

Outline

The Data

- 3 cultural archives
- data structures

The Problem

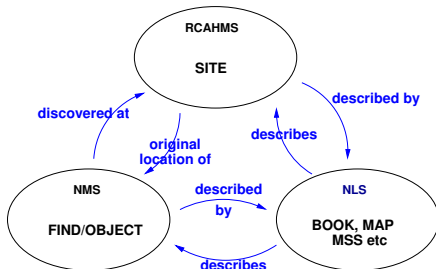
- formulating a “good” query
- understanding the results

Approach and Methods

- finding a new representation of the data
- interaction with user and presentation of results
- issues outwith scope

Overview of data

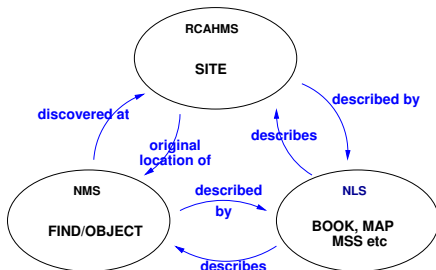
- 3 datasets: RCAHMS, NLS, NMS
- Similar topics and vocabulary: archaeology, Scottish history
- Unexploited relationships



- *NB This is my agenda, not necessarily theirs*

Overview of data

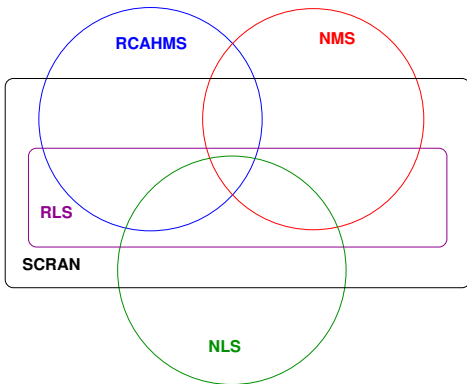
- 3 datasets: RCAHMS, NLS, NMS
- Similar topics and vocabulary: archaeology, Scottish history
- Unexploited relationships



- *NB This is my agenda, not necessarily theirs*

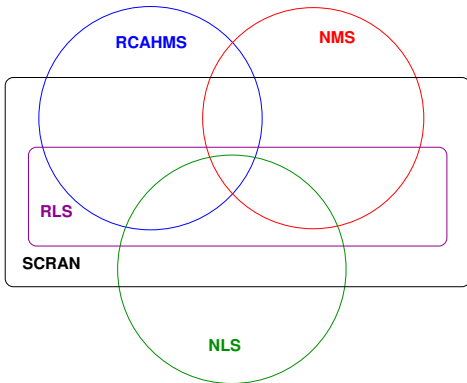
SCRAN and RLS

- SCRAN - founded by NMS, RCAHMS, SMC
- RLS - lead bodies NLS, NAS (and SCRAN)
- *Hundreds* of other contributors: museums, galleries, local history societies, individuals, ...



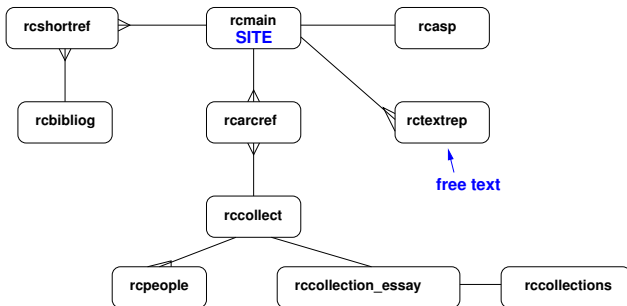
SCRAN and RLS

- SCRAN - founded by NMS, RCAHMS, SMC
- RLS - lead bodies NLS, NAS (and SCRAN)
- *Hundreds* of other contributors: museums, galleries, local history societies, individuals, ...



RCAHMS data

- 250,000 site records + 750,000 associated archive items
- Held in Oracle RDBMS
- Originated as text notes and index cards, from 1908 onwards

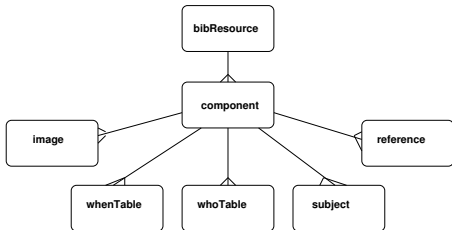


NLS data - SCRAN records

- Core NLS data: bibliographic records in MARC format
- What I have:
 - data prepared for RLS - SCRAN format
 - 10,000 “bib resource” records

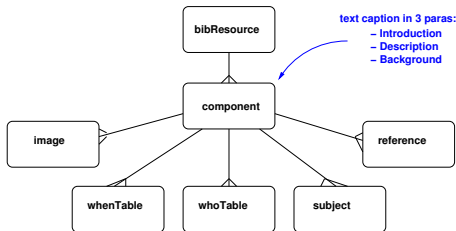
NLS data - SCRAN records

- Core NLS data: bibliographic records in MARC format
- What I have:
 - data prepared for RLS - SCRAN format
 - 10,000 “bib resource” records



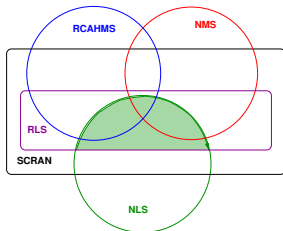
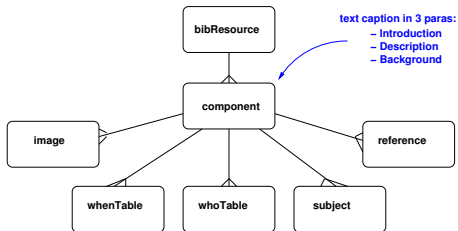
NLS data - SCRAN records

- Core NLS data: bibliographic records in MARC format
- What I have:
 - data prepared for RLS - SCRAN format
 - 10,000 “bib resource” records



NLS data - SCRAN records

- Core NLS data: bibliographic records in MARC format
- What I have:
 - data prepared for RLS - SCRAN format
 - 10,000 “bib resource” records



Demo

SCRAN interface - Bonnie Prince Charlie

NMS data

- 100,000 records
- Archaeological objects and excavation finds
- Text caption with some fixed fields
- Like other datasets - very variable records
- Held in proprietary database package (AdLib)

Background to problem

- In the past:
 - personal visitors
 - expert advisors
 - small target market
- From mid-1990s:
 - big investment in digitisation
 - move to Web availability
 - change of emphasis: general public, not specialists
- Partial solutions: CANMORE (1997), SCRAN (2000)

Background to problem

- In the past:
 - personal visitors
 - expert advisors
 - small target market
- From mid-1990s:
 - big investment in digitisation
 - move to Web availability
 - change of emphasis: general public, not specialists
- Partial solutions: CANMORE (1997), SCRAN (2000)

Background to problem

- In the past:
 - personal visitors
 - expert advisors
 - small target market
- From mid-1990s:
 - big investment in digitisation
 - move to Web availability
 - change of emphasis: general public, not specialists
- Partial solutions: CANMORE (1997), SCRAN (2000)

Good query terms

- Lots of specialist terms: *chambered cairn, long cist, finials, ...*
- Personal names with multiple representations: *Charles Edward Stewart, Alexander 'Greek' Thompson, William Henry Playfair*
- SCRAN solution: lots of hand indexing
- CANMORE: pick lists on some fields
- Ease of use v flexibility
- Experts don't always agree on terminology
- Many thesauri: TMT, AAT, ULAN, TGN, MDS, TGM, LCSH

Good query terms

- Lots of specialist terms: *chambered cairn, long cist, finials, ...*
- Personal names with multiple representations: *Charles Edward Stewart, Alexander 'Greek' Thompson, William Henry Playfair*
- SCRAN solution: lots of hand indexing
- CANMORE: pick lists on some fields
- Ease of use v flexibility
- Experts don't always agree on terminology
- Many thesauri: TMT, AAT, ULAN, TGN, MDS, TGM, LCSH

Good query terms

- Lots of specialist terms: *chambered cairn, long cist, finials, ...*
- Personal names with multiple representations: *Charles Edward Stewart, Alexander 'Greek' Thompson, William Henry Playfair*
- SCRAN solution: lots of hand indexing
- CANMORE: pick lists on some fields
- Ease of use v flexibility
- Experts don't always agree on terminology
- Many thesauri: TMT, AAT, ULAN, TGN, MDS, TGM, LCSH

Examples

- Search for:
 - *pit enclosure* - CANMORE (RCAHMS): 290 hits, SCRAN: unknown term (offers search on *pit* or on *enclosure*)
 - *hill fort* - SCRAN: 81 hits, CANMORE: 0 (but *fort* finds 1805)
 - *Ogams*:
- Different record sets on SCRAN for *Bonnie Prince Charlie* and *Young Pretender*

Examples

- Search for:
 - *pit enclosure* - CANMORE (RCAHMS): 290 hits, SCRAN: unknown term (offers search on *pit* or on *enclosure*)
 - *hill fort* - SCRAN: 81 hits, CANMORE: 0 (but *fort* finds 1805)
 - *Ogams*:
- Different record sets on SCRAN for *Bonnie Prince Charlie* and *Young Pretender*

Examples

- Search for:
 - *pit enclosure* - CANMORE (RCAHMS): 290 hits, SCRAN: unknown term (offers search on *pit* or on *enclosure*)
 - *hill fort* - SCRAN: 81 hits, CANMORE: 0 (but *fort* finds 1805)
 - *Ogams*:
 - *ogam inscribed* - SCRAN: 1, CANMORE: 26
 - *ogham inscribed* - SCRAN: 0, CANMORE: 4
 - thesaurus preferred spelling? *Ogham*
- Different record sets on SCRAN for *Bonnie Prince Charlie* and *Young Pretender*

Examples

- Search for:
 - *pit enclosure* - CANMORE (RCAHMS): 290 hits, SCRAN: unknown term (offers search on *pit* or on *enclosure*)
 - *hill fort* - SCRAN: 81 hits, CANMORE: 0 (but *fort* finds 1805)
 - *Ogams*:
 - *ogam inscribed* - SCRAN: 1, CANMORE: 26
 - *ogham inscribed* - SCRAN: 0, CANMORE: 4
 - thesaurus preferred spelling? *Ogham*
- Different record sets on SCRAN for *Bonnie Prince Charlie* and *Young Pretender*

Examples

- Search for:
 - *pit enclosure* - CANMORE (RCAHMS): 290 hits, SCRAN: unknown term (offers search on *pit* or on *enclosure*)
 - *hill fort* - SCRAN: 81 hits, CANMORE: 0 (but *fort* finds 1805)
 - *Ogams*:
 - *ogam inscribed* - SCRAN: 1, CANMORE: 26
 - *ogham inscribed* - SCRAN: 0, CANMORE: 4
 - thesaurus preferred spelling? *Ogham*
- Different record sets on SCRAN for *Bonnie Prince Charlie* and *Young Pretender*

Demo

Carved stones, personal names: SCRAN, CANMORE, CANTRIP

Interpretation of Results

- Even if the query is good:
 - too many hits
 - often too disparate
 - how do they relate to each other?
- Summary or some interpretation would help

Methods

- Techniques from overlapping disciplines:
 - Information Retrieval
 - Information Extraction
 - Question Answering
 - Knowledge Engineering
- Specific tools:
 - NER
 - Automatic ontology building
 - Supervised machine learning - where possible

Methods

- Techniques from overlapping disciplines:
 - Information Retrieval
 - Information Extraction
 - Question Answering
 - Knowledge Engineering
- Specific tools:
 - NER
 - Automatic ontology building
 - Supervised machine learning - where possible

Does NER help?

*Indexing over noun phrase compounds does not improve performance significantly... probably because the compound terms are generally not shared between the query and the source documents.
(Karen Sparck Jones, 1999)*

Does NER help?

Indexing over noun phrase compounds does not improve performance significantly... probably because the compound terms are generally not shared between the query and the source documents.
(Karen Sparck Jones, 1999)

Problem:

insufficient overlap between query terms and indexed documents

Does NER help?

Indexing over noun phrase compounds does not improve performance significantly... probably because the compound terms are generally not shared between the query and the source documents.
(Karen Sparck Jones, 1999)

Problem:

insufficient overlap between query terms and indexed documents

improve NER indexing
(SEER project)

Does NER help?

Indexing over noun phrase compounds does not improve performance significantly... probably because the compound terms are generally not shared between the query and the source documents.
(Karen Sparck Jones, 1999)

Problem:

insufficient overlap between **query terms** and **indexed documents**

construct a better
query

improve NER indexing
(SEER project)

NER issues

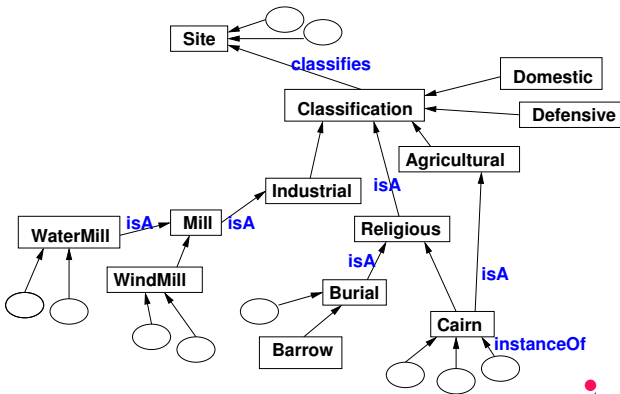
- Recognising key terms and classifying them
- Inferring relations
- Spotting co-reference
- Example: SEER markup of RCAHMS text (demo)

Ontology building 1 - exploit thesauri

- Get graph of classes from Thesaurus of Monument Types

Ontology building 1 - exploit thesauri

- Get graph of classes from Thesaurus of Monument Types

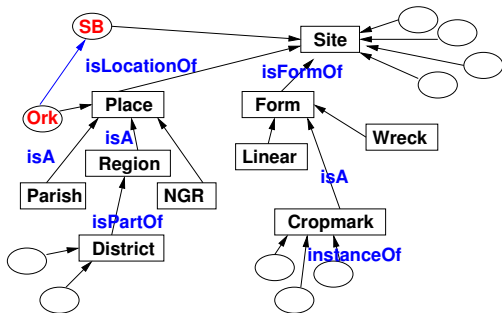


Ontology building 2 - exploit DB structure

- Convert attributes (fields of tables) to class relations
- Link to previous tree

Ontology building 2 - exploit DB structure

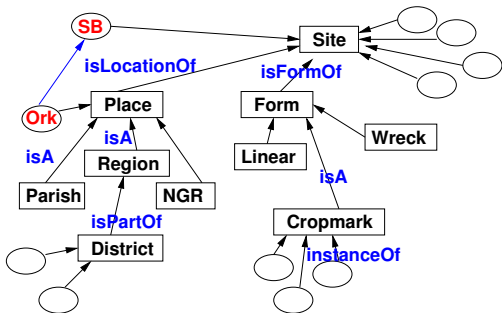
- Convert attributes (fields of tables) to class relations
- Link to previous tree



```
instanceOf("Skara Brae", Site)
instanceOf("Orkney", Place)
isLocationOf("Orkney", "Skara Brae")
```

Ontology building 2 - exploit DB structure

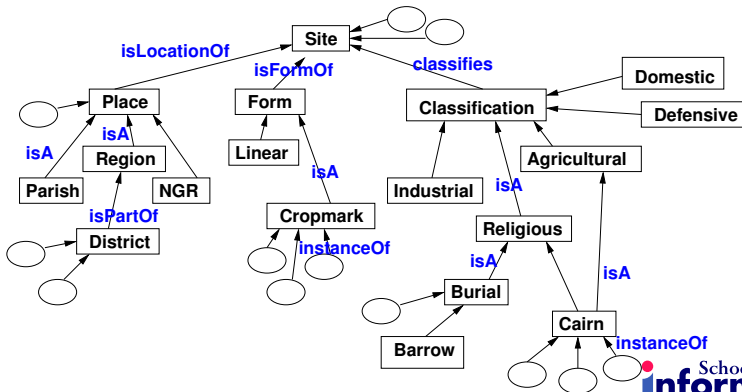
- Convert attributes (fields of tables) to class relations
- Link to previous tree



```
instanceOf("Skara Brae", Site)
instanceOf("Orkney", Place)
isLocationOf("Orkney", "Skara Brae")
```

Ontology building 2 - exploit DB structure

- Convert attributes (fields of tables) to class relations
- Link to previous tree



Ontology building 3 - infer relations from text

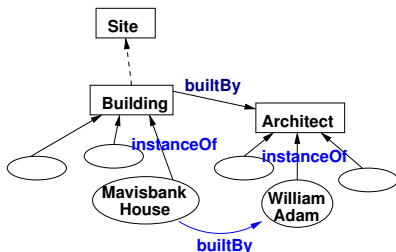
- Subject – Verb – Object construction → instance relation

- Storage: RDF, OWL? In database?
- Link between instance nodes and parent documents? Multiple occurrences of *instanceOf*(“William Adam”, Architect)? More than 3 in tuple? Reification?

Ontology building 3 - infer relations from text

- Subject – Verb – Object construction → instance relation

"William Adam was the architect of Mavisbank House."



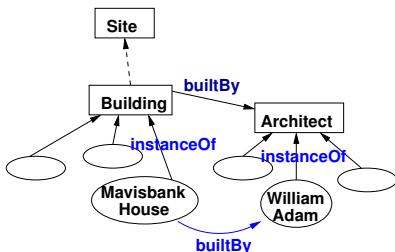
```
instanceOf("Mavisbank House", Building)
instanceOf("William Adam", Architect)
builtBy(Building, Architect)
builtBy("Mavisbank House", "William Adam")
```

- Storage: RDF, OWL? In database?
- Link between instance nodes and parent documents? Multiple occurrences of *instanceOf("William Adam", Architect)*? More than 3 in tuple? Reification?

Ontology building 3 - infer relations from text

- Subject – Verb – Object construction → instance relation

"William Adam was the architect of Mavisbank House."



```
instanceOf("Mavisbank House", Building)
```

```
instanceOf("William Adam", Architect)
```

```
builtBy(Building, Architect)
```

```
builtBy("Mavisbank House", "William Adam")
```

- Storage: RDF, OWL? In database?
- Link between instance nodes and parent documents? Multiple occurrences of *instanceOf*(*"William Adam", Architect*)? More than 3 in tuple? Reification?

At query time

Problem:

insufficient overlap between **query terms** and **indexed documents**

construct a better query

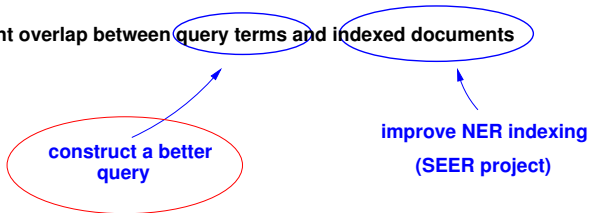
**improve NER indexing
(SEER project)**

- Assess user's query
- Link it to ontology node(s) (by force if necessary!)
- Dialogue with user to create "better" query

At query time

Problem:

insufficient overlap between **query terms** and **indexed documents**



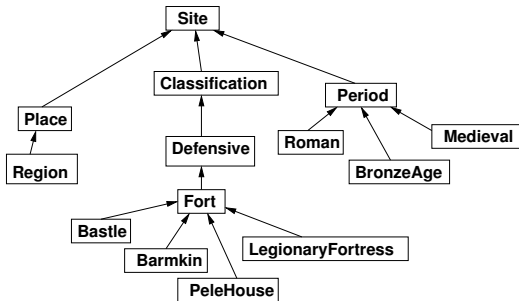
- Assess user's query
- Link it to ontology node(s) (by force if necessary!)
- Dialogue with user to create "better" query

Using the ontology

- Query: *fort*
- *Period = Bronze Age(352) or Roman(234) or medieval(60)*
Location = Dumfries(78) or Grampian(380) or Scotland(1805)

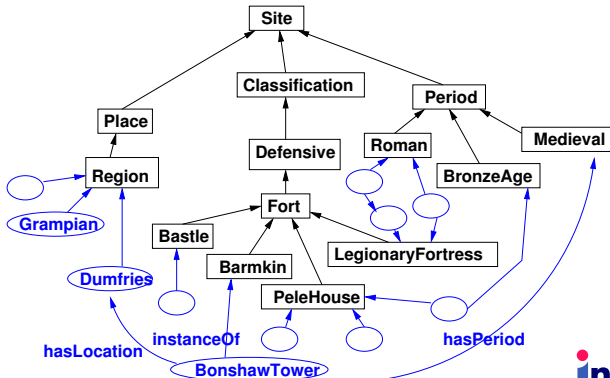
Using the ontology

- Query: *fort*
- *Period = Bronze Age(352) or Roman(234) or medieval(60)*
- *Location = Dumfries(78) or Grampian(380) or Scotland(1805)*



Using the ontology

- Query: *fort*
- *Period = Bronze Age(352) or Roman(234) or medieval(60)*
- *Location = Dumfries(78) or Grampian(380) or Scotland(1805)*



Some options for query processing

- Option A:
 - use subclass terms from ontology for query expansion
 - then TF-IDF over index that includes ontology terms
- Option B:
 - take query nodes and their cliques (or their parent docs?)
 - find candidate documents
 - do clustering over candidates into pre-defined sets
 - invite choice of cluster and return hits from it
- Option C:
 - build summary report based on count of instances attached to query nodes or their subclass nodes
 - offer selection of individual documents from relevant nodes
 - define ranking over documents to determine selection offered

Some options for query processing

- Option A:
 - use subclass terms from ontology for query expansion
 - then TF-IDF over index that includes ontology terms
- Option B:
 - take query nodes and their cliques (or their parent docs?)
 - find candidate documents
 - do clustering over candidates into pre-defined sets
 - invite choice of cluster and return hits from it
- Option C:
 - build summary report based on count of instances attached to query nodes or their subclass nodes
 - offer selection of individual documents from relevant nodes
 - define ranking over documents to determine selection offered

Some options for query processing

- Option A:
 - use subclass terms from ontology for query expansion
 - then TF-IDF over index that includes ontology terms
- Option B:
 - take query nodes and their cliques (or their parent docs?)
 - find candidate documents
 - do clustering over candidates into pre-defined sets
 - invite choice of cluster and return hits from it
- Option C:
 - build summary report based on count of instances attached to query nodes or their subclass nodes
 - offer selection of individual documents from relevant nodes
 - define ranking over documents to determine selection offered

Outwith scope (but interesting)

- Include Web search to inform query?
- Natural Language Generation from ontology relations - break the tie to original documents
- Use for translation – eg Scots Gaelic (*cf* M-PIRO project)
- How to update the data representation!

Outwith scope (but interesting)

- Include Web search to inform query?
- Natural Language Generation from ontology relations - break the tie to original documents
- Use for translation – eg Scots Gaelic (*cf* M-PIRO project)
- How to update the data representation!

Outwith scope (but interesting)

- Include Web search to inform query?
- Natural Language Generation from ontology relations - break the tie to original documents
- Use for translation – eg Scots Gaelic (*cf* M-PIRO project)
- How to update the data representation!

Summary

- 3 datasets: explore robustness of methods across them
- Build new data representation including semantics
- Dialogue with user to produce “good” query
- Present data summary, not just list of hits
- What I haven’t mentioned: evaluation