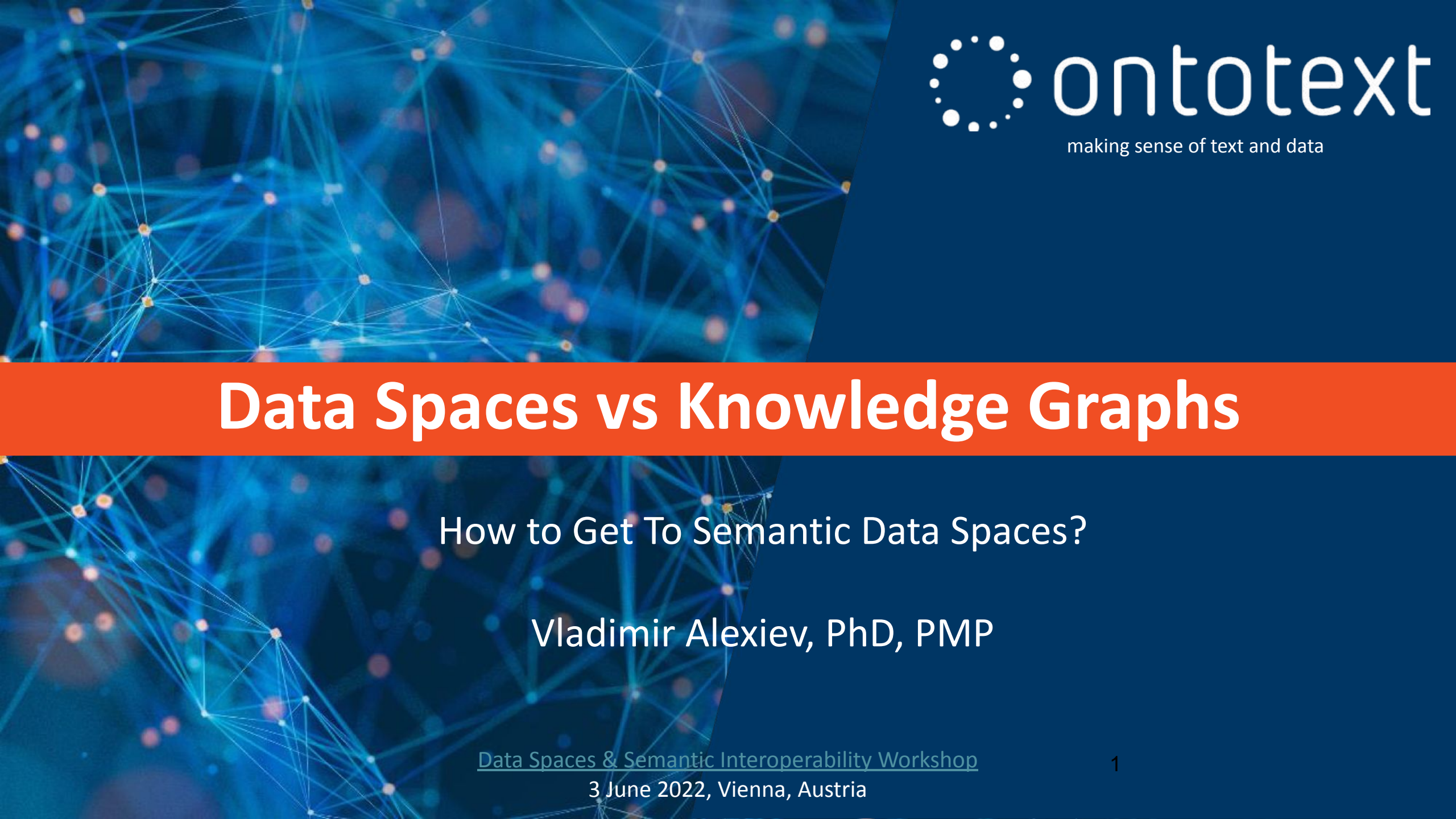


ontotext

making sense of text and data

Data Spaces vs Knowledge Graphs

How to Get To Semantic Data Spaces?

Vladimir Alexiev, PhD, PMP

Data Spaces & Semantic Interoperability Workshop
3 June 2022, Vienna, Austria

Presentation Outline

- Ontotext Intro
  - Clients and Research Projects
  - Building Knowledge Graphs
- Data Spaces
  - Semantic Metadata
  - Why not Semantic Data?
  - Why not Linked Data?
- Use of semantic technologies in industries
  - Why not Linked Data?
- Semantizing industrial data models
  - Polyglot Modeling
  - Hybrid Storage
- Food Safety Market project (FSM)

Ontotext Introduction

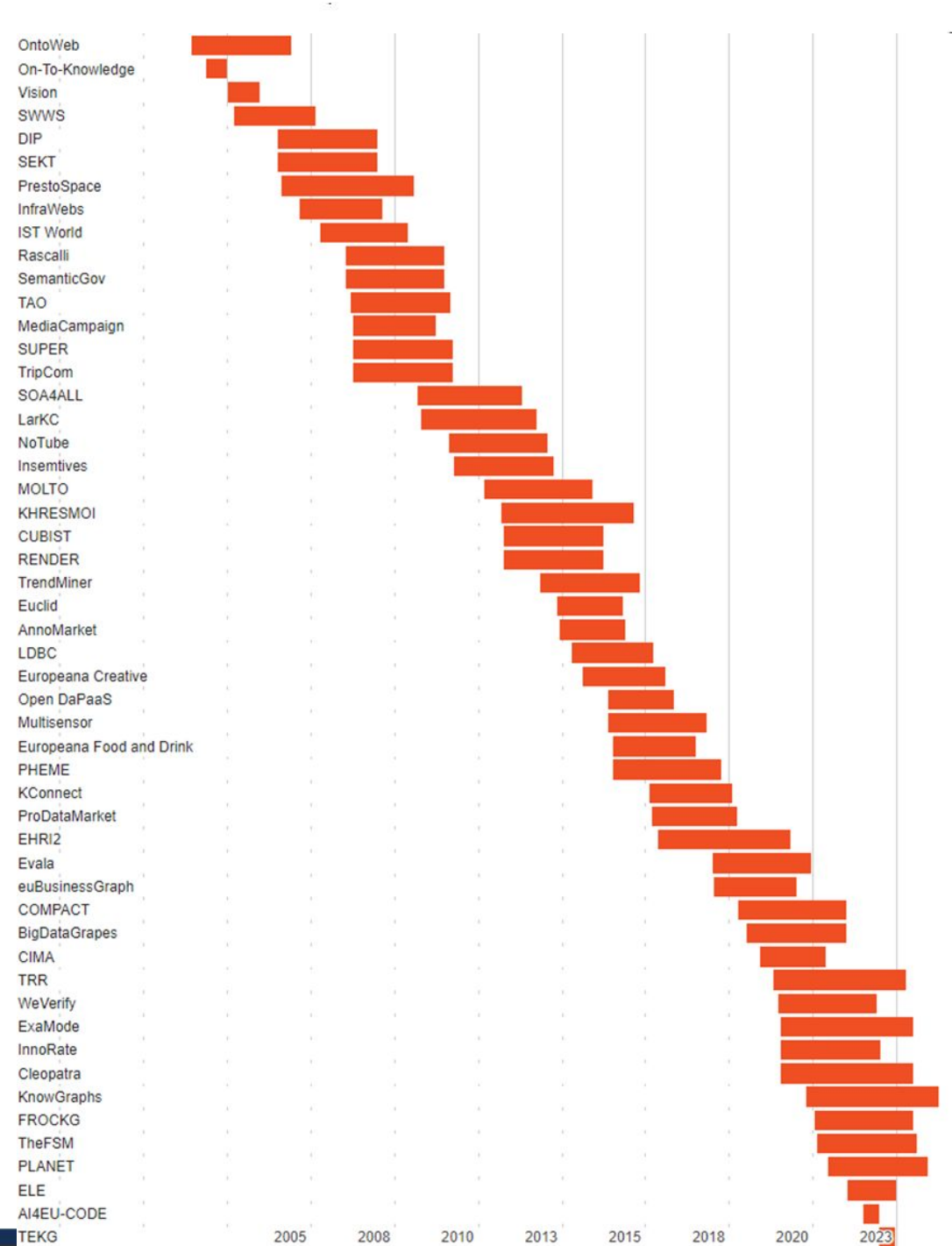
- **Leader**
  - ✓ Semantic Technology Vendor since 2000
  - ✓ Part of Sirma Group: 500 Person, Listed on the Bulgarian Stock Exchange (SKK), part of SOFIX
- **Profitable and Growing**
  - ✓ Global: 80% of revenue from London and New York
  - ✓ Clients: Stellantis (PSA), Schneider Electric, Johnson Controls, Statnett, EDF, ENTSO-E, S&P, BBC, Financial Times, JPMC, UK Parliament, Fujitsu, …
  - ✓ Ontotext just acquired for 30 MEUR
- **Innovator**
  - ✓ Attracted over 15 MEUR in innovation funding (40 projects)
  - ✓ Bulgaria's most successful participant in EU research projects
- **Partners**
  - ✓ Wide partner network: Eccenca, SWC, Synaptica, VocBench, Metaphactory, Enterprise Knowledge…
  - ✓ W3C, STI2, Linked Data Benchmark Council

ontotext

Ontotext EU research projects

Verticals (research and commercial)

- **Industrial data**
  - ✓ Building Information Management
  - ✓ Energy and electricity
  - ✓ Transport & logistics
  - ✓ Data markets and data spaces
- **Other**
  - ✓ Financial services and insurance
  - ✓ Companies, transactions, economics
  - ✓ Healthcare and life sciences
  - ✓ Media and publishing
  - ✓ Fact checking and fighting desinformation
  - ✓ Linguistics and text analysis
  - ✓ Government



Building Knowledge Graphs in 10 Steps

1 Clarify your business & expert requirements

2 Gather and analyze relevant data

3 Clean data to ensure data quality

4 Create your semantic data model

5 Integrate data with ETL or virtualization

6 Harmonize data via reconciliation, fusion and alignment

7 Architect the data management and search layer

8 Augment your graph via reasoning, analytics and text analysis

9 Maximize the usability of your data

10 Make your KG easy to maintain and evlove

- **Webinar:** Knowledge Graphs: 5 Use Cases and 10 Steps to Get There
- **Video:** Building Knowledge Graphs in 10 Steps

ontotext

Data Spaces

- **A huge EU effort and investment on industrial data sharing**
  - ✓ "AI without data is like a human without air"
  - ✓ Key factor for industry digitalization
  - ✓ Contributing to competitiveness, COVID economic mitigation, the Green Deal, weaning off Russian fossil fuels
  - ✓ Associations: BDVA, DAIRO, IDSA, GAIA-X, ADRA
- **Ingredients**
  - ✓ Legal framework: incentivise sharing, while preserving sovereignty
  - ✓ IDSA reference architecture
  - ✓ EU Data Spaces projects and programs
  - ✓ Relevant Horizon Europe calls
  - ✓ Digital Europe: data spaces, innovation hubs (DIH), testing facilities (TEF), training
- **Uses semantic technologies for metadata**
  - ✓ But not (yet) for data

Semantic Metadata in Data Spaces

- **IDSA architecture is based on semantic specifications and ontologies:**
  - ✓ Datasets and related FAIR metadata
  - ✓ Dataset descriptions: topical, temporal, geographic coverage, statistics
  - ✓ Licenses
  - ✓ Participants: owners, users
  - ✓ Access rights
  - ✓ Appropriate use and commercial agreements
  - ✓ etc
- **References**
  - ✓ IDSA Reference Architecture Model. Version 3.0, April 2019
  - ✓ The International Data Spaces Information Model – An Ontology for Sovereign Exchange of Digital Content. ISWC 2020

Why not Semantic Data?

- **Few if any data spaces use semantic data**
  - ✓ You get to the data, but what if each owner provides it in a different format?
  - ✓ You still face the typical data integration challenges
  - ✓ Standardization is at the data access level, not at the data interoperability level
- **Harmonization of data models is left to individual industries**
  - ✓ (and often does not happen)
  - ✓ Even when the respective industry has some semantic models (see later)

Why not Linked Data?

- **Even fewer data spaces use Linked Data (LD) principles**
  - ✓ IDSA Connector architecture is based on data **transfer** (then centralization)
  - ✓ LD principles are based on data **sharing** (distribution and federation)
- **What is the difference?**
  - ✓ A data copy becomes obsolete the moment it's transferred
  - ✓ How much data to transfer at once (what are the business entities?)
  - ✓ Difficult questions: "Where's the latest version of this data? Who is mastering it?"
  - ✓ LD principles provide data that is up-to-date on-demand
  - ✓ With LD, data mastering, responsibility, entity scope are simple and clear
- **Why does it matter?**
  - ✓ Improved efficiency of data provisioning and use, timeliness and locality
  - ✓ "Data Sovereignty" not just in a legal but also a technical sense
  - ✓ Imagine a network of distributed semantic stores, access controlled and collaborating
  - ✓ Wait: such a thing exists: SOLID!

Uptake of Semantic Technologies in Industries

- **Semantic uptake in various industries**
  - ✓ Product Classifications and Catalogs: eClass semantization, Wikidata
  - ✓ Smart Manufacturing: Plattform Industrie 4.0: RAMI, AdminShell
  - ✓ Electricity: IEC Common Information Model (CIM), ENTSO-E CGMES and Market Transparency
  - ✓ Oil and gas: ISO 15926, CFIHOS, etc
  - ✓ Energy efficiency: DABGEO family, SEAS, SEMANCO, OpenADR, etc
  - ✓ Transport and Logistics: GS1 WebVoc, GS1 EPCIS 2.0
  - ✓ Architecture and Construction: IfcOwl, IfcWod, LBD ontologies, Bricks, semantic asset management, object type libraries, etc
  - ✓ Smart cities, Geospatial and Cadaster
  - ✓ Food and Agriculture: FAO (AgroVoc etc), GODAN, crop ontologies, Open Food Facts, etc
- **But most don't follow Linked Data principles**
- **Horizontal efforts**
  - ✓ Industrial IoT: WoT, Thing Description
  - ✓ SAREF and its extensions (SAREF4ENER, SAREF4BLDG…)

Product Classifications and Catalogs

- **Standard Data Models**
  - ✓ ISO 13584-42 and IEC 61360-2 (parts library, PLIB)
  - ✓ IEC 62656 (parcelized ontology model, POM)
- **Industrial catalogs**
  - ✓ eCl@ss, EU CPV, UNSPSC, GS1 GPC, IEC CDD, ECALS, NAMUR, RosettaNet, PFI, eOTD, RNTD, BMEcat, bSI bSDD, COBie
- **Semantization**
  - ✓ eClass semantization effort
  - ✓ Wikidata captures a number of the above: EU CPV, GS1 GPC, etc
  - ✓ Local efforts, eg cpv.data.ac.uk

Industry 4.0: RAMI, AdminShell

- **Asset Administration Shell (AAS)**
  - ✓ Allows incorporating important industrial data exchange standards: OPC UA, AutomationML, Collada, eCl@ss.
  - ✓ Schema definitions in UML, XML schema, JSON schema, RDFS
  - ✓ Data renditions as XML, JSON, RDF
- **Relies on data copying not LD**
  - ✓ Reference data is copied over into blank nodes
  - ✓ Uses string identifiers (eg IRDI) not URI/URLs

```
aas_submodel:submodelElement [
  a aas:Property;
  rdf:subject <type/1/1/F13E8576F6488342/Manufacturer>;
  aas_referable:idShort "Manufacturer";
  rdfs:label "Manufacturer";
  aas_property:category aas_category:CONSTANT;
  aas_hasKind:kind aas_modelingKind:INSTANCE;
  aas_hasSemantics:semanticId [
    a aas:Reference;
    aas_reference:key [
      a aas:Key;
      aas_key:index "0"^^xsd:integer;
      aas_key:type aas_keyElements:GLOBAL_REFERENCE;
      aas_key:local "false"^^xsd:boolean;
      aas_key:value "0173-1#02-AAO677#002";
      aas_key:idType aas_identifierType:IRDI]];
    aas_key:value "Company GmbH"];
```

ontotext

Electricity

- **Common Information Model (CIM): a number of IEC standards**
  - ✓ IEC 61970 Energy management system API
  - ✓ IEC 61968 Application integration at electric utilities - System interfaces for distribution management
  - ✓ IEC 62361 Interoperability in the long term
  - ✓ IEC 62357 Seamless Integration Reference Architecture
  - ✓ IEC 62056 COmpanion Specification for Energy Metering (COSEM)
  - ✓ IEC 62746 Systems interface between customer energy mgmt sys and power mgmt sys
- **Especially important for European single market (ENTSO-E)**
  - ✓ IEC 62325: Energy Market Communication (Exchange)
  - ✓ IEC 61970-600-1: CGMES Structure and rules
  - ✓ IEC 61970-600-2: CGMES Exchange profiles specification
- **Renditions**
  - ✓ UML (IEC 61970-301 etc), RDFS (IEC 61970-501), OWL (IEC 61970-505), RDF-XML (IEC 61970-552).

EIC and IRIs vs UUIDs

- **Energy Identification Codes (EIC)**
  - ✓ They identify all important entities in the electricity domain:
    - areas (control, scheduling, synchronous, market, bidding zones, etc),
    - system operators (TSOs and DSOs),
    - market players (exchanges, traders),
    - electrical assets (power plants, generators, transmission lines, substations, loads)
  - ✓ Not just electricity but other energy (eg gas: ENTSO-G)
  - ✓ Not just EU but global
- **Yet, current CIM**
  - ✓ Does not use IRIs for entity identification, but temporary UUIDs
  - ✓ Used only as an exchange format, not Linked Data
  - ✓ Non-standard RDF XML in handling of model graphs (used for differential models)
- **New IEC 61970 parts in development**
  - ✓ JSON-LD exchange
  - ✓ RDF shapes
  - ✓ permanent UUIDs
  - ✓ But still no LD principles: no conception of distributed mastering and federation of CIM data

Transport and Logistics

- **EPC Tag Data Standard (TDS) identifiers for logistic chain objects:**
  - ✓ Product types (GTIN), batches/lots (LGTIN), individual products (SGTIN), organizations (PGLN), locations (SGLN), documents and their types (GDTI), individual assets such as vans and sensors (GIAI), returnable assets such as palets (GRAI), logistical units (SSCC), shipments (GSIN), consignments (GINC)
  - ✓ Barcodes and RFIDs
- **GS1 WebVoc**
  - ✓ Extension of schema.org, adds many terms relevant to trade and logistics
- **GS1 EPCIS 2.0**
  - ✓ "Object visibility" events (from scanning barcodes or capturing RFIDs)
  - ✓ Sensor readings, certificates; WebVoc extensions
  - ✓ XML, JSON, JSONLD; XSD, JSON schema, JSON-LD context, ontology, SHACL
  - ✓ Ontotext contributed significantly (thanks to H2020 FSM and PLANET)
- **GS1 Digital Links**
  - ✓ From TDS to web info, including logistics master data as LD
  - ✓ Transitioned from paper barcodes to RFIDs to linked data!

Architecture and Construction

- **Architectural designs and construction plans: IFC (ISO 16739-1)**
  - ✓ Defined in ISO 10303-11.2 EXPRESS
  - ✓ Renditions: XML schema, JSON schema, OWL ontology; payload: STEP (text), XML, JSON, RDF, HDF5
  - ✓ IfcOwl carries heavy EXPRESS baggage → various simplifications such as IfcWod
- **Semantic standards**
  - ✓ LBD ontologies
  - ✓ Bricks/Haystack
  - ✓ NL COINS schema
  - ✓ NL NTA 8035 Semantic Data Modeling in the Built Environment
  - ✓ NL NEN 2660 Rules for information modelling of the built environment
  - ✓ ISO 21597 Information Containers (ICDD)
  - ✓ CEN 17632 Semantic Modeling and Linking (SML)
- **Standards we hope will have semantic rendition**
  - ✓ Data Templates, Object Type Libraries, Specification Libraries, Product Catalogs, Common Data Environments
- **Hot topics:**
  - ✓ Semantic asset management
  - ✓ Decentralized semantic stores (SOLID)
  - ✓ Amberg consulting company: "AECO data is too complex **not** to use semantic approaches"

"Legacy" Industrial Data Models

- **Any established industry comes with tons of data standards**
  - ✓ Product Catalogs (eClass and IEC CDD) are defined in ISO OntoML
  - ✓ Architectural data (IFC) is defined in ISO EXPRESS
  - ✓ Electrical CIM is defined in UML, from which RDFS and SHACL are derived using profiles
  - ✓ ACORD insurance standard is defined as JSON Schemas
- **Therefore need to systematically semantize and reuse data standards**
  - ✓ Various approaches for XSD to RDFS/OWL
  - ✓ OMG Ontology Definition Metamodel defines mappings from UML to ontologies
  - ✓ OMG EXPRESS Metamodel defines a mapping from EXPRESS to UML
  - ✓ But nowadays software engineers often use light-weight approaches like JSON Schema

Polyglot Modeling

- **Technology-independent models, understandable by SMEs**
  - ✓ From these, generate various tech artefacts:
    - Diagrams, documentation
    - RDFS, OWL, SHACL, SHEX, JSON-LD Context and Frame,
    - JSON schema, Elastic object model, etc etc etc
  - ✓ "Write once. Use many. Creative laziness encouraged" [RAML]
- **Examples:**
  - ✓ FHIR: domain model to XML, JSON, JSONLD; XSD, JSON Schema, SHEX
  - ✓ LinkML: YAML-based models to various technical artefacts incl. JSON Schema, ontology, SHEX
  - ✓ Ontotext SOML: YAML-based model to SHACL and GraphQL (exposes KG parts to GraphQL)
  - ✓ Schema Salad: YAML-based model to JSON Schema and RDFS
  - ✓ A.ML and CloudInformationModel: YAML to RDFS, SHACL, SQL, R2RML
  - ✓ RAML (RESTful API Modeling Language): YAML to APIs (ala OpenAPI specifications)
- **Many of these are in YAML**
  - ✓ YAML-LD work group started (part of JSON-LD CG)
  - ✓ yaml-ld#19 is about polyglot modeling

Hybrid Storage

- **No need to convert everything to RDF triples!**
  - ✓ Use the most appropriate storage for each purpose
  - ✓ Integrate at the engine level to do cross-storage queries
- **Examples:**
  - ✓ Full-text and faceting/aggregation: connectors to Lucene, SOLR, Elastic
  - ✓ Large-scale documents/objects: Mongo connector, JSON→JSONLD→RDF
  - ✓ Geospatial info: GeoSPARQL (asWKT and asGML literals)
  - ✓ Relational database virtualization: ONTOP, SPARQL to SQL
  - ✓ BI integration: SPARQL views exposed as relational tables, JDBC connector, SQL to SPARQL
  - ✓ Kafka connector (in and out of KG)
  - ✓ GraphQL querying of KGs (ONTO Semantic Objects, TQ GraphQL, Communica, etc)
  - ✓ Time series data (experimenting)
  - ✓ Scientific, engineering, AECO data: HDF5 connector (upcoming)

TheFSM
The Food Safety Market

SME-powered industrial data platform to boost the competitiveness of European food certification

BIG DATA VALUE PUBLIC-PRIVATE PARTNERSHIP

TheFSMeu | @TheFSMeu | TheFSM H2020 | www.foodsafetymarket.eu

Agroknow | ontotext | UBITECH ubiquitous solutions | AGRIVI | tracelabs originTrail Core Developers

universität wien | WAGENINGEN UNIVERSITY & RESEARCH | TÜV AUSTRIA HELLAS | TÜV AUSTRIA ROMANIA | VALORITALIA | TÜV AUSTRIA CYPRUS

TheFSM has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 871703.

ontotext

Thank you!

Contact: vladimir.alexiev@ontotext.com

ontotext