Linked Data and RDF

COMP60421

Sean Bechhofer

sean.bechhofer@manchester.ac.uk

Building a Semantic Web

- Annotation
  - Associating metadata with resources
- Integration
  - Integrating information sources
- Inference
  - Reasoning over the information we have.
  - Could be light-weight (taxonomy)
  - Could be heavy-weight (logic-style)
- Interoperation and Sharing are key goals

6

Linked Data*

- Linked Data or the Data Web is about using the Web to connect related data that wasn't previously linked.
- The intention is that we move from a web of *documents* to a web of *data*
  - The Web as database

- The Linked Data approach builds heavily on RDF.

*Linked data slides based on material from Ian Davis and Tom Heath:
http://www.slideshare.net/iandavis/30-minute-guide-to-rdf-and-linked-data

Database tables

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

Rows represent "things"

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

Columns represent "properties"

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

Intersections represent properties of things

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

Graphical Representation

book —— title ——▶ The Rotters' Club

more generally:

subject —— property ——▶ value
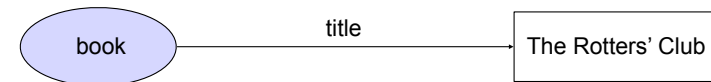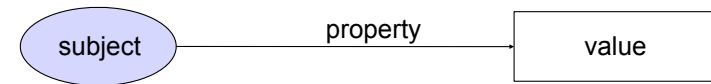
Selecting multiple properties

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

Multiple properties graphically

Relations between "things"
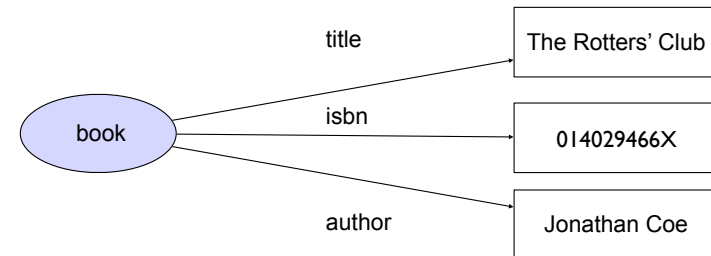
	title	The Rotters' Club
book	isbn	014029466X
	author	Jonathan Coe
publisher		
publisher	name	Penguin Books
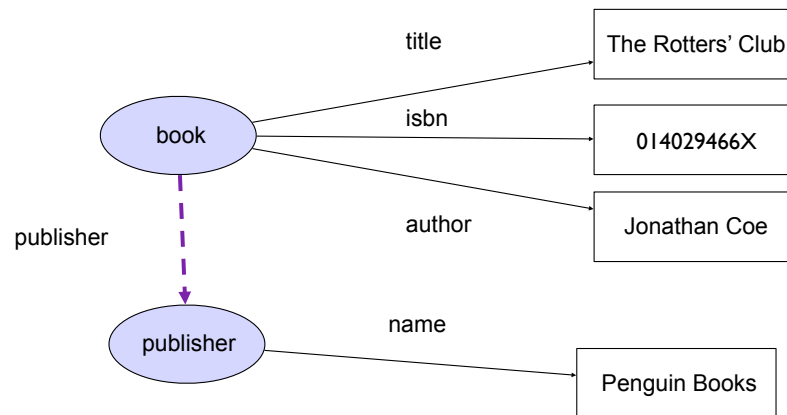
Identification

- We need to be able to identify things globally and uniquely.
- URIs provide this capability
- Key to Linked Data is the use of URIs, specifically http:// URIs.

URIs in graphs

URIs as names for nodes

http://example.com/person/176 → http://example.com/name → Jonathan Coe

URIs as names for relations
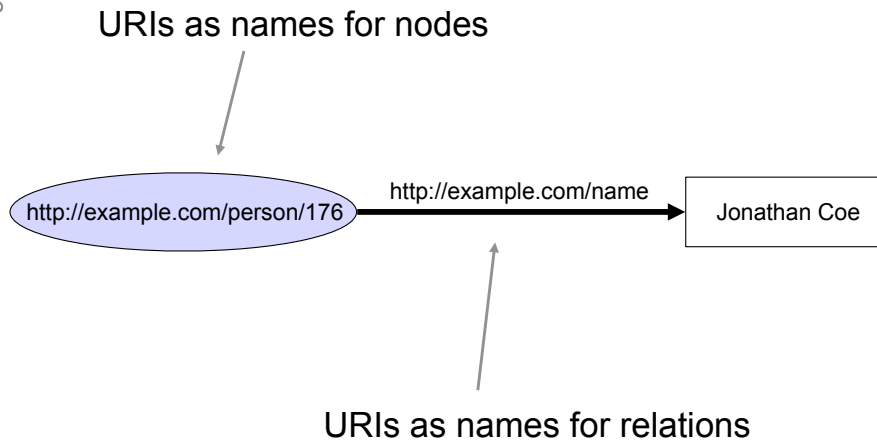
URIs and naming

- URIs identify the things we are describing.
- If two people create data using the same URI, the assumption is that they are describing the same thing.
- Merging/integrating data then becomes easy
  - Although introduces issues of URI control.

Graph Merging
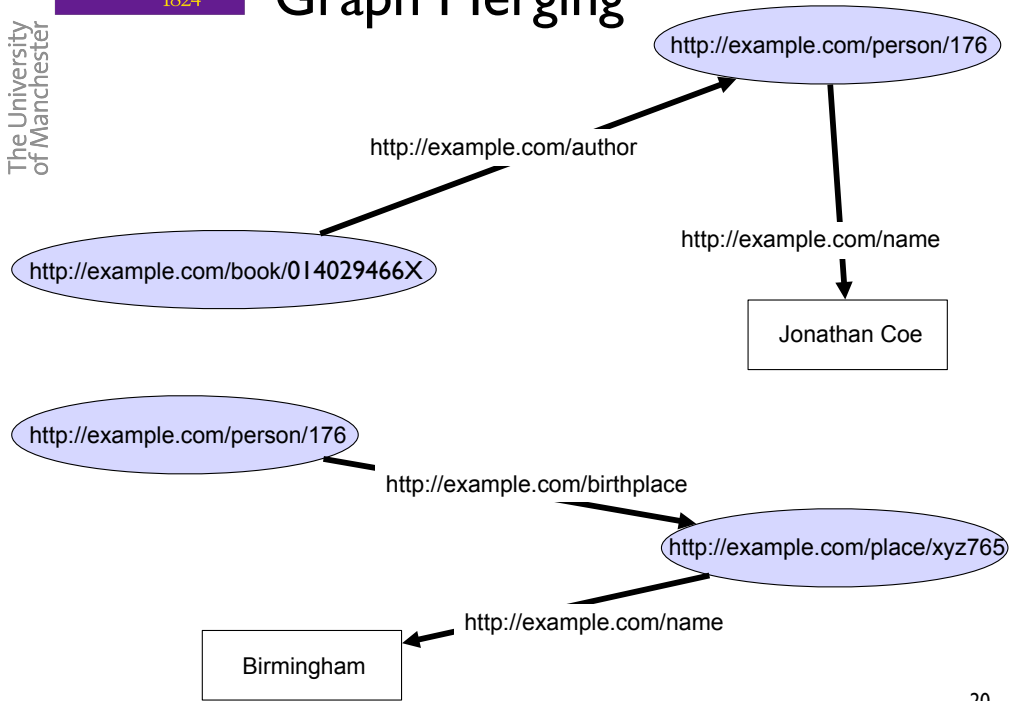
Graph Merging
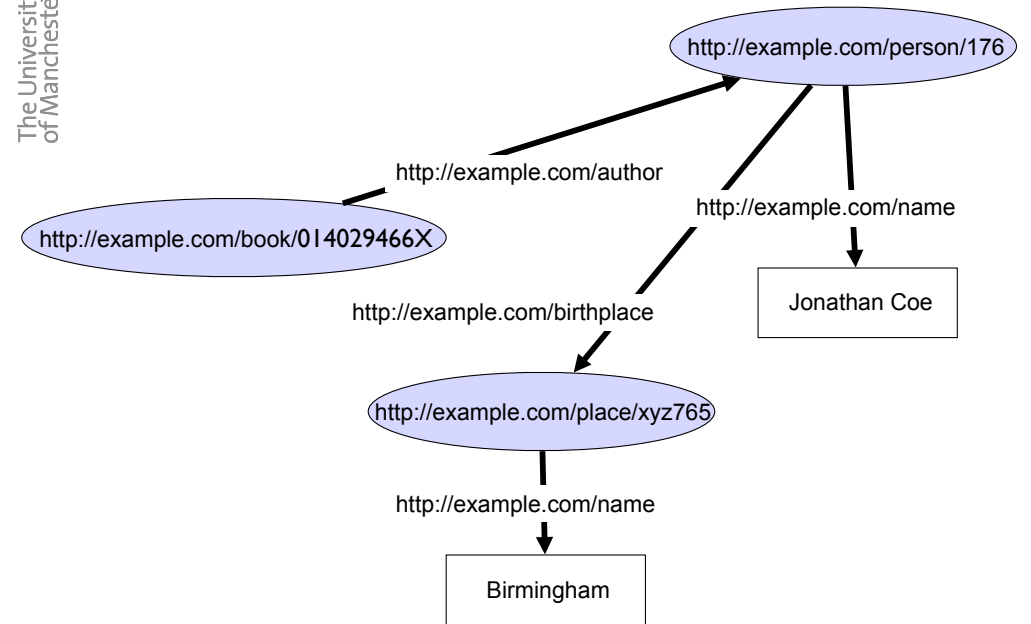
URIs are active

- URIs can be more than just names -- they can be dereferenced, and information can be retrieved.
- In particular, we can lookup the URIs in a graph and potentially retrieve more information about the URI.
- "Follow your nose" navigation
- Information should be returned in appropriate, machine readable formats (e.g. another graph)

Linked Data Principles

1. Use URIs as names for things
2. Use http URIs so that those names can be dereferenced.
3. When a URI is looked up, provide useful information
4. Include statements that link to other URIs so that more information can be discovered.

- Common infrastructure facilitates construction of applications.
  - Largely browsers up to now….
- Other guidelines relating to connecting documents with the data that describes them.
  - Use of content negotiation to supply "appropriate" representations
  - Use of microformats/RDFa to publish data

RDF

- RDF stands for Resource Description Framework
- It is a W3C Recommendation
  - http://www.w3.org/RDF
- RDF is a graphical formalism (+ concrete syntax)
  - for representing metadata
  - for describing the semantics of information in a machine-accessible way
- Provides a simple data model based on triples.

- Allows us to represent relationships between things.
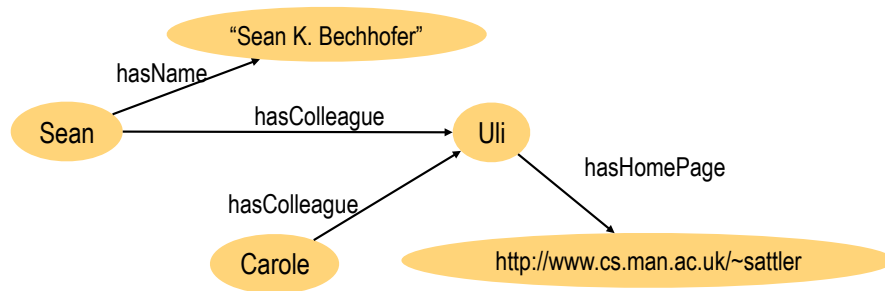
The RDF Data Model

- Statements are <subject, predicate, object> triples:
  - `<Sean,hasColleague,Uli>`
- Can be represented as a graph:
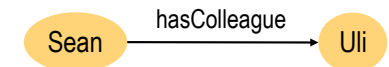
- Statements describe properties of resources
- A resource is any object that can be pointed to by a URI
- Properties themselves are also resources (URIs)

Linking Statements

- The subject of one statement can be the object of another
- Such collections of statements form a directed, labeled graph

- Note that the object of a triple can also be a "literal" (a string)

RDF Syntax

- RDF has an XML syntax that has a specific meaning:
- Every `Description` element describes a resource
- Every attribute or nested element inside a `Description` is a `property` of that `Resource`
- We can refer to resources by URIs

```
<Description about="some.uri/person/sean_bechhofer">
  <hasColleague resource="some.uri/person/uli_sattler"/>
  <hasName rdf:datatype="&xsd;string">Sean K. Bechhofer</hasName>
</Description>
<Description about="some.uri/person/uli_sattler">
  <o:hasHomePage>http://www.cs.mam.ac.uk/~sattler</o:hasHomePage>
</Description>
<Description about="some.uri/person/carole_goble">
  <o:hasColleague resource="some.uri/person/uli_sattler"/>
</Description>
```

What does RDF give us?

- A mechanism for **annotating** data and resources.
- Single (simple) data model.
- Syntactic consistency between names (URIs).
- Low level **integration** of data.

- The Linked Data/Web of Data approach.

28

Querying RDF: SPARQL

- RDF provides us with a way of representing information as a graph
- SPARQL allows us to query this information

  http://www.w3.org/TR/sparql11-overview/
- Provides a query language and the description of a protocol for interacting with SPARQL "endpoints" via HTTP

```
PREFIX etree:<http://etree.linkedmusic.org/vocab/>
PREFIX mo:<http://purl.org/ontology/mo/>
PREFIX event:<http://purl.org/NET/c4dm/event.owl#>
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX timeline:<http://purl.org/NET/c4dm/timeline.owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?artist WHERE
{
?art rdf:type mo:MusicArtist.
?art skos:prefLabel ?artist.
}
```

24

RDF(S): RDF Schema

- RDF gives a formalism for meta data annotation, and a way to write it down in XML, but it doesn't give any special meaning to vocabulary such as subClassOf or type
  - Interpretation is an arbitrary binary relation
- RDF Schema extends RDF with a schema vocabulary that allows you to define basic vocabulary terms and the relations between those terms
  - Class, Property
  - type, subClassOf
  - range, domain

RDF(S)

- These terms are the RDF Schema building blocks (constructors) used to create vocabularies:
  - `<Person,type,Class>`
  - `<hasColleague,type,Property>`
  - `<Professor,subClassOf,Person>`
  - `<Carole,type,Professor>`
  - `<hasColleague,range,Person>`
  - `<hasColleague,domain,Person>`
- Semantics gives "extra meaning" to particular RDF predicates and resources
  - specifies how terms should be interpreted

What does RDF(S) give us?

- Ability to use simple schema/vocabularies when describing our resources.
- Consistent vocabulary use and sharing.
- Basic inference

- Note that RDF is a data model. There are many ways of serialising this data:
  - RDF/XML
  - Turtle
  - N3
  - json-ld

Problems with RDF(S)

- RDF(S) is too weak to describe resources in sufficient detail
  - No localised range and domain constraints
    - Can't say that the range of hasChild is Person when applied to Persons and Elephant when applied to Elephants
  - No existence/cardinality constraints
    - Can't say that all *instances* of Person have a mother that is also a Person, or that Persons have exactly 2 parents
  - No transitive, inverse or symmetrical properties
    - Can't say that isPartOf is a transitive property, that hasPart is the inverse of isPartOf or that touches is symmetrical
- Difficult to provide reasoning support
  - No "native" reasoners for non-standard semantics
  - May be possible to reason via FO axiomatisation

OWL

- OWL: Web Ontology Language
- Extends existing Web standards
  - Such as XML, RDF, RDFS
- Is (hopefully) easy to understand and use
  - Based on familiar KR idioms
- Of "adequate" expressive power
- Formally specified
  - Possible to provide automated reasoning support
- But you already know all this…

Linked Data Benefits

- Separation of data from formatting and presentational aspects
- Self-describing data. Applications encountering unfamiliar vocabularies can dereference and access definitions
- Simplified data access via HTTP and RDF.
  - Heterogeneity of Web APIs
- Open
  - Applications not implemented against fixed set of data sources.
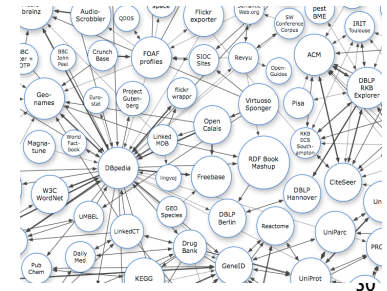
RDF/Linked/Open Data

- dbpedia
  - RDFised version of wikipedia
  - Scraping structured information from info-boxes.
  - Quality?
- Government Data
  - https://data.gov.uk/organogram/cabinet-office
- Open Data Institute
- BBC
- GeoNames
  - Geographical data
  - Lat/long, postal codes etc.
- LCSH
  - SKOS

dbpedia

wikidata

BBC

BBC

Library of Congress

etree.linkedmusic.org

Live Music Archive Linked Data

This server provides access to Linked Data that describes the audio held in the Internet Archive's Live Music Archive (also sometimes known as "etree"). The metadata from etree has been converted to RDF and is exposed through a SPARQL endpoint along with browsable pages.

The dataset contains information describing over 100,000 performances by 4,000 artists including 1,600,000 individual tracks, each of which may be available in a number of formats.

Sample Resources

- The artist *Mogwai*: http://etree.linkedmusic.org/artist/422accc0-4aac-012f-19e9-00254bd44c28.
  - Mogwai Live at the Forum on 16-10-1999: http://etree.linkedmusic.org/performance/mogwai1999-10-16.flac16
  - The Forum venue (as played by Mogwai on 16-10-1999): http://etree.linkedmusic.org/venue/mogwai1999-10-16.flac16
- The artist *Calexico*: http://etree.linkedmusic.org/artist/422f010-4aac-012f-19e9-00254bd44c28
  - Calexico Live at the Serenadenhof on 11-08-2009: http://etree.linkedmusic.org/performance/calexico2009-08-11.flac
- The artist *Andrew Bird*: http://etree.linkedmusic.org/artist/422e57d0-4aac-012f-19e9-00254bd44c28
- VoID metadata describing the dataset: http://etree.linkedmusic.org/etree

Vocabularies

The dataset makes use of a number of vocabularies including:

- The Music Ontology: http://purl.org/ontology/mo/
- The Similarity Ontology: http://purl.org/ontology/similarity/
- The Event Ontology: http://purl.org/NET/c4dm/event.owl

Links
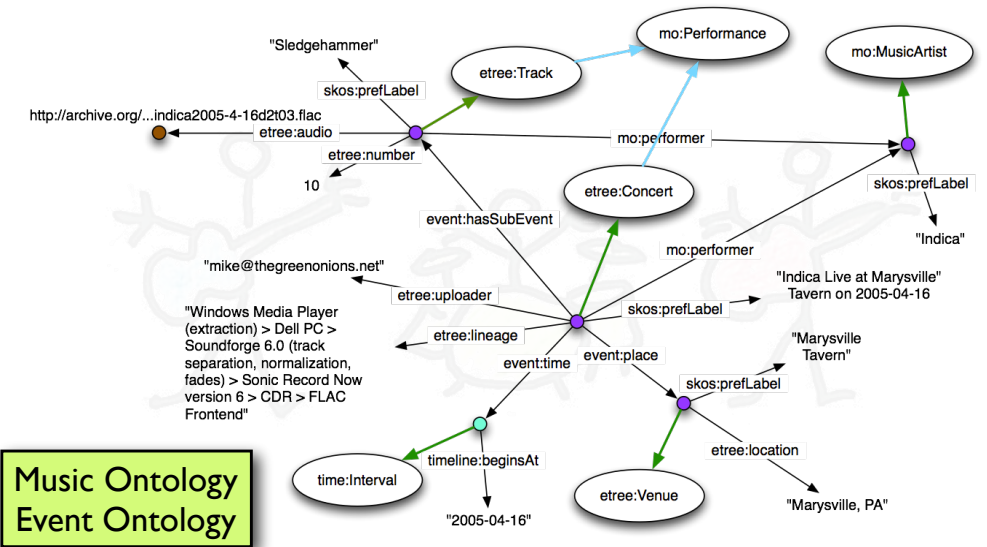
Artists are linked to entries in MusicBrainz, while locations of performances are linked to entries in Geonames and last.fm using a number of methods. Any such alignments are represented explicitly as similarities.

SPARQL Endpoint

A SPARQL endpoint for the data is available at http://etree.linkedmusic.org/sparql. Some sample queries are shown below. Note that as this is currently an experimental service, query timeouts are limited.

- All artists in the collection.
- Artists in the collection with mappings to MusicBrainz.
- Artists and the locations of recorded performances.

37

Modelling

"Sledgehammer"

mo:Performance
mo:MusicArtist
etree:Track
skos:prefLabel

http://archive.org/...indica2005-4-16d2t03.flac
etree:audio
mo:performer
skos:prefLabel
"Indica"

etree:number
10

event:hasSubEvent
etree:Concert
mo:performer

"mike@thegreenonions.net"
etree:uploader
skos:prefLabel
"Indica Live at Marysville Tavern on 2005-04-16"

"Windows Media Player (extraction) > Dell PC > Soundforge 6.0 (track separation, normalization, fades) > Sonic Record Now version 6 > CDR > FLAC Frontend"
etree:lineage
event:place
"Marysville Tavern"

event:time
skos:prefLabel

time:Interval
timeline:beginsAt
etree:Venue
etree:location
"Marysville, PA"

"2005-04-16"

**Music Ontology
Event Ontology**

38

Similarities

Similarity Ontology
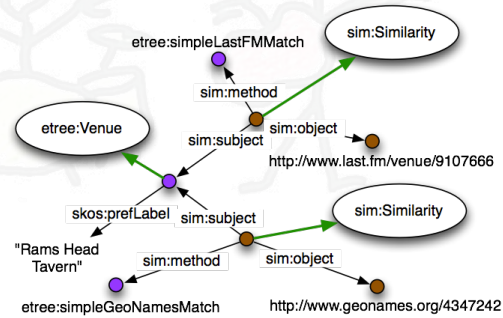
Big Picture
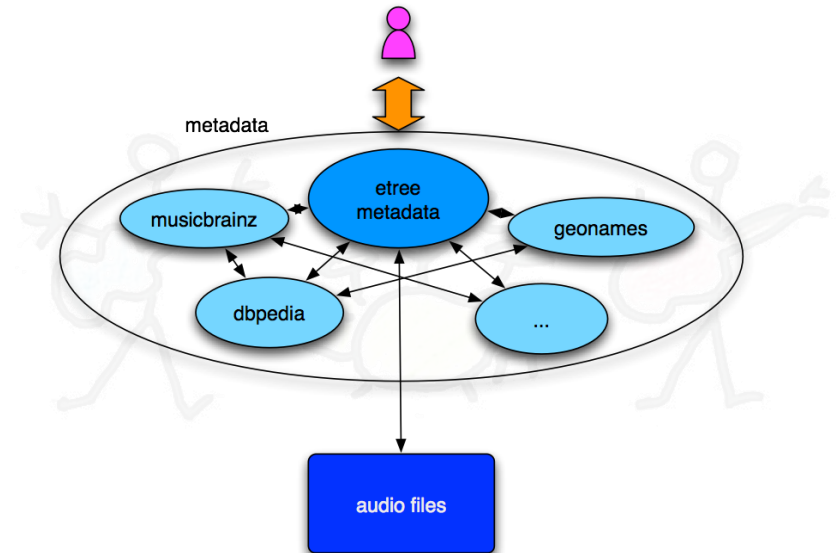
LD in Use

- Five Stars of Open Linked Data
  - http://inkdroid.org/journal/2010/06/04/the-5-stars-of-open-linked-data/
  - ★ Make data available
  - ★ Make it available as structured data
  - ★ Use non-proprietary formats
  - ★ Use URLs to identify things
  - ★ Link your data
- Costs and Benefits
  - http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

Issues with Linked Data

- Identity and co-reference
  - Management of identities
  - How do we handle the fact that different URIs may be used to refer to the same things?
  - Use of owl:sameAs may be too strong (can result in all information, including annotations, metadata etc.) being merged.
- Visualisation
  - Big Fat Graph
- Versioning
  - Version information in URLs?
  - Versioning at architectural level (Memento)
  - How does versioning play with a "follow your nose paradigm"?
- Querying
  - Distributed query across data sets
  - LD applications tend to use an "extract, transform, load" approach.

Issues with Linked Data

- A focus on *data*
  - Vocabularies used to facilitate integration
  - Little *deep* semantics.
  - "Big O vs little o"
    - Role of SKOS and RDF(S)
- Scalability
- A focus on mechanisms for *data publication* rather than *consumption*
  - Lots of work on "recipes", mangling relation sources into RDF etc.
  - What do you actually *do* with the stuff?
  - End user applications
    - Smart cities
  - *Build it and they will come….???*

Do you need them all?

1. Use URIs as names for things
2. Use http URIs so that those names can be dereferenced.
3. When a URI is looked up, provide useful information
4. Include statements that link to other URIs so that more information can be discovered.