

# Relation extraction and graph databases — work in progress

27 July 2006

# Outline

Revisions to SEER NE classes

Revisions to database schema

Candidate predicates

Integrating text relations with rest of graph

## NER/NEC is preliminary step

- SEER-RCAHMS corpus — 1546 annotated files
- 9 NE classes, sub-divided into *Types* and *Subtypes*
- entity nesting: [[[Edinburgh] University] Library]
- summary of markup:

▶ <http://homepages.inf.ed.ac.uk/s0233752/blog/refs/standoff.html>

## NER/C results with SEER-RCAHMS corpus

Entity Category	Precision (%)	Recall (%)	F1
LINK	0.00	0.00	0.00
ORG	97.12	95.30	96.20
PERIOD	78.60	67.14	72.42
PERSNAME	91.93	88.76	90.32
PLACE	79.13	87.88	83.27
REFERENCE	0.00	0.00	0.00
SITE	73.42	55.68	63.33
SIZE	92.64	93.37	93.00
TIMEX	93.72	95.21	94.46
Overall	86.18	83.83	84.99

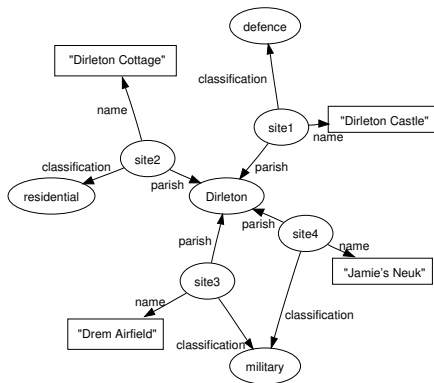
# Revisions

- Drop *Types* and *Subtypes*
- Drop SIZE class
- Split SITE - SITENAME, SITETYPE
- Expand some categories (ORG, CO-REF/LINK)...
- ... restrict some (DATE, REFERENCE)
- ... add new ones (ARTEFACT, ADDRESS)
- New list:
  - ORG, PERSNAME, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, REFERENCE, CO-REF
- Revised annotation guidelines drafted

# Automatic translation from RDBMS to RDF

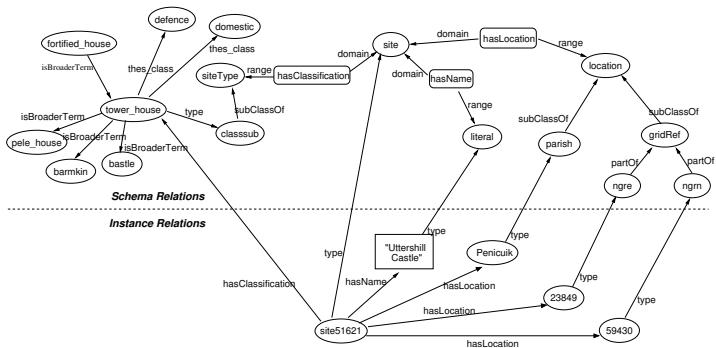
SITE

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



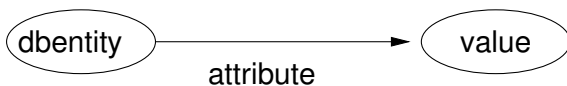
- The devil is in the detail...
- See, eg, [Berners-Lee(2006)]
- Interest growing: <http://esw.w3.org/topic/RdfAndSql>

# Moving complexity into the schema



- Group similar attributes/predicates together
  - eg *location*: {parish, grid reference, street name,...}
  - top level: *loc, ind, id, sitename, date, classn, agent, desc*
  - each with up to 4 sub-categories
- Now simplifying even further

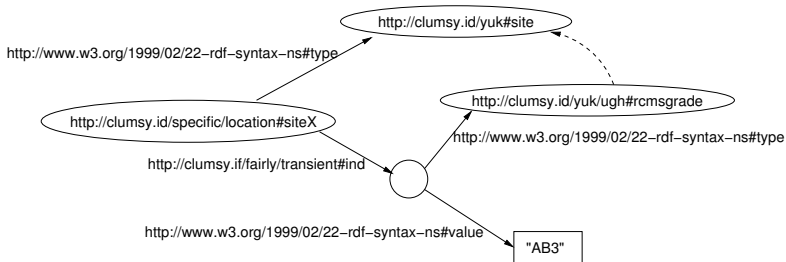
## Really Dreadful Framework



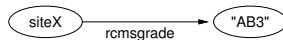
- What do we have on the right hand side?
- (site, location, Edinburgh) - *Edinburgh* has built-in semantics
- (site, rcmsgrade, AB3) - *AB3* is only meaningful in context
- Mixture of URIs and literals?
- What if a literal acquires “meaning” and needs its own “identity”?
- RDF tries to put semantics into data

## Zooming in on the graph

- Everything's either a literal or a URI
- Use bnodes to get round (some of) the URI ghastliness

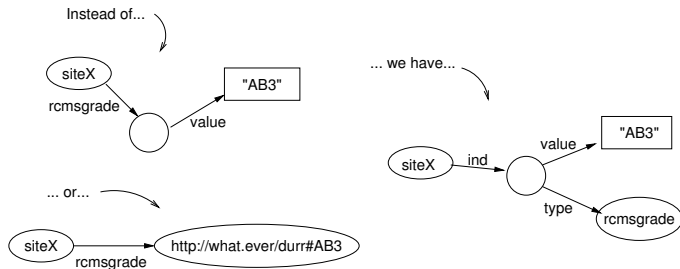


We need this for every triple from the database, of the form:



## Revised design

- Note *design* — no longer automatic
- The database fields become **classes** instead of **properties**
- Even *Edinburgh* becomes a literal hanging off a bnode
- Note also: we're a long way from [Berners-Lee(2006)]



## Basic plan for determining predicate set

- Analyse SEER-RCAHMS corpus by POS tag — eg verb frequencies
- Cluster the candidates into groups — predicate classes
- Annotate for relations:
  - classification task
  - predicate recognition
  - predicate classification
- Attempt with heuristics; compare with machine learning

# Analysing for candidate predicates

- Used chunker to find Verb Groups in SEER-RCAHMS corpus
- Listed by frequency:  
[▶ http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck3.freq](http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck3.freq)
- Converted to lower case and lemmatised:  
[▶ http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck2.freq](http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck2.freq)
- 861 candidates

## Clustering verbs

- Using NGD — Normalised Google Distance, [Cilibrasi and Vitanyi(2004)]

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y) - \log f(x, y)\}}{\log M - \min\{\log f(x), \log f(y)\}}$$

- Experiment with top 21 terms:

▶ <file:///home/kate/phd/relPreds/ngdExp.html>

- Using WordNet

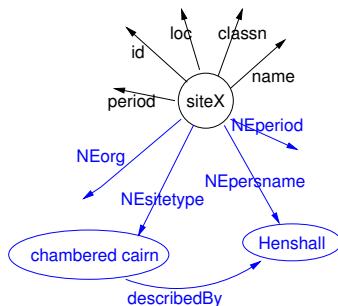
- find synset of candidate verb, and second hypernym
- build inverted index: all candidates falling in second hypernym synset
- perhaps should use antonyms?
- comparison with NGD experiment:

▶ <file:///home/kate/phd/relPreds/preds8.out>

- by hand: ▶ <file:///home/kate/phd/relPreds/vg21.byhand>
- Possible alternative: pairwise comparison using `wordnet::similarity`

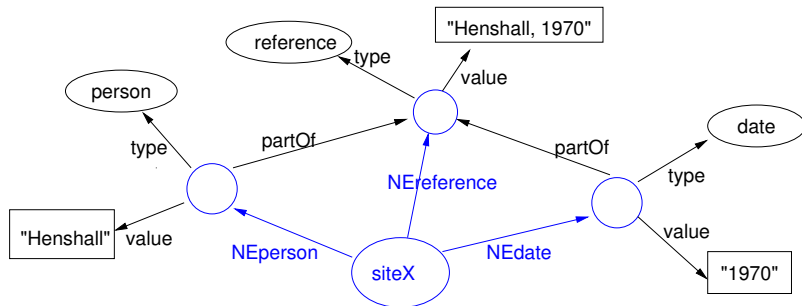
## Tie NEs to database parent

- For RCAHMS, parent is *site* node
- Separate set of NE predicates
- Text relation  $\rightarrow$  (NE, predicate, NE) triple



## Nested entities

- Connect all NEs to parent site — flatten the nesting
- Use “partOf” to relate inner and outer entities



# References



Tim Berners-Lee.

Relational Databases on the Semantic Web.

Internet note, 2006.

URL <http://www.w3.org/DesignIssues/RDB-RDF.html>.  
v 1.22 2006/02/01 (originally published September 1998).



Rudi Cilibrasi and Paul M. B. Vitanyi.

Automatic meaning discovery using Google.

Internet, 2004.

URL <http://arxiv.org/abs/cs.CL/0412098>.