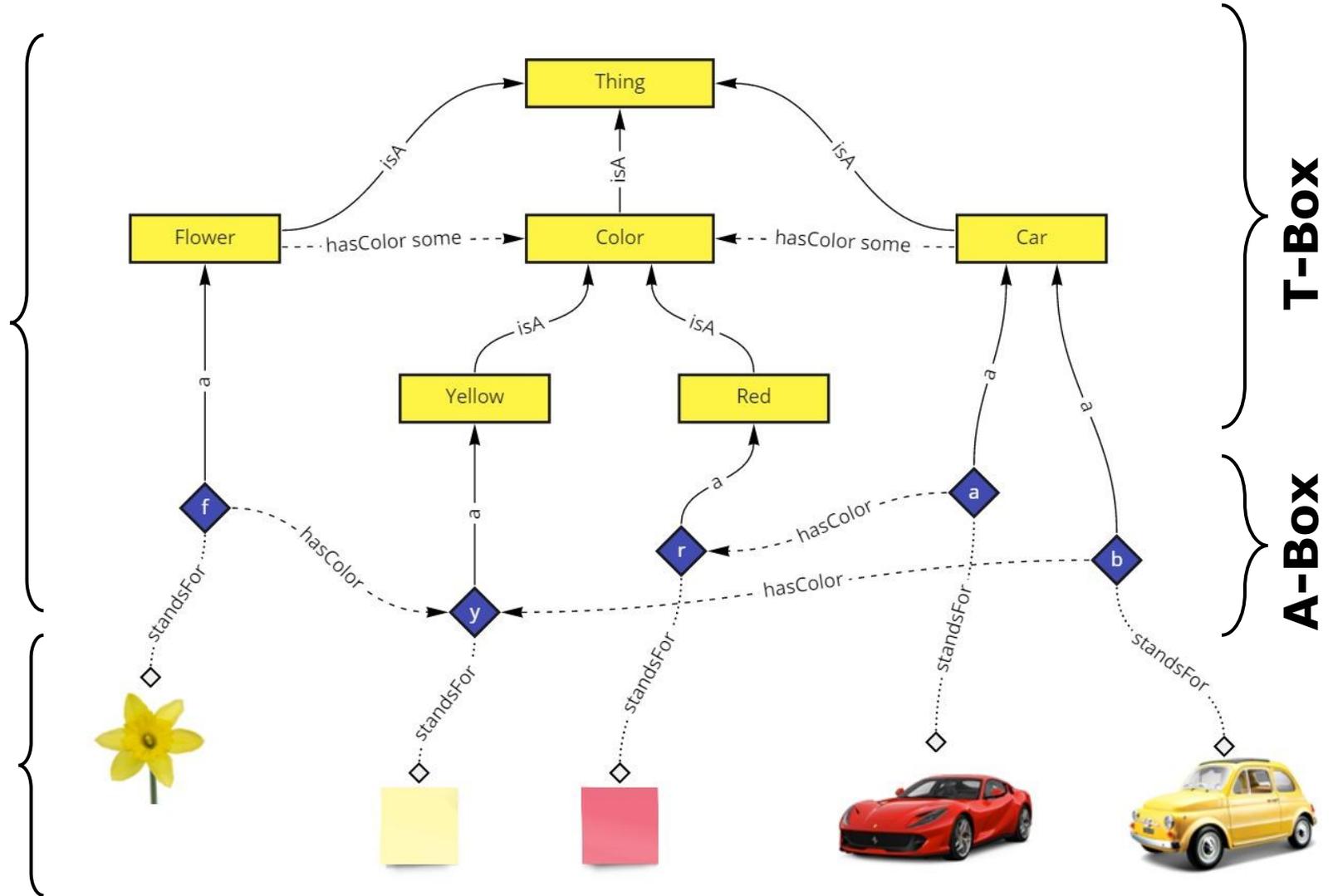
OWL 2 Ontology in Flesh & Bones

An OWL 2 Ontology is formally expressed in a persistent form by axioms declarations following a **specific syntax** (e.g. ASCII file with Turtle syntax)

https://www.w3.org/TR/owl2-syntax/#Introduction
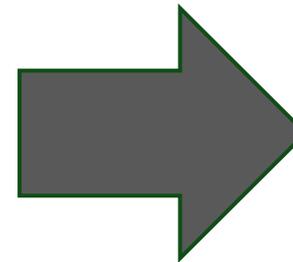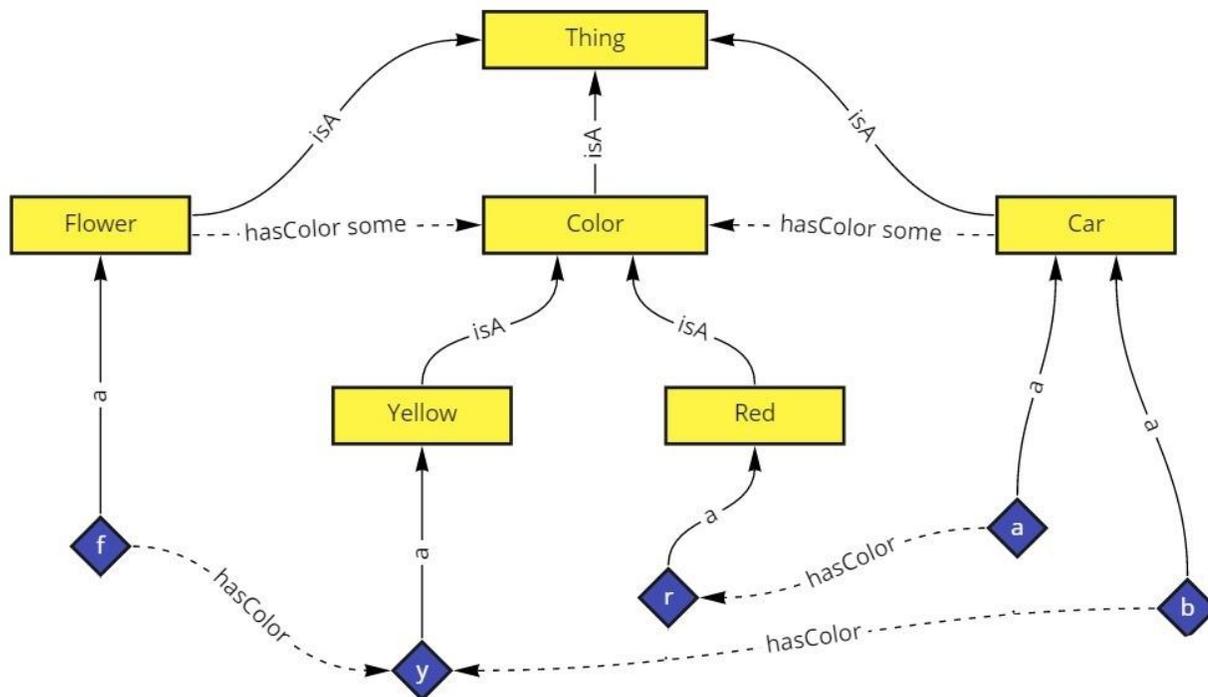
```
@prefix : <http://www.semanticweb.org/emanuele/ontologies/example#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@base <http://www.semanticweb.org/emanuele/ontologies/example> .

<http://www.semanticweb.org/emanuele/ontologies/example> rdf:type owl:Ontology .

###  http://www.semanticweb.org/emanuele/ontologies/example#hasColor
:hasColor rdf:type owl:ObjectProperty ;
          rdfs:subPropertyOf owl:topObjectProperty ;
          rdfs:range :Color .

###  http://www.semanticweb.org/emanuele/ontologies/example#Car
:Car rdf:type owl:Class ;
     rdfs:subClassOf [ rdf:type owl:Restriction ;
                       owl:onProperty :hasColor ;
                       owl:someValuesFrom :Color
                     ] .

###  http://www.semanticweb.org/emanuele/ontologies/example#Color
:Color rdf:type owl:Class .

###  http://www.semanticweb.org/emanuele/ontologies/example#Flower
:Flower rdf:type owl:Class ;
        rdfs:subClassOf [ rdf:type owl:Restriction ;
                          owl:onProperty :hasColor ;
                          owl:someValuesFrom :Color
                        ] .

###  http://www.semanticweb.org/emanuele/ontologies/example#Red
:Red rdf:type owl:Class ;
     rdfs:subClassOf :Color .

###  http://www.semanticweb.org/emanuele/ontologies/example#Yellow
:Yellow rdf:type owl:Class ;
        rdfs:subClassOf :Color .

###  http://www.semanticweb.org/emanuele/ontologies/example#a
:a rdf:type owl:NamedIndividual ,
            :Car ;
   :hasColor :r .

###  http://www.semanticweb.org/emanuele/ontologies/example#b
:b rdf:type owl:NamedIndividual ,
            :Car ;
   :hasColor :y .

###  http://www.semanticweb.org/emanuele/ontologies/example#f
:f rdf:type owl:NamedIndividual ,
            :Flower ;
   :hasColor :y .

###  http://www.semanticweb.org/emanuele/ontologies/example#r
:r rdf:type owl:NamedIndividual ,
            :Color .

###  http://www.semanticweb.org/emanuele/ontologies/example#y
:y rdf:type owl:NamedIndividual ,
            :Color .
```

The Need for a Materials Ontology

In 2018 several European practitioners in Materials Science under the **governance** of the **European Materials Modelling Council** (https://emmc.eu/) expressed the need to develop a **knowledge framework** consistent with **scientific principles and methodologies** to complement the **existing physical-mathematical approach.**

The **Elementary Multiperspective Material Ontology (EMMO)** is an ontology that has been developed as foundation for such knowledge framework.

From Elementaries to Materials
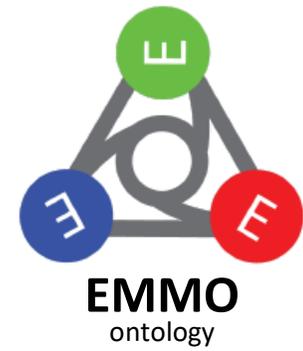
EMMO 1.0.0 achieves complete multiscale representational capabilities, embracing and relating all entities from elementary particles to continuum, including physical models (e.g. QFT, Schrodinger, classical electrodynamics)
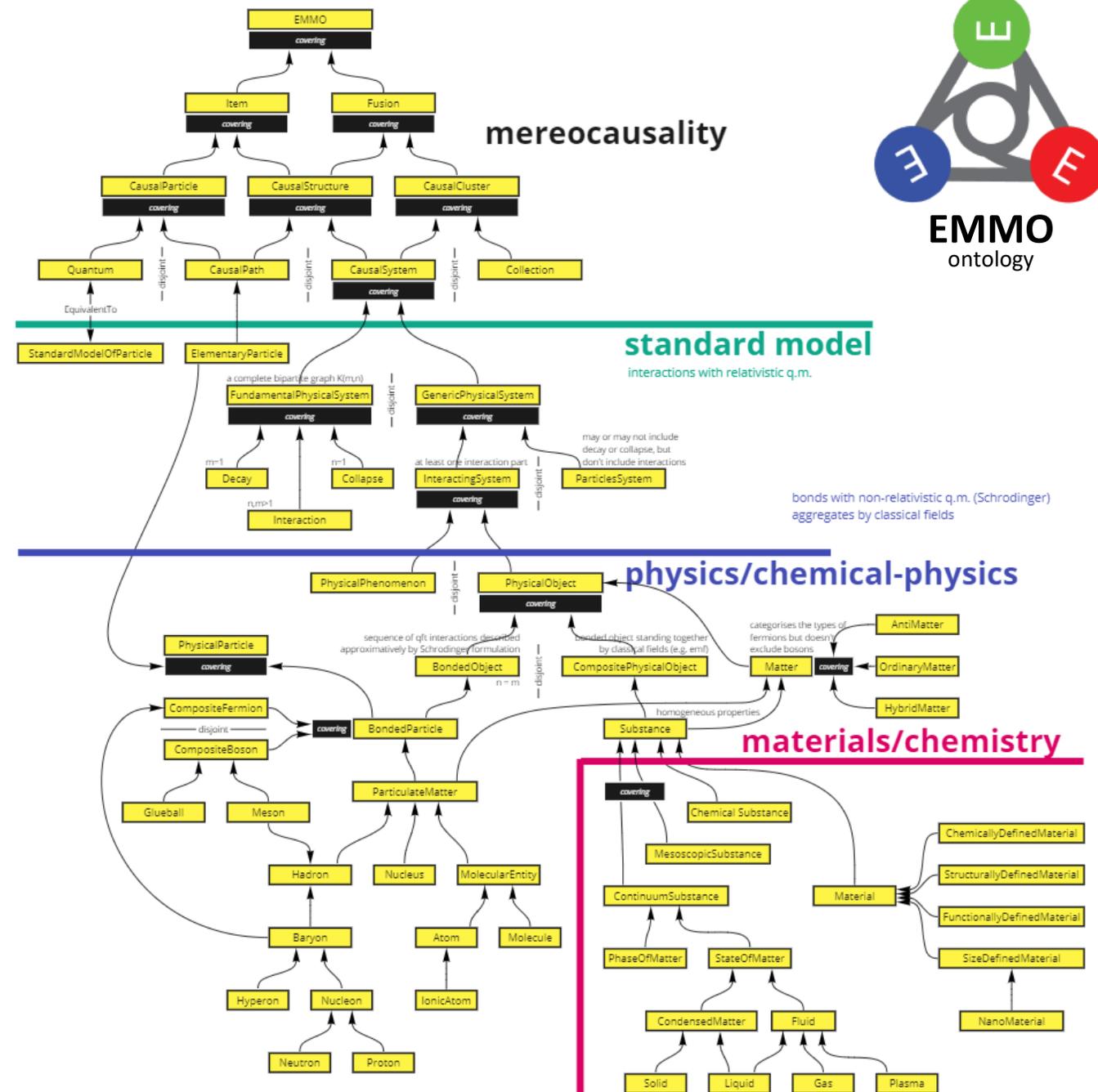
Life is not Data but State of Things!

BattINFO
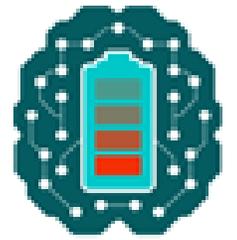
Battery Ontology
- Based on **EMMO**
- Representing batteries cells, materials, components, methods and properties
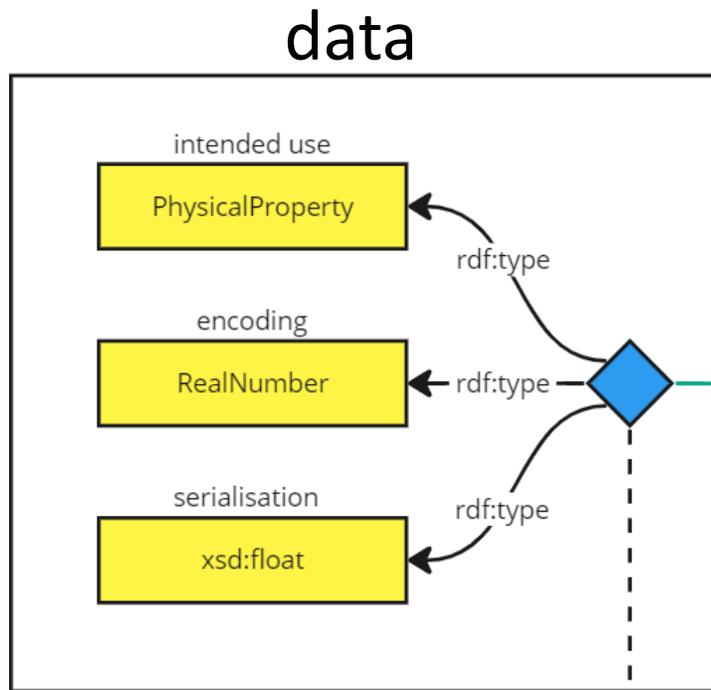
https://big-map.github.io/BattINFO/
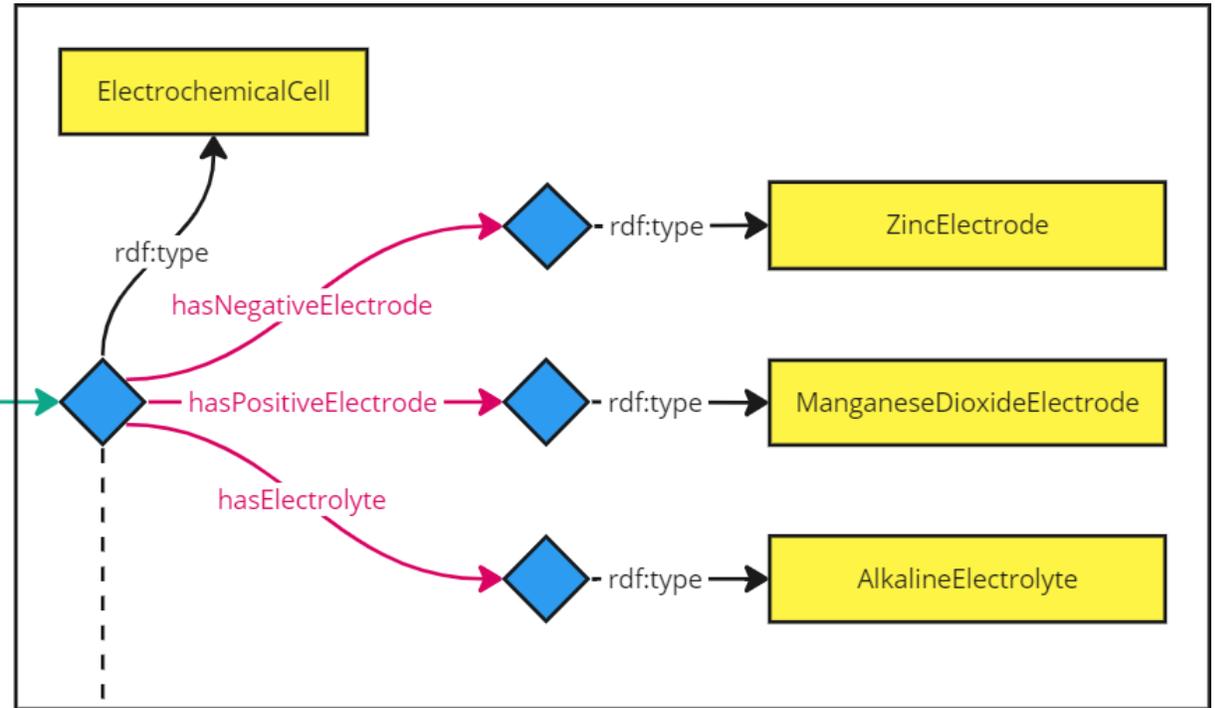
Life is not Data but State of Things!
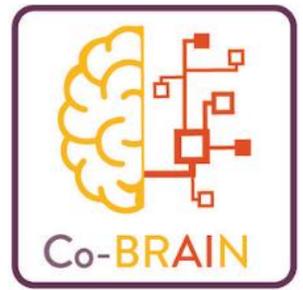
4D mereocausal representation of a **Thermal Spraying process** in **CoBRAIN** project connecting all the relevant entities using mereocausality relations, to document the **overall state of things**.
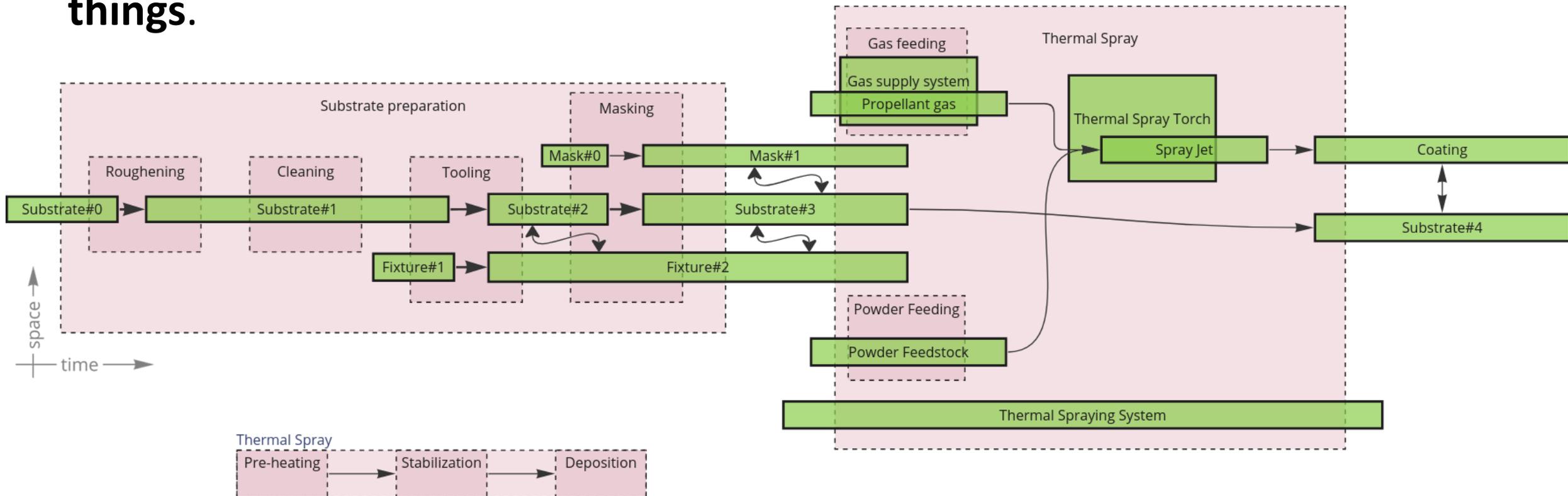
Life is not Data but State of Things!
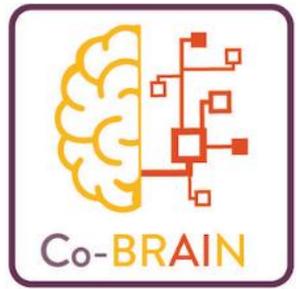
In **CoBRAIN** we document each sample coming from thermal spraying run according to **characterisation results** (e.g., microindentation, abrasion, SEM, RAMAN) or **modelling** (e.g., CFD of thermal spraying, splat formation, miscrustructure, DFT).
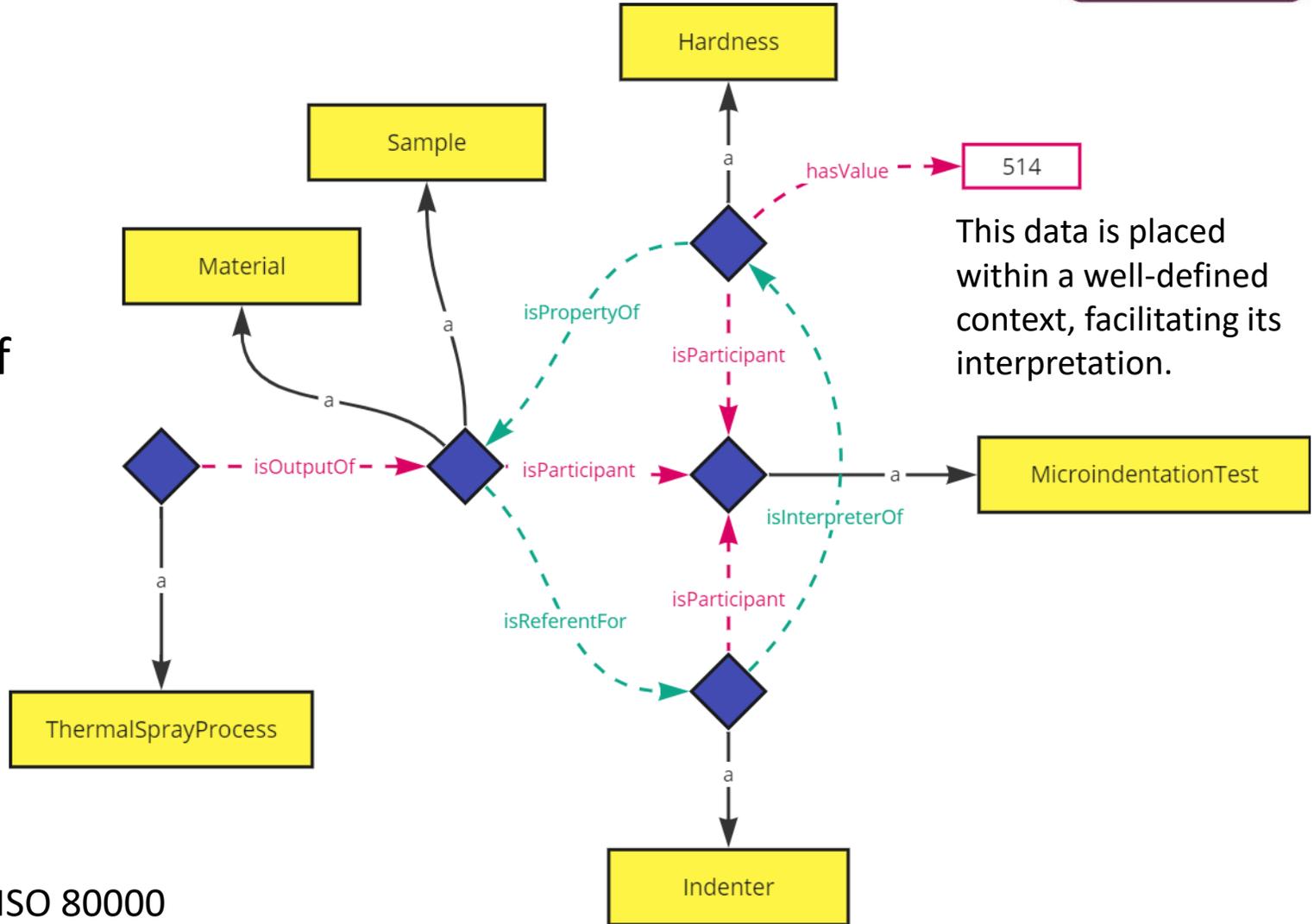
This data is placed within a well-defined context, facilitating its interpretation.

Vocabulary of **Metrology**, **SI Unit** System, **ISQ** ISO 80000
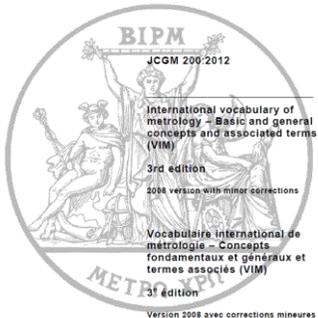
Ontologising Fukushima Accident

The Fukushima Daiichi Accident

Technical Volume 1/5
Description and Context of the Accident

The NEO ontology has been used to represent the knowledge expressed in the **IAEA Fukushima Daichi Accident** official report in the form of knowledge graph.

with Ilaria Maria Paponetti (UNIBO)

Ontologising Fukushima Accident

The NEO ontology has been used to represent the knowledge expressed in the **IAEA Fukushima Daichi Accident** official report in the form of knowledge graph.

with Ilaria Maria Paponetti (UNIBO)
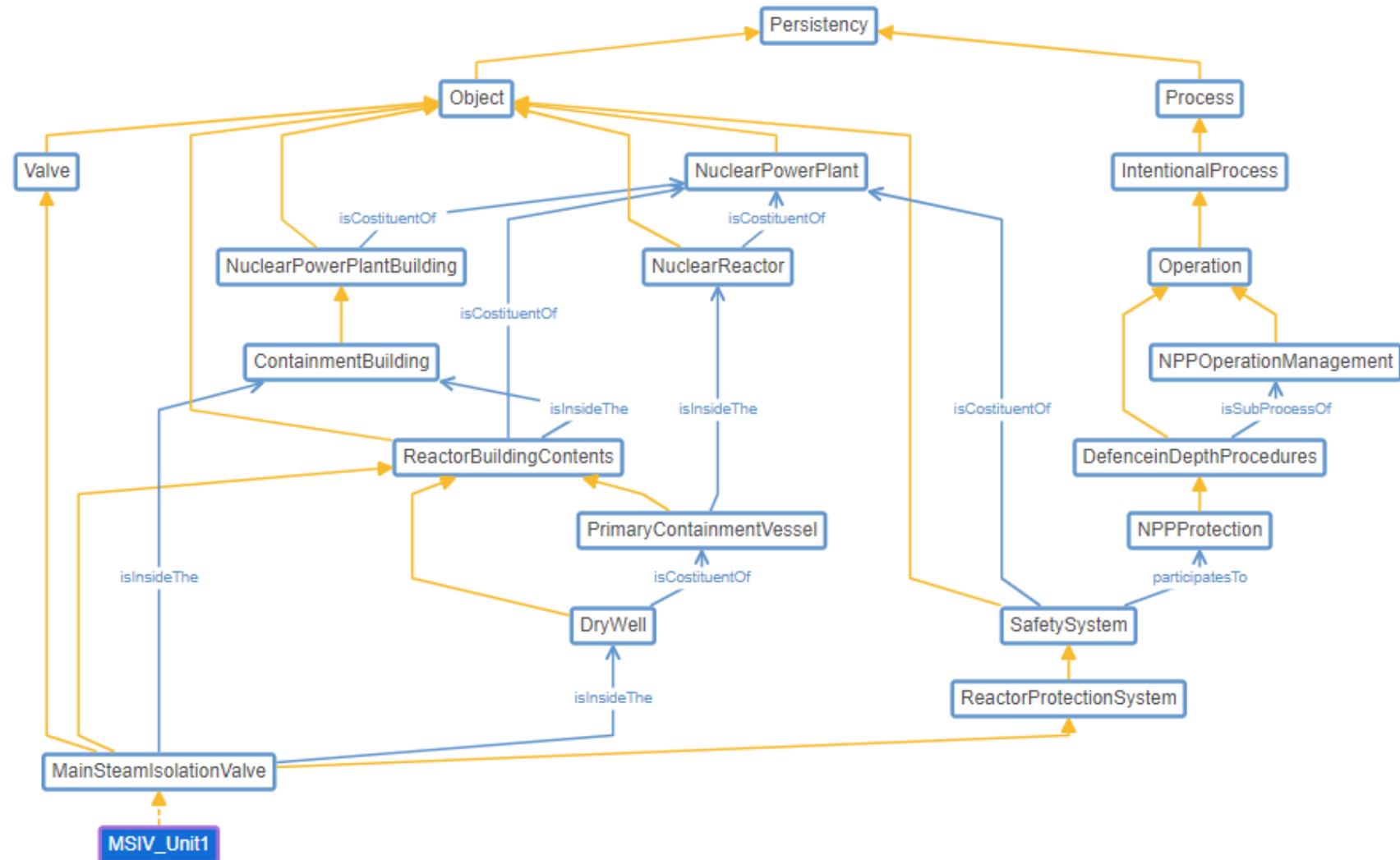
Keywords are not Enough!!!

- The assumption that a simple **keywording of the data** is sufficient to provide some level of **FAIR-ness** is simply wrong!!!

- **Ontologies** provide a **formal context** to facilitate the interpretation of the data by the users (e.g. human beings, AI tools)

- Some ontologies are **better than others** in representing knowledge of a specific domain

- The **EMMO** has been designed for **applied sciences**, characterization, modelling, and supported by EMMC and EMCC.
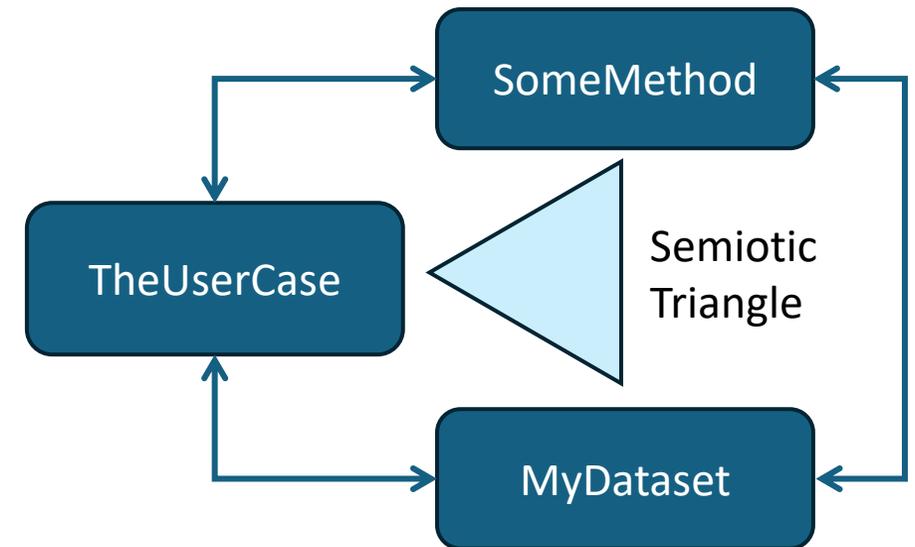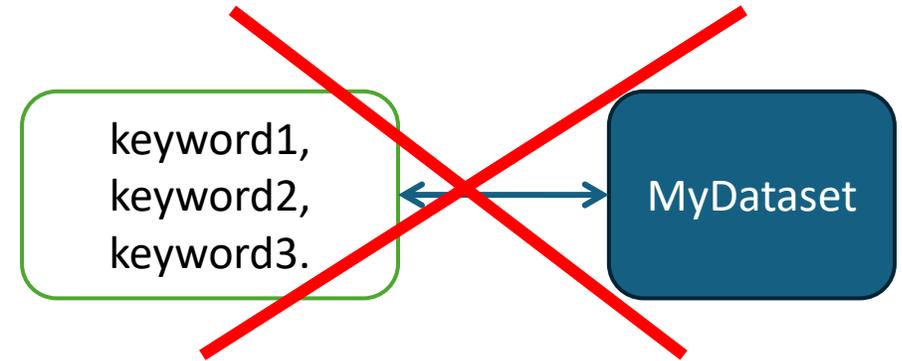
The Path to (Ontological) Wisdom

Data as **syntactic** structures (formats)

Data as **signs** for real entities (keywords, taxonomy)

Data relations mimic real entities relations, a.k.a. **state of things** (ontology)

Discovering distant data relations thanks to ontological representation of related states of things

Explaining distant data relations thanks to ontological representation of related states of things
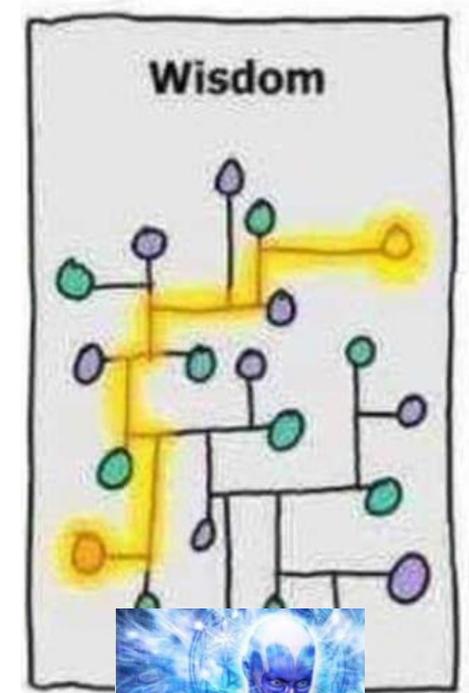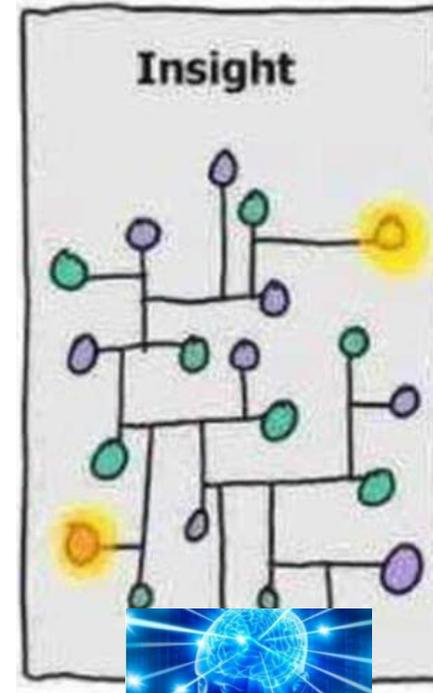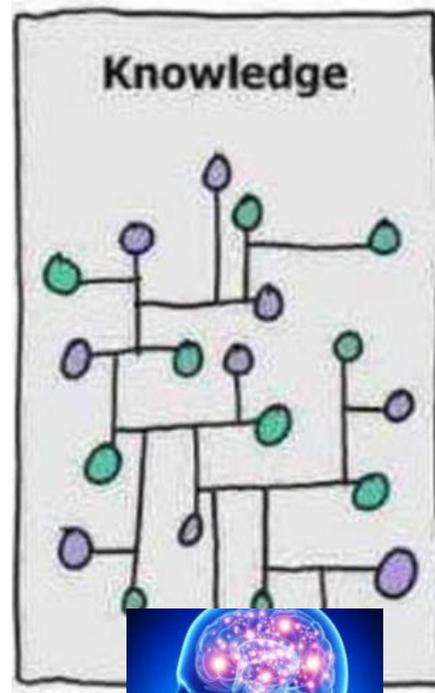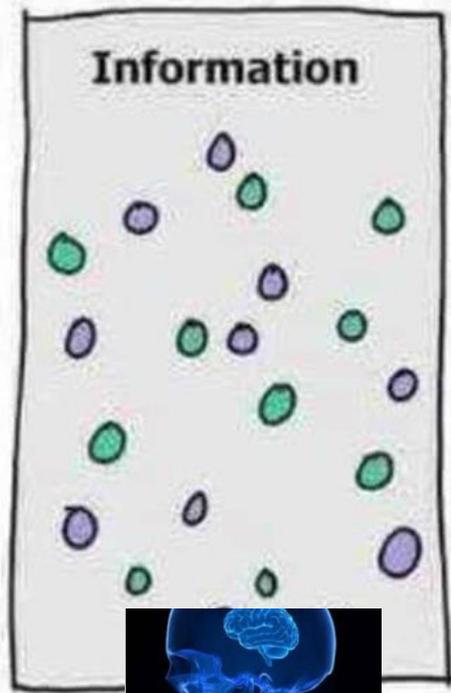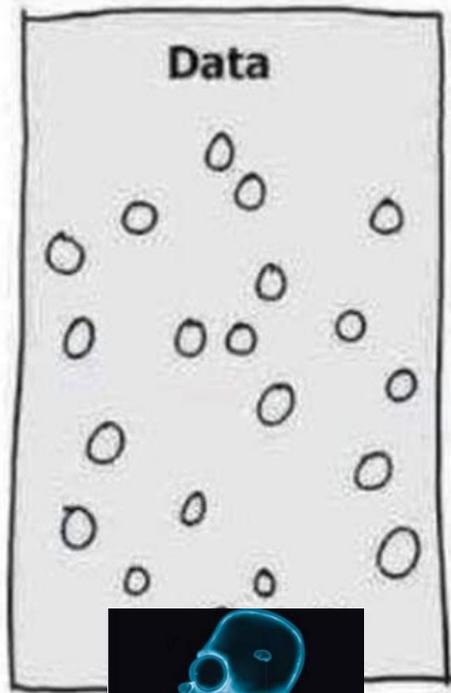
Summarising...
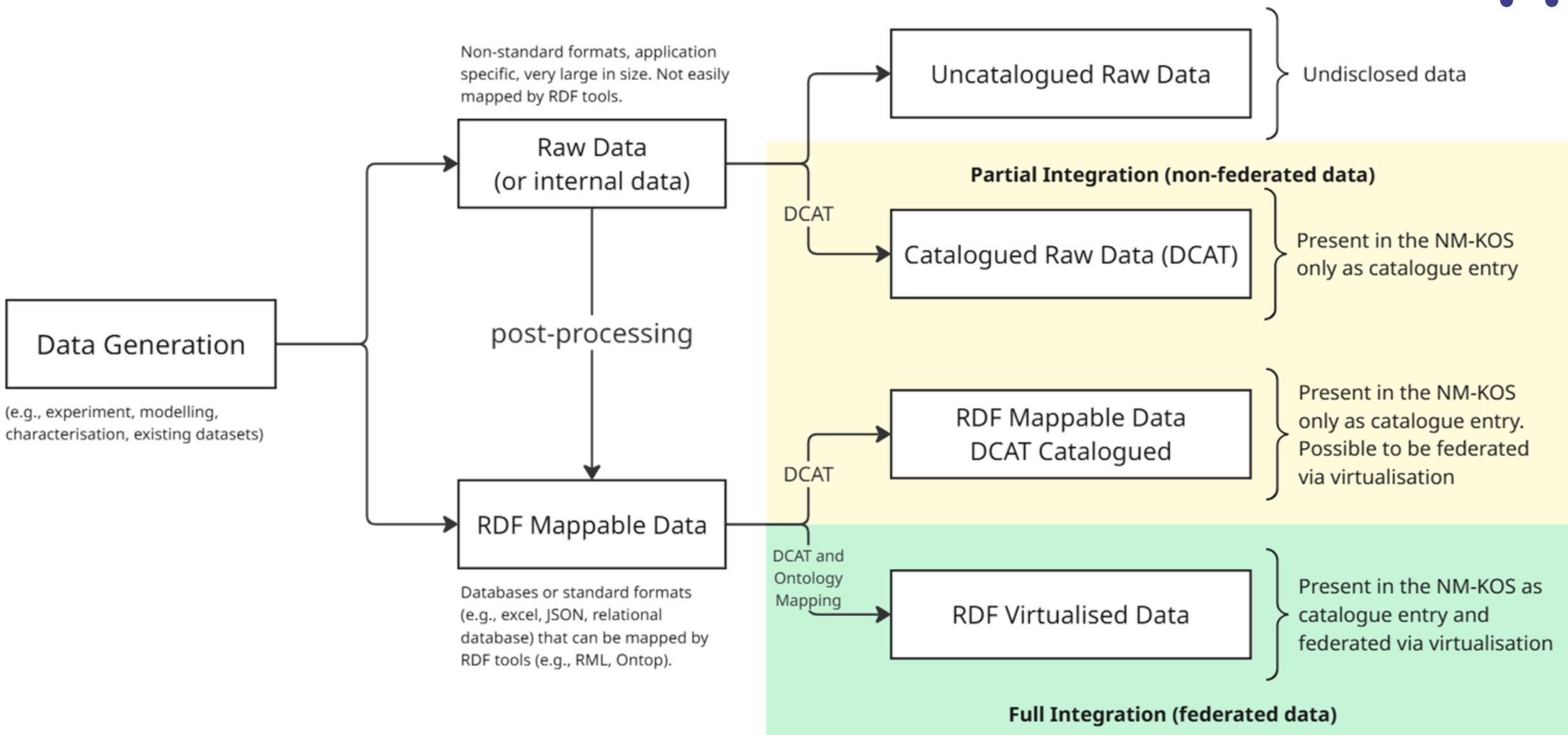
Summarising...

Data producers must assess their data generation activities and categorise their data in terms of raw internal data or RDF mappable data:

1. In case of **raw internal data**, it is to be distinguished between:
   a. **Uncatalogued raw data**, that have no immediate value outside the producer premises because of their specificity, or
   b. **catalogued raw data**, that are not RDF mappable, but may be reused by other users (e.g., reproducibility of results). These datasets must be documented with DCAT metadata.
2. In case of **RDF mappable data**, the owners are still required to document their datasets with DCAT metadata. Then they may opt for:
   a. **partial integration** (mandatory for all projects), where are expected to setup and populate a robust, standards-compliant relational or non-relational databases that enable reliable, long-term data access and facilitate semantic interoperability (in a local or a third party trusted repository). Alternatively, to setup a datasets repository, using formats that are compatible with RDF tools for data mapping and virtualisation. While this approach does not expose internal datasets, it allows metadata to be semantically linked to the NM-KOS vocabulary (e.g., through references to shared themes or ontology terms, enabling filtered search and thematic organization). Or for
   b. **full integration** (mandatory for RL1 projects, as per RL1 work programme), where are expected to develop an application ontology datasets mapping, setup a local SPARQL endpoint virtualising their data.

Summarising...

'As open as possible, as closed as necessary'

In Horizon Europe, research data should be made open access by default and licensed under the latest version of **CC BY** (attribution required) or **CC 0** (public domain), or equivalent.

However, it is recognized that data should be 'as open as possible, as closed as necessary', and **exceptions** can be made when providing open access to data:

- Is against the beneficiary's legitimate interests, including regarding commercial exploitation;

- Is contrary to any other constraints, such as data protection rules, privacy, confidentiality, trade secrets, Union competitive interests, security rules, intellectual property rights or;

- Would be against other obligations under the Grant Agreement.

In such cases, data can be kept restricted, closed or under embargo, but beneficiaries **must explain in the DMP the legitimate exception(s) under which they choose to restrict access to (some of the) research data**.

Summarising...

Publish or preserve

In the final phase of your research project, you stipulate who can *use* your data and how others can gain access to it. You need to consider if and how and where you will publish or preserve your data.

The specifics on who can access the data and how should be included, not only in your publications, but also in your published metadata.

From a FAIR perspective, access to data does not mean making data open. Instead, you have to specify who can access the data under what conditions or whom to ask for permission to access the data.

If the data are personal or otherwise sensitive, they should only be preserved, if there is a legal basis; and protocols for access have to be carefully considered.

Example 1
At the end of the research project **the data will stay at my institute** so that new data can be added all the time.
I provided DCAT-AP metadata and a PID for them through Zenodo. Other researchers can apply for access, and, if approved, they can send a model to analyse the local data. They will never get physical access to the raw data.

Example 2
At the end of the research project the data, that have been semantically mapped to the EMMO ontology, are openly published in an external **trusted repository** with SPARQL access (my institution have no such facilities), so that they can be federated by the NM-KOS or other initiatives.

Example 3
At the end of the research project the data are stored in **my institution SQL database** and available through a webpage. I provided DCAT-AP metadata and a PID for them through Zenodo. I also mapped my database to an EMMO application ontology so that, besides the SQL endpoint, I have also a SPARQL endpoint to make them RDF-accessible and federated in the NM-KOS. However, I decided to expose to generic SPARQL access only a portion of data, since for the more sensitive data I prefer to rely on my local database login authorisation.

"No Panic" Slide

There are three levels of foreseen partners engagement:

CONNECT-NM RL1 Projects (full integration): that are expected to contribute to the NM-KOS development proactively, as foreseen by the RL1 work programme. Partners of these projects are expected to be engaged to the development of all the NM-KOS layers and contribute with their datasets to the creation of an initial network of knowledge bases.

CONNECT-NM Non-RL1 Projects (partial integration): most of them that are expected to develop datasets that reach at least the level of data documentation foreseen by the DMP, with a documentation that depicts a conceptual mapping with existing ontologies, without a formal mapping. It is expected that, following RL1 results, other RLs projects will be motivated to invest resources towards a full virtualisation of their datasets.

Non-CONNECT-NM Data Partners: it is expected that such partners could be motivated in joining the NM-KOS after the establishment of a sufficiently large repository so that federation and interoperability benefits could be demonstrated and appreciated by the overall community. In the meantime, such partners can be involved with simple DCAT-based dataset documentation, without actual integration and virtualisation.

Ontologies Mapping RDF-friendly PID DCAT

RDF-friendly PID DCAT

???

"No Panic" Slide

General advice and guidance will be provided by WP4 partners, suggesting the best strategies for the implementation of semantic technologies solutions, as part of the CONNECT-NM T4.2 Guidance and follow up of projects within the research line activities.

Handbook and Toolkit to facilitate partners understanding and creation of DCAT metadata documentation for their datasets will be provided by WP4 in the next months.

Resources for the practical implementation of the necessary steps to reach full or partial integration with the NM-KOS must be **provided by data owners**. In the case of CONNECT-NM partners, resources must be allocated in the RLs projects for such task.

The EMMO development team, as part of the **EMMC Focus Area Digitalisation & Interoperability** will provide support throughout the entire process of creating the ontologies required for semantic mappings, according to its mission.

https://emmc.eu/focus-areas/digitalisation-interoperability/

Useful Readings

How to comply with Horizon Europe mandate
for Research Data Management
https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm

A deep dive into FAIR data
https://howtofair.dk/

Semantic Web for the Working Ontologist
https://dl.acm.org/doi/book/10.1145/3382097

https://doi.org/10.5281/zenodo.15504436

https://doi.org/10.5281/zenodo.15625581

WP4 Contacts

Partners who wish to participate to the CONNECT-NM NM-KOS initiative should contact the WP4 team at the following emails or through the project website:

- **Michal Pecelerowicz (NCBJ, WP4 Leader):** michal.pecelerowicz@ncbj.gov.pl

- **Emanuele Ghedini (UNIBO, WP4 Advisor):** emanuele.ghedini@unibo.it

- **Ilaria Maria Paponetti (UNIBO):** ilaria.paponetti2@unibo.it

- **Barbara Ferrucci (ENEA):** barbara.ferrucci@enea.it

- **Maria Perez-Hernandez:** mperez@fi.upm.es

- **Julian Arenas Guerrero (UPM):** julian.arenas.guerrero@upm.es

Thank you!

emanuele.ghedini@unibo.it