

Tethering Cultural Data with RDF

Kate Byrne

School of Informatics, University of Edinburgh

Jena User Conference, 10th–11th May 2006



Outline

Project overview

- the nature of the data, and motivation for the project

Data translation

- transforming the source data into a graph structure

Data accessibility

- guiding user queries over graph data; providing context
- presentation of results, tailored to user's needs



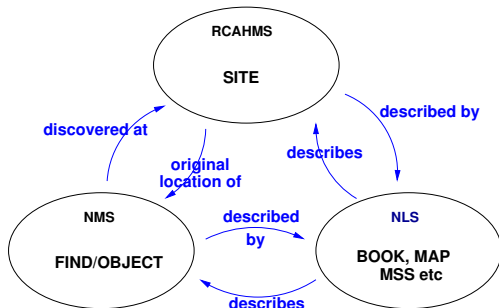
Overview: the project in a nutshell

- Cultural heritage data — fixed fields, free text, thesauri
- Query problems:
 - complex data structures, specific to each collection
 - specialist domain terminology
 - limited access to free text
- Goal: query application (*Tether*) that will guide non-experts
- Two key elements of proposal:
 - *relation extraction*: two-place predicates from free text
 - *graph database*: combine all data in one simple format



Source Data

- 3 datasets: RCAHMS, NLS, NMS
- Similar topics and vocabulary: archaeology, Scottish history
- Unexploited relationships



Project overview

- the nature of the data, and motivation for the project

Data translation

- transforming the source data into a graph structure

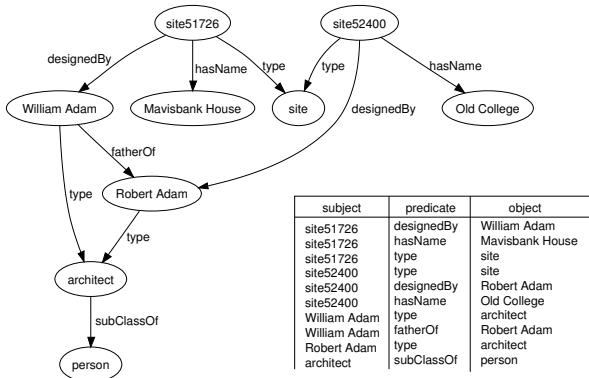
Data accessibility

- guiding user queries over graph data; providing context
- presentation of results, tailored to user's needs



Data transformation - overview

- Graph of *Subject, Predicate, Object* triples
- Represented in RDF, in Jena triple store with MySQL backing
- Using RDFS alongside RDF for *subClassOf*, *type*, *domain*, *range* etc
- Probably will use some OWL, SKOS, DC etc



Construction of graph database

- Bottom up approach:
 - instance population first
 - then infer schema relations
- From database fields
- From thesauri — TMT, FISH, SPECTRUM, AAT, LCSH...
- From free text — relation extraction using NLP

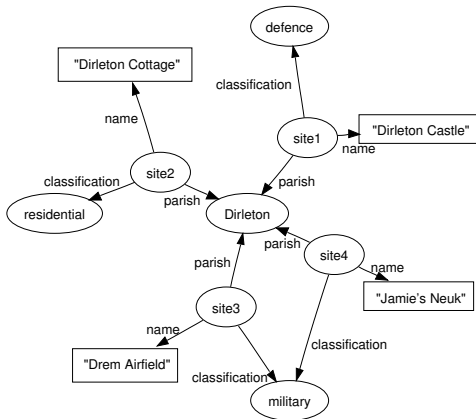


From Database Fields

- Automatic RDBMS to RDF conversion

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



Varying the Automatic Procedure

- Use “entity” nodes including primary key instead of b-nodes
- Ignore empty fields
- Replace concatenated keys with single-column surrogates
- Manual schema design - group similar attributes together
- “Pre-join” tables to eliminate redundant nodes
- Theoretical maximum: 235 million triples (rows x cols)
- Actual: 15.4 million (without schema)



Varying the Automatic Procedure

- Use “entity” nodes including primary key instead of b-nodes
- Ignore empty fields
- Replace concatenated keys with single-column surrogates
- Manual schema design - group similar attributes together
- “Pre-join” tables to eliminate redundant nodes
- Theoretical maximum: **235 million** triples (rows x cols)
- Actual: **15.4 million** (without schema)

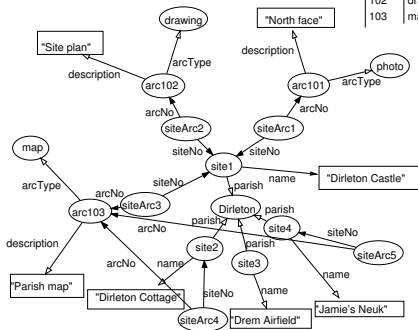


Eliminating redundant nodes

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

siteArcNo	siteNo	arcNo
1	1	101
2	1	102
3	1	103
4	2	103
5	4	103

arcNo	arcType	description
101	photo	North face
102	drawing	Site plan
103	map	Parish map

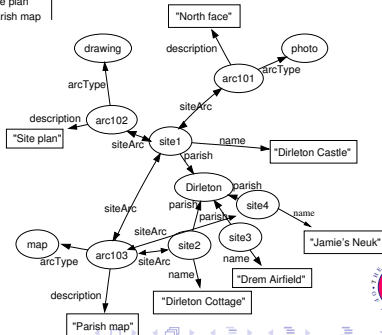
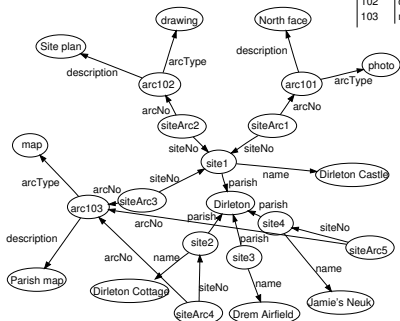


Eliminating redundant nodes

SITE		
siteNo	name	parish
1	Dirleton Castle	Dirleton
2	Dirleton Cottage	Dirleton
3	Drem Airfield	Dirleton
4	Jamie's Neuk	Dirleton

SITE-ARC		
siteArcNo	siteNo	arcNo
1	1	101
2	1	102
3	1	103
4	2	103
5	4	103

ARCHIVE		
arcNo	arcType	description
101	photo	North face
102	drawing	Site plan
103	map	Parish map

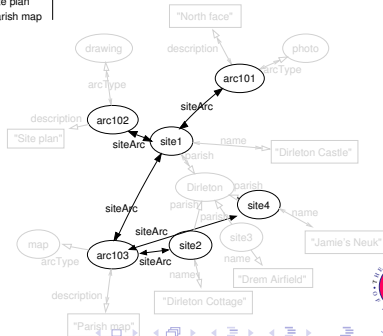
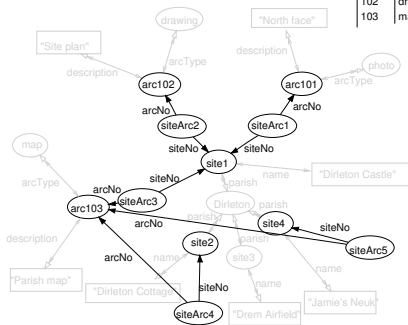


Eliminating redundant nodes

SITE		
siteNo	name	parish
1	Dirleton Castle	Dirleton
2	Dirleton Cottage	Dirleton
3	Drem Airfield	Dirleton
4	Jamie's Neuk	Dirleton

SITE-ARC		
siteArcNo	siteNo	arcNo
1	1	101
2	1	102
3	1	103
4	2	103
5	4	103

ARCHIVE		
arcNo	arcType	description
101	photo	North face
102	drawing	Site plan
103	map	Parish map



Eliminating redundant nodes

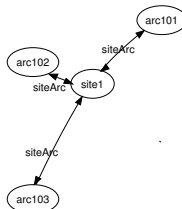
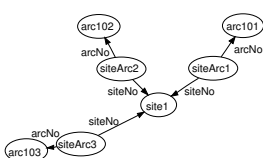
SITE ↓

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

SITE-ARC		
siteArcNo	siteNo	arcNo
1	1	101
2	1	102
3	1	103
4	2	103
5	4	103

ARCHIVE ↓

arcNo	arcType	description
101	photo	North face
102	drawing	Site plan
103	map	Parish map



Some Practical Results

- Using test subset of RCAHMS data, and non-RDF format:
 - 40,000 sites + 100,000 child records — **50MB**
 - translates to 1.3 million triples
 - held as simple triples (not RDF) — **100MB** (incl indexing)
 - *translation doubles physical size*
- Using most of RCAHMS dataset, and RDF:
 - 250,000 sites with 750,000 archive items — **380MB**
 - translates to 15.4 million triples, without schema relations
 - held as RDF in Jena — **2.7GB**
 - *RDF translation multiplies physical size by 7*
 - URIs are bulky



Some Practical Results

- Using test subset of RCAHMS data, and non-RDF format:
 - 40,000 sites + 100,000 child records — **50MB**
 - translates to 1.3 million triples
 - held as simple triples (not RDF) — **100MB** (incl indexing)
 - *translation doubles physical size*
- Using most of RCAHMS dataset, and RDF:
 - 250,000 sites with 750,000 archive items — **380MB**
 - translates to 15.4 million triples, without schema relations
 - held as RDF in Jena — **2.7GB**
 - *RDF translation multiplies physical size by 7*
 - URIs are bulky



Relation Extraction from Text — NLP

- Grammar-based approaches:
 - hand-written rules, machine learning, or combination
 - problems: volume of data, lack of annotated data
- “Brute force” statistical associations:
 - frequently co-occurring terms
 - poor results in my experiments
 - (*information with RCAHMS; 6 inch with building*)
- Planned method:
 - pre-processing to find key **Named Entities** (NEs)
 - simple grammar-based rules to combine them into relations
 - evaluate against machine learning with annotated data



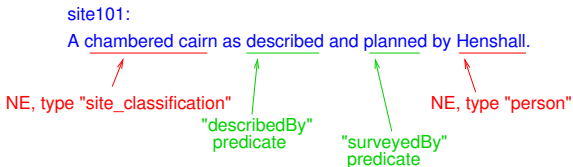
Relation Extraction Step 1 — Entity References

1. Package free text fields as individual documents
2. Tokenise, POS tag, chunk
3. NE recognition and classification
 - (F-scores average 85%)
4. Co-reference resolution
 - (“C R Mackintosh”, “Rennie Mackintosh” etc)
5. Translate each entity to canonical form



Relation Extraction Step 2 — Construct Relations

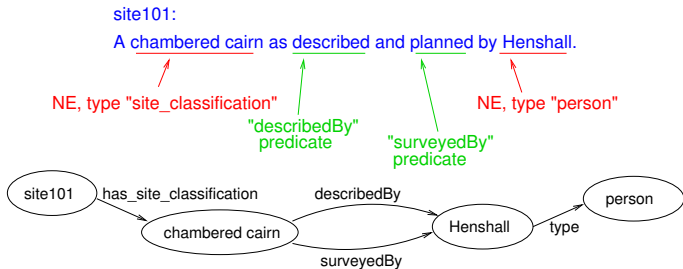
6. Identify candidate predicates (frequency analysis on POS tags)
7. Clustering step (using word distance measure)
8. Build triples:



9. Merge with rest of graph database

Relation Extraction Step 2 — Construct Relations

6. Identify candidate predicates (frequency analysis on POS tags)
7. Clustering step (using word distance measure)
8. Build triples:



9. Merge with rest of graph database

Project overview

- the nature of the data, and motivation for the project

Data translation

- transforming the source data into a graph structure

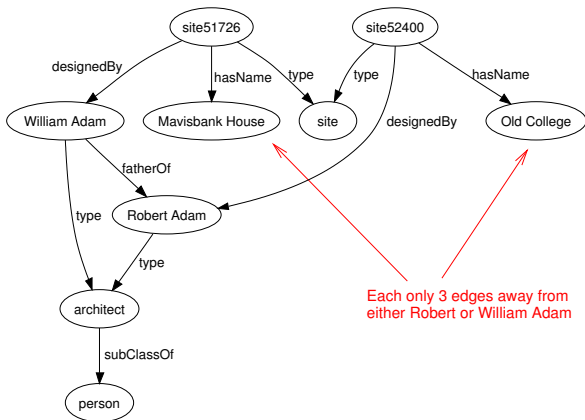
Data accessibility

- guiding user queries over graph data; providing context
- presentation of results, tailored to user's needs



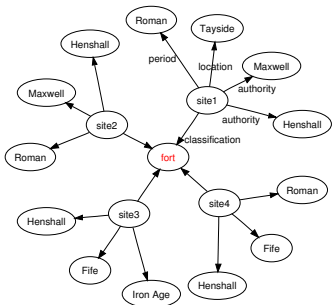
Graph vs RDBMS Queries (1)

- Detecting implicit relationships



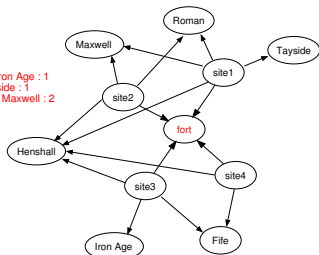
Graph vs RDBMS Queries (2)

- Summarising under multiple headings



Summary for "fort":

Period – Roman : 3, Iron Age : 1
Location – Fife : 2, Tayside : 1
Authority – Henshall : 4, Maxwell : 2

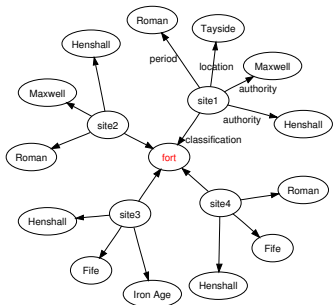


- find collections of nodes by predicate label (eg all *period* ones)
- cluster values within those collections (*n Roman, m Iron Age*)

- Therefore need small set of predicates

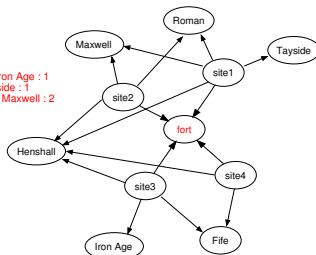
Graph vs RDBMS Queries (2)

- Summarising under multiple headings



Summary for "fort":

Period – Roman : 3, Iron Age : 1
Location – Fife : 2, Tayside : 1
Authority – Henshall : 4, Maxwell : 2



- find collections of nodes by predicate label (eg all *period* ones)
- cluster values within those collections (*n Roman, m Iron Age*)

- Therefore need small set of predicates

Limiting the Size of the Predicate Set

- Can do automatically for text relations
 - predicates chosen by POS analysis, followed by clustering
- Manual task for database attributes
 - database column names are arbitrary
- Group similar attributes/predicates together
 - eg *location*: {parish, grid reference, street name,...}
 - top level: *loc, ind, id, sitename, date, classn, agent, desc*
 - each having up to four sub-categories

Query Experiments

- Compare performance, RDBMS vs RDF, for standard queries
- Compare scope:
 - RDF queries that are difficult in RDBMS (implicit relations)
 - queries that are difficult in SPARQL (node degree)



Tailored Presentation of Results

- Natural Language Generation to produce descriptive text
- Pilot project using *M-PIRO* system¹ on RCAHMS data
- Generates sentences from data fields + language model
- Also features:
 - multi-lingual output
 - selection of user profiles
 - personalised text: compares current item to ones seen earlier

¹*Isard et al., 2003*

M-PIRO authoring tool --- /home/kate/mpiro/cd/MPIRO-authoring-v4.4/raahms.mpiro

File Options Help

Search

USER TYPES DATABASE LEXICON


Data Base

- Basic-entity-types
 - agent
 - classification
 - event
 - find
 - location
 - period
 - site
 - agriculture
 - defence
 - defined-by-form
 - domestic
 - industrial
 - recreation
 - religion-and-ritual
 - burials
 - chambered-cairn
 - blackhammer
 - cuween-hill
 - dwarrie-stane
 - holm-papa-westray
 - knowe-of-yarso
 - maes-howe
 - midhowe
 - quoyness
 - taversoe-tuick
 - unstan
 - wideford-hill
 - cist
 - henge-stone-circle
 - modern-era-religion
 - standing-stone
 - transport-related

Data-types

Language-independent fields of "taversoe-tuick"

Fields	Fillers
images	<rc-taversoe-tuick.jpg>
mapno	hy42nw
parish	rousay-and-egilsay
classgrp	
classsub	chamb-cairn-clsub
rcmsgrade	I
othstat	guardian
period	neolithic
associated-find	
council	orkney
survey	taversoe46



This site is another place used for burials, situated in Rousay and Egilsay. "Taversoe Tuick", an Orkney-Cromarty Bookan-type cairn, situated on the slope of the hill. Up to 1898 it appeared to be a small heather-covered knoll about 4' high, but in that year part of the upper chamber was exposed and access was gained to the intact chamber and passage below. Like the previous site, Taversoe Tuick dates from the Neolithic. It is a chambered cairn, which means a Neolithic burial monument comprising a stone-built chamber within a mound of stones. The council area is the Orkney Islands. In status Taversoe Tuick is a Guardianship site and it is located in HY42NW. It was surveyed in 1946.

No notes

adult Preview 21

Summary

- Framework:
 - Relation extraction from text
 - Translation of mixed content to triples graph
- Practical query application:
 - search criteria tailored for each query
 - user not expected to specify criteria — picks from those offered
 - context of results provided

