IMEKO TC6 International Conference on Metrology and Digital Transformation
September 19 – September 21, 2022, Berlin, Germany

AN INTRODUCTION TO LINKED DATA AND THE SEMANTIC WEB

Clifford Brown [a], *Daniel Hutzschenreuter*, [a, *], *Julia Neumann* [a]

[a] Physikalisch-Technische Bundesanstalt (Department 9.4 Metrology for Digital Transformation), Braunschweig and Berlin, Germany
* Corresponding author. E-mail address: daniel.hutzschenreuter@ptb.de

Abstract – For a sustainable transformation of processes and services from metrology into a digital age, data needs to become more comprehensible to machines. Linked Data and tools from the Semantic Web are promising developments helping to reach this goal. We give an introduction to relevant concepts and tools that metrologists are being increasingly confronted with in the scope of digital transformation.

Keywords: linked data, semantic web, overview, metrology

1. INTRODUCTION

In 2008 Tim Berners-Lee gave a seminal and visionary presentation on the future of information in the form of data on the web [1]. In this presentation he discussed what has been referred to as the third wave in the development of the web, or Web 3.0 (Web 1.0 being static pages; Web 2.0 being active pages driven by user request). In this third wave, which Berners-Lee referred to as the Semantic Web, information data would be the driving force for change. Up to that point in time, a user could see downloaded information on the screen, typically in the form of a table, but had no way of getting that data other than to physically copy it from the screen onto paper and then copy again into a computer. (There were ways of "scraping" data from the original HTML but this was non-trivial and not available to the general public).

Berners-Lee's idea was to make data easily available to the general user. People would be able to have instant access to the data they downloaded, almost like opening a spreadsheet file of data on a computer. Then, invariably, users would want to combine this downloaded data with other data, either their own, or with other downloaded data. However, to combine data from two different sources requires full knowledge of both datasets. The user needs to know the details of the data; the information data alone does not provide this knowledge. The user needs the information behind the data, or the data behind the data, otherwise known as metadata [2]. If the original provider of the data has done a good job, they would provide this information e.g., units of measure (and uncertainty) and a full description of what the data is providing a "measure of". Again, Berners-Lee's idea was that the metadata would be provided along with the information data.

So how can the core data and all the relevant metadata be provided? Even before that, how can the originator of the data store all this information? The immediate answer might be a database of some form. Unfortunately, there is immediately a problem. Standard databases are great for storing information about systems that do not change, or where the types of information do not change significantly over time. This situation is very rare (even when standard databases are used), and totally unsuitable for the ever evolving and dynamic nature of data and associated metadata in science. A storage system that can evolve with this ever-increasing knowledge resource is required. The solution is Linked Data storage. The concept behind Linked Data storage is very simple. Instead of many complex interlinked tables of data, typically referred to as a database schema for the standard database approach, Linked Data storage in principle only needs one table which has only three columns.

The logic behind these three columned tables, or "triplestores" as it is referred to is based on Description Logic [3]. The three parts of the triple concept are made up of: "a subject"; "a predicate"; and "an object". Consider Table 1 which shows a table of information about some children and their physical attributes and interests. It can be un-pivoted to create the equivalent triples shown in Table 2.

Table 1. Child information table

	Age (Years)	Height (m)	Interest
Fred	6	0.9	Dinosaurs
Alice	8	1.2	Aircraft

Table 2. The triple table

Subject	Predicate	Object
Fred	has Age	6
Fred	has Height	0.9
Fred	has Interest	Dinosaurs
Alice	has Age	8
Alice	has Height	1.2
Alice	has Interest	Aircraft

Both tables 1 and 2 contain the same core information simply organised in a different way. Table 1 contains 6 pieces of information; table 2 contains 6 entries. Notice how the column headings (Predicates in the triple table), and the row identifiers names (Subjects in the triple table) have now become part of the table.

However, you will see that metadata is missing. The Age and Height information have lost their units of measure.

Moreover, something else has been lost; this is the fact that the Name of the child in the Table 1 was the unique identifier for each row. This is important information and needs to be replaced. Table 3 shows how this metadata is provided using the triple format.

Table 3 Metadata for Child Information Table

Subject	Predicate	Object
Age	has Unit	Year
Height	has Unit	meter
Child	has Attribute	Name
Child	has Attribute	Age
Child	has Attribute	Height
Child	has Attribute	Interest

The Child concept was important in the original table and continues to be important in the "triples" way of viewing the data. This is a fundamental point. Conceptualisation is important in the world of standard database design, as it largely defines the important tables in schema design. It is even more important in the Semantic Web paradigm. As we have moved away from a multi-table design to, in principle, a single-table design, we must be clear about our concepts because they will all be stored together.

As we have seen the "triples" storage, linked-data format is a very powerful tool for storing information. Today, the concept is defined within the Resource Description Framework (RDF) [4].

2. THE SEMANTIC WEB TOOL STACK

At the beginning of development in the Semantic Web, the RDF [4] and Extensible Markup Language (XML) [5] were the first building blocks. Since then, it has been continuously developed and enhanced by main standards organisation for the internet which is known as the World Wide Web Consortium (W3C). Figure 1 provides an outline of today's components in the Semantic Web tool stack we are going to address in our overview. The structure of the tool stack is inspired by a similar figure from Tim Berners-Lee. Each of the components will be briefly introduced in the following subsections.

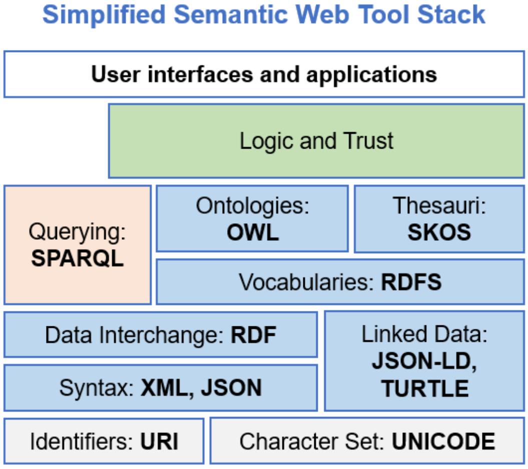


Fig. 1. Outline of the Semantic Web Tool Stack.

Machine-interoperable data represents one of the most demanding requirements to create value in support of today's digital transformation of processes and businesses. The Semantic Web components described in this work are important vehicles for improving a machine-to-machine communication of data that will be part of many digital metrological applications over the medium- and long-term.

2.1. Identifiers and Character Sets: URI and UNICODE

Each piece of data, its relation to other pieces of data, and underlying defining schemes and semantics are provided with worldwide unique identifiers. These identifiers build the anchor points for establishing a highly interconnected data network (web) where machines can reason on data and implicitly infer hidden knowledge. Typically, Uniform Resource Identifiers (URIs) are used [6].

All underlying formats for syntax, linked data, taxonomies, thesauri, and ontologies are typically represented by human readable characters in text form. The Semantic Web uses UNICODE (Universal Code Character Set) encoding [7] and its implementation by UTF-8 (Unicode Transformation Format – 8 bit) to ensure an unambiguous exchange between the text form and its binary encoding (e.g., for data storage or data transmission).

2.2. Syntax: XML and JSON

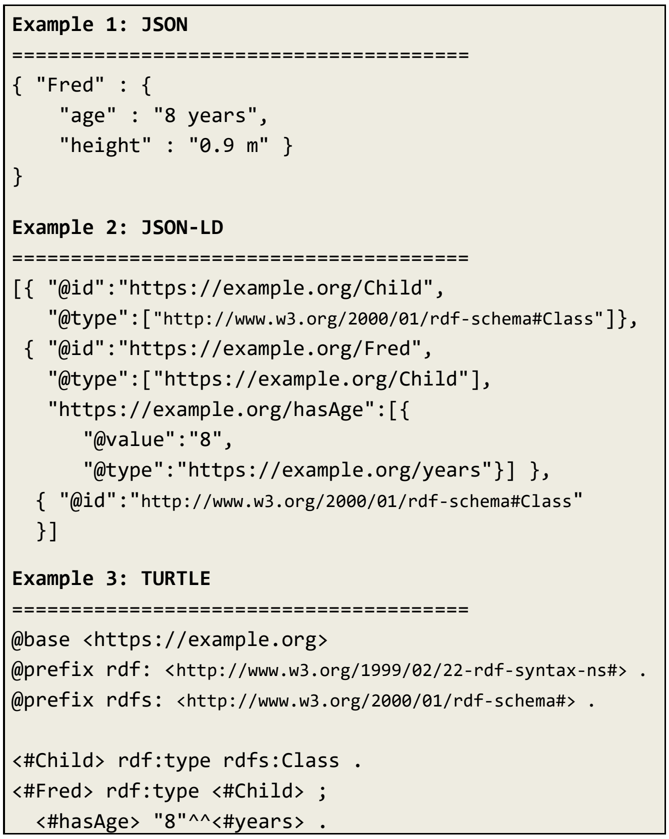
XML and JSON (Java Script Object Notation) [8] are two leading formats with a well-defined syntax for writing data in a way suitable for machine-reading and data exchange. Both allow the representation of data in a tree-like structure, where all nodes (root, branches, and leaves) have a name (called element tag). Figure 2 shows a JSON example providing information for 'Fred'. JSON is preferred today by many developers for Web programming interfaces as it is simpler and faster than XML. Beside traditional relational data bases (e.g., SQL), newer document data bases are established on XML and JSON structured data. Early implementations of RDF were based on XML, while current work is primarily applying JSON and even more lightweight formats (see Section 2.4).

2.3. Data Interchange and Vocabularies: RDF and RDFS

Basic aspects of RDF were already described at the introduction. Thus, we will briefly discuss vocabularies provided by the RDF Schema (RDFS) for modelling Subject-Predicate-Object triples following common principles [9]. RDFS introduces a specification of classes and properties. In our example from Table 2, further lines could be added defining a) that 'Child' is a RDF class (Child rdf:type rdfs:Class), and b) that 'Fred' is an instance of class Child (Fred rdf:type Child). Note, that 'rdf:type' is a property used as a predicate. Classes and their instances can be used for all three components of RDF triples.

2.4. Linked Data: JSON-LD and TURTLE

In computer science, Linked Data, or sometimes graph data, refers to structured data that is linked with other sources of data following the goal of making the data more meaningful for queries by machines. The information being linked is commonly providing metadata describing the data. The attribution of IRIs is fundamental to establishing linked data.



```
Example 1: JSON
======================================
{ "Fred" : {
    "age" : "8 years",
    "height" : "0.9 m" }
}

Example 2: JSON-LD
======================================
[{ "@id":"https://example.org/Child",
   "@type":["http://www.w3.org/2000/01/rdf-schema#Class"]},
 { "@id":"https://example.org/Fred",
   "@type":["https://example.org/Child"],
   "https://example.org/hasAge":[{
      "@value":"8",
      "@type":"https://example.org/years"}] },
  { "@id":"http://www.w3.org/2000/01/rdf-schema#Class"
  }]

Example 3: TURTLE
======================================
@base <https://example.org>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<#Child> rdf:type rdfs:Class .
<#Fred> rdf:type <#Child> ;
  <#hasAge> "8"^^<#years> .
```

Fig. 2. JSON, JSON-LD and TURTLE example for child ‘Fred’ who is 8 years old and 0.9 metres high.

JSON-LD is the JSON based serialisation of linked data. With the JSON syntax as a basis, it introduces a fixed set of node names (element tags starting with symbol ‘@’) to establish a description of the context (the meaning) of the data [10].

Example 2 in Figure 2 shows JSON-LD data where the context attributes ‘@id’ and ‘@type’ are used to link the information that ‘Fred’ is a child and what his age is. Furthermore, ‘@id’ identifies an IRI.

TURTLE (Terse RDF Triple Language) is a lightweight implementation for RDF [11] and linked data. Its syntax is based on a minimalistic set of controlling characters making it suitable for larger data sets. Example 3 in Figure 2 is the TURTLE encoding equivalent to the previous JSON-LD example for ‘Fred’. The ‘@prefix’ attribute links to relevant data and metadata. A line is always starting with the subject (e.g., <#Fred>), followed by the predicate and the object separated by a blank space. The end of a triple is marked by a dot. A subject can have multiple predicate-object associations using the semicolon as separator.

Both, JSON-LD and TURTLE allow linking of controlled vocabularies (e.g., thesauri and ontologies) and inclusion of data type primitives from XML (e.g., xs:double, xs:boolean). It is frequently found in real-world data.

2.5. Knowledge Representation: SKOS and OWL

Having in place RDF and linked data still leaves open how to best use these tools allowing machines to understand the logic and meaning behind the data (Semantic). The layer of knowledge representation provides means for filling this gap.

SKOS (Simple Knowledge Organisation System) provides a framework for defining ‘concepts' and relations between these concepts. It improves the portability of data between different domains, sources, and machines [12]. Element ‘skos:Concept’ for instance is an RDF object used to identify RDF subjects as concepts (e.g., child or measuring instrument are concepts). The predicate ‘skos:closeMatch’ provides the means to define two concepts to be very similar (e.g., a measuring instrument is a close match to a measuring device). SKOS can be used to implement Thesauri that list terms in groups of synonyms and related concepts.

Ontologies traditionally refer to outcomes from a harmonisation process in philosophy grouping all being things and basic structures of our reality in one unambiguous system of knowledge [13]. In computer science, an ontology stands for a formal representation system of knowledge in a specific domain. The Semantic Web includes **OWL** (Web Ontology Language) as a framework for implementing such ontologies [14]. It defines entities (e.g., ‘owl:class’ or ‘owl:objectProperty’), expressions (e.g., owl:equivalenClass), and logical axioms.

Data based on RDF and the Semantic Web is typically not only using a single tool for interlinking data, metadata, and knowledge, but attributes from RDFS, SKOS, OWL and many more resources are extensively combined to provide domain-specific data.

2.6. Queries: SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) is a harmonised language for realising queries through linked and RDF data and metadata [15]. It fully exploits the benefits of the Semantic Web concept allowing to describe queries with search criteria bringing together data and metadata from different domains.

3. EXAMPLES USING SEMANTIC WEB TOOLS

This section lists some examples for data using Semantic Web approaches with relevance in measurement science. Due to the limited space descriptions are kept very brief.

QUDT (Quantities, Units, Dimensions, and Types) is an implementation of an ontology for entities and properties of unit and quantity systems [16]. The most recent version (2.1) is encoded by TURTLE using attributes from OWL, RDF, RDFS, SKOS, and other vocabularies.

Semantic information in sensor networks were investigated in [17] giving a comprehensive use-case for a combination of various schemes and ontologies from sensor networks, digital representation of quantities and units of measure, as well as mathematical properties.

A FAIR dataset publication example from EMPIR project Met4FoF [18] uses schemes for SI-based metrological data, and publications, as well as ontologies for quantities and sensors to describe metadata encoded in JSON. The aim is a FAIR data publication allowing better reuse of the date with machine-learning.

The TURTLE metadata model from NFDI4Ing, an engineering consortium within national German research data infrastructure (NFDI), shows an extensive example for an interoperation of twenty different data schemes and ontologies [19].

FIWARE, the worldwide leading (open source) ecosystem for Smart Cities has found Linked Data to be the technology of choice to address the issue of data integration in massive systems of heterogeneous sensors and actors [20]. All data and metadata are based on JSON-LD, where its use follows the ETSI NGSI standard (Next Generation System Information).

4. DISCUSSION

A usage of Linked Data and tools from the Semantic Web to encode data and metadata describing its context offers a great potential in the future to enable machines to better interpret and interoperate data from different sources. It can help to break up today's silos of proprietary data that exist within organisations and at external sources. A combination of data from different sources and domains is very promising with respect to facilitating new research findings from application of methods from data science and machine-learning. Cyber-Physical-Systems and autonomous processes in Industry 4.0 for example envision an extensive communication of and action on data by machines requiring comprehensibility. Furthermore, many examples for FAIR (findable, accessible, interoperable, and reusable) data are based on Semantic Web approaches to achieve interoperability.

While looking at the relevance and benefits of Linked Date and Semantic Web, their challenges for an application should also be considered. The underlying technical framework is very extensive and understanding the concepts, formats, and tools is not an easy task. The number of experts in this domain from metrology is very limited at present and a broader use will need extensive training in the core technology. In addition, expertise is needed to understand differences of data quality from a metrological perspective resulting from different ways of applying the core technologies. Furthermore, Linked Data can combine many different terminologies and data models across domains making it quite difficult to understand and requires new approaches for software interfaces to parse the data. It may be difficult for some organizations to access such data if policies are in place preventing the interconnection to external data and terminology resources.

The Semantic Web is an attempt to solve extremely difficult problems of interoperability. Fundamental to providing a solution to these issues is the problem of identity. Another key tenet of the Semantic Web is the openness to all voices which is fundamental to the operation of science, the "Open World Assumption" (OWA).

5. CONCLUSION

Linked Data and tools form the Semantic Web provide promising technologies for representing metrological data that is suitable for use by machines and allow better comprehension of knowledge within the data and a machine-driven inference of new knowledge. The RDF concept can help to create a flexible data storage and transmission environment capable of highly heterogeneous data contents that are still interoperable and ready for the application of First Order Logic.

Major drawbacks hindering the immediate use of Linked Date and Semantic Web tools are its high complexity and technically very abstract concepts. Enabling their usage in metrology requires not only a signification amount of training to gain expertise, but also necessitates a significant change in how companies and applications consume data when compared to traditional less interoperable data models.

Finally, implementations of data and semantics serving metrology have very different quality concerning an ability to interoperate measurement data in today's applications. Future research is needed to find sustainable quality requirements.

ACKNOWLEDGMENTS

We would like to thank our colleague Anupam Prasad Vedurmudi from PTB for his valuable contribution to this review of Linked Data and the Semantic Web. Furthermore, we like to thank the anonymous reviewers for their work.

REFERENCES

[1] T. Berners-Lee, J. Hendler, O. Lassila, *The Semantic Web*, Scientiffic American, 2001.

[2] Gartner IT Glossary: https://www.gartner.com/en/information-technology/glossary, (accessed 08/2022)

[3] Bader F., et al.; "The Description Logic Handbook: Theory, Implementation and Applications"; Cambridge University Press.

[4] Ressource Description Framework https://www.w3.org/TR/PR-rdf-syntax/, (accessed 08/2022)

[5] Extensible Markup Language (XML) 1.0 (Fifth Edition), https://www.w3.org/TR/xml/ (accessed 08/2022).

[6] Internationalized Resource Identifiers (IRI), https://en.wikipedia.org/wiki/Internationalized_Resource_Identifier, (accessed 08/2022).

[7] UNICODE, https://home.unicode.org/, (accessed 08/2022

[8] Java Script Object Notation, https://www.json.org/json-en.html, (accessed 08/2022).

[9] RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/, (accessed 08/2022).

[10] *JSON-LD 1.1, A JSON based Serialization for Linked Data*, https://www.w3.org/TR/json-ld11/, (accessed 08/2022

[11] RDF 1.1 Turtle, Terse RDF Triple Language, https://www.w3.org/TR/turtle/, (accessed 08/2022).

[12] SKOS Simple Knowledge Organization System Primer, https://www.w3.org/TR/skos-primer/, (accessed 08/2022).

[13] Wikipedia article on Ontology: https://en.wikipedia.org/wiki/Ontology, (accessed 08/2022)

[14] OWL 2 Web Ontology Language Primer (Second Edition), https://www.w3.org/TR/owl2-primer/, (accessed 08/2022).

[15] SPARQL Query Language for RDF, https://www.w3.org/TR/rdf-sparql-query/, (accessed 08/2022).

[16] Quantities, Units, Dimensions, Types Ontology, https://www.qudt.org/, (accessed 08/2022).

[17] M. Gruber et al., Semantic Information in Sensor Networks: How to Combine Existing Ontologies, Vocabularies and Data Schemes to Fit a Metrology Use Case, in: proceedings of IEEE International Workshop on Metrology for Industry 4.0 & IoT 2020.

[18] T. Dorst etal., *Sensor data set of one electromechanical cylinder at ZeMA testbed (ZeMA DAQ and Smart-Up Unit)*, 2021, https://doi.org/10.5281/zenodo.5185953

[19] Metadata4Ing, *An ontology for describing the generation of research data within a scientific activity*, https://w3id.org/nfdi4ing/metadata4ing/1.0.0/, (accessed 08/2022).

[20] FIWARE, https://www.fiware.org/, (accessed 08/2022).