

Having Triplets – Holding Cultural Data as RDF

Kate Byrne, School of Informatics, University of Edinburgh

18th September 2008



Outline

Background

- a (very) brief history of Information Management
- RCAHMS and The National Monument Record of Scotland

RDB2RDF Conversion

- why would you want to?
- how do you do it?
- problems with the basic procedure
- a schema for cultural heritage

Including Thesauri

- using SKOS
- integrating with archive content

Milestones in Information Management

- 1450s:
 - Gutenberg invents movable type printing
 - Aside: 2008 marks quincentenary of printing in Scotland.
See <http://www.500yearsofprinting.org/>
- 1950s:
 - first “modern” computers come into use
- 1970s:
 - the Internet emerges
 - database management systems begin to be widely used
- 1990s:
 - Tim Berners-Lee invents the World Wide Web (1990)
 - Tim Berners-Lee proposes the **Semantic Web** (May 1994)
 - first version of **RDF** becomes W3C Recommendation (1999)

The Semantic Web is Old!

- 1994 was 14 years ago – aeons in web time
- Not reached critical mass – not enough triples
- Is RDF part of the problem?
- Cultural organisations are often “early adopters”
 - lots of interest in converting data for the Semantic Web
 - but not many full, live systems

RCAHMS

The Royal Commission on the Ancient and Historical Monuments of Scotland

- Founded in February 1908
- <http://www.rcahms.gov.uk/>
- One of Scotland's 6 National Collections



- Mission –
 - **survey** the built environment
 - **maintain a record** of buildings and archaeological sites
 - **promote understanding** of the material

RCAHMS Collections

- Several million items in various collections –
 - field data, measured survey, oblique aerial survey
 - architects' plans, photographs, manuscripts, maps, *etc*
 - aerial photographic repository for Scotland, plus TARA
- National Monument Record of Scotland (NMRS)
- available online – **Canmore**
- Snapshot available for research –
 - 270,000 site records, each with text document
 - 750,000 archive item records
 - *Tether* uses all this data: warts and all
 - annotated corpus:
 - 1500 documents
 - Named Entities and Text Relations

Users

- **Professional:** archaeologists, architects, planners, developers
- **Semi-specialist:** local history societies, further education
- **Non expert:** casual interest, tourism, schools

Background

- a (very) brief history of Information Management
- RCAHMS and The National Monument Record of Scotland

RDB2RDF Conversion

- why would you want to?
- how do you do it?
- problems with the basic procedure
- a schema for cultural heritage

Including Thesauri

- using SKOS
- integrating with archive content

Why Convert? – The Hidden/Invisible/Deep Web Problem

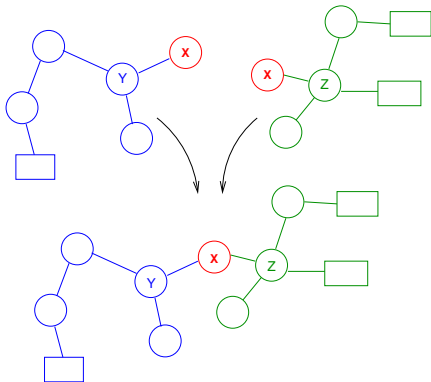
- Most data is (still) in databases, especially “good” data:
 - carefully curated datasets, built over decades/centuries
 - like NMRS
- Web crawlers can't see inside databases –
- – unless you “expose” individual search results:
- http://www.rcahms.gov.uk/pls/portal/canmore.newcandig_details_gis?inumlink=8019
- The Semantic Web lets you put your data “out there”

Why Convert? – Linked Data

- Related information:
 - NAS & GRO: births, deaths, marriages – *Scotland's People*
 - RCAHMS: sites from Neolithic to now – *Scotland's Places*
 - NMS: excavation finds, cultural objects
 - NLS: bibliographic material supporting all of it
- Interconnecting relational databases is hard:
 - you need to know the schema in detail
 - with SQL, you **cannot** query without knowing attribute name
 - with SPARQL, you can
 - security issues
 - complex networking protocols – not http
 - whereas RDF was designed for data linking...

Dataset Linking in RDF

- Same resource node appears in two graphs?
ie same URI
- – graphs are automatically linked



Alternatives to Conversion

- You don't *have* to instantiate the database as RDF:
 - query SQL database using SPARQL – eg SquirrelRDF, R2D2
 - virtual RDF graph interface for relational dbs – eg D2RQ
 - hybrid “middleware” database engine – eg Virtuoso
- Saves duplicating data – but more work at query time
- All need the database schema available
- Principles same as for full conversion

Background

- a (very) brief history of Information Management
- RCAHMS and The National Monument Record of Scotland

RDB2RDF Conversion

- why would you want to?
- **how do you do it?**
- problems with the basic procedure
- a schema for cultural heritage

Including Thesauri

- using SKOS
- integrating with archive content

How to Convert DBs - W3C Guidance

- RDF has been around since 1999...
- W3C Sem Web FAQ: <http://www.w3.org/RDF/FAQ#reldb>:

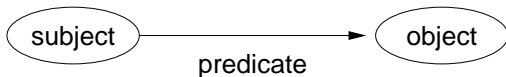
“How do I export my data from a Relational Database?”

This is one of the active areas of R&D, and no final answer is yet available.”

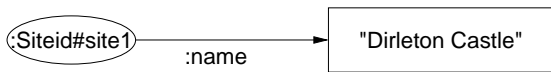
- W3C RDB2RDF Inclubator Group due to report in 2009
- <http://esw.w3.org/topic/Rdb2RdfXG/StateOfTheArt>
- Lots of tools emerging: D2R, Virtuoso, Squirrel etc.
- Can the process be automatic?
- *Should* it be automatic?

RDF Basics

- “Facts” expressed as subject–property–object triples:



- “Resources”: nodes or arcs with URIs
- Resource nodes can be subject – so can link triples together
- Literals can only be the object – no new links possible



@prefix : <<http://www.ltg.ed.ac.uk/tether/>> .

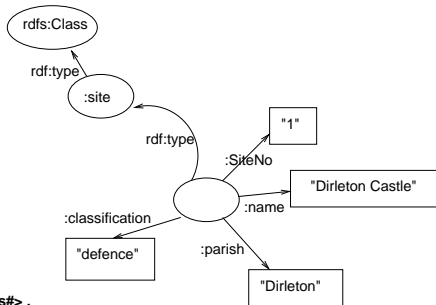
A Simple Example – One Database Record

“Table as Class; Column as Predicate”

SITE

siteNo	name	parish	classification
1	<i>Dirleton Castle</i>	<i>Dirleton</i>	<i>defence</i>
2	Dirleton Cottage	Dirleton	residential
3	Drem Airfield	Dirleton	military
4	Jamie's Neuk	Dirleton	military

@prefix : <http://www.ltg.ed.ac.uk/tether/> .
 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .



How Many Triples?

- One triple for every database cell
- $\Rightarrow total = rows \times columns$
- For subset of NMRS used: **235 million**
- – then add schema
- Can we do reasoning over this many? **No.**
- Can we even do SPARQL queries? **Barely.**
- All those URIs – data bloat is huge
- Do we just wait for more powerful machines?
- ...or do we try to make the conversion more efficient?
- There are a lot of inefficiencies in the basic procedure

Background

- a (very) brief history of Information Management
- RCAHMS and The National Monument Record of Scotland

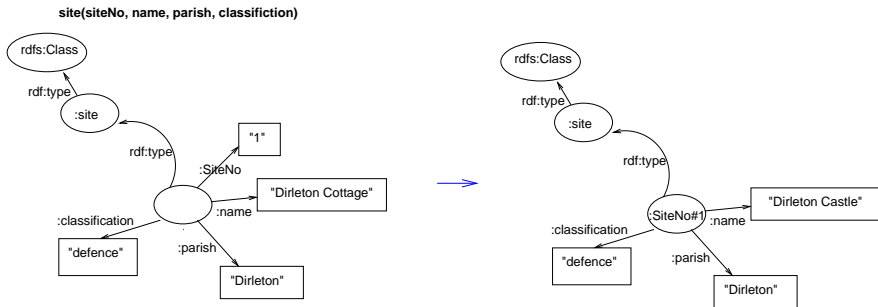
RDB2RDF Conversion

- why would you want to?
- how do you do it?
- **problems with the basic procedure**
- a schema for cultural heritage

Including Thesauri

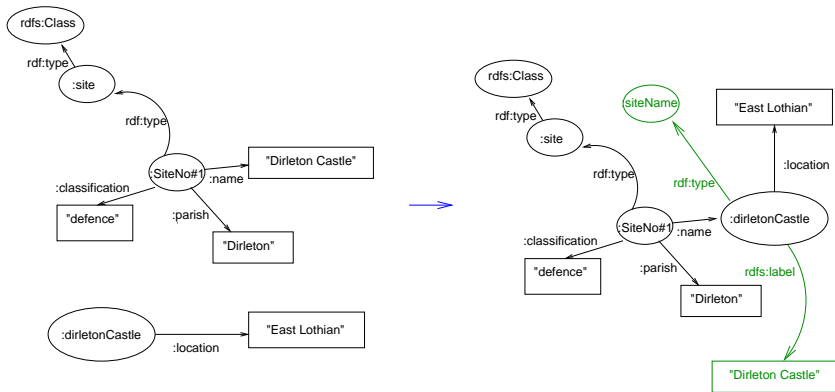
- using SKOS
- integrating with archive content

To Bnode or Not to Bnode?



- duck typing – is it good data management?
- primary keys: important data items need direct reference
- needs schema knowledge

Literals or Resources?



- each unique resource needs a type and a label

Literals or Resources?

- Avoid literals!
- Graph is sterilised at literals – no further links
- Encode database values as URIs
- Some unlikely URIs:
 - PhotoDesc – '#5: 6"x4" neg, B&W'
 - `http://www.ltg.ed.ac.uk/tether/Photodesc#%235:
%206%22x4%22%20neg%2C%20B%26W`
 - # and & need special care
- What are these URIs for?

Lots More Issues...

- URI generation:
 - is `http://www.example.com/place/edinburgh` the same resource as `http://www.example.com/city/edinburgh`?
- Remove nulls – but beware of OWA/CWA issues
- Decode all coded values
- Relational database joins produce redundant key-to-key links:
 - can be pruned if they have no local attributes
 - saves one triple per row of larger table in joins
- How much do we save (on projected 235 million triples)?
 - overall total needed in *Tether*: **22 million**

Background

- a (very) brief history of Information Management
- RCAHMS and The National Monument Record of Scotland

RDB2RDF Conversion

- why would you want to?
- how do you do it?
- problems with the basic procedure
- a schema for cultural heritage

Including Thesauri

- using SKOS
- integrating with archive content

Designing the Schema

- Standard approach duplicates data:
 - database source information encoded in RDF predicate
 - database source information appears in target node
 - database source information becomes RDF Class name
- Instead, design **attribute set** – keep it compact
- Design **class hierarchy** – put database metadata here *only*

Attribute Set and Class Hierarchy

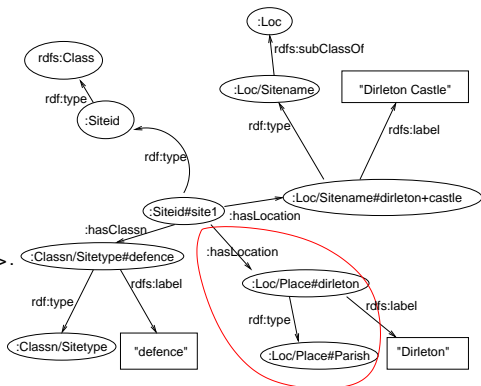
- *Tether* attribute schema is generic:
 - Who? What? Where? When? model
 - *:hasAgent*, *:hasAgentRole*, *:hasClassn*, *:hasDesc*, *:hasLocation*, *:hasPeriod*, *:hasId*, *:hasFlag*
 - Including event modeling: *:hasEvent*, *:hasPatient*, *:partOf*
 - Plus “local” predicates needed to relate classes (9)
 - Standard vocabularies used for framework (RDF, RDFS, OWL)
 - Tiny predicate set – but covers cultural heritage needs
- Class hierarchy encodes characteristics of particular dataset
 - 15 top level classes – generic for cultural heritage
 - tree structure of subclasses covers NMRS data
 - only 60 classes needed
 - automatic RDB2RDF translation would generate hundreds

Visualisation of *Tether* Design

SITE

siteNo	name	parish	classification
1	Dirleton Castle	Dirleton	defence
2	Dirleton Cottage	Dirleton	residential
3	Drem Airfield	Dirleton	military
4	Jamie's Neuk	Dirleton	military

@prefix : <http://www.ltg.ed.ac.uk/tether/> .
 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .



Background

- a (very) brief history of Information Management
- RCAHMS and The National Monument Record of Scotland

RDB2RDF Conversion

- why would you want to?
- how do you do it?
- problems with the basic procedure
- a schema for cultural heritage

Including Thesauri

- using SKOS
- integrating with archive content

Converting Domain Thesauri to RDF

- **Monument Type Thesaurus** and **Object Type Thesaurus**
- Both are Scottish variants, based on EH published thesauri
- EH model (CRM-EH) in turn linked to CIDOC-CRM
 - See STAR project [Binding et al., 2008]
- Well-structured data; easy to convert to RDF:
 - Monument Types: 17,000 triples
 - Object Types: 4,000 triples

Using SKOS (Simple Knowledge Organisation System)

- Using a subset of SKOS classes and predicates:
 - Classes: `ConceptScheme`, `ConceptScheme`
 - Predicates: *broader*, *related*, *prefLabel*, *altLabel*, *scopeNote*, *inScheme*, *hasTopConcept*
- SKOS distinguishes **Concepts** from **labels** –
 - non-preferred thesaurus terms are modelled with `altLabel`...
 - ...so they are RDF literals, not resources
 - cannot be linked forwards to their preferred terms
 - local *prefTerm* predicate used

Integrating Thesaurus with Archive Content

- Options considered:
 1. keep content and thesaurus separate; link on *skos:exactMatch*
 2. as above but link with *owl:sameAs*
 3. generate thesaurus node within the content graph
 4. place existing content classification nodes within thesaurus
- Option 4 chosen
 - same concept, so same URI
 - the resource is a *:Classn/Sitetype* instance
 - to convert to 1 or 2, just add back triples that were omitted

Summary

- Need to expose RDB data as RDF to join Semantic Web
- Conversion process has pitfalls
- RDF design needs as much care as RDB design
- Automatic processes generate millions of redundant triples
- Ten-fold reduction in graph size possible
- Integration of domain thesauri demonstrated





Binding, C., May, K., and Tudhope, D. (2008).

Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM.

In Proceedings of European Conference on Digital Libraries (ECDL08), volume LNCS, pages 280–290, Aarhus, Denmark. Springer.