Australian Government
Australian Digital Health Agency

Towards Knowledge Graphs for National Healthcare Connectivity

Australian Digital Health Agency

WhitePaper

August 2023

Table of Contents

Metadata

Publisher: Australian Digital Health Agency
Author: Renato Iannella
Contact: renato.iannella@digitalheath.gov.au
Status: DRAFT
Issued: 18 August 2023
ISBN: 1 74064 700 9

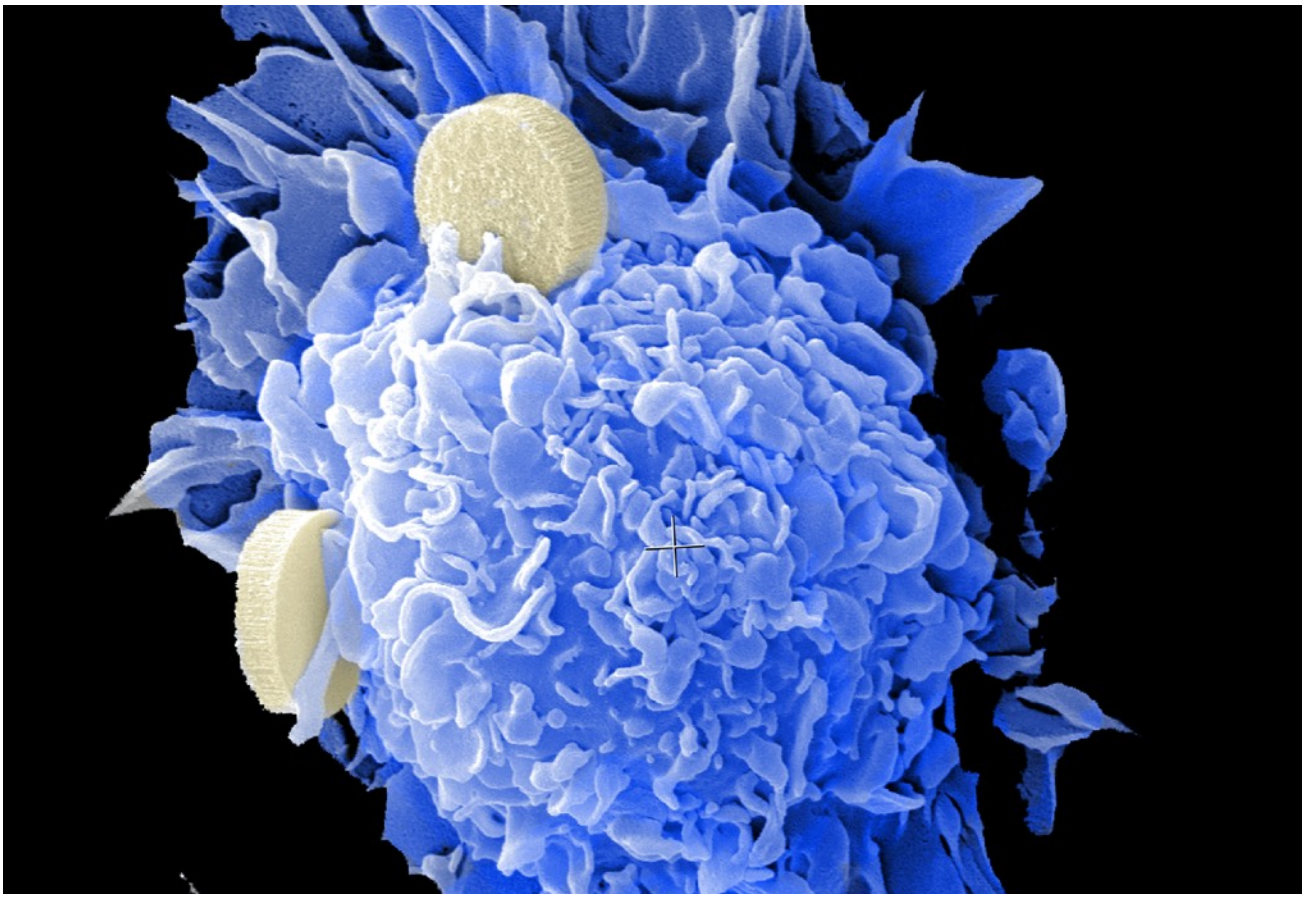

Photo by National Cancer Institute

Introduction

The healthcare sector is a global producer and consumer of data. As a safety-driven sector, the adoption of new data-enhancing technologies is a purposefully planned and careful process, often challenged by the argument that past approaches to data need to be supported in the future. The rate at which the sector moves towards adopting new technologies may miss generational opportunities to adopt contemporary and modern advances that are commonly embraced by other industries and sectors.

This whitepaper examines the current state of data standards interoperability in the healthcare industry and looks at a pathway to a future of enhanced data capabilities. Specifically, the use of Knowledge Graph approaches and technologies and the long term impact these will have on defining the strategic future for connected data services. The move towards knowledge graphs provides an opportunity for the sector not only to mature the technology roadmap, but to establish a solid, consistent, and safe platform for future data services in the healthcare sector.

Technology changes in healthcare are typically informed and championed by new sector drivers and capabilities, together with a future long term vision.

For example, the Australian National Digital Health Strategy Roadmap[1] includes the following priority action outcomes:

- Information sharing at transition of care - moving the health system towards real time data exchange.
- Create connected digital solutions - drive integration across care settings to increase the information available to the whole healthcare ecosystem.
- Expand information available and shared - connect health information sources to national platforms and empowering the consumer to take control of their health journey.
- Plan for emerging data sources and technology - governments, industry, and healthcare providers actively prepare for and embrace innovations and cutting-edge technologies.

The Australian National Healthcare Interoperability Plan[2] further emphasises the sharing of healthcare information as one of its five priority areas underpinned by connective-ness through interoperability across the ecosystem. The Interoperability Plan describes the future state of healthcare sharing to include free flowing information across jurisdictions transparently meeting consumer consent expectations.

Another Interoperability Plan priority area is Identity which is described as mainly focused on healthcare identifiers for consumers, providers, organisations, and services. Such identifiers are fundamental to the healthcare ecosystem to correctly identify and discover the correct entities and facilitates important connections between these entities. These identifiers are well-managed, stable, unique, normative, governed, and (sometimes) resolvable.

The UK Reshaping Health with Data[3] policy framework includes a number of visions:

- Ensure the data architecture underpinning the health and care system can easily work together to make more effective and efficient use of data.
- Members of the public and their care teams will have access to timely, high-quality data to improve care quality and inform choices about their care and support.

More specifically, the visions includes principles such as the importance of validation of healthcare data to improve data quality, the use of identifiers to help discover information more consistently, the use of common vocabularies to help build a single version of the truth, and data models that correctly define entities and are translatable into executable code.

The Canadian Interoperability Roadmap[4] describe their common 10-year vision to include:

- A modernised health system built on connected care - so that health information will move with them through the system, ensuring no patients fall through the cracks.
- Access to health information for all Canadians - to ensure seamless transitions of care across the system.

---

[1] https://www.digitalhealth.gov.au/about-us/strategies-and-plans/national-digital-health-strategy-and-framework-for-action

[2] https://www.digitalhealth.gov.au/about-us/strategies-and-plans/national-healthcare-interoperability-plan

[3] https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data

[4] https://www.infoway-inforoute.ca/en/component/edocman/6444-connecting-you-to-modern-health-care-shared-pan-canadian-interoperability-roadmap/view-document

- Digital health innovation by and for Canadians - work towards setting a foundation that enables the use of innovative and emerging technologies.

The Canadian approach for the building blocks of advanced interoperability include:

- Data Foundation and Portability Framework that will drive toward a single, extensible approach, encompassing a reference-able data model and guidance around semantic data components.
- A Health Data Content Framework that will support a very predictable syntactical and semantic digital transformation for data exchange.
- Consistent Data Semantics for personal health information that flows from one provider to another to be unambiguously understood, with a highlight of the current lack of support for multi-lingual terminology.

To summarise, the national plans align on future healthcare information that needs to be shared more broadly, and connected more often, and identifiable more widely, and real-time availability, to enable an inclusive and comprehensive healthcare ecosystem.

Healthcare Challenges

For the healthcare sector, where information can be complex and inter-related, knowledge graphs can provide a roadmap to deliver solutions for many of the sectors's challenges. Some of these include[5]:

- The integration of clinical vocabularies using a common framework.
- The flexibility of ontologies to represent declarative knowledge and their ability for growth and adaptation to new clinical paradigms (such as patient-centered health care).
- Transforming clinical guidelines into basic reasoning structures and formal logic rules for consistent and effective automation.
- Knowledge filtering tools at the point-of-care for real-time access to the only required information for the subject of care.
- Data analytics that supports data-driven healthcare that can be transform evidence-based knowledge into automate-able clinically-safe outcomes.

[5] Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges. Artificial Intelligence In Medicine, 2019. https://doi.org/10.1016/j.artmed.2019.101713

Interoperability Horizons

Healthcare information is, and should be, the central focus for any national clinical architecture. As information technologies and mechanisms evolve and mature over time, there are critical points in time to reflect on new approaches to healthcare information. These approaches reflect paradigm shifts in wider industries as well as the supporting technologies and infrastructure.

Figure 1 below shows the past, current, and future evolutions of healthcare information interoperability as a series of horizons. Each interoperability horizon covers a wide time span, but typically around a 10 to 20-year time period and are also heavily overlapping periods as the sector requires significant migration periods between approaches. Currently, we are somewhere towards the end of horizon 1 and the beginning phase of horizon 2.

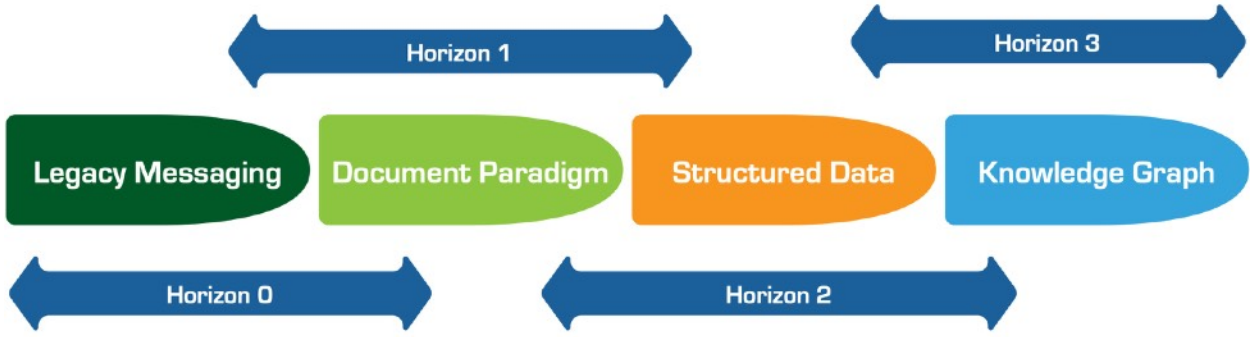

Figure 1 - Interoperability Horizons

Horizon Zero saw the first instantiation of healthcare information interoperability with a protocol-based messaging paradigm. As the first approach, **HL7 V2** enabled the transfer of data across healthcare systems utilising obtuse technical encoding schemes to express segments of healthcare information. Despite its age and complexity, HL7 V2 is still used significantly through-out the healthcare sector.

Horizon One introduced the notion of documents to mimic and represent the then current-state physical objects that were created and exchanged by healthcare providers. **HL7 Clinical Document Architecture (CDA)** provided the information standard for these template documents. The document-like structure supported the current state artefacts with different levels of document complexity (from attachments to embedded terminology) and conformance.

Horizon Two recognised the need to be more data-centric and a move away from the limited (and complex) document-like structures. **HL7 Fast Healthcare Interoperable Resources (FHIR)** was developed as a data exchange standard and provides succinct (stand-alone and reusable) structured information "resources" for common healthcare entities (eg patient, encounter, medication etc). In addition, FHIR defines the infrastructure for encoding (XML, JSON), querying and transport of FHIR data.

Horizon Three proposes a new approach where information follows a knowledge graph model. A knowledge graph is a data space that represents all information as connected nodes and edges which enables the relationships between the entities to become a key focus. The graph can easily be traversed

to discover complex relationships across disparate entities. The key benefit in the graph approach is that the patient can be the central focus and all related data (eg an encounter or prescription) will directly emanate from this central entity (as opposed to the patient being previously associated with numerous sets of structured data resources as in FHIR).

Other benefits of knowledge graphs is that they provide a richer context and semantic understanding of health information by representing entities and their relationships, enabling both humans and machines to better interpret and interact with the data they model.

Across all the first three Horizons, terminology has remained rather consistent in their approaches for defining vocabularies (eg **LOINC** and **SNOMED CT**) as these have been developed in parallel to information exchange models. In Horizon 3, to support the knowledge graph approach, terminologies will need to be harmonised with open data vocabulary mechanisms (such as **W3C SKOS**). For most, this will be a reasonable transformation process, and for others - like SNOMED CT - this will open up significantly more opportunities as SNOMED CT is formally based on a semantic model.

Interoperability is defined by the Healthcare Information and Management Systems Society (HIMSS) as "The ability of different information systems, devices and applications (systems) to access, exchange, integrate and cooperatively use data in a coordinated manner, within and across organisational, regional and national boundaries, to provide timely and seamless portability of information and optimise the health of individuals and populations globally."

A Deeper Dive

To get a better understanding of the different approaches across the last 3 horizons, consider the conceptual view in Figure 2, which shows the Document Paradigm (left), Structured Data (middle) and Knowledge Graph (right) approaches.

In the context of national healthcare platforms, such as My Health Record (MHR), the Document paradigm is the current state. Clinical Documents (such as Shared Health Summaries and Discharge Summaries) are stored with a specific Consumer (the patient identified by their Individual Healthcare Identifier IHI). In reality, some parts of each Document that contain "atomic" data are extracted into a separate repository (but mainly used for indexing purposes) and some also contain a PDF version of the original document. However, the primary consumer experience is tailored towards displaying views of their clinical documents as this is the core entity available.

In the Structured Data paradigm, the focus now is on "data resources" that represent some healthcare entity (as a set of discrete properties). In FHIR, for example, there are over 150 such "Resources" (from Allergy Intolerance to Vision Prescription). Each of these structured data resources are snapshots of events or activities that has occurred and associated with a patient via their IHI. The resources have granular data expressed with properties (eg name, dose, expiration date) some of which are coded from terminologies. That is, they represent an entity in a formal terminology (such as SNOMED CT or LOINC etc). The mechanism to represent these coded entities is specific to FHIR but in essence are localised properties inside each resource.

In the Knowledge Graph paradigm, the focus is on connections between entities (the nodes) and relationships (the edges) with other entities. This is the classic 'triple model' (subject, predicate, object) that underpins formal ontology models. In the knowledge graph, all entities and relationships are

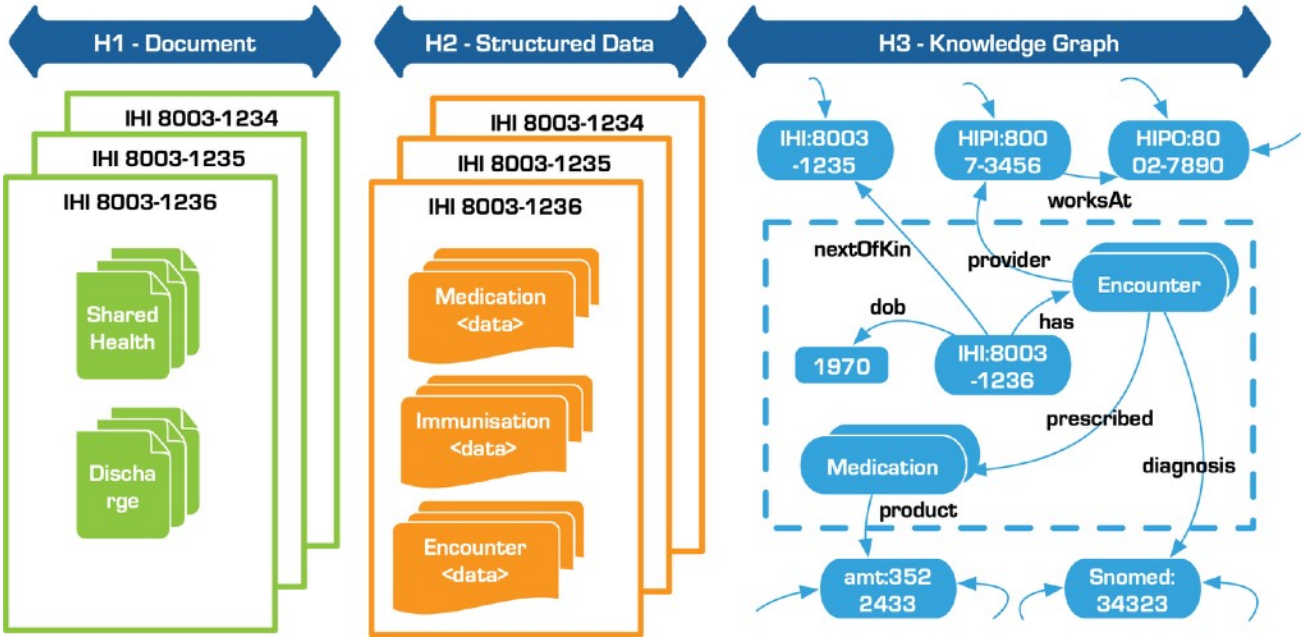

Figure 2 - Horizons Comparison

represented by globally unique identifiers, as this ensures the integrity of the graph, as well as enabling it to grow and expand beyond local limitations.

As shown in Figure 2, there is an entity representing the central Consumer with relationships to entities directly linked to the patient (such as *Encounters*, and *Medications*). Each of these entities also have relationships to other entities, such as the *Encounter* with the *diagnosis* property to a SNOMED CT concept, the *prescribed* property to a *Medication* resource with link to an **AMT** concept, and a connection to the *provider* (an identified Provider entity). The fundamental difference here is that these entities are the same entity for all other relationships over the entire graph (ie, they will have many incoming relationships). This means the graph integrity is strengthened as everything about, for example, *snomed:34323* will point to the same entity. The dashed lines shows the sub-graph related to the specific consumer (*ihi:8003-1236*) but this is fully connect to the entire knowledge graph.

These future Horizons represent opportunities to address the drivers behind national healthcare connective-ness and provide a future roadmap for the evolution of clinical information for the Australian healthcare ecosystem. Specifically, these capabilities are addressed:

- Sharing - the move towards modular information entities and common data exchange patterns
- Identifiable - the move towards the ubiquitous use of common, global, resolvable identifier systems
- Connected - the move towards knowledge graph models
- Real-time - the move towards direct data linkages across repositories

Each of these individual capabilities above are sufficient for healthcare information, but all are necessary for the future state of healthcare stakeholders.

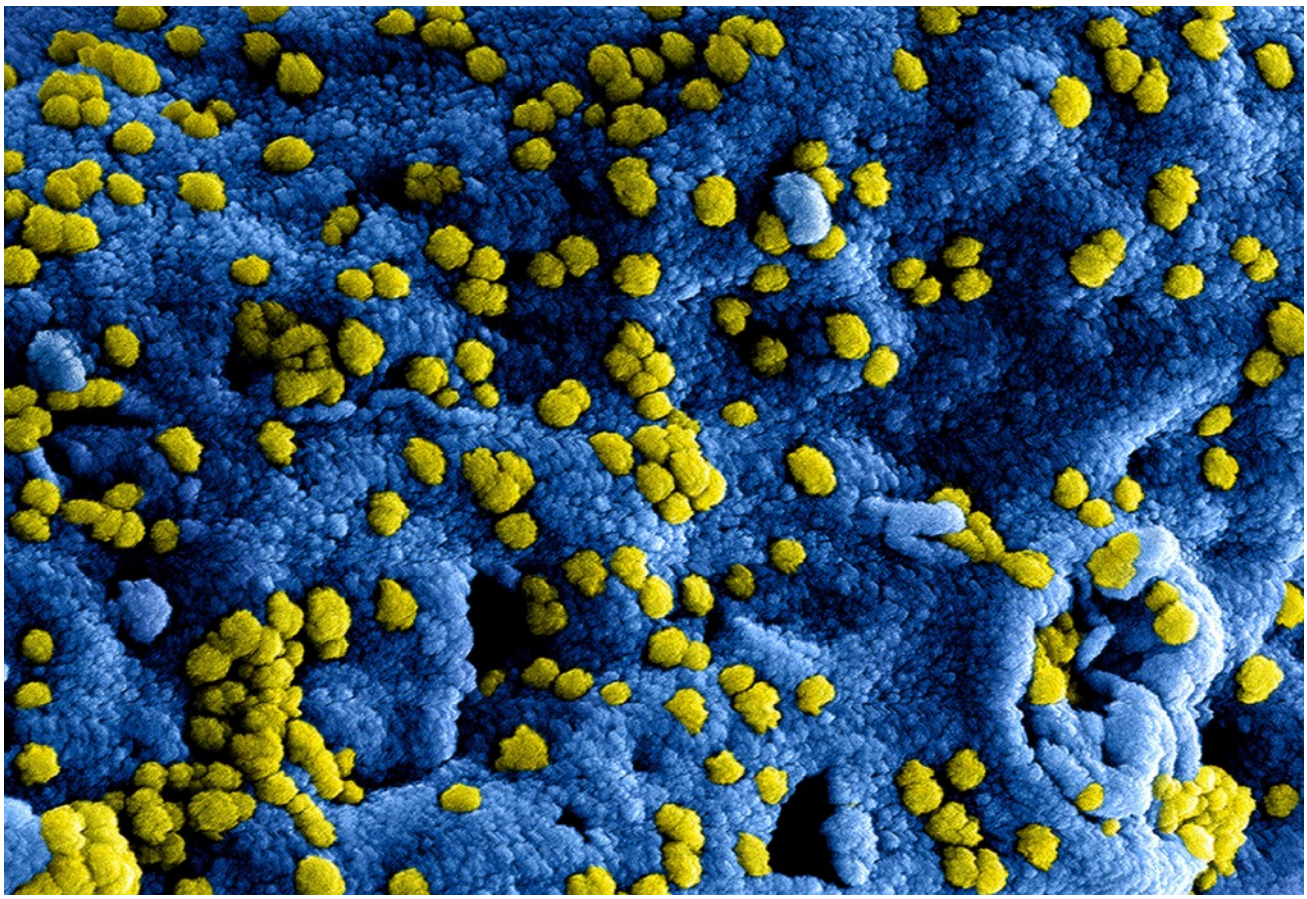

Photo by CDC

Knowledge Graphs

Knowledge graphs are information systems with specific fundamental characteristics that provide advantages and benefits over traditional approaches. The core repository is a network graph model and represents real-world information with nodes and edges. Knowledge graphs are themselves a paradigm shift for data management, as they provide improved ways for unified data access, flexible data integration, and enhanced data analytics models. Leading IT analysis firms, such as Gartner[6], IDC[7], and Forrester[8] have all predicted this new wave of change.

Like most technologies there are variations in the knowledge graph sector when it comes to information approaches and interoperability. Knowledge Graphs are classed as either "property graphs" or "directed edge-labeled graphs" (the latter being more colloquially referred to as a "semantic graph"). Conceptually they are similar but there are fundamental differences with semantic graphs targeting interoperability, integration, and sharing across boundaries[9]. Both types allow a graph to be modelled and nodes/edges defined, but semantic graphs utilise the W3C Semantic Web formalisation of Classes

[6] https://www.gartner.com/smarterwithgartner/gartner-top-10-data-analytics-trends

[7] https://www.idc.com/getdoc.jsp?containerId=US46433020

[8] https://www.forrester.com/report/use-knowledge-graphs-to-manage-knowledge-instead-of-data/RES179193

[9] https://flur.ee/fluree-blog/rdf-versus-lpg/

(nodes) and Properties (edges). This means that a consistent and computable information model is the basis for the entire knowledge graph as well as the identity of all entities is based on URIs to support Linked Data.

The second fundamental difference is that semantic graphs use the industry standard SPARQL query language and TURTLE/JSON-LD serialisation. This means interoperability and exchange across and between semantic graphs is fully supported but property graphs are vendor dependent.

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation
Tim Berners-Lee, Inventor of the Web

Semantics is the ability of computer systems to exchange data with unambiguous, shared meaning enabled by machine computable logic, inferencing, knowledge discovery, and data federation between information systems[10]. The **W3C Semantic Web** is a family of fundamental technologies that provides the framework for semantic data interoperability and includes information modelling, vocabularies, ontologies, formal reasoning, and query protocols. The Semantic Web is grounded in mathematically complete logic which provides the assurance for the inferencing outcomes.

The semantic knowledge graph benefits include:

- Expressibility - the semantic data model (based on ontologies and vocabularies) allow humans and computers to develop, represent, visualise, and interpret data in an unambiguous manner.
- Interoperability - the range of technical specifications are all based on open industry standards and linked data being a fundamental approach for cross-graph linkages.
- Performance - the graph data model (triples) supports highly optimised and efficient data stores capable of storing and querying “trillions of triples”[11].
- Integration - the graph model can easily accommodate disparate data sources and provide a single harmonised view across enterprise repositories.
- Conformance - the graph can be constrained to shapes that match business rules, community requirements, and industry regulations.
- Independence - the semantic graph model is the same as what is stored in the graph database, as there is no need to map to an implementation model. This also increases the portability of knowledge bases across vendor platforms.
- Virtualisation - knowledge graphs can exist across heterogeneous platforms and behave as a single graph

There are number of unique characteristics to semantic graph expressibility:

- Ontologies define the classes (types of things) and properties (characteristics of things) and the relationships between them.
- Both class and properties are first-class entities and independent from each other, and are also both hierarchical.
- Properties have defined semantic roles and directions.

[10] https://en.wikipedia.org/wiki/Semantic_interoperability

[11] https://info.cambridgesemantics.com/hubfs/Trillion%20Triples%20Whitepaper%20-%20GQE%20Benchmark%20FINAL%20Dec%206.pdf

- Every class/property and every instance of a Class (an individual) has globally unique identity.
- The sameness and class membership (of individuals) and equivalence (of class/properties) can be automatically resolved.
- Ontologies are self-describing by using this sameness/equivalence feature and referring to common open ontologies.
- Automated reasoning services can infer new statements of fact based on the formal model underpinning ontologies.

These defining characteristics are necessary requirements for semantic systems. They represent a significant uplift in information model expression and semantic automation capabilities above existing structured systems and methodologies.

Industry Uses

A use case of Knowledge Graphs comes from the financial sector in the form of The Financial Industry Business Ontology[12] (FIBO). FIBO is supported by large organisations (US Department of Commerce, Deutsche Bank, Goldman Sachs etc) with a mission to develop, maintain, and promote a machine-readable and unambiguous data standard that enables understanding of the financial terminology, cross-system federation and aggregation of data to improve effectiveness of decisions, to improve efficiency in regulatory reporting and to fast-track the adoption of advanced analytical capabilities for financial services. FIBO covers over 3,000 classes and over 600 property relationships and heavily uses the SKOS vocabulary framework.

The BBC developed multiple related ontologies and technical architecture to host the FIFA World Cup and London Olympics sports information and future news services[13]. They found that:

- A triple-store provides a concise, accurate and clean implementation methodology for describing domain knowledge models.
- A semantic graph approach provides ultimate modelling expressivity, with the added advantage of deductive reasoning.
- SPARQL simplifies domain queries, with the associated underlying graph schema being more flexible than a corresponding SQL/RDBMS approach.
- Combining the triple approach with dynamic atomic data as an architectural foundation simplifies the publication of content as "open linked data" between systems and across the wider cloud.

Wikipedia is the largest online community-developed encyclopaedia of knowledge. The DBpedia project leverages this global source of knowledge by extracting structured information from Wikipedia and creating the DBpedia Ontology[14] covering the many domains and the individuals represented in Wikipedia in various languages.

The ontology currently covers 768 classes which form a subsumption hierarchy and are described by 3,000 different properties and currently contains about 4,233,000 individual instances. The DBpedia ontology is often used as a ‘core ontology’ for other ontologies to link and reuse concepts as well as a source of individuals to form equivalence relationships.

[12] https://spec.edmcouncil.org/fibo/

[13] https://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html

[14] https://www.dbpedia.org/resources/ontology/

QIMR Berghofer (based at the Royal Brisbane Hospital) is a medical research institute specialising in cancer, infectious disease, mental health, and chronic disorders. Like most research laboratories, they use a Laboratory Information Management System (LIMS) to manage their research data. Traditional LIMS, however, are built on relational databases and their brittle infrastructure can't keep pace with rapidly changing data and hypotheses. QIMR Berghofer's choose a new LIMS based on a vendor graph database[15] allowing constant and fast iteration on data models, data definitions, and types of data. The lab's data management can now dynamically respond to the needs of the research, instead of contorting the study to fit in the inflexible data model.

Vendors

The vendors of semantic Knowledge Graphs products range from the major cloud providers (AWS Neptune[16]) to industry stalwarts (Oracle Graph[17] and MarkLogic[18]) to bespoke knowledge graph companies (Ontotext GraphDB[19] and Stardog[20]). There are also related vendors that focus on the development and governance of ontologies and (SKOS-based) vocabularies for knowledge graph platforms (TopQuadrant[21] and PoolParty[22]).

Health Related Graphs

One of the additional benefits of a knowledge graph approach is the publication of open graphs for consumers of health related data and data linking opportunities. A number of examples include:

- The KEGG Disease database[23] - A collection of disease entries such as cancers, immune system diseases, neurodegenerative diseases, cardiovascular diseases, and metabolic diseases where known disease genes are highlighted. It also contains infectious diseases with interacting molecular networks of both pathogens and humans.
- Hetionet[24] combines information from 29 public databases and contains 47,031 nodes and 2,250,197 edges that cover anatomical structures, signs and symptoms, adverse drug reactions, complex diseases, molecular functions, and protein-coding human genes.
- STRING[25] is a database of known and predicted protein-protein interactions that includes direct (physical) and indirect (functional) associations that stem from computational prediction and from knowledge transfer between organisms. The STRING database currently covers 67,592,464 proteins from 14,094 organisms.

---

[15] https://www.stardog.com/company/customers/qimr/

[16] https://aws.amazon.com/neptune/

[17] https://docs.oracle.com/en/database/oracle/oracle-database/19/rdfrm/#Oracle®-Spatial-and-Graph

[18] https://www.marklogic.com/product/platform/

[19] https://www.ontotext.com/products/graphdb/

[20] https://www.stardog.com/platform/

[21] https://www.topquadrant.com/

[22] https://www.poolparty.biz/

[23] https://www.genome.jp/kegg/disease/

[24] https://het.io/about/

[25] https://string-db.org/

- HealthECCO[26] supports science and research in the fields of medicine and biology in the search for cures for diseases such as Covid-19 and diabetes. The platform is a knowledge graph that integrates a growing number of different but related data sets. The key advantages that enabled this are the flexibility, extensibility and scalability of graph technologies. While the concepts behind data integration are not new, graph technology now renders possibilities that have failed in previous attempts. They also provide a public GraphQL API on top of the knowledge graph to support the development of additional applications or integration of third party tools making the data interoperable.

Knowledge Graph Framework

Knowledge graphs offer a significant new capabilities and platforms to support the complete "data management" lifecycle in non-traditional approaches. Domains can drive new knowledge graph implementation by articulating and establishing their sectors semantic model, terminologies, business rules and conformance requirements, and delivery of capabilities .

These are manifested in four key technology paradigms (as shown in Figure 3):

- Ontologies - the domain semantic models that represents the concepts and relationships
- Vocabularies - the domain terminology that represents the vocabularies, taxonomies, and managed terms and values.
- Shapes - the domain rules that represents the business processes, conformance requirements, and regulatory standards for data.
- Query - the platform infrastructure to capture and deliver data requests to meet a number of capabilities and functions.

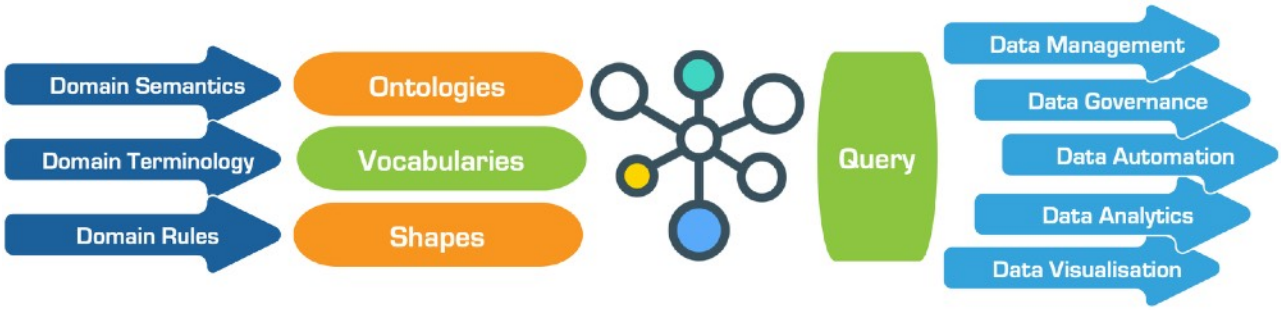

Figure 3 - Knowledge Graph Framework

A typical knowledge graph platform supports multiple capabilities, some of which include:

- Data Management - managing the core functions of knowledge capture and dissemination including searching, browsing, context extraction, and unified access.
- Data Governance - managing master data, and the lifecycle stewardship of ontologies, terminologies, and rules.

[26] https://healthecco.org/

- Data Automation - managing data transformations, data quality, data monitoring, and data integrations.
- Data Analytics - managing data analysis, entity resolution, and machine learning.
- Data Visualisation - managing the visualisation of graph data and relationships.

Ontologies

Ontologies are building blocks of knowledge graphs. They represent the formal explicit description of concepts in a domain of discourse. Ontologies are more formally-based than other modelling approaches (eg UML) to support computable outcomes (such as reasoning). In essence, an ontology describes the entities (classes) and property (relationships) describing various features and restrictions over the concepts. An ontology together with a set of individual instances of classes constitutes a knowledge base.

W3C has developed three standards for knowledge representation (RDF, RDF Schema, OWL) and these are generally used as a single language for ontology modelling. As shown in Figure 4, the core capabilities of the ontology specifications are harmonised around classes and properties, with individuals being instances of classes.

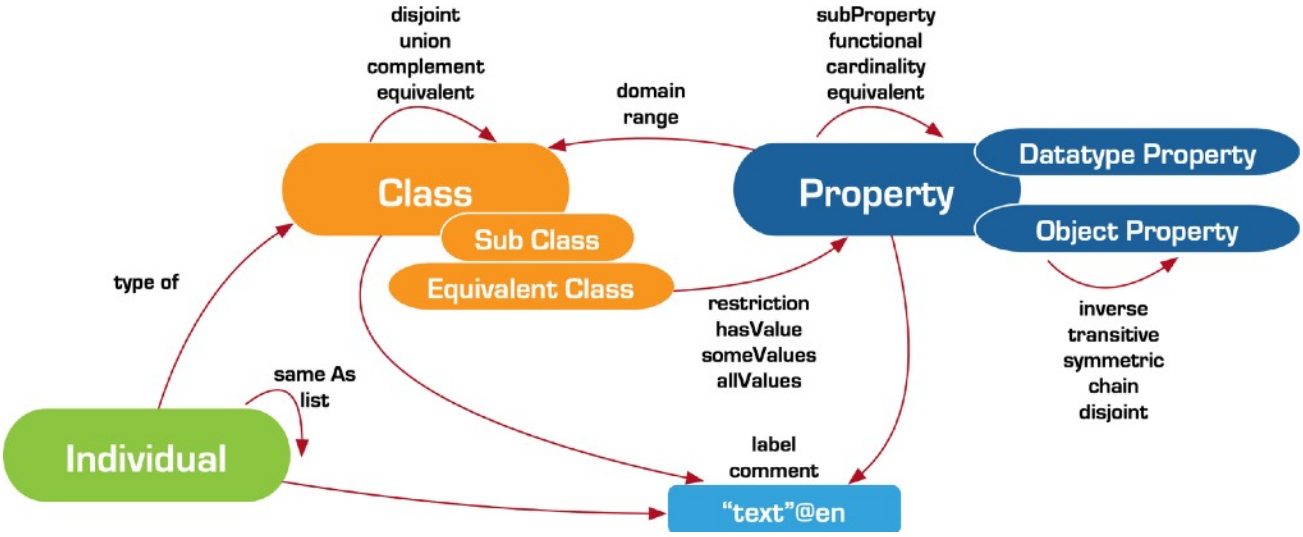

Figure 4 - RDF/OWL Ontology Model

Classes can be declared in numerous ways from explicit definitions to a more complicated set of restrictions rules over properties (and values) that can be automatically inferred. Properties can also have an inherit set of semantics, such as inverse and transitivity that can be modelled in the ontology to match the domain needs. Overall, the features and power of the ontology sets it apart from all other structured modelling approaches. The ontology is the key provider of semantics to the knowledge graph platform.

The ubiquitous SNOMED CT is an exemplar of a clinical ontology and has been represented in the semantic OWL language[27]. This means that the SNOMED clinical relationships are computable as OWL relationships, resulting in consistent, and safe, reasoning outcomes. For example, the "myocardial infarction" class is defined with necessary object property relationships with "associated morphology" and "finding site" (each with specific individual values). This enables entailment of individuals with those conditions, as well other disorder subtypes (such as "ischemic heart disease").

[27] https://confluence.ihtsdotools.org/display/DOCOWL

There is a trend of adopting ontologies for enabling semantic interoperability that facilitate health information exchange. The most important role offered by ontologies in this context is the ability to allow technology agnostic methods of communicating the meaning of concepts used across the healthcare sector., Ontologies have proved to be more dynamic than other methods used as we move towards achieving more comprehensive levels of semantic interoperability.

From a health care perspective, ontologies[28] can be used to maximise:

- meaning that can be inferred from coded data;
- different granularities of data (of words and coding);
- the ability to cope with temporal change in definitions, clinical practice and fluctuation; and
- structural changes (system studies, e.g. encounters, health professionals, governance and privacy).

The biomedical community has also been active in developing numerous ontologies[29]. An example is the Vaccine Ontology which is a community-based biomedical ontology in the domain of vaccine and vaccination. The Vaccine Ontology standardise vaccine types and annotations, integrates various vaccine data, and supports computer-driven reasoning. This ontology defines nearly 7,000 classes and 233 properties with 167 individuals (see Figure 5 for a partial view).

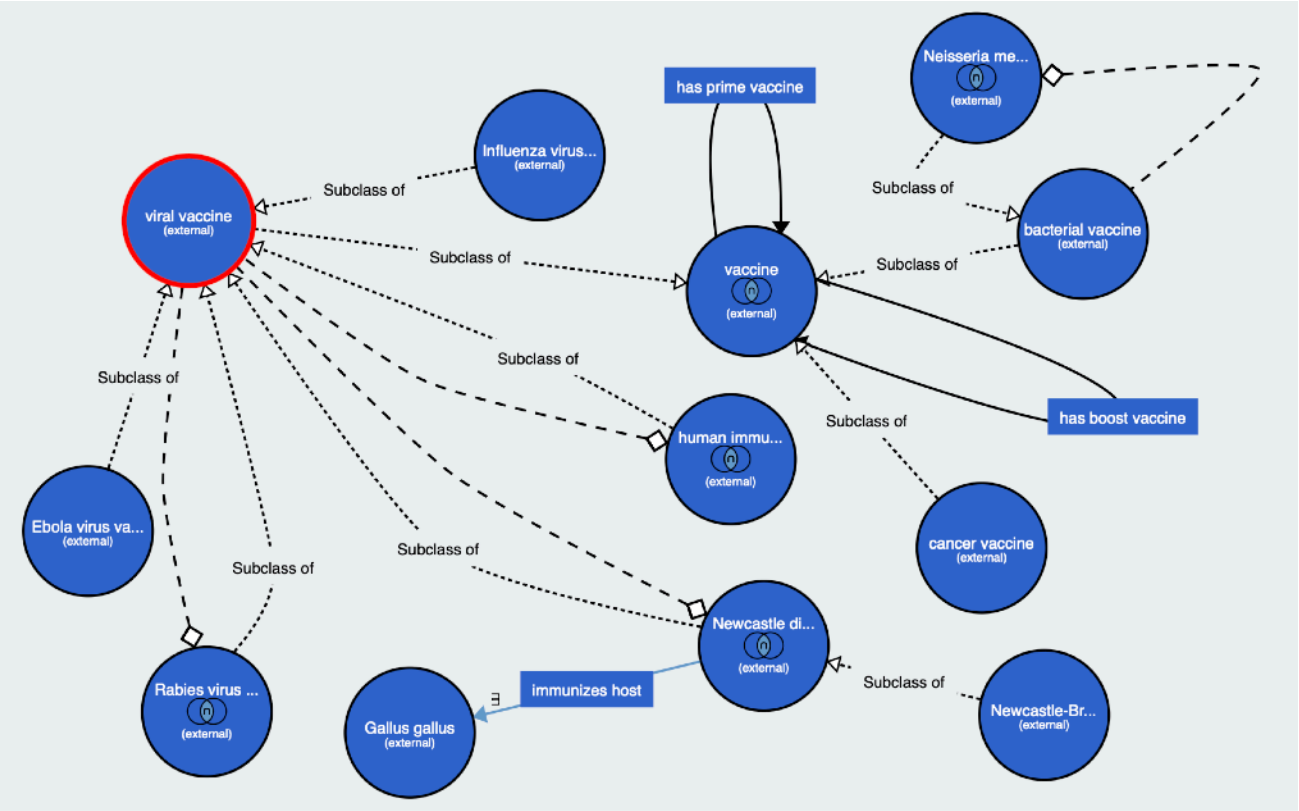

Figure 5 - The Vaccine Ontology (partial)

As an example of the power of ontologies, the "Rabies Virus" (shown in the bottom left of Figure 5) is defined as a "viral vaccine" class which also has the property "immunizes against microbe" with the

[28] Using ontologies to improve semantic interoperability in health data, H. Liuanaye (et al), Health & Care Informatics, V.22 N.2 https://informatics.bmj.com/content/22/2/309.long

[29] https://bioportal.bioontology.org/

value “Rabies lyssavirus” (as a URI) from the National Center for Biotechnology Information (NCBI) taxonomy (shown as the diamond incoming relationship). This class is automatically inferred from the matching individuals in the knowledge graph.

A recent paper[30] describes why the biomedical domain has been one of the early adopters of graph databases, enabling more natural representation models and better data integration workflows, exploration and analysis facilities. Specifically:

- Graphs provide more natural modeling of many-to- many relationships.
- Graph-oriented query languages provide more intuitive means for writing complex network traversal and graph algorithm queries than table-oriented ones like SQL.
- The schema-less/optional grants flexibility.
- Graph databases present higher performance for relationship-centric searches.

Another paper[31] discusses a number of biomedical and clinical use cases that can be addressed using graph and knowledge graph-based approaches:

- Prediction of the likelihood of an existing edge between two nodes based on the entirety of the knowledge. For example, prediction of edge likelihood of drug compound and patient to predict personal risk of adverse reaction or links between two diseases posing a high comorbidity risk.
- Identification of highly connected regions within a real-world data graph that can identify patients with a high similarity, e.g., patients with a certain disease subtype.
- Prediction of the likelihood of a patient node being assigned a label based on the entirety of their medical data. For example, patient node gets assigned a disease risk label.
- The inherent connected representation in graphs allows for the easy traversal of the graph to identify pieces of information that are separated by several nodes. When combining patient data with terminological knowledge, this allows for complex queries, e.g., identification of all patients based on a medical condition and its subtypes.

[30] An overview of graph databases and their applications in the biomedical domain, Timo´n-Reina (et al), Database, 2021 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8130509/pdf/baab026.pdf

[31] From Data to Wisdom: Biomedical Knowledge Graphs for Real-World Data Insights, K. Hansel (et al), Journal of Medical Systems, 2023 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10191934/

Vocabularies

Vocabularies are a fundamental part of knowledge graphs as they define the terminology used by the information models and represents the domain's scope of terms and values. In healthcare, vocabularies have been a part of the information fabric for decades. Classic terminologies such as **ICD-10**, LOINC, and SNOMED CT have been used extensively across the healthcare information domain. Each plays a specific role and supports clinical and administrative insight into describing healthcare interactions. Additionally, new "values sets" have been defined by healthcare information exchange standards.

These individual healthcare vocabularies all have different underlying structures and representation formats. This leads to interoperability blockers and requires multiple translation mechanisms, which are costly to develop and maintain.

The international community develop the **W3C SKOS** (Simple Knowledge Organisation System) standard specifically to address this issue. The SKOS model is used for defining, sharing and linking controlled vocabularies, taxonomies, thesauri and business vocabularies. As shown in Figure 6, the "concept" is the fundamental element of SKOS which can be part of general schemes, and more specific ordered collections.

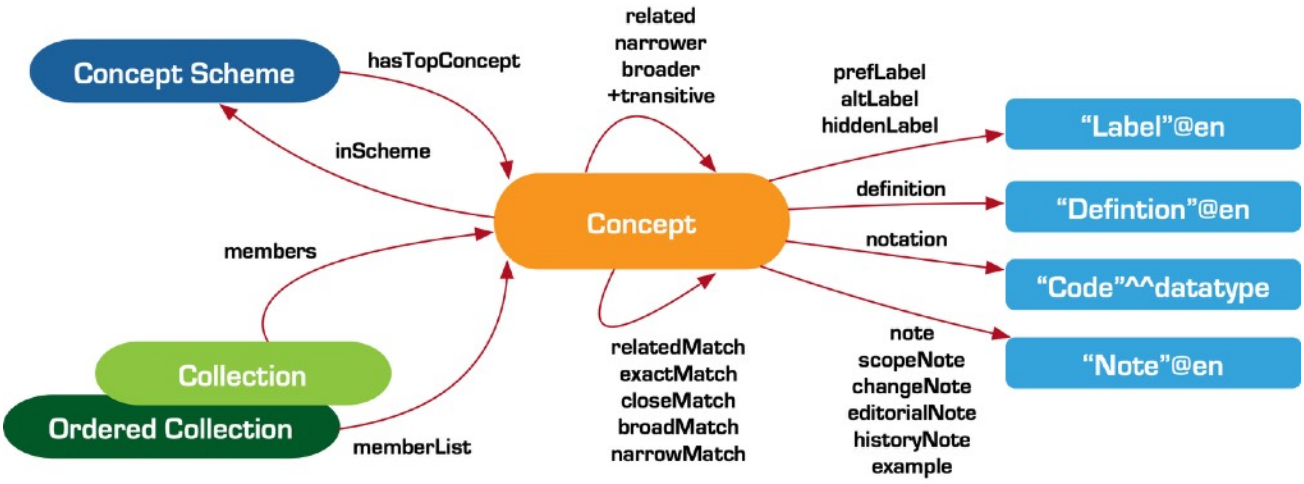

Figure 6 - W3C SKOS Information Model

A SKOS concept has many relationship properties describing it in context with other members of a scheme, such as broader/narrower and includes transitive relationships. In addition a SKOS concept has many matching relationships for mapping to similar concepts in other schemes. Each SKOS concepts has annotation properties for describing various labels, definitions, notation, and notes. A key feature of SKOS is the use of language tags on all annotations enabling it to support diverse multilingual consumers.

SKOS is built on the W3C Semantic Web stack and can also be extended in terms of new (or refined) property relationship semantics. Because of its benefits and global standardisation, SKOS is one of the most used approaches for vocabularies in Knowledge Graph platforms.

SKOS is used by many large institutions, such as the US Library of Congress[32] and the Getty Institute[33] to represent their communities vocabularies and support the "linked data" opportunities for discovery and reuse.

The Australian Education sector has developed the Schools Online Thesaurus (ScOT)[34] which is a controlled vocabulary of terms used in Australian and New Zealand schools. It encompasses all subject areas as well as terms describing educational and administrative processes. The thesaurus links non-preferred terms to curriculum terms. It also relates terms in a browsable structure.

Why is SNOMED CT both an Ontology and a Vocabulary? The traditional and common use of SNOMED CT is as a Clinical Terminology so fits well as a Vocabulary with hierarchical term structures. But SNOMED CT also expresses additional semantics with Classes and Properties that are used to define computable clinical relationships.

An example of use of SKOS in healthcare is the "Set of Patient-Centered Outcome Measures for Pregnancy And Childbirth"[35] developed by the International Consortium for Health Outcomes Measurement (ICHOM). The objective of this vocabulary was to define a minimum, internationally appropriate set of outcome measures for evaluating and improving perinatal care with a focus on outcomes that matter to women and their families.

Another set of SKOS vocabularies was developed by the Virus Outbreak Data Network (VODAN)-Africa. It is focused on improving health data analysis, under the regulatory provisions of the country, and strengthening national capacities for health data analytics as well as the use of health data at the point of care. They developed a number of SKOS vocabularies[36] the covered terms that are used within outpatient department registration, diseases, and health facilities.

[32] https://id.loc.gov/

[33] https://www.getty.edu/research/tools/vocabularies/lod/

[34] https://vocabulary.curriculum.edu.au/

[35] https://bioportal.bioontology.org/ontologies/ICHOM-PROMS-PCB

[36] https://vodan-totafrica.info/page.php?i=7&a=vocabulary

Photo by Brano

Shapes

A fundamental part of the Semantic Web is support for the Open World Assumption (OWA). The OWA makes it clear that if information is not known, then it is not false and applies when an information system has incomplete information. For example, consider a patient's clinical history system. If the patient's clinical history does not include a particular allergy, it would be incorrect (and unsafe) to state that the patient does not suffer from that allergy. It is unknown if the patient has that allergy, unless more information is given to disprove the assumption.

Modelling in ontologies supports the OWA approach as this is the closest to how the real-world operates. Typically ontologies are designed to have a broad scope in terms of modelling a domain and to support reasoning outcomes. This means that an ontology encapsulates many scenarios of real-world usage, and is not designed to constrain the views of the domain.

To support more fine-grained constraints over ontology scenarios, the **W3C SHACL** (Shapes Constraint Language) was developed to enable rules to be specified over the "shape" of the graph. SHACL is used to define classes together with constraints on their properties and comes with several built-in types of constraints such as cardinality, data types and allowed vocabulary values. SHACL can also define more complex kinds of constraints for almost arbitrary validation conditions. SHACL validation can verify whether the graph data instance fulfils the constraint rules described by the ontology model.

Many different SHACL rules can be developed that cover the same aspects of the ontology. This means that different business use cases can be supported without having to redefine (and manage) changes to

the overarching ontology. This separation of ontology (model) from data conformance (rules) is an advantage of the knowledge graph platforms.

Consider the example ontology shown in Figure 7. This ontology shows the Encounter class having property relationships to the subject of care, when the encounter occurred, the reason, priority and the diagnosis from the encounter.

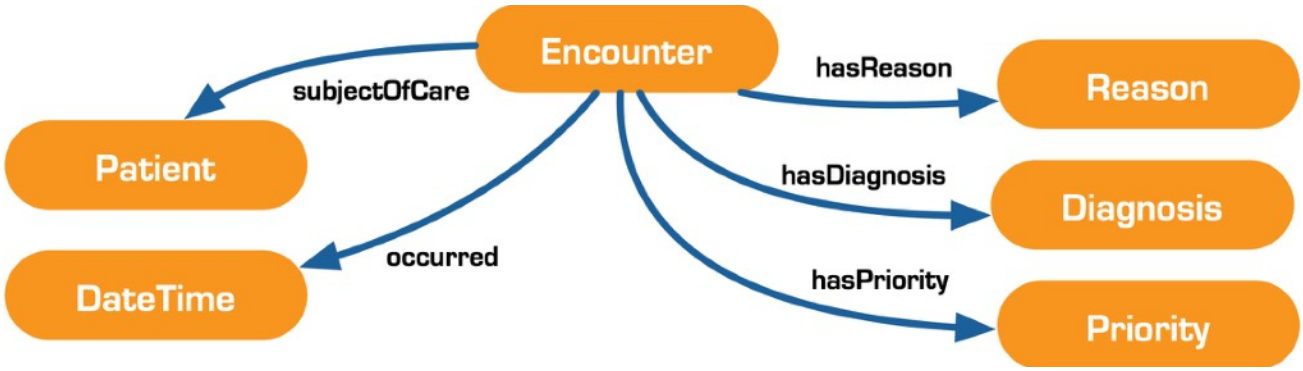

Figure 7 - Example Ontology

From this ontology, two specific business conformance rules can be created (with some common rules):

SHACL Rule #1: General Encounter	SHACL Rule #2: Priority Emergency Encounter
Must be an Encounter type	
Must have only one subjectOfCare property that links to an individual of type Patient	
Must have only one occurred property of type dateTime	
Must have only one hasReason property that uses values only from the HL7 Reason vocabulary	Must have only one hasReason property that uses the specific value of <http://hl7.org/reason/emergency>
Must have only one hasDiagnosis property that uses values only from the SNOMED vocabulary	
	Must have only one hasPriority property that uses some values from the HL7 Priority vocabulary or the ICHOM Triage vocabulary

The SHACL rules are themselves expressed as a graph and can be used by knowledge graph platforms to verify that data meets specific business rules before being committed to the graph repository. A SHACL validation engine takes as input the graph to be validated and a graph containing SHACL shapes declarations and produces a validation report, also expressed as a graph. SHACL has the ability to specify a severity level of validation results, but also the ability to return suggestions on how data may be fixed if the validation result is raised.
The advantage of using SHACL shapes is that the business rules can be expressed in the same language as the knowledge model and transparency is significantly improved.

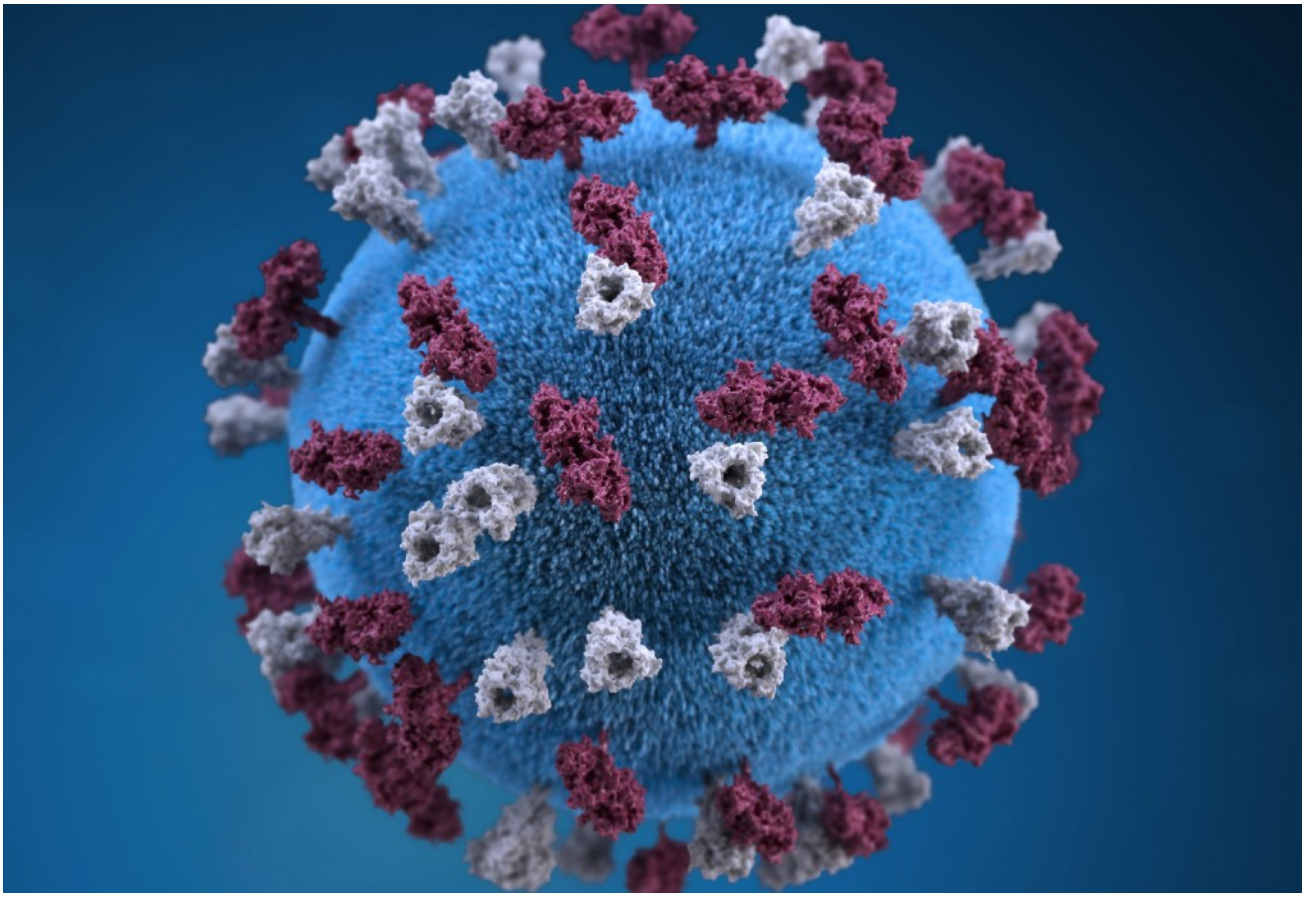

Photo by CDC

Query

The query interfaces exposes the inputs and outputs from a knowledge graph platform. The **W3C SPARQL** query language is the international standard for semantic knowledge graphs. SPARQL is an optimised query language for graph triples. SPARQL supports the key functions for data query and update and supports working with the ontology reasoning outcomes, including functions and rules for results ordering, transformations and manipulation. SPARQL also supports federated queries (for virtualisation) where a single query can be executed on multiple graphs and the results harmonised.

SPARQL is an extremely powerful query language and some platforms also provide support for the **GraphQL** query language. GraphQL is a graph-oriented API that typically is used as the front-end API for developers. GraphQL is simpler and provides query and mutation services with a schema that provides a complete and understandable description of the data in the graph. Most platforms provide GraphQL as a common API which is then bound to SPARQL and to any other APIs supported by the platforms. This means the utility of GraphQL as a common integration layer across not only graph data, but other forms as well (structured, relational etc).

Case Study: Pharmaceutical

The Allotrope Foundation[37] is a consortium of pharmaceutical companies formed in 2013 to revolutionise the way they acquire, share and gain insights from scientific data through the application of community-derived data standards consistently delivered via an extensible technology framework. The Allotrope Foundation manages and enables a community-led effort to standardise ontologies, vocabularies and rules commonly used in scientific data and experiments. The foundation choose the ontology approach as it promoted interoperability across heterogeneous data systems, exploited the relative stability of natural language (not code dependent), and the ease at extendibility of models.

In the current laboratory ecosystem, different customers can present a widely diverse set of requirements and the Allotrope Framework means a reduction in the complexity and diversity of customer requirements both now and in the future. A consolidated set of standards flexible enough to accommodate more dimensionality, detail or complexity to the data, while still supporting legacy data, removes that cost and reduces the friction of introducing new innovations to a mature market. The opportunities the semantic foundation offers opens the door to a whole new world of need and opportunity for innovation and new solutions in the data lifecycle.

Figure 8 shows a representation of the ontology entity model. Their approach utilised the rise of many other existing ontologies to build (more quickly) their framework of knowledge:

- Basic Formal Ontology[38] (BFO) to represent the categories that are shared across a broad range of domains
- W3C SKOS for defining vocabularies and taxonomies
- QUDT vocabulary[39] of units of measurement and quantities
- W3C PROV[40] for provenance information
- W3C CUBE[41] for describing multi-dimensional data, such as statistics

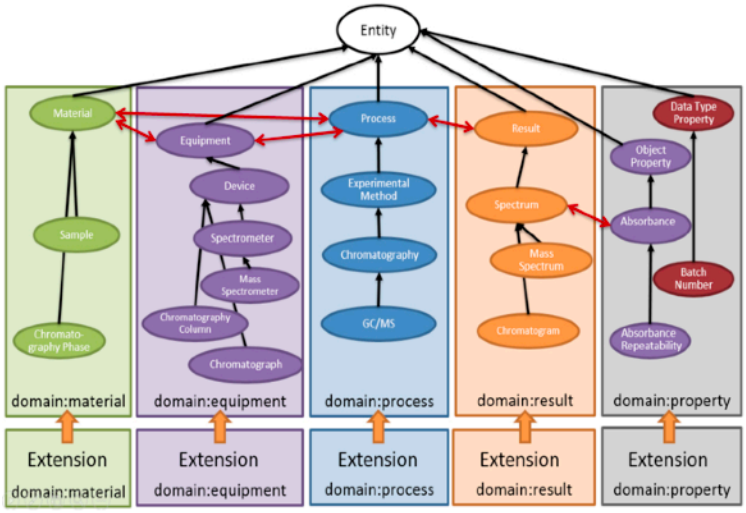

Figure 8 - Allotrope Model

[37] https://www.allotrope.org/

[38] https://basic-formal-ontology.org/

[39] https://www.qudt.org/pages/HomePage.html

[40] https://www.w3.org/TR/prov-primer/

[41] https://www.w3.org/TR/vocab-data-cube/

- PAV ontology[42] for defining different roles of the agents contributing content
- W3C ORG[43] ontology for describing organisations
- W3C SHACL for business rules and validation

The combined data model of the Allotrope Foundation Ontology (AFO) covers over 60 pharmaceutical domain areas from Automated Reactors to X-Ray Powder Diffraction. The AFO has over 3,000 classes and 300 properties as shown in Figure 9. The AFO is also published on the BioPortal of Ontologies[44]. One of the advantages of this is that BioPortal provides mapping services across ontologies. As an example, Figure 10 shows the mapping of some concepts between AFO and SNOMED.

The AFO also utilises W3C SHACL to define conformance rules for data. For example the "Conductivity Shape" defines three property constraints for any entity of type "Conductivity":

- Max and min cardinality of one
- Must use a QUDT numeric value of type xsd:decimal
- Must use a QUDT unit of qudt:SiemensPerMeter

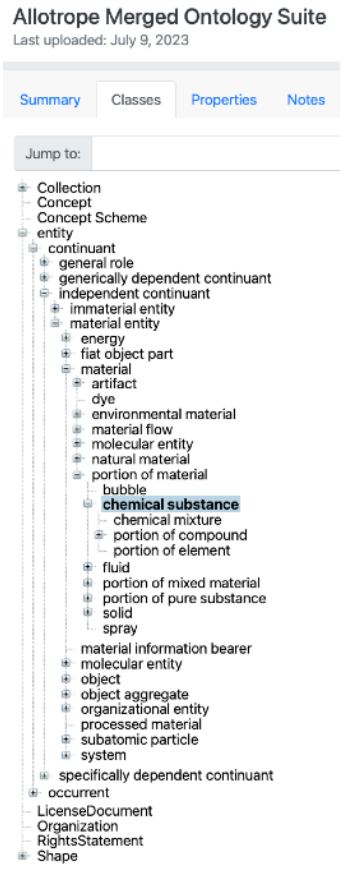

Figure 9 - Allotrope Classes (Partial)

ALLOTROPE MERGED ONTOLOGY SUITE	SNOMED CT
Concentration	concentration
Injection device	injection device
Gas	gas
Gas	gas
Pump	pump
Pressure	pressure
Ambient temperature	ambient temperature
Fluid	fluid
Nephelometry	nephelometry
Balance	balance
Mass	mass
Object	object
Optical density measurement	optical density measurement
Suspension	suspension

Figure 10 - AFO to SNOMED Mapping

[42] https://pav-ontology.github.io/pav/

[43] https://www.w3.org/TR/vocab-org/

[44] https://bioportal.bioontology.org/ontologies/AFO

Case Study: HealthDirect

Healthdirect[45] is an Australian government-funded virtual health service that provides access to health advice and information to help people manage their health and connect them to the right care at the right time. HealthDirect's services include a national health services directory of practitioners, COVID-19 information, pregnancy and child information, aged care information, ambulance triage, and general health symptoms. All of this content is linked to the Australian Health Thesaurus[46] (AHT), which is a SKOS vocabulary of terms and concepts used by these services.

The AHT is a controlled vocabulary of medical, health and human services related concepts, organised into a hierarchical structure - based on the Medical Subject Headings (MeSH)[47] controlled vocabulary. Other data, such as synonyms or alternative labels, are also applied to the concepts providing a rich reference source of health and related services information. The thesaurus is integral to Healthdirect's search and discovery capability across its platforms. It is used to improve the search experience (query expansion, results ranking) and for content management purposes (relevant examples, matching symptoms). The AHT vocabulary, as shown in Figure 11, is also directly linked to similar concepts in the Australian Medicines Terminology (AMT), Pharmaceutical Benefits Scheme, and the Therapeutic Goods Administration vocabularies, and medicine images from Medicines Information Pty.

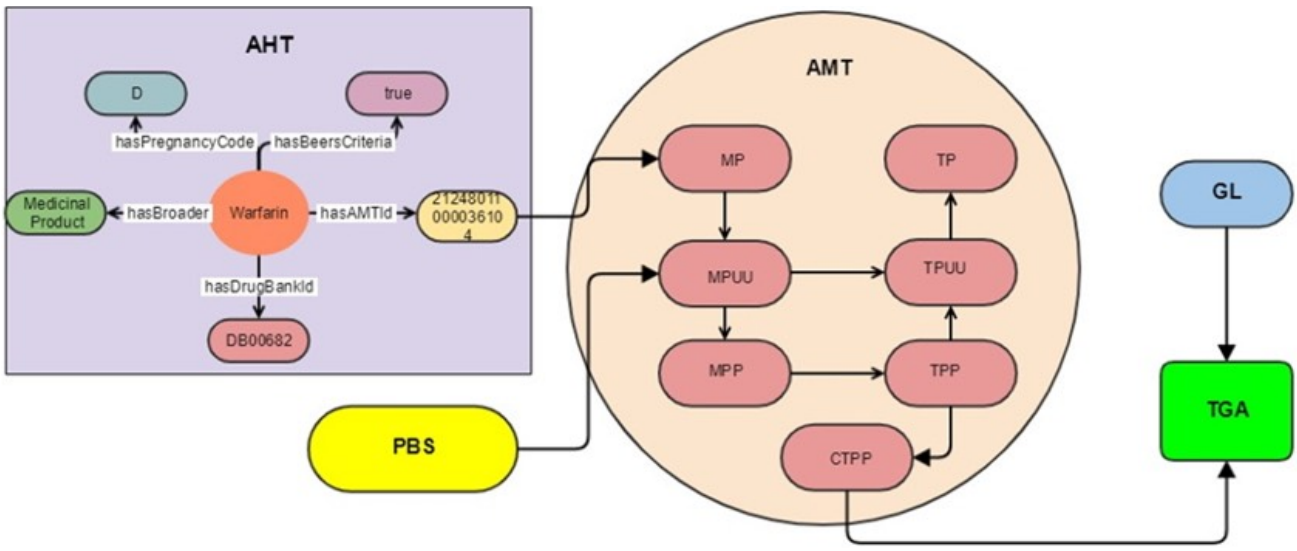

Figure 11 - Australian Health Thesaurus Linkages (HealthDirect supplied)

The AHT contains standard SKOS properties, such as *hasBroader*, but also includes additional extensions, such as *hasBeersCriteria* (see left of Figure 11). The AHT links to the relevant AMT concept

[45] https://www.healthdirect.gov.au/

[46] http://thesaurus.healthdirect.org.au/aht.html

[47] https://www.nlm.nih.gov/mesh/meshhome.html

(see middle of Figure 11), and all the AMT facts are converted from their native format and stored in a graph database.

HealthDirect utlise a vendor solution for managing the AHT SKOS vocabulary as well as the ontologies used in the graph database. Figure 12 shows an example of how one concept (*Asthma*) is described.

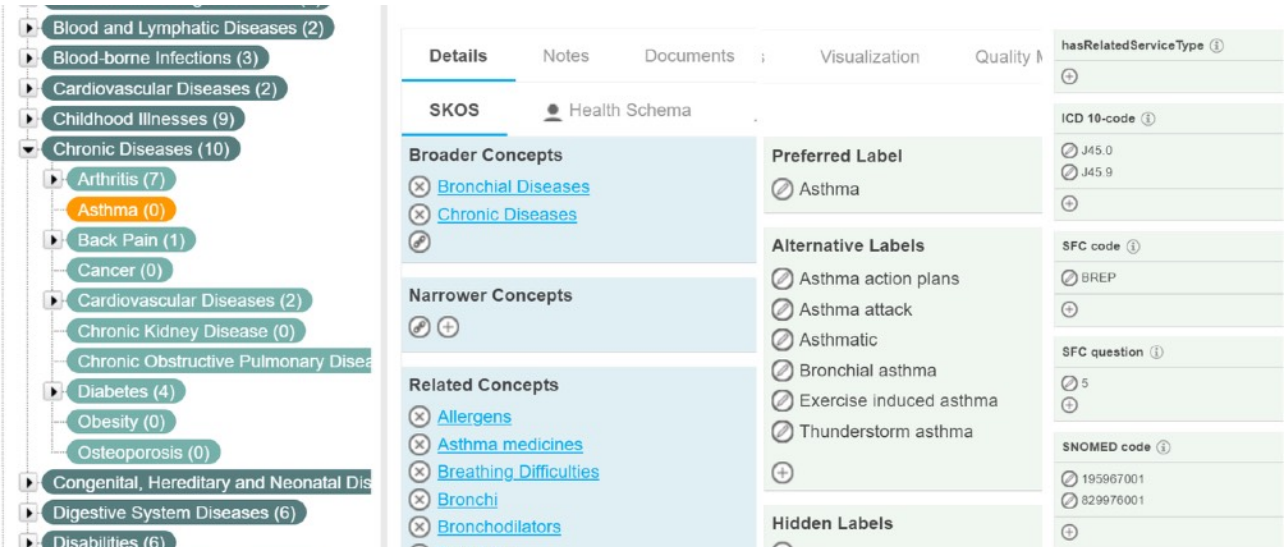

Figure 12 - AHT SKOS View (HealthDirect supplied)

The left column in Figure 12 shows the hierarchy of concepts of the AHT. The central area shows the standard SKOS properties such as broader, narrower, related concepts and the various label properties. The far right shows the extensions to SKOS including mappings to ICD-10 and SNOMED codes.

HealthDirect view knowledge graphs as a fundamental underpinning of their future data fabric process and tools that support:

- Development of metadata, taxonomies and schemas.
- Build and run pipelines to transform data.
- Store and visualise data.
- Create views for data consumption.

Photo by Stephen Foster

National Healthcare Ontology Framework

At the national level, the development of a Healthcare Information Exchange (HIE) would benefit significantly with the knowledge graph approach for two primary reasons. First, the "ontology as a platform" direction provides all the advantages and consistency in terms of semantic modelling and conformance for exchanged healthcare data. Second, the resultant national healthcare ontology will provide the basis for mapping the complexity of data of current state approaches as well as providing a future target state based on a direct graph interface.

Figure 13 shows the national ontology framework that includes a national HIE as the front-end to the ontology platform. The HIE provides legacy and current state access to exchange standards (V2, CDA, DICOM, FHIR, etc) and includes the capability to transform the incoming and outgoing data. This

transformation is primarily mapping between the exchange standard and the national ontology. For example, all the data exchange standards represents Allergy in different structures but the mapping will always be the common concept in the ontology.

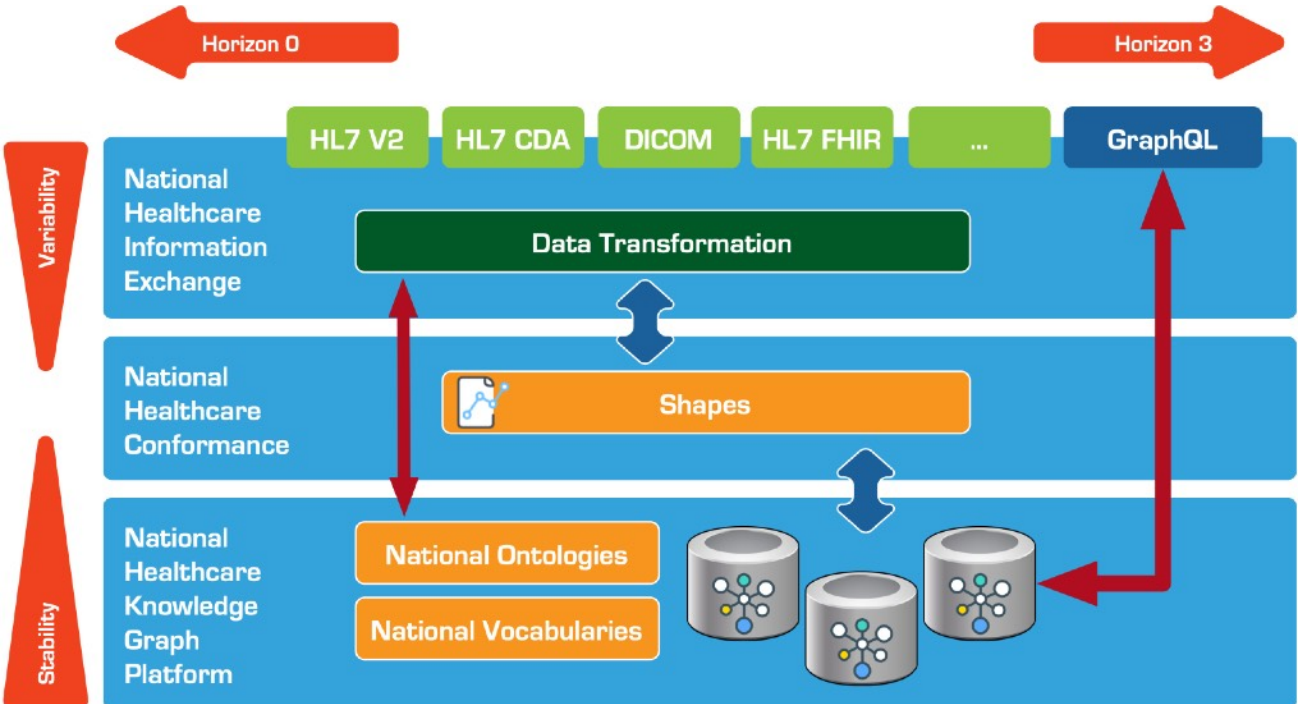

Figure 13 - National Ontology Framework

The HIE will take on the greater role due to the large *variability* of data exchange options in the wider healthcare community. Conversely, the backend Healthcare Knowledge Graph provides a more stable platform as data is harmonised via the healthcare ontology.

As an example of variability, HL7 FHIR has been through 5 versions for the nearly 200 individually defined resources. Each resource has also has 7 levels of maturity. This means there are nearly 700 permutations of FHIR versions that could be supportable. Interestingly, only 14 of the 200 resources are deemed "normative" so there will always be frequent version changes.

The national ontology is a harmonisation and abstraction of the healthcare exchange standards. This also means that the ontology is based on real-world use cases. In addition, the national ontology should be aligned to healthcare conceptual models, such as ISO 13940:2015 System of Concepts to Support Continuity of Care[48] and ISO 12967-2:2020 Health informatics Service Architecture Part 2: Information Viewpoint[49].

The framework includes conformance support with Shapes which are evaluated after the transformation to the common ontology representation. This also means that the business rules for conformance all also harmonised and governed at the ontology layer, not at the various exchange layers.

Finally, as horizon 3 approaches, the healthcare ontology can also be exposed for direct access (with technologies like GraphQL). This will mean that exchanges with healthcare data providers will utlise and benefit from the (single) ontology approach. In addition, the development process for providers will be streamlined and minimised as only one "exchange" standard needs to be supported.

[48] https://www.iso.org/standard/83432.html

[49] https://www.iso.org/standard/71038.html

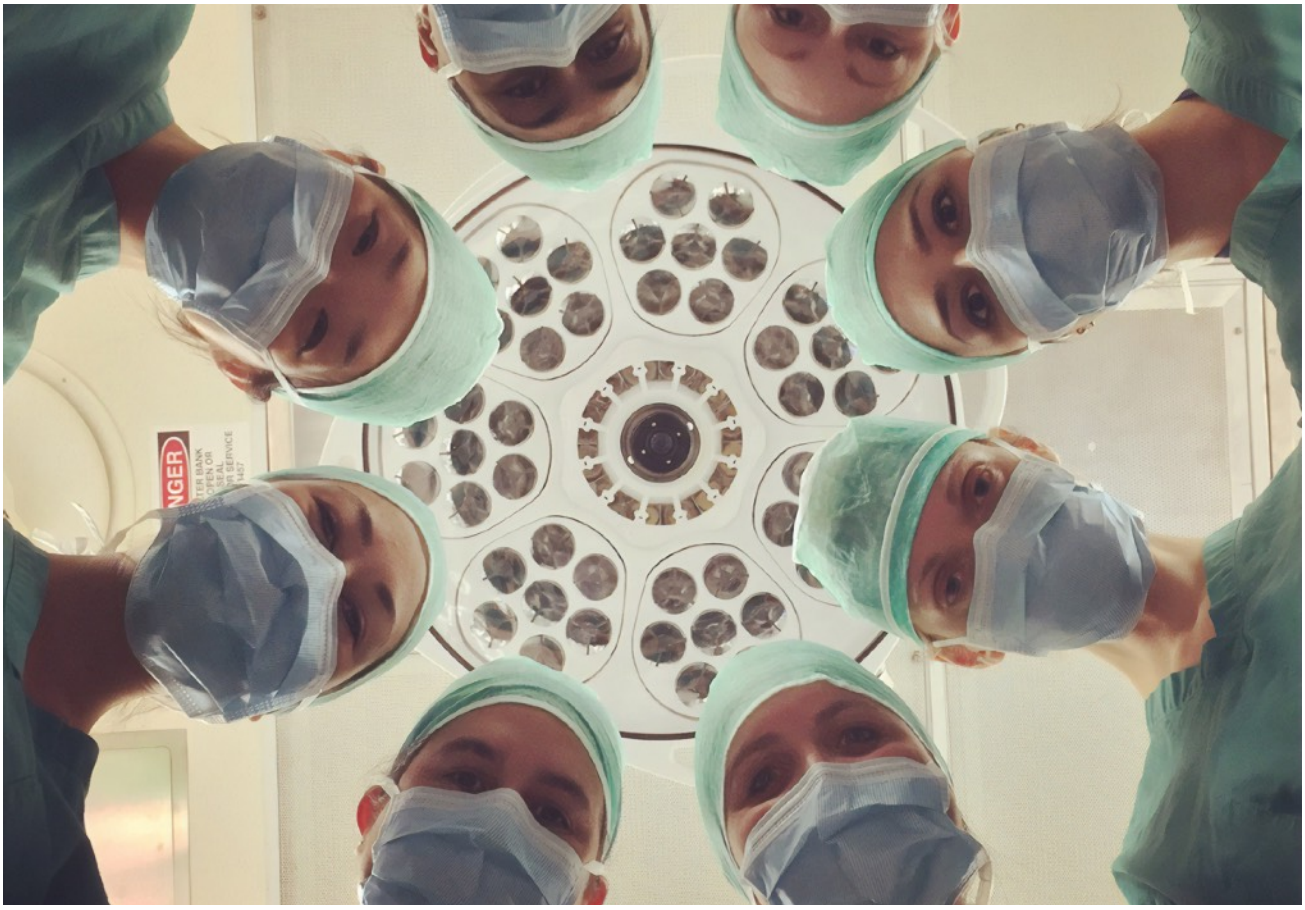

Photo by National Cancer Institute

Summary

The healthcare sector is at the forefront of a massive data transformation era. Modern semantic technologies can drive such industry-wide transformations with new opportunities that challenge past practices with advanced graph-based information services. These technologies can break-down the barriers across the healthcare sector and domains by opening up data and sharing knowledge for better and safer decision management. Knowledge graphs are also critical in supporting future healthcare artificial intelligence deployment as they provide the fundamental "explainability" capability necessary for trusted and transparent machine learning outcomes[50].

The healthcare sector has a long and wide technological gap between the current and future state. The use of semantic technologies will provide the common fabric for connectivity, interoperability and cohesion. However, this needs to be purposely managed and led in this direction otherwise new silos are created for the sector, which it should not accept.

This white paper has presented a new semantic-driven healthcare sector underpinned by mature knowledge graph technologies. The roadmap to this future needs wide sector consensus and planning, and promises to deliver improved opportunities for data exchange, data conformance, and knowledge sharing for safer delivery of healthcare services.

[50] Knowledge Graphs and Explainable AI in Healthcare, E. Rajabi (et al), Information 2022 https://www.mdpi.com/2078-2489/13/10/459

References

AMT - Australian Medicines Terminology is a subset of SNOMED CT that describes medicines in a standardised format

- https://www.healthterminologies.gov.au/

GraphQL - Graph Query Language

- https://graphql.org/

HL7 V2 - HL7's Version 2 Messaging standard and Protocol for Electronic Data Exchange in Healthcare Environments

- https://www.hl7.org/implement/standards/product_section.cfm?section=13

HL7 Clinical Document Architecture (CDA) - Hl7 V3 markup standard that specifies the structure of clinical documents

- https://www.hl7.org/implement/standards/product_brief.cfm?product_id=496

HL7 Fast Healthcare Interoperable Resources (FHIR) - HL7 standard for health care data exchange

- https://www.hl7.org/fhir/

ICD-10 - International Statistical Classification of Diseases and Related Health Problems

- https://icd.who.int/browse10/2016/en

LOINC - International standard for identifying health measurements, observations, and documents

- https://loinc.org/

SNOMED CT - An international clinical terminology designed to record clinical information at the point of care

- https://www.snomed.org/

W3C Semantic Web - A set of formal knowledge representation language standards for semantic interoperability

- https://www.w3.org/TR/rdf-concepts/
- http://www.w3.org/TR/rdf-schema/
- https://www.w3.org/TR/owl2-overview/

W3C SKOS - Simple Knowledge Organization System for expressing the structure and content of vocabularies

- https://www.w3.org/TR/skos-primer/

W3C SHACL - Shapes Constraint Language for validating RDF graphs against a set of conditions

- https://www.w3.org/TR/shacl/

W3C SPARQL -Defines the syntax and semantics of the a query language for RDF graphs

- https://www.w3.org/TR/sparql11-overview/