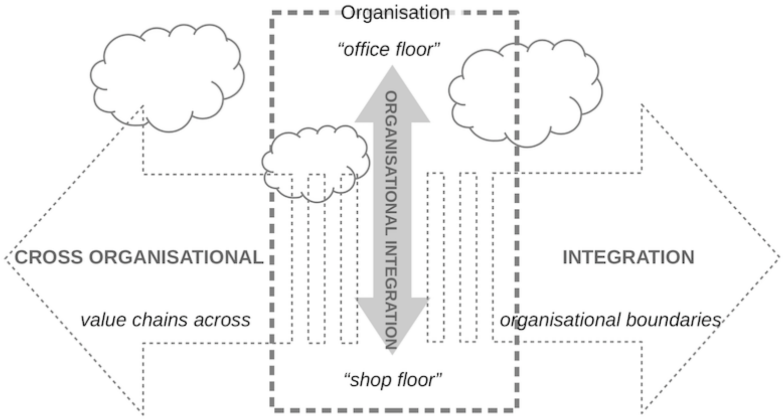
W3C Whitepaper on digitalisation and the role of standards

Dave Raggett, W3C
email: dsr@w3.org

W3C is a world-wide member-driven community devoted to the standardisation of Web related technologies. W3C organised a [workshop on Graph Data in March 2019] and is now seeking to bring stakeholders together in a Graph Standardisation Business Group to gather use cases and requirements across sectors, to guide technical standardisation work, and to coordinate liaisons with other standards development organisations and industry alliances.

Introduction

Data is increasingly important to businesses. The amount of data is rising dramatically along with the challenges for maximising the business value obtained from data assets and processes throughout the enterprise. Older approaches to data are no longer adequate and senior management is now seeking the means to manage data in ways that scale across the enterprise. This digital transformation (*aka* digitalisation) involving the adoption of advanced digital technologies is expected to have a profound impact across many sectors.



Michael Porter: A **value chain** is a set of activities that a business operating in a specific industry performs in order to deliver a valuable product or service for the market

In smart manufacturing, digitalisation is often referred to as Industry 4.0, and seeks to make companies more efficient and capable of supporting rapid reconfiguration to cater for predictive maintenance, bespoke orders and constant changes in market conditions.

Traditional means for collecting data using paper forms and spreadsheets are giving way to new approaches that are more amenable to data integration and data management. This emphasises the need for accurate information on the meaning of data. On a small scale this can be held in people's heads, or written down in supporting documents. However, this doesn't scale as employees change roles or the documentation is no longer up to date. In addition, software used to process data is likely to embody implicit assumptions about the data. This knowledge is hidden in the software and hard to maintain. The high cost of updating the software acts as an inhibitor when it comes to adapting to inevitable changes in business requirements, working with partners and integration with other datasets.

This points to the need to collect and maintain machine interpretable metadata about distributed heterogeneous information systems and processes throughout the enterprise. Such metadata includes the data types, master data, and information about the provenance, data quality, agreed terminology and so forth.

The Rise of Graph Data

Graph Data, Property Graphs and Knowledge Graphs: Graphs, with nodes (*aka* vertices) and labelled directed links (*aka* edges), are a very flexible way of representing data and metadata (information about data). Graphs can be reasoned over with rules, and processed with efficient graph algorithms. **Property Graphs** are a form of graphs in which you can annotate both nodes and links, e.g. to indicate temporal, spatial, provenance, and data quality information. **Knowledge Graph** is a term for graphs that describe concepts and relationships, often in the form of taxonomies. Knowledge Graphs were popularised by Google as a basis for smart search results, and are very relevant to businesses for enterprise wide data management, providing a convenient point of access to metadata from many data sources. Knowledge Graph can be considered as a more accessible equivalent to the term "ontology".

Traditional approaches to storing data include manually filled out paper forms, Word documents, PDF documents, comma separated value (CSV) files, spreadsheets and SQL/RDBMS databases in which data is held as a set of tables. Graph data models enable new use cases and in certain cases different approaches to existing ones. Using graph data either by itself or in conjunction with existing approaches has the following benefits:

- Easy and intuitive expressions of queries saving developer time
- Fast performance for graph traversal in complex graph queries
- Well suited to integrating data from heterogeneous sources
- Well suited for situations where the data model is evolving
- Better suited than noSQL stores when there are links within data

Graph data can be managed as an entity by itself or integrated as part of existing systems, e.g. using relational tables, JSON, XML, external tables, etc. Whilst RDF has an extensive suite of standards, there are unfortunately no common or standard languages or APIs used by commercial property graph databases at this time. Today, there are several offerings and initiatives proposed to address this: the Apache Tinkerpop APIs and Gremlin language; the work underway in ISO/IEC on extensions to the SQL standard to support access to Property Graphs, as well as initiatives like GQL, a proposed query language for property graph databases. W3C is seeking to standardize an interchange framework for integration across graph databases that makes it easier to agree on the meaning of data between suppliers and consumers.

There has been a rapid expansion in the number of Graph Database vendors:[1]

[1] *See :* https://www.bloorresearch.com/technology/graph-databases/

CRAY DATASTAX FRANZINC. GRAKN.AI

Graph Base MEMGRAPH Microsoft neo4j

Objectivity ontotext Oracle and Spatial Graph OrientDB

redislabs HOME OF REDIS Stardog TigerGraph TROVARES seriously scalable

and many more …

RDF as the foundation for graph data

RDF is W3C’s framework for Graph Data, and is associated with a mature suite of standards that have been widely adopted. RDF starts with binary directed labelled relationships, e.g. Mary loves Jack with Mary as the subject, loves as the predicate and Jack as the object. RDF features URIs for global identifiers for concepts and predicates. Where these URIs are based upon HTTP, the identifiers may be dereferenceable to obtain further information in a variety of formats, e.g. HTML, RDF/XML, Turtle, N3 and JSON-LD. Existing standards include exchange formats, schema languages (e.g. OWL), query/update with SPARQL, graph data validation (SHACL), and a wide range of vocabularies for different purposes.

Enabling Open Markets of Services with the Web of Things

The Internet of Things (IoT) has the potential to provide huge streams of data, along with the need for associated metadata and integration with many other information services. Unfortunately, the IoT suffers from fragmentation, and this is holding back its potential. W3C is addressing this with the Web of Things – an abstraction layer that connects suppliers and consumers of services, independently of the underlying IoT technologies and standards. Our approach allows suppliers to expose digital twins to consumers as software objects with properties, actions and events. Applications can thus interact with things independently of their location. Each thing is identified with a URI as basis for rich metadata (in JSON-LD) describing the interaction model as well as the context in which it resides, e.g. the kind of sensor, the units of measure, and what the sensor is measuring. The approach supports processing at the network edge, in the cloud, and in between.

The Web of Things Working Group has advanced the Thing Description specification to Candidate Recommendation status. To realise the potential for open markets of services further work is needed to clarify the requirements and to develop additional standards where they are needed, for example:

- Open markets require open standards[2] for suppliers and consumers to find each other and work together
- Standards for describing services, e.g. different kinds of sensors and actuators, and what they measure or control (this will involve the need for many vocabularies)
- Standards for describing the software interfaces and data formats (the Thing Description specification)
- Standards for terms and conditions for service contracts (including liability)
- Standards for payments (W3C has some work in progress on payments)
- Standards for security and for enabling trust between suppliers and consumers, taking into account the regulatory framework and legal recourse

Not all of this is in place – we need your help! The Web of Things is closely coupled to W3C's work on RDF and graph data in respect to the role of data and metadata to support services across the Internet.

Sentient Web = digital twins + graphs + AI/ML

The **Sentient Web** is defined as ecosystems of services with **awareness** based upon sensors and information services, and **reasoning** based upon graph data and rules together with graph algorithms and machine learning. This definition subsumes the Web of Things and stresses the importance of reasoning[3], artificial intelligence and machine learning.

Work Currently in Progress

W3C has ongoing work for JSON-LD, an exchange format for RDF based upon JSON, on updates for metadata for Dataset Exchange, and on the use of JSON-LD to describe the interaction model for the Web of Things. New work is under consideration for base directionality of RDF string literals. Work still in the incubation phase includes the Easier RDF initiative which seeks to make graph data and rules easier for the average developer. This includes ideas for bridging the gap with Property Graphs, and opportunities for using rule languages in place of traditional application software, with the means to efficiently support distributed sources of data and metadata. In addition, W3C has many Community Groups working on topics related to graph data.

Looking further out

Work on big data and analytics has focused on ways to identify patterns and to extract value. This frequently involves statistical techniques. By contrast, much of the work on ontologies for the Semantic Web has focused on logical deduction. This faces difficulties in

---

[2] See OpenStand's principles for open standards as jointly defined by IEEE, ISOC, W3C, IETF and IAB.

[3] In practice, reasoning may be based on software for graph traversal and manipulation, rather than formal logical deduction.

respect to incomplete, uncertain and inconsistent data that is likely to include errors. A limited work around is to apply rules of thumb for cleaning data prior to further processing.

To more effectively address the challenges of real-world data, we need to embrace rational forms of reasoning that exploit prior knowledge and past experience. This involves the need to blend symbolic knowledge (graph data and rules) with computational statistics, i.e. sub-symbolic knowledge, drawing upon decades of work in the field of Cognitive Psychology and related disciplines.

This can then be used to enable machine learning of knowledge graphs through induction, and rule sets through heuristics and reinforcement learning techniques. Such approaches will be essential to scale up to large vocabularies and rulesets that would be impractical to maintain manually, given the inevitable evolution of requirements as a consequence of constantly changing business conditions.

Another important use case is for mapping data between different vocabularies. At a large enough scale, it is inevitable that different communities will develop their own vocabularies, using different requirements and different terminology. As such, for open markets of services that span different communities, there will be a need to be able to map data between different vocabularies.

This can be likened to translating between different human languages in that the precise meaning of terms may depend on the context. In addition, when moving from a weak vocabulary to a strong vocabulary, there is a need to address missing data. For machine translation of human languages, statistical approaches have proven their worth, and it is likely that they will likewise be useful for mapping data between different vocabularies.

Whilst this work is still at an early stage, there are opportunities for benefitting from collaboration in respect to sharing use cases and requirements, data sets and open source software.

Opportunities for the Graph Standardisation Business Group

The intended aim of the Graph Standardisation Business Group is to gather use cases and requirements across sectors, to guide technical standardisation work, and to coordinate liaisons with other standards development organisations and industry alliances. The group could conduct some technical work if needed to incubate ideas prior to transfer to the W3C standardisation track.

It is expected that the requirements for graph data standards are likely to be similar across different sectors, but this needs to checked, and it is anticipated that the Business Group will form a separate task force for each sector under consideration. The potential sectors include, but are not limited to smart cities, transport and connected vehicles, smart buildings, smart energy, smart manufacturing, smart healthcare and smart homes.

W3C is in a unique position to provide a cross sector perspective on graph data in respect to integration across multiple ecosystems and standards, using Web technologies as a means to connect heterogeneous distributed information sources. The Business Group will be able to gather business use cases and requirements, and to use these to ensure that technical standardisation work in other groups is properly justified in terms of these business requirements. Another important role will be to establish a shared vision through liaisons across standards development organisations and industry alliances.

Next steps

In the short term, the plan is to work on finding interested parties and collaborating on a Business Group charter. This charter will set out the aims of the Business Group and the patent commitments in the case that the charter covers collaboration on technical work, with a view to transferring it to W3C Working Group(s) for standardisation.

Please contact Dave Raggett <dsr@w3.org> if you would like to know more.