Introducing the Semantic Technologies

Why Resource Descriptions Framework (RDF) and Ontologies

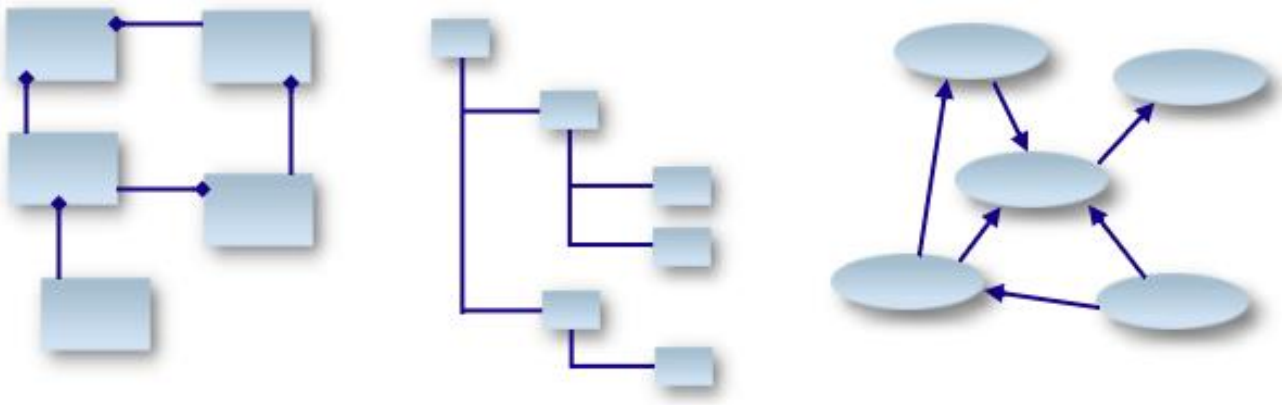
The *Semantic Web* is a *Web 3.0 web technology* – a way of linking data between systems or entities that allow for rich, self-describing interrelations of data available across the globe on the web.

In essence, it marks a shift in thinking from publishing data in human readable HTML documents to machine readable documents. That means that machines can do a little more of the thinking work for us.

Today, much of the data we get from the web is delivered to us in the form of *web pages* – HTML documents that are linked to each other through the use of *hyperlinks*. Humans or machines can read these documents, but other than typically seeking keywords in a page, machines have difficulty extracting any meaning from these documents.

Resource Description Framework (RDF) is a common acronym within the semantic web community because it is one of the basic building blocks for forming the web of semantic data. What it defines is a type of database called a **graph database**.

Introducing the Graph Database



Data is generally stored in relational databases. This has been a suitable model for the last few decades as it enables reasonable computers to store the data and allow searching. Upside is that each piece of data is only stored in one place and each piece of data is atomic.

The downside is that the database tables have to be developed in advance usually from entity relational diagrams, the tables don't naturally relate to reality, and it is hard to link various databases together as required, especially if across different systems. A more natural representation for the Internet (and Web) is the network or graph model. Data items are defined as nodes and the

relationships defined as the arcs. A graph can represent anything and allow different pieces of disparate data to be related to each other (see dog and cat example below).

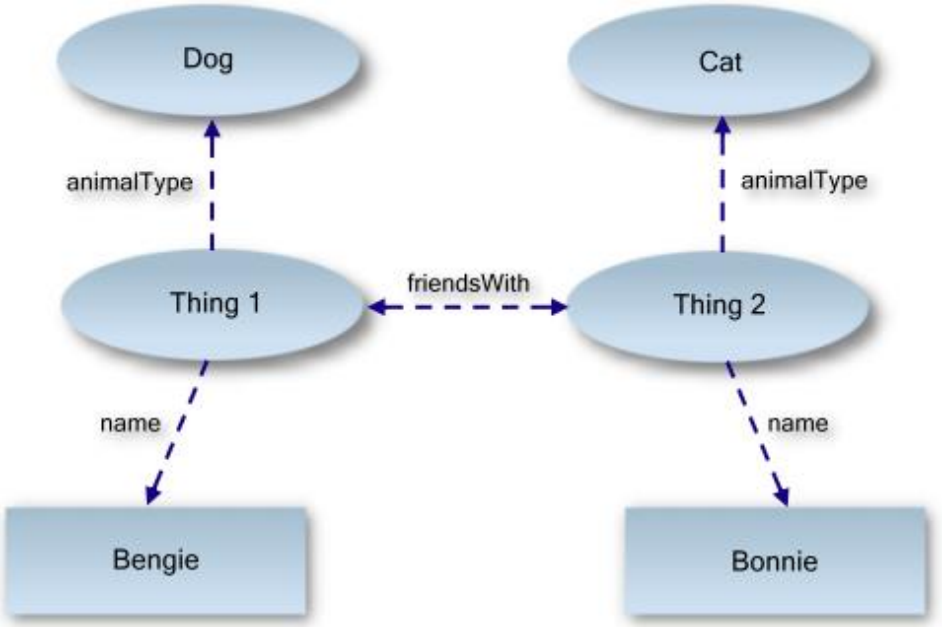
Implied Relationships through RDF

It's easiest first to look at a series of statements about how things relate to each other and to visualise these as a graph before looking at how these relationships might be expressed in RDF.

Look at the following statements describing the relationship between a dog (called Bengie) and a cat (called Bonnie).

- ➔ Bengie is a dog
- ➔ Bonnie is a cat
- ➔ Bengie and Bonnie are friends.

Using these three simple statements, let's turn this into a data graph.



The relationships implied by this graph are fairly intuitive, we can see that our two *things* – identified by "Thing 1" and "Thing 2" – have the *properties* **name**, **animalType** and **friendsWith**.

From this, we can see that "Thing 1" is named Bengie and "Thing 2" is Bonnie. "Thing 1" is a dog and "Thing 2" is a cat. And finally, both are friends with each other (implied by the *friendsWith* property pointing in both directions).

RDF and triples are a way of defining a network as the triple <subject, predicate, object> defines two nodes (subject, object) and the link (predicate).

Spatial data currently held in relational databases can be converted to triple stores and managed with software such as Fuseki (an HTTP interface to RDF data).

Current relational databases can be made into virtual triple stores as well. Triple stores can be queried using SPARQL (SQL for triple stores). Importantly, each element of a triple can be a URI (IRI now to deal with different languages, URIs being mainly in English), allowing further distribution of data and definitions. For example, if a predicate is called “near”, the IRI can point to a location

where the concept is defined. It may be the Euclidean distance between two points (spatial) or the distance between people in a family tree.

Building Complex Relationships

Extra links can be added on the fly without the need to redefine databases. For spatial data (parcels in a cadastre) assumes the norm is one person owns one parcel. It is easy to add links to show ownership of many parcels by one person, multiple people owning one parcel etc. Such changes can be made on the fly by the user as required, and there is no need for a data supplier to redesign databases to accommodate such changes.

RDF really comes into its own when an end user tries to link information from multiple sources that have been independently designed. Current systems would require this information to be mapped and the outputs put into interoperable formats such as OGC web services and processes developed to manage system and service changes.

This sort of information interchange across incompatible, independently designed systems takes time, money and human semantic interpretation of the different datasets and schemas. This process would need to occur with any other services and knowledge linkages. It will always require humans to understand the meaning of the data and agree on common formats.

Of importance to the semantic web, RDF enables access to knowledge and rules, as well as the data allowing sophisticated user defined operations to occur, again without the data supplier having to configure systems specifically for a user. Ontologies and rules allow high level queries and processing to occur by many users on the fly that is currently not possible.

Big data and cloud based processing by clusters of processors can realise the semantic web, RDF and triple stores. Hadoop (invented by Yahoo! and used by Facebook) stores data as <key, value> pairs and can be used to represent RDF triples. Fast querying methods (Pivotal HD; Cloudera Impala) mean that Hadoop is becoming an alternative to Oracle and other large database systems.

Ontologies Assist with Modelling

So how do we model information from diverse sources? Firstly, the sites need to apply a common, standard vocabulary to describe its data that is contextually consistent. For example, the term 'admin boundary' should mean the same thing for both sites, as should the term 'locality'.

This may be done by the two sites adopting the same **base ontology**, or a common vocabulary, for expressing the meaning behind the data it exposes, and publishing that data on a queryable endpoint so that the two sites can communicate with each other across the web.

The cross-domain knowledge sharing need not just apply to websites, but also within the knowledge bases built by *organisations*. Semantic web technologies are not restricted to applications or information published on the web.

Although there may be a little more groundwork required when first setting up a semantic database, the benefits for ease of cross-domain integration from across the globe and the time saved and ideas gained from doing so are highly significant.

Standard vocabularies, or formal ontologies representing terms within a domain of knowledge, are already available freely from various organisations dedicated to creating standard vocabularies for a range of subjects – for example media terms, or biomedical terms, or scientific terms.

Some examples:

- Dublin Core Metadata Initiative (DCMI) – creates ontologies for a range of subjects, particularly focusing on common, every day terms and terms important in media
- Friend Of A Friend (FOAF) – focuses on developing a standard vocabulary/ontology for social networking purposes
- OpenCyc – an ontology of everyday, common sense terms.

Dr David McMeekin from our Program 3 team released a report on **ontologies, vocabularies and various tools;** an environmental scan of tools for ontologies and vocabularies. David's report is the first comprehensive compendium of tools in this area.

A copy of the report can be downloaded here.

Supporting Organisational Reform

Publishing and consuming Linked Data allows for cooperation without coordination. Data publishers may effectively cooperate to produce – individually – data sets that may be reused and recombined by unknown third parties. There is no need for Linked Data publishers to coordinate efforts. Use of the Linked Data guidelines (which include a mandate to reuse existing vocabularies and to publish details of new vocabularies used) is sufficient.

Future proofing technology choices in a rapidly changing IT landscape is prudent. No organisation should be beholden to one vendor to store, access or analyse organisational data. Indeed, through leveraging modern RDF data exchange formats and international Internet standards, organisations have every opportunity to avoid vendor lock-in permanently.

The CRCSI Spatial Infrastructures Program team will be providing our partners with a range of short papers to support the knowledge transfer of research concepts throughout 2016 – 2017.