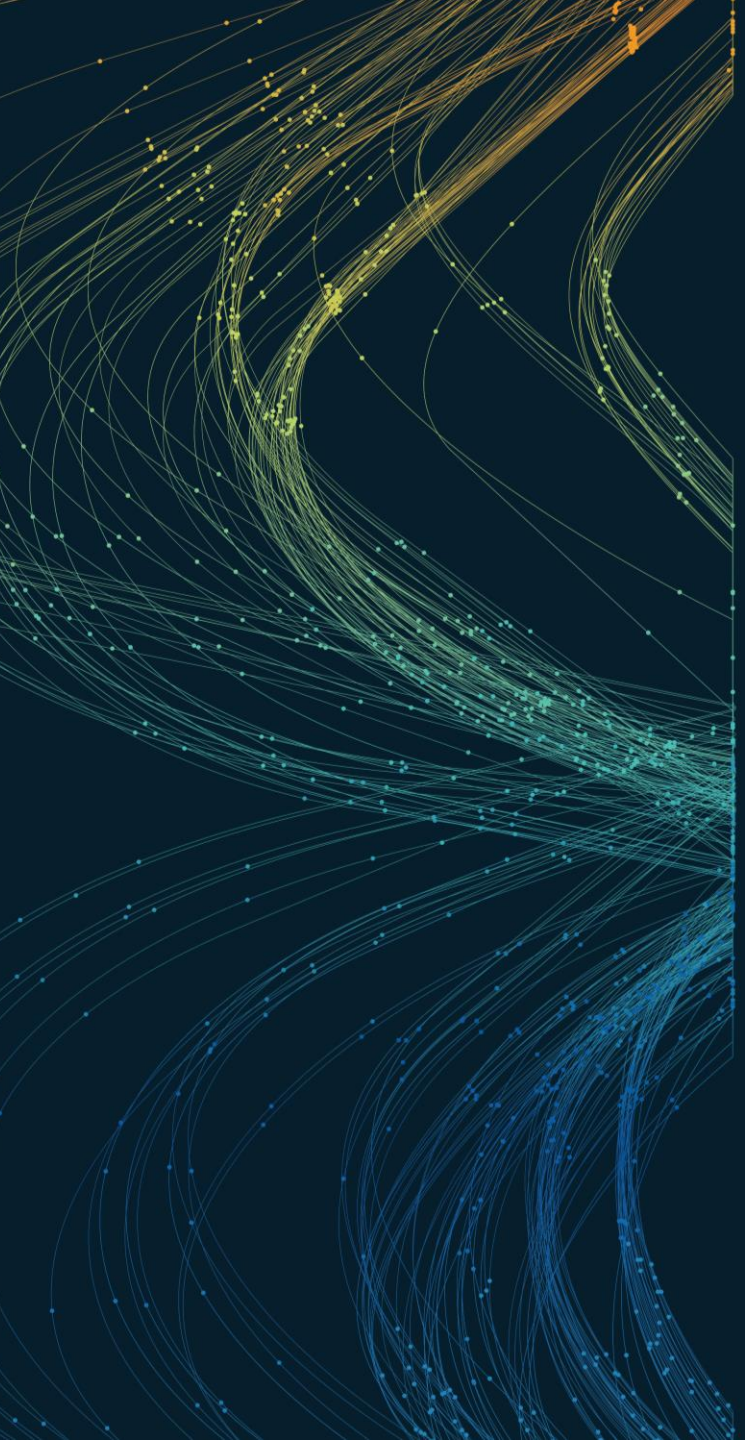
15-16 March 2023

# ENDORSE 2023 – Key takeaways

Link: https://op.europa.eu/en/web/endorse



# CONTENT

# 1. Towards data spaces

# 1a.
# Semantic Interoperability of Legal Compliance: *From GDPR to Data Spaces*

Following the GDPR in establishing responsible data practices, the European strategy for Data creates an unprecedented vision of a single market based on data sovereignty. It is accompanied by the Data and Data Governance Acts, and the establishment of Data Spaces. However, along with benefits, these regulations also bring new challenges related to their effective enforcement.

In this presentation, Dr. Pandit outlines a paradigm shift in how legal requirements and enforcement are approached by using semantic vocabularies.

***Speakers:***

**Harshvardhan J. Pandit,** Assistant Professor, ADAPT Centre, Dublin City University, Ireland Towards Data Spaces

European Commission

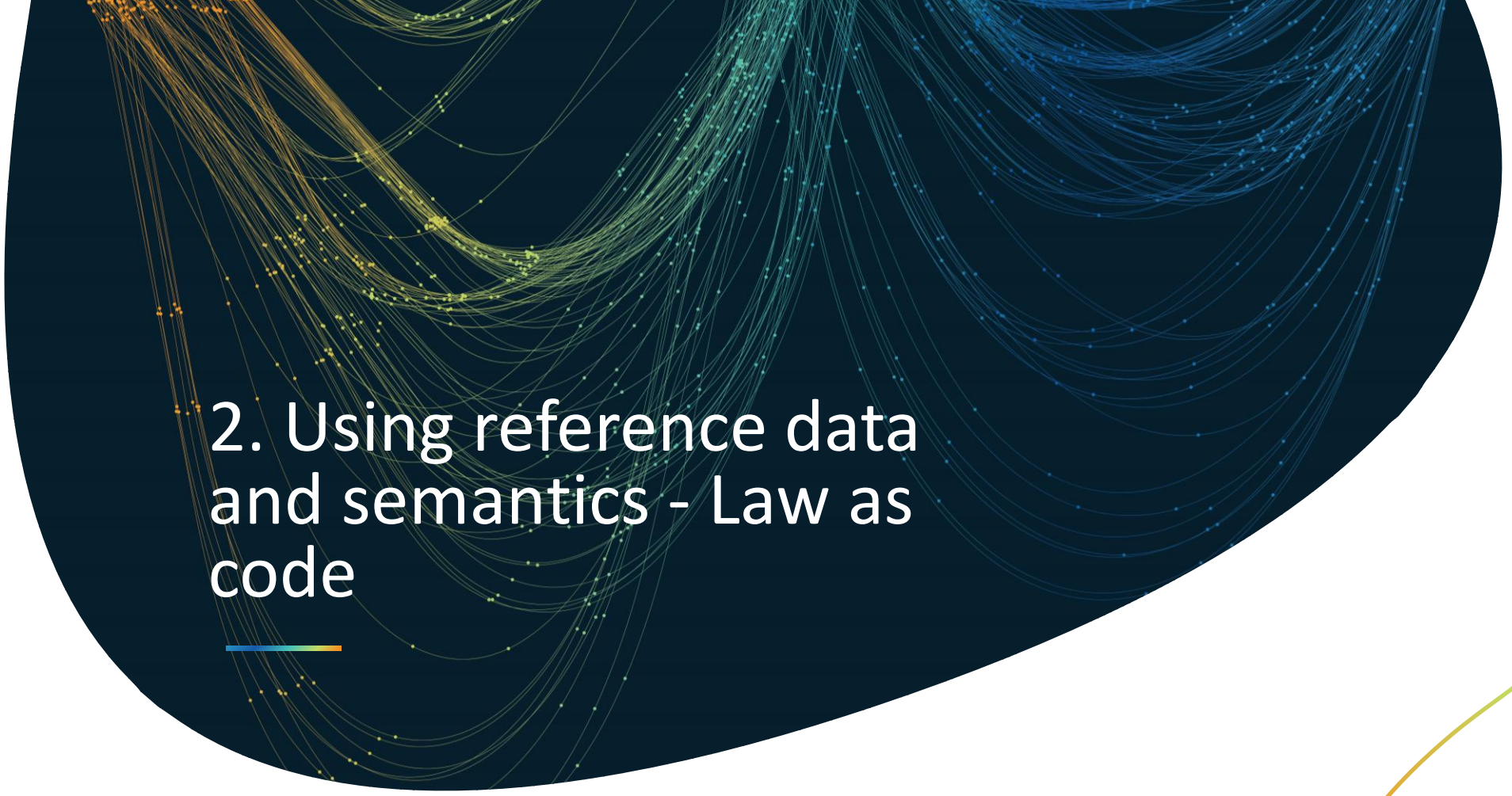
# 1a. Semantic Interoperability of Legal Compliance: From GDPR to Data Spaces

- **The EU is building a repertoire of "data spaces", each with their own corresponding laws.** For example, the GDPR applies to personal data, the Digital Services Act and the Digital Markets Act apply to digital services, and the AI Act applies to the use of AI.
    - This **may seem like a 'jigsaw', whereby legislation fits into one (correct) solution**. However, the **reality is more like a 'lego' - there are many different solutions, but we need to figure out what pieces we are working with and how to put them together.**

- **The goal is to create effective compliance mechanisms that are based on proven information practices and standards, and that can be automated**.
    - To achieve this goal, we need: (1) **automated solutions** to validate information (which should be **machine-readable**); (2) **interoperable specifications** for legal metadata; and (3) **consistency in legal mechanisms**, meaning a shared interpretation of what the law says and how it is to be implemented.

- The Data Privacy Vocabularies and Controls Community Group (DPVCG) is working on a rich set of vocabularies called the **Data Privacy Vocabulary (DPV)**. For the moment, these are **based on the GDPR and provide a taxonomy** for important terms such as: data controllers, data-use purposes, legal basis, etc.
    - The **real value of the DPV is promoting a shared understanding**. In doing so, it allows the creation of a central information model that can then be used for compliance related tasks.

- **Future solutions will have to work across different laws, using modular and simple vocabularies** that are focused on the specific tasks, operations, and obligations prescribed by the various regulations. This set-up ought to make them more **adaptive to legislative changes.**



# 2. Using reference data and semantics - Law as code

European Commission

# 2a.
# AKN4EU for consolidation.
## *Is the cat dead or alive? Akoma Ntoso 'MultipleVersions' as a solution to identify relevant legal clauses*

*This presentation illustrated the way in which detailed metadata, a well-defined document structure, and sophisticated automation could be used to make EU consolidated Acts more user-friendly, transparent and accessible.*

***Speakers:***

**Žilvinas Bubnys,** Head of consolidation and summaries of EU law sector at Publications Office of the European Union, Luxembourg

**Maria Kardami,** Information manager at Publications Office of the European Union, Luxembourg

# 2a. AKN4EU for consolidation. Is the cat dead or alive?
## Akoma Ntoso ‘MultipleVersions’ as a solution to identify relevant legal clauses

- **Consolidation involves combining the original legislative text and all subsequent amendments into a single consolidated text**. Each year, the EU produces 2,000 new versions of Acts in all 24 EU languages, which amounts to over 3.5 million pages of consolidated text.

- **Machine processing is used to handle the large volume of text**. In turn, this **requires clear and unambiguous amending instructions**.
- While **amending language can vary, four main types** of textual changes have been identified: replacement, addition, deletion, and renumbering. Once these are classified, the consolidation system can take the necessary action.
    - However, further **complexity is created by the different dates stipulated by amending acts**. These can be dates related to publication, application, transposition, or conditional application dates using phrasing such as 'once conditions are met' or 'upon notification'.

- **Due to this multitude of dates, consolidated texts would often be made available only after the modifications became applicable**. This meant that the usefulness of such documents depended on the user's perspective (i.e., whether they need to know about upcoming modifications).
- To tackle this situation, the Publications Office **implemented the multipleVersions concept.** This meant **creating XML files that combined in a single document all possible versions of a legal Act. Information is thus provided on:**
    - the **text of the original basic Act;**
    - the **text of each and every modification**, as prescribed in subsequent amending Acts and Corrigendums;
    - **detailed metadata about these modifications, including dates and date-types** (i.e., enactment, application, et. al.).

# 2a. AKN4EU for consolidation. Is the cat dead or alive?
## Akoma Ntoso ‘MultipleVersions’ as a solution to identify relevant legal clauses

- The **AKN4EU specification customizes the Akoma Ntoso XML standard to the specificities of EU legislation.** The **multipleVersions XML file follows these standards.** As such, it is **structured as follows:**
    - First, **an identifier for each amending Act is provided, alongside the relevant dates and date-types**.
    - Second, **the different dates are grouped in separate temporal groups, based on their amending Act .**
    - Finally, **the actual textual modifications are linked to the relevant temporal group** (i.e., making the correlation between textual changes and different dates).

- The **outcome of the aforementioned process is a singleVersion file** being generated, with the help of XSLT transformation technology, by providing two main parameters (date-value and date-type). For instance, **a user could ask to see the consolidated version of a legal text as applicable at a certain date**.

- This endeavour was **facilitated by acknowledging the need for semantic interoperability,** as well as by **using the open XML format, controlled lists of values, a defined set of controlled vocabularies, and unique ELI identifications**. Moreover, they are **preparing an ontology in order to disseminate the consolidation metadata as open linked data**.

# 2b.
# From Regulation to Knowledge

The Norwegian Maritime Authority needed a new supervision system that would allow all stakeholders involved in ensuring vessel compliance to interact more effectively. The new system utilises machine-readable linked data and Natural Language Processing techniques to discover context and relationships within regulations, which are then transformed into Resource Description Frameworks (RDF). This presentation provided an overview of said approach.

***Speakers:***

**Veronika Heimsbakk,** Managing AI Engineer, SME Semantic Technologies - Insights & Data, Capgemini, Norway

# 2b. From Regulation to Knowledge

- **The Norwegian Maritime Authority (NMA) was looking for a replacement for their supervision system that determines compliance with mandatory maritime statutes and regulations**. The project involved **automating the issuing of personal certificates for sailors**; **providing machine-readable, linked data reflecting current and historic regulation**; i**mplementing tools for enrichment of regulatory requirements**; and **enabling NMA domain experts to maintain the knowledge and metadata themselves**.

- The project **started out manually**, but the team found that using **Natural Language Processing (NLP)** to identify context, concepts, and relationships **was faster and more accurate** than manual interpretation.

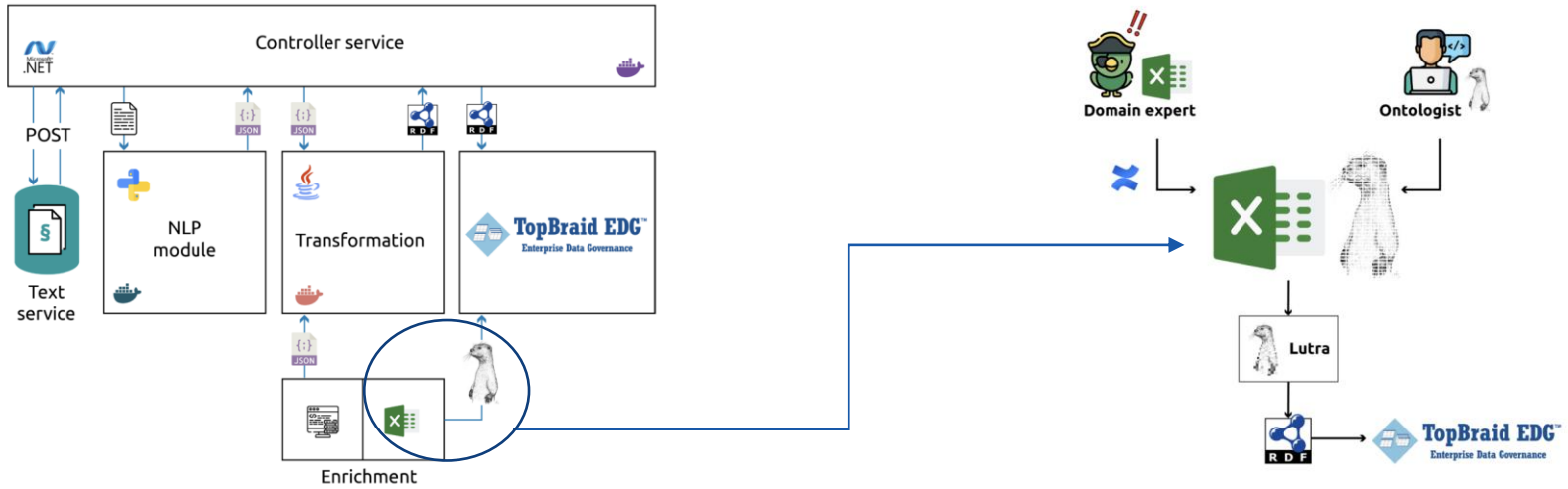
- **The pipeline from regulation to knowledge graph** involves:
  - First, a t**ext service that converts the regulation into plain text**;
  - Second, an **NLP module that identifies keywords and context** (said keywords refer to the legal scopes and characteristics of a regulatory requirement);
  - Third, a **Resource Description Framework (RDF) transformation** which is implemented in Java with RDF4J. From there, the finished graph is passed on to the **graph database** of choice, in this case, **TopBraid EDG.**

  ...

# 2b. From Regulation to Knowledge

...

- In addition to this automated pipeline, the project team has also provided **enrichment service**s. As such, domain experts can also access:
  - A **web service containing a form where vessel inspectors can fill in their criteria** for inspection;
  - **Excel spreadsheets describing code systems.**
    - This enrichment is to be **done in OTTR via a manual process**. An **ontologist creates a template for serialization and the domain expert populates** said template. Then, the **spreadsheet is transformed into an RDF, using the Lutra implementation of OTTR, and fed into TopBraid.**
- Finally, a **control service** oversees the communication between these microservices.



# 2b. From Regulation to Knowledge

- **The finished model of the regulatory requirement is produced using the Shapes Constraint Language (SHACL)**. This was judged to be better for legal knowledge because it does not allow incomplete information and is able to describe the different legal scopes in greater detail.

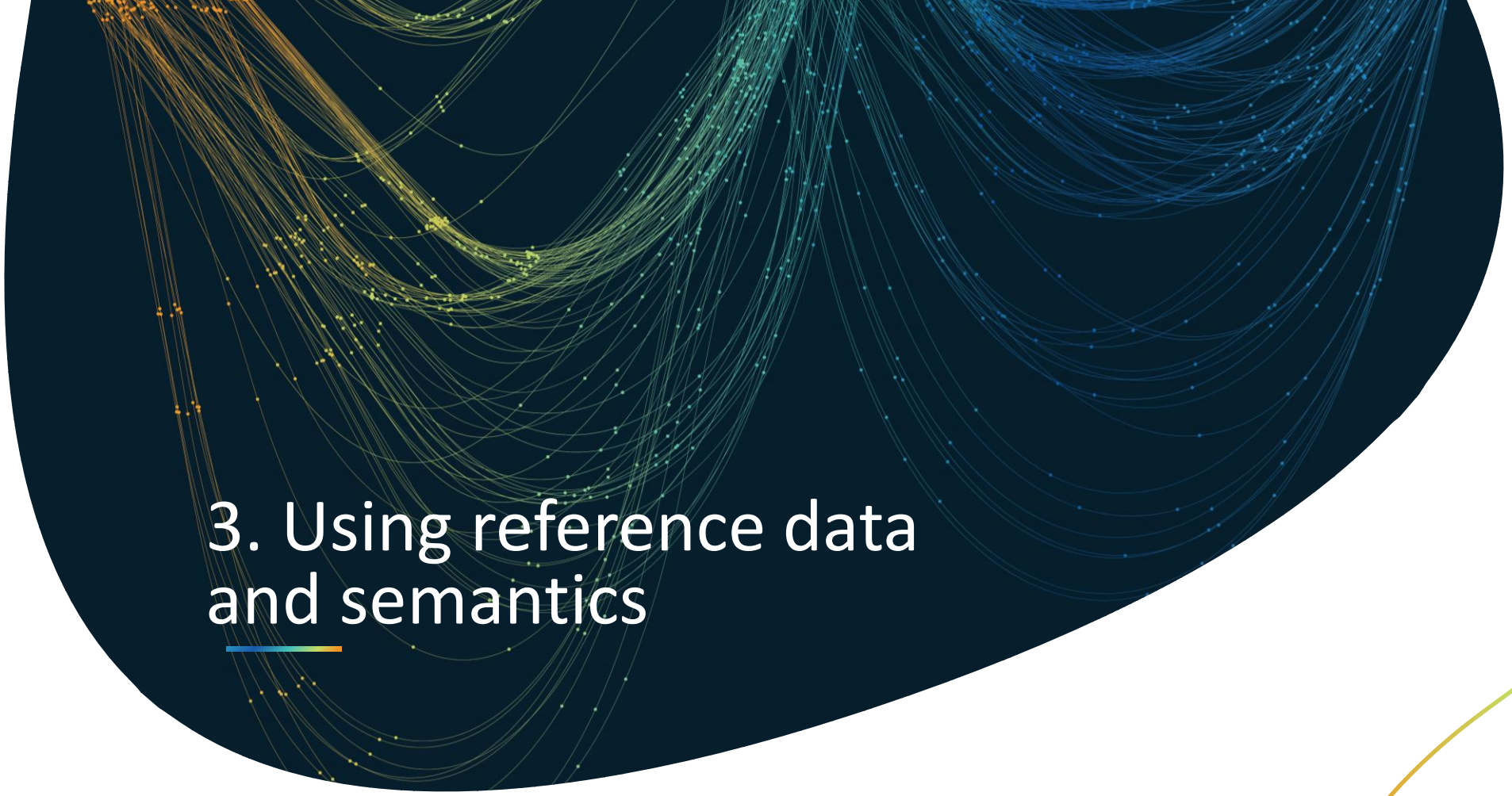
- **The knowledge graph consists of a TBox (OWL ontologies) and an ABox (Data Graphs)**.
  - **The TBox includes:** vessel types, trade areas, education, general concepts, a SFI Foundation Specification classification system;
  - **The ABox includes**: regulatory requirements (modelled in SHACL), personal certificates (modelled in SHACL for validation purposes, to automate the issuing of the certificates); checkpoints and activities (NMA vocabulary) .

- **Data governance was facilitated through the adoption of TopBraid EDG**. This tool provides: a form-view for browsing resources and editing nodes; a workflow with comparison reports and archive; as well as complete control of the ontology lifecycle management. Versioning, graph history, and graph statistics are likewise provided.
  - **Governance was further facilitated through**: (1) the adoption of **Enterprise Data Governance (EDG)** in dev/test/prod environment running on Azure; (2) **property grouping the different kinds of legal scopes and characteristics**; and (3) a **SPARQL library** including production-ready queries for backend applications to extract data and use for different purposes.

# 2b. From Regulation to Knowledge

- **Challenges** included: **maintaining the code bas**e (and making sure the client can do the same without input from contractors); **discovering relationships through named entity recognition models**; ensuring **client comfort with semantic technologies**; and **creating proper documentation** of every aspect.

- **Lessons** learnt included: the **importance of close cooperation between client domain experts and ontologists**; the **time-saving benefits of Natural Language Processing**; the **importance clients attach to editing possibilities in a low-code user interface**; and the **suitability of SHACL for the law** domain.



# 3. Using reference data and semantics

# 3a.
# Sharing reference data with Re3gistry

The Re3gistry software was initially developed in 2013, by the Joint Research Council (JRC), to support the INSPIRE Registry. In 2020, it was released as a generic open-source project due to its high reusability potential. The presentation demonstrated how the Re3gistry software can be utilised to swiftly make reference codes accessible to the public, all while ensuring a formal content management process that is effective even in complex governance settings.

***Speakers:***

***Juan Martin Pelegrina,*** *Bilbomatica, in service contract with the European Commission Joint Research Centre*

***Lorena Hernandez,*** *Project Officer, Joint Research Centre - European Commission*

***Jordi Escriu,*** *Scientific Project Officer, Digital Economy Unit, Joint Research Centre - European Commission*

# 3a. Sharing reference data with Re3gistry

- **Re3gistry is an open-source solution that works as a container of reference codes, allowing administrators to share and manage them**. It has **several functionalities**, such as **providing a central access point**, **enabling the use of labels and reference code descriptions**, and **allowing information to be retrieved by machines through its REST API.**

- **Version 2 was released in the summer of 2022**. One of its key features is a **web-application that helps users access the reference codes.**
    - Other **key features** include: an **interactive management interface**, **multilingual content support**, **free-text search**, **multiple authentication options**, the **availability of several formats (**JSON, XML, RDF, etc.), a **simplified workflow**, the **possibility to search for items with specific status**, **open-API**, **RSS feed**, and **Docker installation**.

- Re3gistry was **originally developed in 2013 to support the implementation of the INSPIRE Registry and** has **since been used by various public and private organizations**, including **European national agencies in several Member States and EFTA countries**. The Publications Office is also **considering its use in the publication of Corporate Reference Data Sets.**

# 3a. Sharing reference data with Re3gistry

- The **latest release of Re3gistry  is v2.4.1**, which is available on GitHub and Joinup. A **minor update, v2.4.2,** is expected to be launched soon, while a **major update, v2.5.0, is scheduled for June 2023**. The project has also been **included for Google Summer of Code 2023.**

- Overall, Re3gistry is **a useful tool which:**
    - **helps users avoid common mistakes**, such as misspellings;
    - **facilitates internationalisation** via the use of multilingual labels;
    - **ensures semantic interoperability** by allowing the exchange of data between system and applications;
    - and **increases the value of reference codes** by making them easier to use and reuse.

# 3b.
# Simplifying the reuse of concepts across organisations

The National Data Management (NaDB) is a programme launched by the Swiss Federal Council to create the foundation for the implementation of the Once Only Principle. Within this program, the Swiss Federal Statistical Office (FSO) is in charge of developing and maintaining an Interoperability Platform (IOP). One of the aims of the platform is to provide a centralized metadata repository that facilitates the reuse of standardized nomenclatures. This presentation focused on the technical and organizational aspects needed to promote the harmonisation and reuse of standards.

***Speakers:***

***Fabian Santi,*** *Interoperability specialist, Swiss Federal Statistical Office (SFSO), Switzerland*

# 3b. Simplifying the reuse of concepts across organisations

- The **National Data Strategy in Switzerland mandates the Federal Statistical Office (FSO) to build, maintain, and deploy an interoperability platform containing the metadata describing the Swiss data landscape**.
  - The platform is not only intended for use by the federal government but also by different regional offices and even private companies.
  - The **FSO has two roles within the platform: contributor of different datasets, standards, and nomenclatures, and builder and developer of the platform**.

- The presentation identified **three building blocks for harmonisation: technical aspects, governance, and process model.**

- On the **technical** side**:**
  - The platform uses a **DCAT-Compatible Metadata Catalogue.** The metadata is **structured and machine friendly.** Moreover, it also includes **descriptions of datasets that are not open government data.**
  - **Nomenclatures on I14Y** are likewise employed. There are **collections of designations and technical terms** from legally binding instruments. **Different organisations using the platform can describe different nomenclatures**, some are solely for description while others represent binding standards within the domain.
  - Additionally, the platform uses **structural metadata on I14Y to provide information about the internal structure of data collections.**

# 3b. Simplifying the reuse of concepts across organisations

- On the **governance** side:
  - Each institution onboarded onto the platform must have a **Data Steward who is responsible for and controls data harmonisation.**
  - The **Local Data Steward works on the institutional level,** coordinating with local data owners, custodians and IT, to ensure that **data is stored according to the standards and metadata specifications.**
  - At the **federal level, the Swiss Data Steward** oversees the **communication between Local Data Stewards and the federal government**, **signs off on and promotes the right standards**, and **upholds the quality of metadata.**

- Two **standardisation process models** are available:
  - The **normal process involves forming a working group, developing a standard, collecting feedback from stakeholders, getting a decision from data owners, and publishing said decision on the Interoperability Platform**.
  - However, the hope is to **promote a process where institutions are made to adhere to a certain standard through the power of example**. Essentially, the idea is that when **different institutions describe their data sets on the Interoperability Platform, they may use different formats** for the same piece of information. By making **these differences visible on the platform,** the hope is that **institutions will be motivated to follow the standard used by the majority.** This would promote standardisation and harmonisation, making it easier for institutions to share data with each other.

# 3c.
# Driving a successful Digital Strategy - 360-degree Application of FAIR principles for Data Harmonisation and Standardisation

The Pharma Industry has adopted FAIR (findable, accessible, interoperable, and reusable) Data, but mostly in isolation from a digital strategy for healthcare and life sciences. This presentation covers how to implement a 360-degree FAIRification of data and required capabilities for reference data services.

***Speakers:***

***Martin Romacker,*** *Data & Analytics Roche Innovation Center Basel*

# 3c. Application of FAIR principles for Data Harmonisation and Standardisation

- Roche have been working on a project to ensure that **data is FAIR - Findable, Accessible, Interoperable and Reusable**. Having FAIR data means that **scientists and data consumers can find, access, and understand the data without the need for the data owner**. Additionally, **machines can automatically find and semantically use the data (i.e., the data is machine actionable).**

- There is a **continuum** running from unstructured data to FAIR strict data:
    - **Unstructured data**: This refers to free text documents that are not easily searchable;
    - **Semi-structured data**: This type of data has some structure, but it is not fully spelled out. There may still be free text in the document.
    - **Structured data**: This type of data has a defined schema or structure. However, it may not be FAIR because the elements do not convey the semantics in terms of global, unique, persistent, and resolvable identifiers. This makes it more difficult to work with.
    - **FAIR relaxed data**: This type of data has a defined schema and controlled vocabularies.
    - **FAIR strict data**: This type of data has a schema definition that is ontology-driven and has a GUPRI (global, unique, persistent, resolvable identifier) for data elements and values. This makes it easily searchable and **machine-actionable** insofar as it can be integrated into knowledge-driven data management value chains without any further transformation.

- Only **FAIR strict data is machine actionable**; the **other categories are merely machine readable.**

# 3c. Application of FAIR principles for Data Harmonisation and Standardisation

- The **Scientific Interoperability Hub** is a reference system that offers three different products: **terminology management, semantic dataset definition, and conceptual modelling (ontologies).**
    - These products are **designed to be FAIR and support FAIRification.** They **ensure semantic interoperability and high data quality.**
    - The products **serve as reference data for standardised terminologies, metadata, and conceptual models, which allows for semantically linking internal and external data assets for data acquisition and integration.**
    - They **support more than 100 productive applications across all Roche functions and sites,** and **guarantee currency and ongoing support.**

- Among other, the **Data Harmonisation Service** offers: c**ustom tailored terminologies**, **dataset models and ontologies**; as well as, **maintenance and enhancement of standards**; **shared semantics (conceptual models)**; and **technical support**.

- Overall, **in order to have a successful and valuable digital strategy, machine readability is insufficient; we need machine-actionable data. The application of FAIR strict principles is mandatory for achieving this.**


See you soon...
in one of our upcoming BLSI events


interoperable europe

innovation ∞ govtech ∞ community

Stay in touch

EU-legal-interoperability@ec.europa.eu

https://joinup.ec.europa.eu/collection/better-legislation-smoother-implementation