Data Spaces vs Knowledge Graphs

How to Get To Semantic Data Spaces?

Vladimir Alexiev 0000-0001-7508-7428
Chief Data Architect, Ontotext Corp (Sirma AI)

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

Abstract

EU invests heavily in Data Spaces (DS) as a mechanism to enable commercial data exchange and therefore industry digitalization and proliferation of Data Science (DS) and Artificial Intelligence, in particular Machine Learning (ML). While DSs use heavily semantic technologies, that is limited to describing metadata, license agreements, data market participants, etc. I argue that using Linked Data and semantic technologies for the data itself offers significant benefits regarding more efficient data sharing and use, and improvements to ML and DS processes. I give an overview of the state of semantic data sharing in several industrial domains (Product Classifications and Catalogs, Manufacturing Industry, Electricity, Transport and Logistics, Architecture and Construction; and close with a brief overview of technological enablers required for Semantic Data Spaces.

Introduction

The EU is investing heavily in Data Spaces (DS) and related legislation initiatives (eg [EU strategy]) and commercial incentives to facilitate the sharing of data. This covers both commercial and open data, across projects, institutions, cross-organization initiatives, whole industries, and across industries. EC believes that data sharing is the most important factor in enabling industrial digital transformation and the use of Artificial Intelligence (in particular Machine Learning), which is seen as an enabler of better competitiveness, contributing to EU's Green Deal goals and to post-COVID economic recovery.

DSs already use semantic technologies heavily in describing essential components of data sharing: **datasets** and related **metadata**, licenses, participants, users, access rights, use and commercial agreements, etc. [IDS RAM] is based on semantic web specifications and ontologies.

However, few if any data spaces use semantics to represent the **data itself**, and fewer still use **Linked Data** (LD) principles. IDSA connectors are based mostly on data **exchange** and the harmonization of data models is left to individual industries (and often does not happen).

In most cases **semantic harmonization** based on ontologies and shapes, and **sharing and federation** based on LD principles can offer significant benefits in terms of efficiency of data provisioning and use, timeliness and locality of information. While IDS RAM and the EU interpret Data Sovereignty mostly as a legal term (where is data hosted, and how to ensure the rights of data owners are protected), I believe that Sovereignty also has important technical connotations: who masters what data, how to ensure that data always uses the most recent version of referenced data, where to find that latest version.

LD principles dictate that the up-to-date Version of Record of a fine-granularity piece of data will always be found at a certain IRI. **Knowledge Graphs** (KG) based on semantic web and LD have been used extensively in the last 5 years to enable the creation of holistic and comprehensive knowledge bases in certain domains. Prominent semantic web conferences include [ISWC] since 2001 and [ESWC] since 2004. Semantic web conferences with industrial focus [i-SEMANTiCS] started in 2004-2009. KG are featured as one of the most prominent topics at these conferences in the last 4-5 years, and dedicated KG conferences [KGC] appeared in 2019.

KGs built on open data are well-known in life sciences, chemistry, biodiversity, agriculture, climate science, cybersecurity, bibliography (science KGs) etc. But many commercial KG projects in a variety of areas have also been implemented (commercial semantic web companies like Ontotext see unflagging interest in this topic), eg:

- Information: Google KG, Pinterest KG
- Products and shared services: Amazon product graph, AirBnB KG, Uber KG
- Media and publishing: BBC, Financial Times, OUP
- Financial and insurance: Wells Fargo, Capital One, Refinitiv, Pitney Bowes, JP Morgan Chase, Chubb, SIX Swiss Exchange

I believe that by using LD principles and semantic technologies not just for metadata but also for the actual data, EU DS can reap significant benefits. In addition to more efficient data sharing and use, this includes improvements to Machine Learning and Data Science processes. The topics of Machine Learning and Knowledge Graphs receive increasing amounts of research [Google Scholar].

Industrial Examples

This section is a brief overview of data sharing in several industrial domains.

Product Classifications and Catalogs

Interoperable information about products is of crucial importance for automating e-commerce and manufacturing data flows. Such data includes product classifications, their parts and characteristics (attributes). Standards that define data models for representing products include:

- ISO 13584-42 and IEC 61360-2 (parts library, PLIB)
- IEC 62656 (parcelized ontology model, POM)

National, international or vertical classifications include eCl@ss, EU CPV, UNSPSC, GS1 GPC, ECALS, NAMUR, RosettaNet, PFI, eOTD, RNTD, BMEcat, bSI bSDD, COBie.

Unfortunately, many of these classifications don't have commonly accepted semantic representations and not even stable URLs. Some examples:

EU CPV (Common Procurement Vocabulary) includes an 8-digit taxonomy with 10,250 nodes (as of 4 June 2020) and 904 Supplementary codes. They don't have official URLs but stable URLs are available at data.ac.uk, eg http://cpv.data.ac.uk/code-24111700.html (nitrogen) and http://cpv.data.ac.uk/code-JA14.html (WAN). A simple semantic representation based on SKOS is available, eg http://cpv.data.ac.uk/turtle/code-24111700.ttl.

GS1 GPC (Global Product Classification) includes a taxonomy with 6073 nodes (categories) and 4 hierarchical levels "Segment>Family>Class>Brick" and an attribute/value system with 13284 attributes and values (as of 18 Dec 2020). See extensive discussion of GPC's structure at Wikidata: property P8957 and Property_proposal/GS1_GPC_brick_code#GPC_Scope_and_Structure.

GPC doesn't have a semantic representation. The official GPC browser https://gpc-browser.gs1.org/ doesn't show individual URLs. There is a test site with individual URLs, eg https://mh1.eu/gpctest/50260000: Vegetables (Non Leaf) - Unprepared/Unprocessed (Fresh).

eCl@ss is an industrial initiative with over 4k client companies that cooperates with 80 leading industrial and standardization organizations including ISO, IEC, CEN, buildingSmart International, DIN, ETIM, Applia, etc. The eCl@ss classification is in its version 12 and includes 19k classes and several thousand properties. **IEC CDD** (Common Data Dictionary) includes a number of product classifications and catalogs, including IEC 61987 Process automation, ISO 23584 optics, IEC 13584 measuring instruments, IEC 60721 environmental declaration, IEC 61360 Electric/electronic components,IEC 62683 Low Voltage switchgear, IEC 62474 Material declarations. Both use the ISO 13584-42 model (PLIB or OntoML). That model is based on idiosyncratic information artifacts such as classification classes, characterization classes, application classes, blocks, aspects, etc. It is far from the real world of economy/manufacturing that has products, product classes, properties, manufacturers, documents, spec sheets, etc; thus doesn't follow [Ontological Realism] as exemplified e.g. by https://schema.org. For identification of classes, properties and elements, they use IRDI instead of IRI that are not resolvable and not permanent as they carry a version number (eg IRDI "0173-1#02-AAO677#002" stands for a property “Manufacturer name”). The IEC CDD License is semi-open and only partial dumps are available; eCl@ss is closed. Neither reuses any ontologies or LOD datasets

Manufacturing Industry

There are various national industry digitization initiatives (Industry 4.0) of which the German Plattform Industrie 4.0 is arguably most advanced. Its main technical achievements are the Reference Architectural Model Industrie 4.0 (RAMI) and the Asset Administration Shell (AAS) model. AAS allows incorporating important industrial data exchange standards such as OPC UA, AutomationML, Collada, and eCl@ss. AAS has schema definitions in UML, XML schema, JSON schema and RDFS; and data renditions as XML, JSON, RDF. The RDF rendition and ontology is presented in [AAS Part 1] Annex G "RDF Schema and Complete Example". However, the RDF rendition of AAS does not follow LD and semantic principles. Rather than web-accessible property definitions, it copies definitions from other standards locally (e.g. below, from eCl@ss), using blank nodes and IRDIs that are not referenceable, e.g.:

```
aas_submodel:submodelElement [
  a aas:Property;
  rdf:subject <http://i40.customer.com/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"];
```

The same copying of definitions is shown in [AAS ECLASS], Fig 18 and Fig 19 "Referencing an ECLASS Property".

The AAS ontology uses many "unnatural" semantic constructs. E.g. in addition to XSD datatypes, it uses individuals representing "legacy" types in the iec61360: namespace, eg STRING_TRANSLATABLE, REAL_MEASURE, REAL_COUNT, REAL_CURRENCY. The ontology is not available as a separate file, but as a text dump in a PDF. There does not seem to be an issue tracker. Its namespace (eg http://admin-shell.io/aas/2/0/ or https://admin-shell.io/aas/3/0/RC01/ does not resolve. It uses SHACL for validation, but also some improper SHACL constructs (eg sh:pattern inside an owl:DatatypeProperty, and sh:pattern ".+", which is better expressed as sh:minLength 1).

Electricity

IEC 61970 Energy management system API defines a Common Information Model (CIM) that is defined as UML (IEC 61970-301 CIM Base and a number of application standards), RDFS (IEC 61970-501), OWL (IEC 61970-505), RDF-XML (IEC 61970-552). CIM defines electrical network equipment and connectivity, Generation facilities, Static Transmission Network Model, Solved Power System State, Dynamics Profile, Diagram (Schematics), etc. It is also the foundation of a number of additional standards, eg:

- IEC 61968: Application integration at electric utilities - System interfaces for distribution management
- IEC 62361: Power systems management and associated information exchange - 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 management system and the power management system

CIM is especially important in the EU due to Europe's single energy market (coordinated by ENTSO-E) as it is the foundation of [CGMES] and market data exchange:

- IEC 62325: Energy Market Communication (Exchange)
- IEC 61970-600-1: CGMES Structure and rules
- IEC 61970-600-2: CGMES Exchange profiles specification

CIM plays a prominent role in the IEC 63200 Smart Grid Architecture Model (SGAM) and the IEC Smart Grid Roadmap and Mapping Tool; also see [EKG].

Global Energy Identification Codes [EIC] exist that identify 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). But despite the existence of EIC, CIM does not use IRIs for entity identification: it uses temporary UUIDs. CIM is used only as an exchange format; its RDF XML is non-standard in its handling of model graphs (used for differential models).

New IEC 61970 parts in development define JSON-LD exchange and RDF shapes, and posit the use of permanent UUIDs. But still, LD principles are not followed: there is no conception of distributed mastering and federation of CIM data.

Ontotext is making inroads in the use of KGs for electricity, see [TEKG] and [TEKG spec].

Transport and Logistics

GS1 is the global standards-setting organization in transport and logistics. In several decades it has made the transition from paper barcodes to RFID to logistics master data as LD. Various GS1 standards have a semantic rendition:

EPC Tag Data Standard (TDS) defines a number of global identifier schemes for objects in the logistics chain, including 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).

GS1 Digital Link defines web-resolvable URLs for TDS identifiers, using either the global GS1 resolver (eg https://id.gs1.org/gtin/9506000134352?linkType=all) or per-company resolvers. About 65 GS1 Link Types are defined that include various kinds of product information. A new link type serves logistics master data in JSON-LD using the GS1 Web Voc (see below).

The Electronic Product Code Information System (**EPCIS**) standard defines "object visibility" events for tracing objects across the logistics chain. The Core Business Vocabulary (CBV) defines a number of nomenclatures to be used in EPCIS. EPCIS 2.0 has a JSON and semantic rendition. Ontotext contributed to the development of GS1 EPCIS 2.0 as follows:

- Contributed issues, bug reports, best practices about publishing ontologies, semantic resolution, etc;
- Created the EPCIS Semantics document;
- Created mappings of EPC/TDS identifiers to GS1 classes, and additions to the GS1 class hierarchy;
- Improvements to the EPCIS ontology and RDF shapes;
- Specific proposals for gs1:CertificationDetails and gs1:MeasurementType;
- Meta-properties to enable the generation of “dual” documentation of the JSON model and RDF (ontological) model

GS1 Web Voc is an extension of schema.org that adds a number of product-specific classes, properties and nomenclatures. However, it shows some problems rooted in its legacy in older EDI standards and XML-centric modeling approaches. For example:

- gs1:Country is a bit confused whether it is a country (gs1:countryCode) or country subdivision (gs1:countrySubdivisionCode)
- Does not allow the expression of any other geo gazetteers except ISO 3166 and ISO 3166-2
- The fields Country.countrySubdivisionCode and PostalAddress.addressRegion are redundant with respect to each other

See WebVoc issues for more examples. More importantly, GS1 has a large number of data standards that need to be unified with each other, and in the process can be modernized with a semantic rendition.

Architecture and Construction

The Architecture, Engineering, Construction and Operation (AECO) community is rapidly becoming a prominent user of LD. There is a thriving Linked Building Data (LBD) W3C community and the [LDAC] workshops have been organized since 2012.

The most prominent standard for describing architectural designs, built assets and construction projects is [IFC] (ISO 16739-1). It is defined in ISO 10303-11.2 EXPRESS and has renditions as XML schema, JSON schema, OWL ontology. IFC payload has renditions as STEP (text), XML, JSON, RDF, HDF5. The "canonical" IFCowl ontology has a heavy EXPRESS heritage and does not represent relations and datatypes in "naturally", leading to complex and heavy RDF representation. That's why a number of alternative IFC RDF representations have emerged with streamlined geometry data, or interfacing to HDF5 for binary storage.

- 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)
- BRICKS and Haystack: industry schemas for describing building assets and IoT devices
- Digital Buildings (Google) and Real Estate Core (Microsoft and collaborators) for facility management and IoT
- Smart Appliance Reference architecture (SAREF) for IoT devices and its extensions SAREF4ENER for energy and SAREF4BLDG for buildings
- ISO 50008 Building energy data management for energy performance, Smart Energy Aware Systems (SEAS) ontology for energy efficiency and smart grid interactions
- Semantic Sensor Networks (SOSA/SSN)
- LBD ontologies: PRODUCT Ontology, PROJECT Management, Properties evaluation (PROPS), Ontology for Property Management (OPM), Building Topology Ontology (BOT), Ontology for Managing Geometry (OMG), Ontology for Geometry Formats (FOG), Geometry Metadata Ontology (GOM), RDF-based geometry (GEOM), Building Product Ontology for assembled products (BPO)

An especially hot topic is semantic Asset Management of buildings and infrastructure. A number of standards in progress are getting a semantic rendition, including Data Templates, Object Type Libraries, Specification Libraries, Product Catalogs, Common Data Environments. The use of decentralized semantic environments for AECO data is also considered [LBD SOLID; Decentralised].

Semantic Integration and Polyglot Data

Here we briefly list some technical enablers for Semantic Data Spaces:

- Polyglot Modeling approaches. Most industries have extensive data exchange standards, which come in a variety of formalisms. For KG integration, it is necessary to harmonize them to semantic formats. Examples:
  - [ODM] defines mappings from UML to ontologies that can repeat the systematic derivation of RDFS/OWL ontologies from UML models as done for Electrical CIM
  - [EXPRESS metamodel] defines a mapping from EXPRESS to UML that can modernize EXPRESS schemas such as IFC
  - FHIR is a technology-independent schema in the healthcare domain that is rendered as XML & XML schema, JSON & JSON schema, RDF & SHEX
  - LinkML is a technology-independent modeling language based on YAML that can generate various technical artfacts, including JSON-Schema, ShEx, RDF, OWL, GraphQL, and SQL DDL
  - Semantic Objects Modeling Language (SOML) is a simple modeling language based on YAML and used in the Ontotext Platform. It can be generated from RDFS/OWL/schema ontologies (owl2soml; also see soml on github) and can generate SHACL shapes, GraphQL schemas, and translate GraphQL queries to SPARQL.

- Knowledge Graph storage, querying and inference using centralized semantic repositories or decentralized approaches (SOLID)
- Connectors and Hybrid Storage technologies, including relational-RDF virtualization, GraphQL querying of KGs, RDF-as-relational querying, connectors from RDF to full-text and faceting engines (Lucene, SOLR, Elastic), Kafka connector, HDF5-SPARQL, etc.

References

- [AAS ECLASS] Modeling the Semantics of Data of an Asset Administration Shell with Elements of ECLASS. Industrie 4.0 whitepaper, 29 Jun 2021
- [AAS Part 1] Asset Administration Shell Part 1 - The exchange of information between partners in the value chain of Industrie 4.0. Version 2.0.1, Nov 2019 and May 2020
- [CGMES] Common Grid Model Exchange Standard (CGMES) Library. ENTSO-E. Accessed 1 May 2022
- [Decentralised] Pattern-Based Access Control in a Decentralised Collaboration Environment. W.Jeroen, R.Taelman, R.Verborgh, P.Pauwels, J.Beetz, E.Mannens. Linked Data in Architecture and Construction (LDAC 2020), CEUR Workshop Proceedings. Dublin, Ireland: CEUR 2636, 2020.
- [DIN 27070] DIN SPEC 27070. Requirements and reference architecture of a security gateway for the exchange of industry data and services. 21 Feb 2020
- [EIC] Energy Identification Codes (EICs). ENTSO-E
- [EKG] Energy Knowledge Graphs to Facilitate Evolution of the European Energy Market, C.Ivanov and V.Alexiev. Presentation at Ontotext Knowledge Graph Forum.
- [ESWC] Extended Semantic Web Conference. DBLP bibliography. Accessed 1 May 2022
- [EU strategy] A European Strategy for data | Shaping Europe's digital future. Accessed 1 May 2022
- [EXPRESS metamodel] Reference Metamodel for the EXPRESS Information Modeling Language. OMG. Version 1.1, 1 May 2015
- [Google Scholar] Machine Learning and Knowledge Graphs. Google Scholar. Accessed 1 May 2022
- [i-SEMANTiCS] International Conference on Semantic Systems (i-SEMANTiCS). DBLP bibliography. Accessed 1 May 2022
- [IDS-RAM] International Data Spaces Association: Reference Architecture Model. Version 3.0, April 2019
- [IFC] Industry Foundation Classes (IFC). buildingSMART International. Accessed 1 May 2022.
- [ISWC] https://dblp.org/db/conf/semweb/index.html, DBLP bibliography. Accessed 1 May 2022
- [KGC] The Knowledge Graph Conference. Accessed 1 May 2022
- [LBD SOLID] Towards a decentralised common data environment using linked building data and the SOLID ecosystem. J.Werbrouk, P.Pauwels, J.Beetz, L.van Berlo. Proceedings of the 36th CIB W78 2019 Conference
- [LDAC] Linked Data in Architecture and Construction. Accessed 1 May 2022.

- [ODM] Ontology Definition Metamodel. OMG. Version 1.1, 2 Sep 2014
- [Ontological Realism] Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies. B.Smith and W.Ceusters. Applied Ontology 5, 139-188. 2010
- [TEKG spec] Transparency EKG Requirements Specification, Architecture and Semantic Model. V.Alexiev, V.Ribchev, M.Chervenski, N.Tulechki, M.Radkov, A.Kunchev, R.Nanov. Ontotext, 14 Apr 2022
- [TEKG] Transparency Energy Knowledge Graph, V.Alexiev. Presentation at Joint INTERRFACE Open Call Projects meeting, 31 Jan 2022