

# STAR

## Semantic Technologies for Archaeological Resources

<http://hypermedia.research.glam.ac.uk/kos/star/>



Arts & Humanities  
Research Council

DANMARKS   
BIBLIOTEKSSKOLE

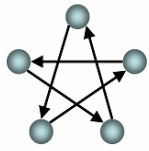


ENGLISH HERITAGE

University of Glamorgan

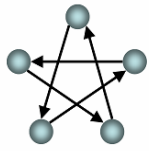
you live, you learn





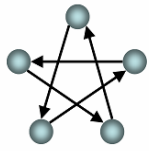
## Project Outline

- **3 year AHRC funded project**
  - Started January 2007, finish December 2009
- **Collaborators**
  - English Heritage
  - RSLIS, Denmark
- **Aims**
  - *“To investigate the potential of semantic terminology tools for widening access to digital archaeology resources, including disparate datasets and associated grey literature”*
  - To demonstrate cross search and browsing at detailed, meaningful level

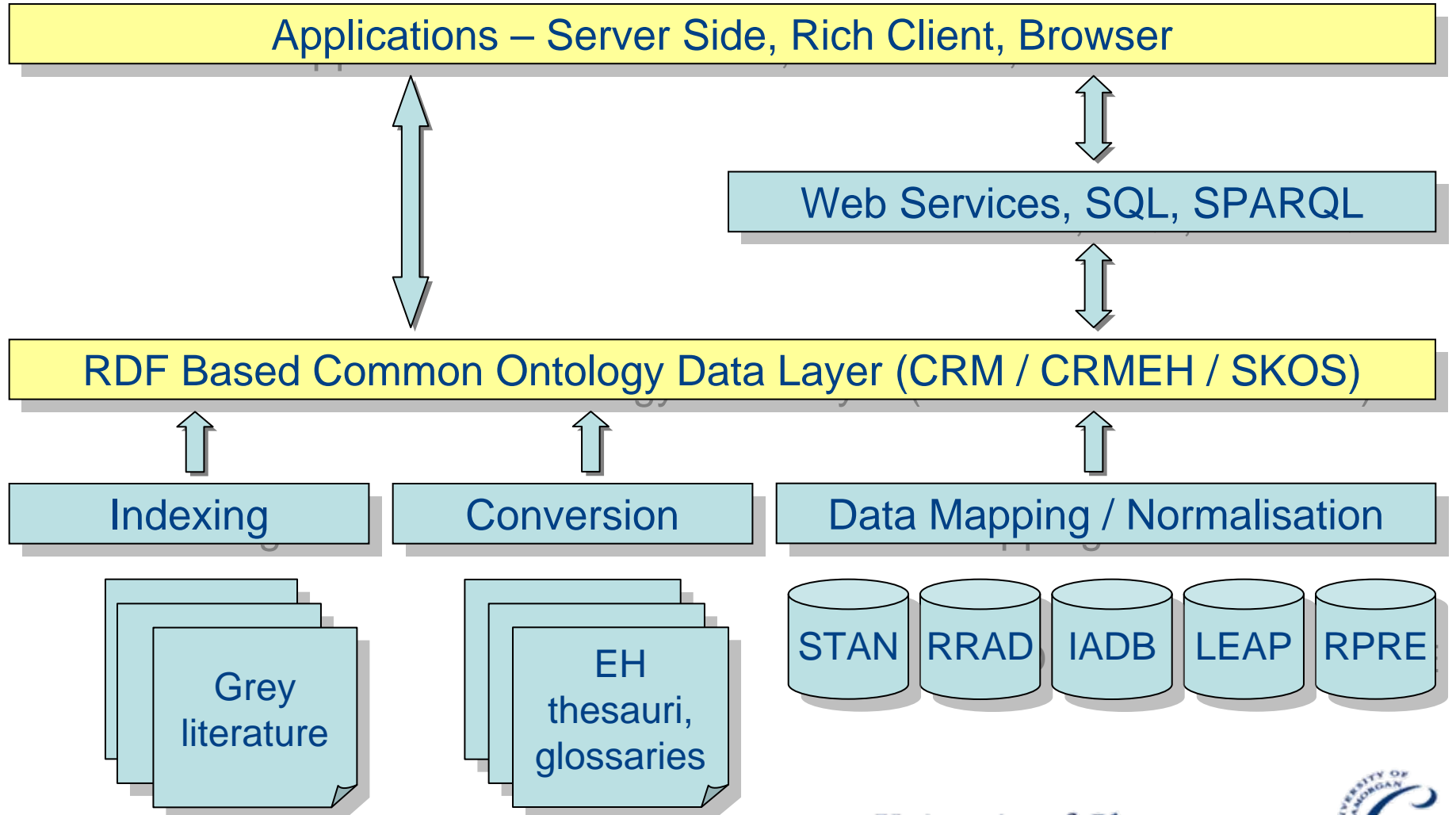


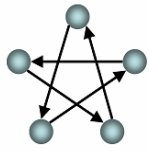
# Introduction

- **Data Modelling (RDF)**
  - CIDOC Conceptual Reference Model (CRM)
  - English Heritage Ontological Model (CRMEH)
  - English Heritage thesauri & glossaries (SKOS)
- **Data Mapping**
  - Mapping issues – domain specific vs. general model
  - Granularity, coverage
  - Data normalisation issues
- **Data Extraction**
  - Custom extraction utility
  - Consolidation to RDF 'triple store' database
- **Pilot Applications**
  - SKOS Web Service / Client applications
  - CRM Web Service / Client applications



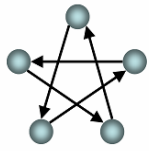
# General Architecture





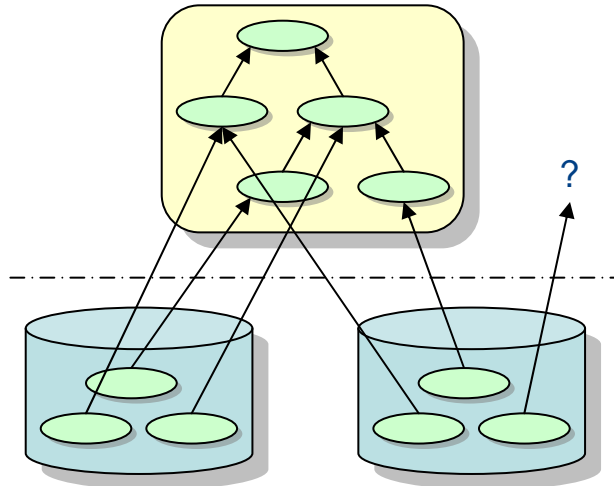
# Data Modelling - RDF

- **CRM** [ <http://cidoc.ics.forth.gr/> ]
  - CIDOC Conceptual Reference Model
  - International standard ISO 21127:2006
- **CRMEH** [ <http://hypermedia.research.glam.ac.uk/kos/CRM/> ]
  - English Heritage Ontological Model
  - Extends CIDOC CRM for archaeological domain
- **SKOS** [ <http://www.w3.org/2004/02/skos/> ]
  - Simple Knowledge Organisation System
  - RDF representation of thesauri, glossaries, taxonomies, classification schemes etc.



# Potential Data Mapping Problems

General schema (CRM)



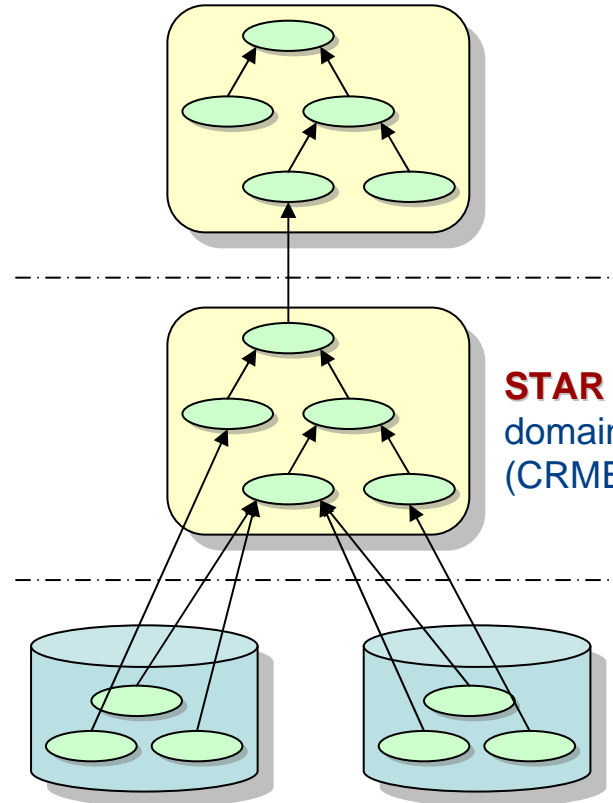
Domain specific datasets

*"...the abstractness of the [CRM] concepts...makes them ambiguous to any human user."*

*"...If several experts specify mappings independently...they will produce incompatible mappings and fail the goal of enabling interoperability."*

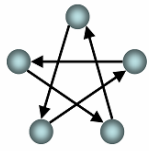
[from **BRICKS** FP6 IP Poster, ECDL2007]

General schema (CRM)



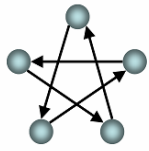
**STAR** approach -  
domain specific schema  
(CRMEH)

Domain specific datasets



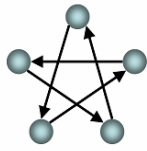
## Data Mapping Exercise

- Mapped DB Field → CRMEH Entity
  - manual process, requires expert domain knowledge
- Needed to Consider
  - Granularity & completeness of mapping
  - Model coverage vs. dataset coverage
  - Event based vs. relational model
    - Events only implicit in mappings and datasets



## Data Extraction - Scope

- Extraction of data to RDF triples
  - 5 archaeological datasets
  - Custom data extraction application
- Conversion of controlled terminology
  - 6 thesauri converted to SKOS
  - 27 glossaries created in SKOS
    - Created based on recording manuals
    - MultiTes XSL transformation to SKOS



# Custom Data Extraction Application

SQL Builder v1.0

Database: RRAD

	Subject	Predicate	Object
Type			
Pre			
Column			
FROM clause			
WHERE clause			

```

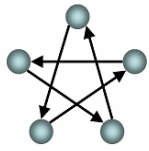
<?xml version="1.0"?>
<rdf:RDF xml:base="http://tempuri/star/base#"
xmlns:crm="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#"
xmlns:crmeh="http://tempuri/star/crmeh#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
<crmeh:EHE0007.Context rdf:about="http://tempuri/star/base#EHE0007.rrad.context.contextno.1">
<crm:P3F.has_note>
<crmeh:EHE0046.ContextNote rdf:about="http://tempuri/star/base#EHE0046.rrad.context.description.1">
<rdf:value>Upper ploughsoil over whole site no Sub-division for the convenience of finds processing '1' contains finds
contexts '3759', '3760' and '3763'.</rdf:value>
</crmeh:EHE0046.ContextNote>
</crm:P3F.has_note>
</crmeh:EHE0007.Context>
<crmeh:EHE0007.Context rdf:about="http://tempuri/star/base#EHE0007.rrad.context.contextno.10">
<crm:P3F.has_note>
<crmeh:EHE0046.ContextNote rdf:about="http://tempuri/star/base#EHE0046.rrad.context.description.10">
<rdf:value>Original recorded coordinates: 0980/0960</rdf:value>
</crmeh:EHE0046.ContextNote>
</crm:P3F.has_note>
</crmeh:EHE0007.Context>
Etc.
    
```

Generated RDF Data

Test SQL

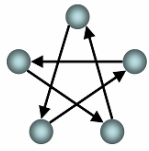
Write RDF...



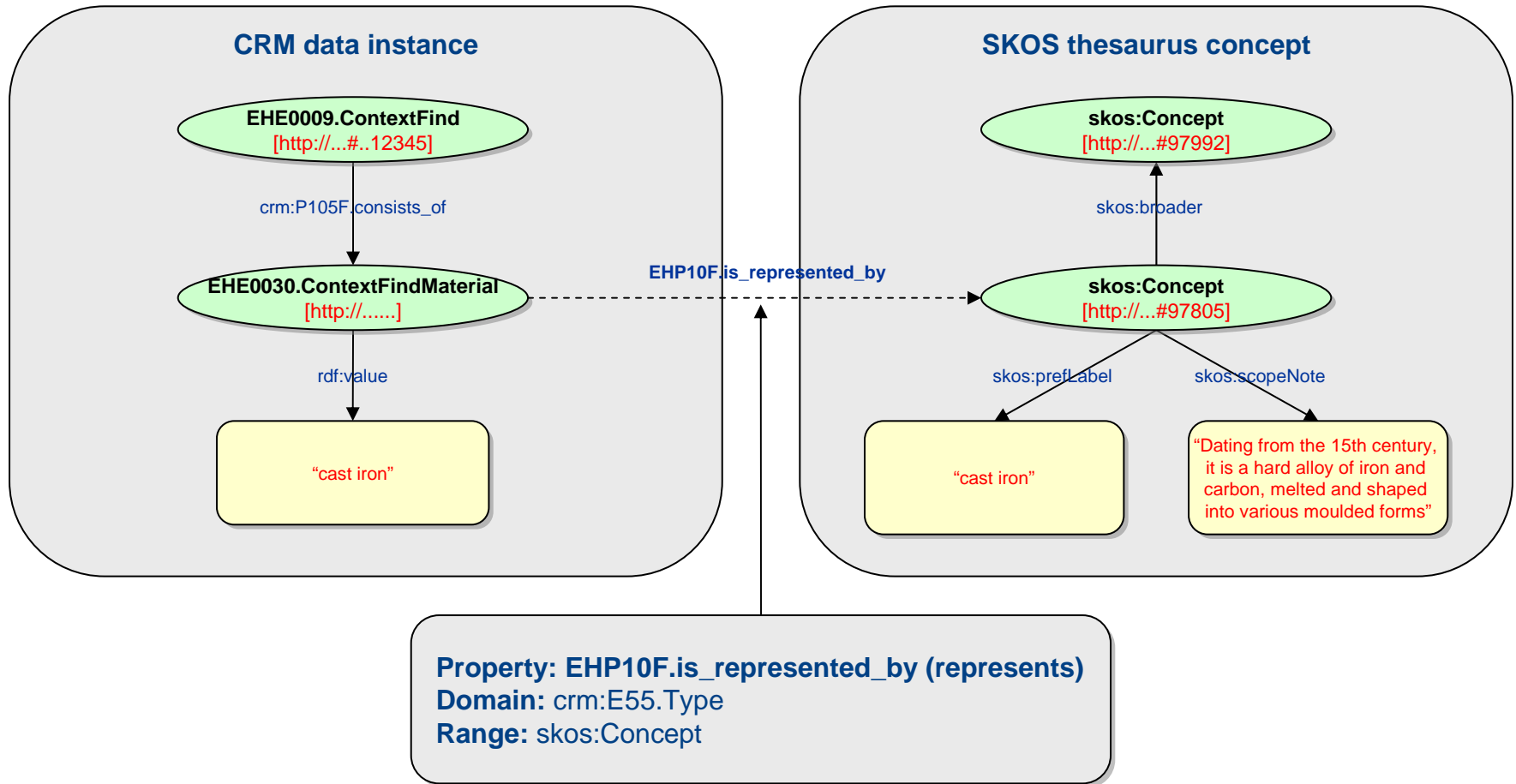


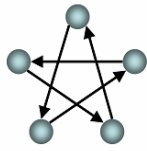
# Data Extraction and Consolidation

- **Data Extraction** – resultant files
  - 6 EH thesauri → 6 RDF files
  - 27 EH glossaries → 27 RDF files
  - 5 archaeological datasets → 305 RDF files
- **Data Consolidation** - RDF “triple store” database
  - 1,148,882 entities, 2,998,005 triples
  - Diverse data held within single DB schema
    - CRM, CRMEH, SKOS, OWL, DC etc.
  - Platform independent import/export
    - RDF-XML, NTriples, Turtle
  - Inbuilt support for SPARQL queries
  - Custom extension for full text search

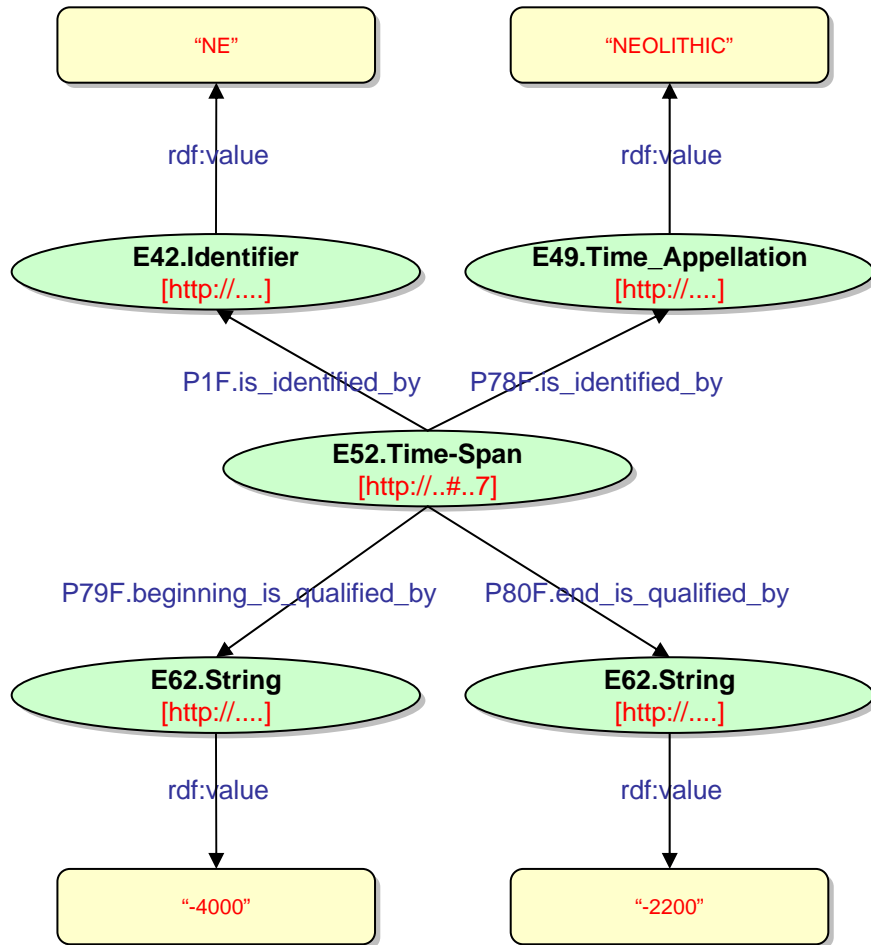


# Linking CRM and SKOS



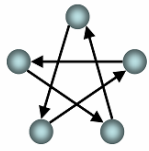


# RCHME Archaeological Periods in CRM

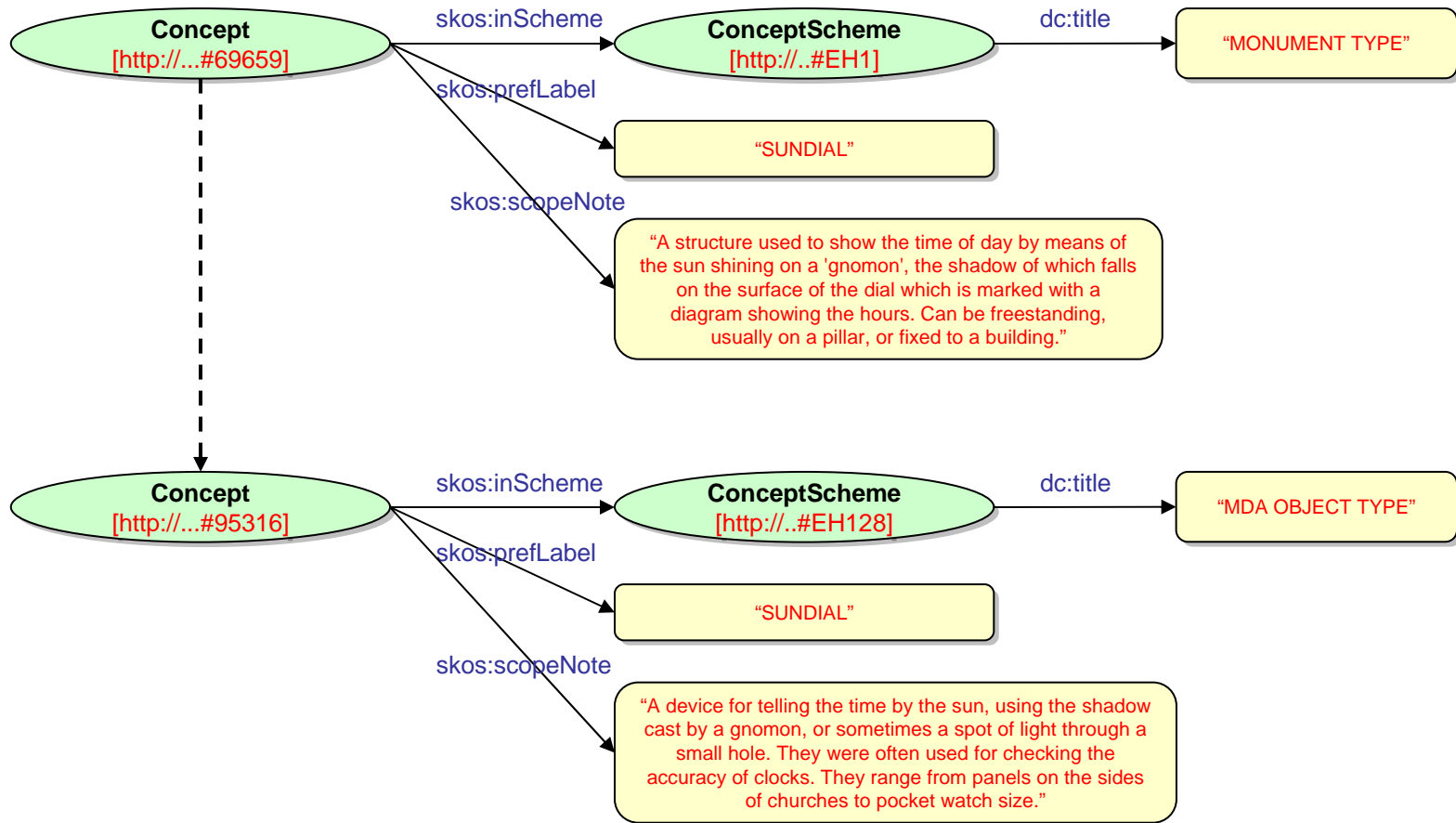


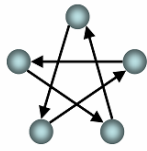
```
<crm:E52.Time-Span rdf:about="&base;E52.rchme.periods.termuid.7">
  <crm:P1F.is_identified_by>
    <crm:E42.Identifier rdf:about="&base;E49.rchme.periods.code.7">
      <rdf:value>NE</rdf:value>
    </crm:E42.Identifier>
  </crm:P1F.is_identified_by>
  <crm:P78F.is_identified_by>
    <crm:E49.Time Appellation rdf:about="&base;E49.rchme.periods.term.7">
      <rdf:value>NEOLITHIC</rdf:value>
    </crm:E49.Time Appellation>
  </crm:P78F.is_identified_by>
  <crm:P79F.beginning_is_qualified_by>
    <crm:E62.String rdf:about="&base;E62.rchme.periods.mindate.7">
      <rdf:value>-4000</rdf:value>
    </crm:E62.String>
  </crm:P79F.beginning_is_qualified_by>
  <crm:P80F.end_is_qualified_by>
    <crm:E62.String rdf:about="&base;E62.rchme.periods.maxdate.7">
      <rdf:value>-2200</rdf:value>
    </crm:E62.String>
  </crm:P80F.end_is_qualified_by>
</crm:E52.Time-Span>
```



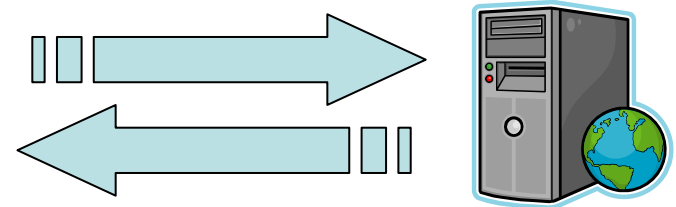
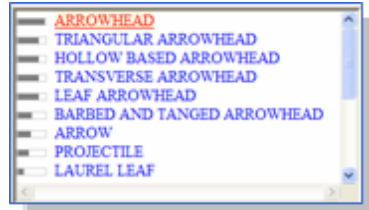
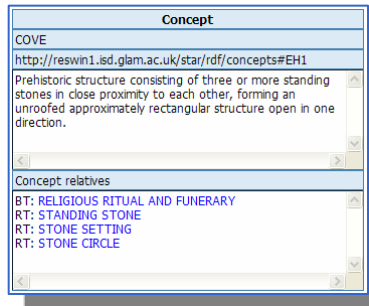
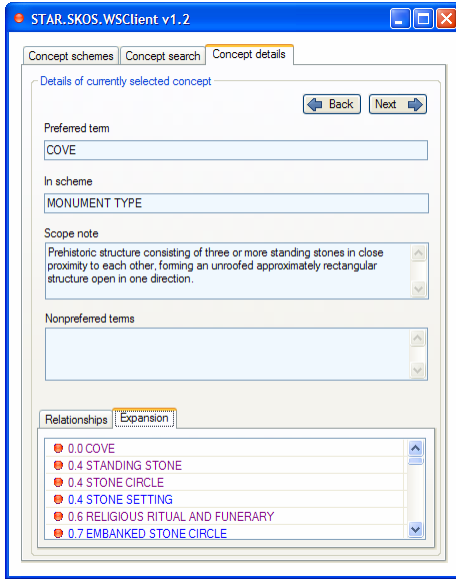


# Relationships Between EH Concepts





# SKOS Web Service and Client Applications



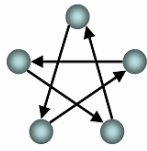
SKOS Web Service

Windows based client application

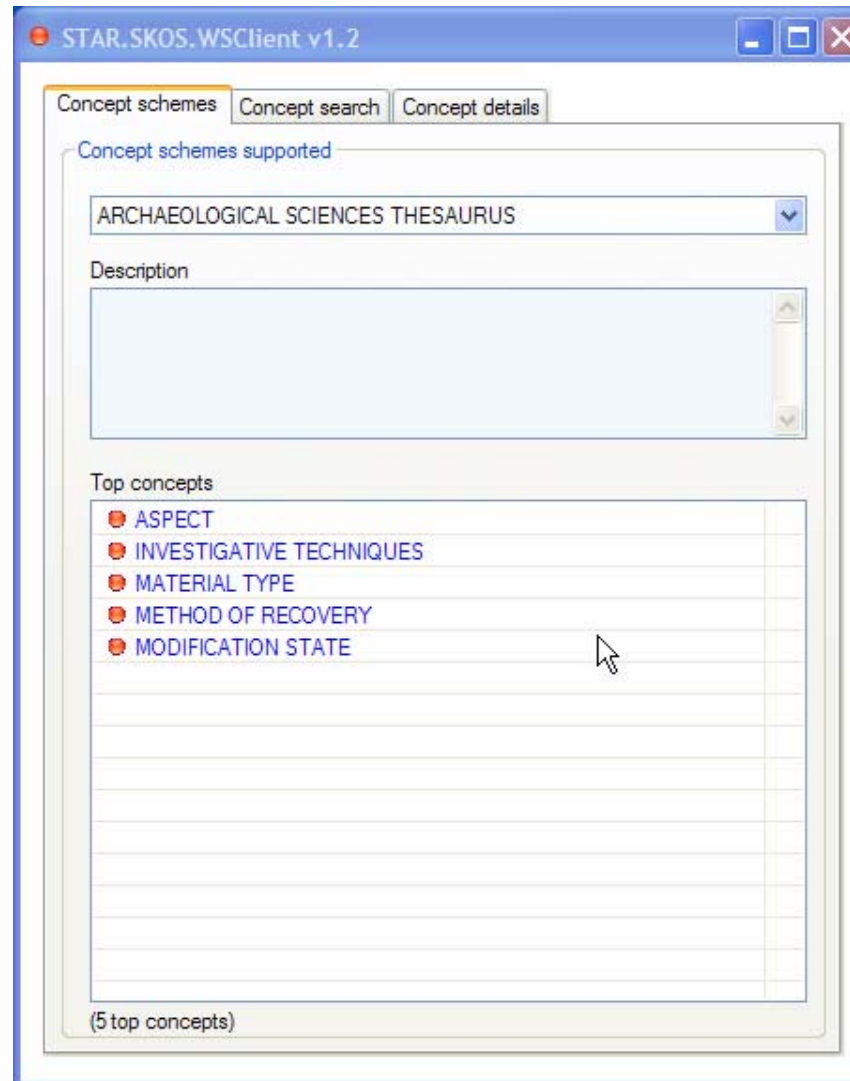
Web browser based components ('widgets')

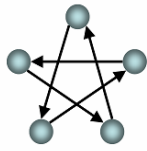
## SKOS Client Applications





# SKOS Client – Windows Application





# SKOS Client - Widgets

**Concept search**



Concept Search

**Concept**

BROOCH

<http://reswin1.isd.glam.ac.uk/star/rdf/concepts#EH128>

Ornament often with a hinged pin and catch, worn fastened to clothing.

**Concept relatives**

BT: JEWELLERY

NT: **ANNULAR BROOCH**

NT: BOW BROOCH

NT: DRAGONESQUE BROOCH

NT: LONG BROOCH

NT: PENANNULAR BROOCH

Concept Details

**Concept schemes**

EVIDENCE

MONUMENT TYPE

TIMELINE THESAURUS (test only)

ARCHAEOLOGICAL SCIENCES THESAURUS

MDA OBJECT TYPE

MAIN BUILDING MATERIALS

**Top concepts**

ANIMAL EQUIPMENT

AGRICULTURE AND SUBSISTENCE

ARCHITECTURE

FURNISHINGS AND FURNITURE

TRANSPORT

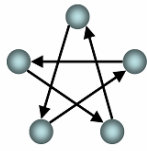
FOOD PREPARATION AND CONSUMPTION

Concept Schemes

- BOW (WEAPON)
- LONGBOW
- ARROW
- PROJECTILE WEAPON
- HUNTING OBJECT
- PROJECTILE
- BLOWPIPE (WEAPON)
- CROSSBOW
- SPEAR

Concept Expansion





## CRMEH Web Service and Client Application

STAR.CRM.WSClient v1.4

Suggest terms: brooch

Suggested terms:

- CRUCIFORM BROOCH
- DISC BROOCH
- DOLPHIN BROOCH
- DRAGONESQUE BROOCH
- HEADSTUD BROOCH
- HOD HILL BROOCH
- Hook Norton Brooch
- KNEE BROOCH
- Lamberton Moor Brooch
- LANGTON DOWN BROOCH
- LONG BROOCH
- NAUHEIM DERIVATE BROOCH
- PENANNULAR BROOCH
- PLATE BROOCH
- POLDEN HILL BROOCH
- Rosette Brooch
- SAUCER BROOCH
- Simple One-Piece Brooch
- SMALL LONG BROOCH

Expanded query for selected term: +("PENANNULAR BROOCH" "BROOCH" "Fibula" "Brooch Spring")

Search: +brooch +(pennanular fantail >nauheim)

Legend:

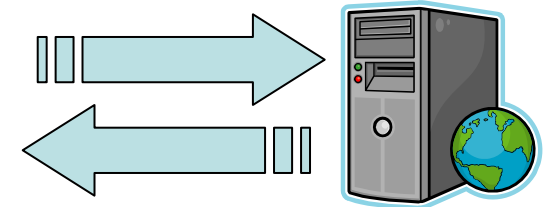
- Raunds Roman
- Raunds Prehistoric
- LEAP Silchester
- Show Colours

Property	Text
value	Nauheim derivative, the spring formed in one piece with the rest of the brooch. Two tur...

Nauheim derivative, the spring formed in one piece with the rest of the brooch. Two turns remain of a four turn spring made in one piece with a plain bow of sub-rectangular section which tapers to a narrow foot bearing an unperforated catch-plate. Type common in South Britain in the mid first century AD.

http://tempuri/star/base#e62.rrad.object.descriptivetext.1567

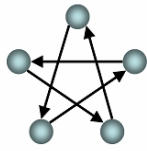
### CRMEH Client Application



### CRMEH Web Service

Search and browse seamlessly across multiple archaeological datasets





# CRMEH Client Application

STAR.CRM.WSClient v1.4

Suggest terms

Go

Suggested terms

Expanded query for selected term

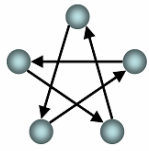
Search Favourites

Go

Legend

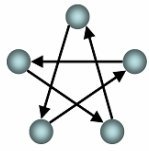
- Raunds Roman
- Raunds Prehistoric
- LEAP Silchester
- Show Colours

Property	Text



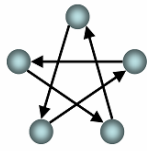
## Next Steps for Database Cross Search

- Query builder – expose support for complex query patterns (SPARQL)
- Closer integration of thesauri into overall search process
- Browser-based application components (AJAX)
- Include further archaeological datasets



# Indexing Grey Literature Documents

- Information Extraction for enabling 'rich', semantic aware indexing of Archaeology fieldwork reports.
- Rule Based method for NER name entity recognition and semantic annotation of terms with respect to the ontological model CRMEH
- Thesaurus and Flat Gazetteer lists to support the task of NER; connecting annotated terms to thesaurus and model classes.
- XML to represent simple and combined annotation structures .

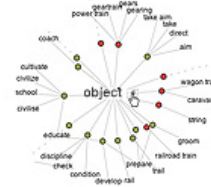


# Information Extraction Framework

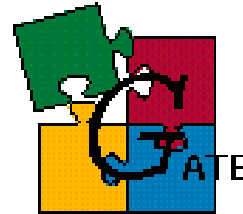
**EH Thesaurus**  
- Object Types  
- Archaeological Periods



**Ontology**  
-CIDOC CRM-EH



**Java Pattern Engine**



**Gazetteer Lists**

**General Architecture for Text Engineering**



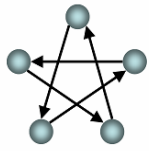
**Corpus Collection**



**<?xml?>**

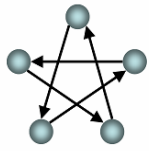
XML structures to represent simple and combined annotation forms





# The Process of Information Extraction

- 3000 terms from EH Thesauri (Periods – Object Types) supported the NER task.
- Flat gazetteer entries to expand on single term annotation
- Linguistic pattern rules for revealing connections between annotations
- Prototype development
  - Initially targeted to E49.Time Appellation entities , expanded to E19.Physical Object entities.
  - 9 Annotation Types Produced in a cascading process using Jape transducers.
  - Attribute assignment to semantic descriptions of annotated terms
  - Nested annotation structures to express connections between terms.



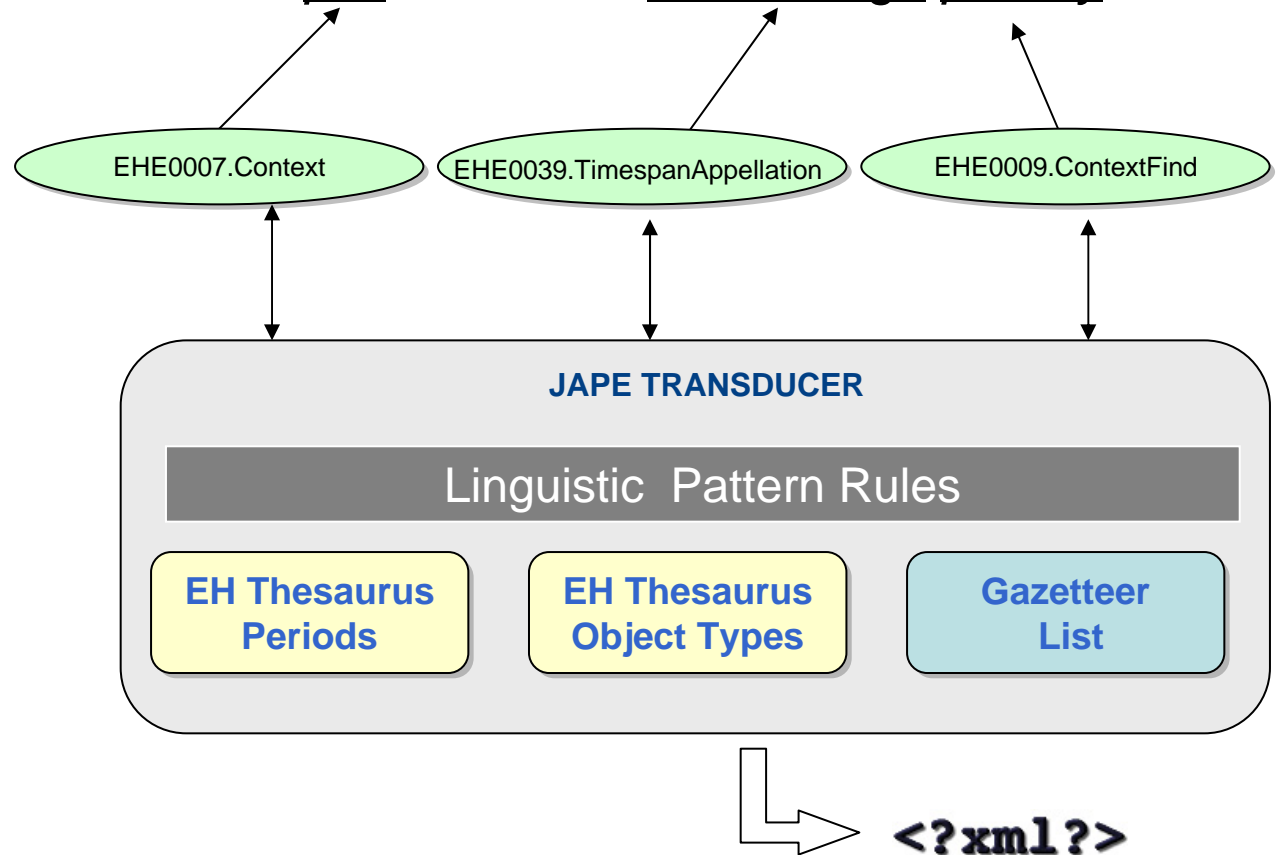
# Linking Text to Semantic Descriptions

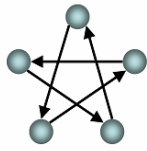
“...eleven of these seventeen pits contained Bronze Age pottery.”

```

({Context} | {Context_Period})
({Token.category == VB} | {Token.category == VBZ} | {Token.category == VBP} | {Token.category == VBN} | {Token.category == VBG} | {Token.category == VBD})
({Archeo_Object} | {Object_Type})

```





# XML Exports

```
<Context_Find gate:gateId="222760"
rule="Context Find">
  <Context
gate:gateId="221765" rule="Context"
>pits</Context>
  contained
  <Archeo_Object
gate:gateId="221710"
rule="ArcheoObject_Simple" >
    <E49_Time_Appellation
gate:gateId="220303" SKOS-
EH="134732" rule="thesaurus
term">Early Bronze
Age</E49_Time_Appellation>
      <Object_Type
gate:gateId="220577" SKOS-
EH="051544" rule="thesaurus
term">pottery</Object_Type>
    </Archeo_Object>
  </Context_Find>
```

Archeo Object				
TERM	ANNO+		GRAMMAR	COUNT
ROMAN COIN	Term	skos	E49_Time_Appellation #text	1
	ROMAN	134738		
	COIN	95423		

**COIN**  
A piece of metal, usually cast, struck or stamped, with a definite value.

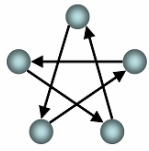
**Broad Term:**  
CURRENCY

**Top Term:**  
CURRENCY

Object_Type:			
TERM	CIDOC-EH	SKOS-EH	COUNT
WALL	Object_Type	96129	5
LEVEL	Object_Type	95353	4
STAFF	Object_Type	96735	2
POTTERY	Object_Type	100055	1
CERAMIC	Object_Type	141190	1
BRICK	Object_Type	96010	1
COIN	Object_Type	95423	1
POST	Object_Type	97616	1
BUCKET	Object_Type	96364	1

**PHP DOM\_XML**  
employed over XML structures to render the annotations and to display thesaurus descriptions





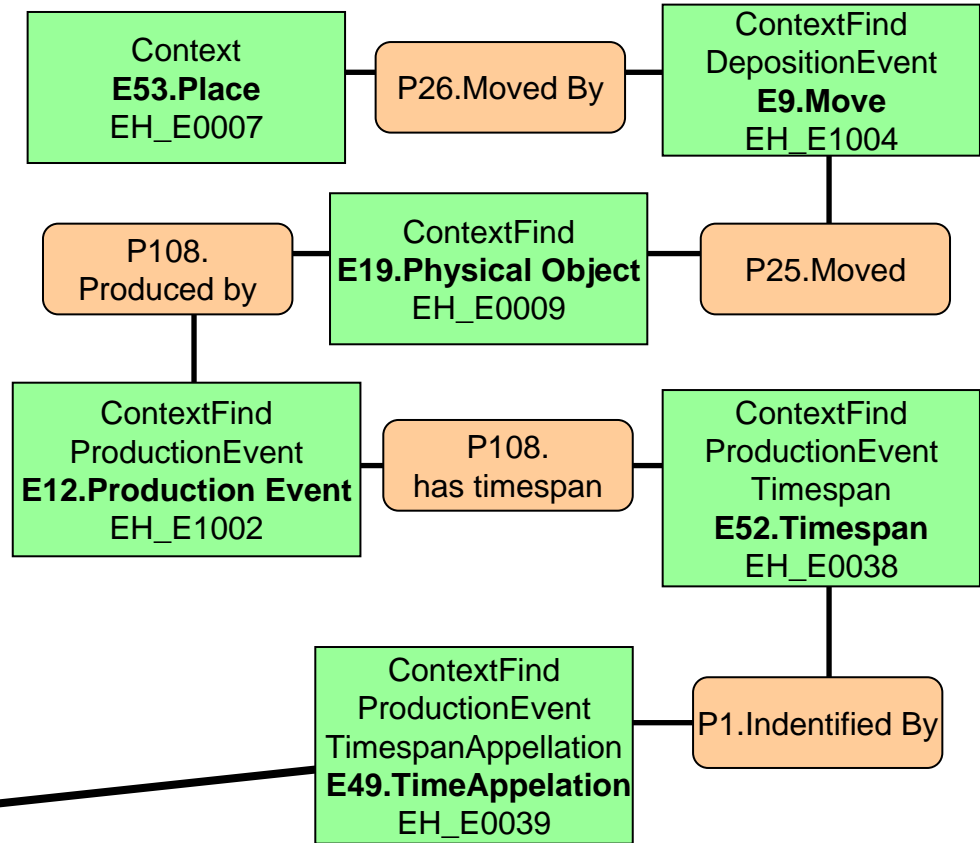
# Model Adaptation Issues

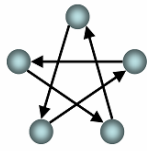
“...eleven of these seventeen pits contained Bronze Age pottery.”

```

<Context_Find>
  <Context>pits</Context>
  contained
  <Archeo_Object>
    <E49_Time_Appellation>Bronze Age
    </E49_Time_Appellation>
    <Object_Type>pottery
    </Object_Type>
  </Archeo_Object>
</Context_Find>

```





## Work in progress

**Andronikos - Excavations on Grey Literature**

Corpus - Excavations  
 Home  
 Early Xrays  
 02\_06  
 26\_05

(1752) - Manor Bam Pernehall Road Archaeological Monitoring Report  
 Heritage Network - 2006  
 Annotated Document: heritage1-12895

Address	Location	Grid Reference	Monument Type
ALTHS - Saunders G OOID - report number 305	Site: Manor Bam, Pernehall Road, Kingsoe County: BEDFORDSHIRE District: BEDFORD Parish: BOUNHURST AND KEYSOE County: ENGLAND	Type: POINT Mcode: TL Easting: 776 Northing: 6274	Type: MONUS Description: ERICK WALL Period: POST MEDIEVAL

**Information Extraction**

E43 Time Appellation				Time Appellation			
TERM	CIDOC-EH	SKOS-EH	COUNT	TERM	ANNO+	GRAMMAR	COUNT
19TH CENTURY	E43_Time_Appellation	134040	4				
MODERN	E43_Time_Appellation	134747	3				
20TH CENTURY	E43_Time_Appellation	134841	1				
ROMAN	E43_Time_Appellation	134730	1				
MEDIEVAL	E43_Time_Appellation	134745	1				
POST-MEDIEVAL	E43_Time_Appellation	134746	1				
DOMESDAY SURVEY	E43_Time_Appellation	136415	1				

Object Type				Andron Object			
TERM	CIDOC-EH	SKOS-EH	COUNT	TERM	ANNO+	GRAMMAR	COUNT
WALL	Object_Type	96129	5				
LEVEL	Object_Type	96363	4				

[andronikos.kyklos.co.uk](http://andronikos.kyklos.co.uk)

- Closer integration between ontological model and Jape rules
- Mapping of annotation structures to RDF triples
- Enriching the process with additional Knowledge resources ie glossaries
- Expand the method to larger corpus collection
- Negation detection and document pre-processing

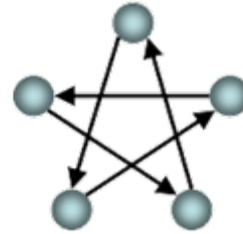
### 3.2. Phase I: Late Neolithic/Early Bronze Age

(Figs. 4, 5 and 6)

Features of this phase mainly occur in a cluster of pits (0162) and this cluster accounts for the vast majority of finds recovered from the entire site. The cluster consists of seventeen pits, all of which had very similar form and fill characteristics. The fills were uniformly similar, with dense quantities of charcoal containing hazel nutshell, burnt bone fragments and Beaker pottery sherds. This suggests that the pits were all open and filled simultaneously from a common source. Eleven of these seventeen pits contained Early Bronze Age pottery. Fifteen also contained worked flint that appears to be contemporary with the pottery.

A further five small pits lay nearby to the north-west (0221, 0223, 0227, 0229 and 0233) and two of these (0223 and 0227) also contained Early Bronze Age material. The appearance of these pits was somewhat different to those in 0162, they are less well defined, loosely scattered and lack the dense charcoal fills of the pits in the 0162 group. Three more pits (0004, 0069 and 0116) which are scattered across the site also contained Early Bronze Age pottery. 0004 is a sizeable feature and lies within another small group of features which also contain Iron Age material. The other two pits, 0069 and 0116 are isolated features but 0116 contained a sizeable assemblage of struck flint.





**STAR**

## Semantic Technologies for Archaeological Resources

<http://hypermedia.research.glam.ac.uk/kos/star/>  
<http://andronikos.kyklos.co.uk>

[avlachid@glam.ac.uk](mailto:avlachid@glam.ac.uk)

[dstudhope@glam.ac.uk](mailto:dstudhope@glam.ac.uk)

[cbinding@glam.ac.uk](mailto:cbinding@glam.ac.uk)



Arts & Humanities  
Research Council



ENGLISH HERITAGE

University of Glamorgan

you live, you learn

