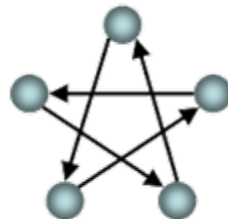


# ***Semantic Annotations in the Archaeological Domain***

Andreas Vlachidis, Ceri Binding, Keith May, Douglas Tudhope

## **STAR** **Semantic Technologies for Archaeological Resources**



Arts & Humanities  
Research Council

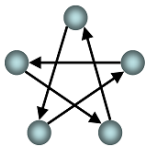


ENGLISH HERITAGE

University of Glamorgan

you live, you learn

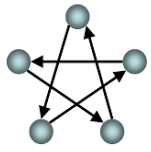




# About This Presentation

- **The STAR project**
  - Aims and Objectives
  - Architecture of Semantic Access to Disparate data sets
  - Adapted Conceptual Models and Knowledge Resources
  - Progress to date and available Web services
- **Semantic Annotations Pathway**
  - The aim of the Research
  - OBIE for rich, semantic indexing
  - Domain Specific Requirements
- **Excavating Grey Literature Documents**
  - General Architecture for Text Engineering (GATE)
  - Rule Based Pattern Matching Approaches
  - 'Gold Standard' Pilot Evaluation
- **Adaptation Issues and Conclusions**
  - Ontological Model Verbosity
  - Prototype Query Builder
  - Prototype Indexing Deployment

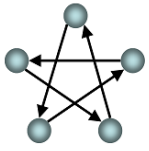




# The STAR Project

- **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

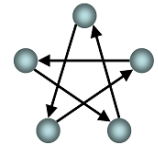




# Conceptual Models and Knowledge Resources

- **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 Organization System
  - RDF representation of thesauri, glossaries, taxonomies, classification schemes etc.

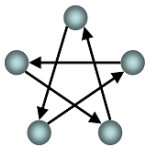




# CIDOC Conceptual Reference Model

- *“The **CIDOC CRM** is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to”* [ <http://cidoc.ics.forth.gr/> ]
- About 80 classes and 130 properties for cultural and natural history
- Intellectual guide to create schemata, formats, profiles  
Extension of CRM with a categorical level, e.g. reoccurring events
- Best practice guide for data integration (mapping)  
Transportation format for data integration / migration /Internet

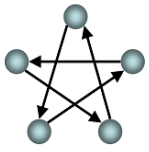




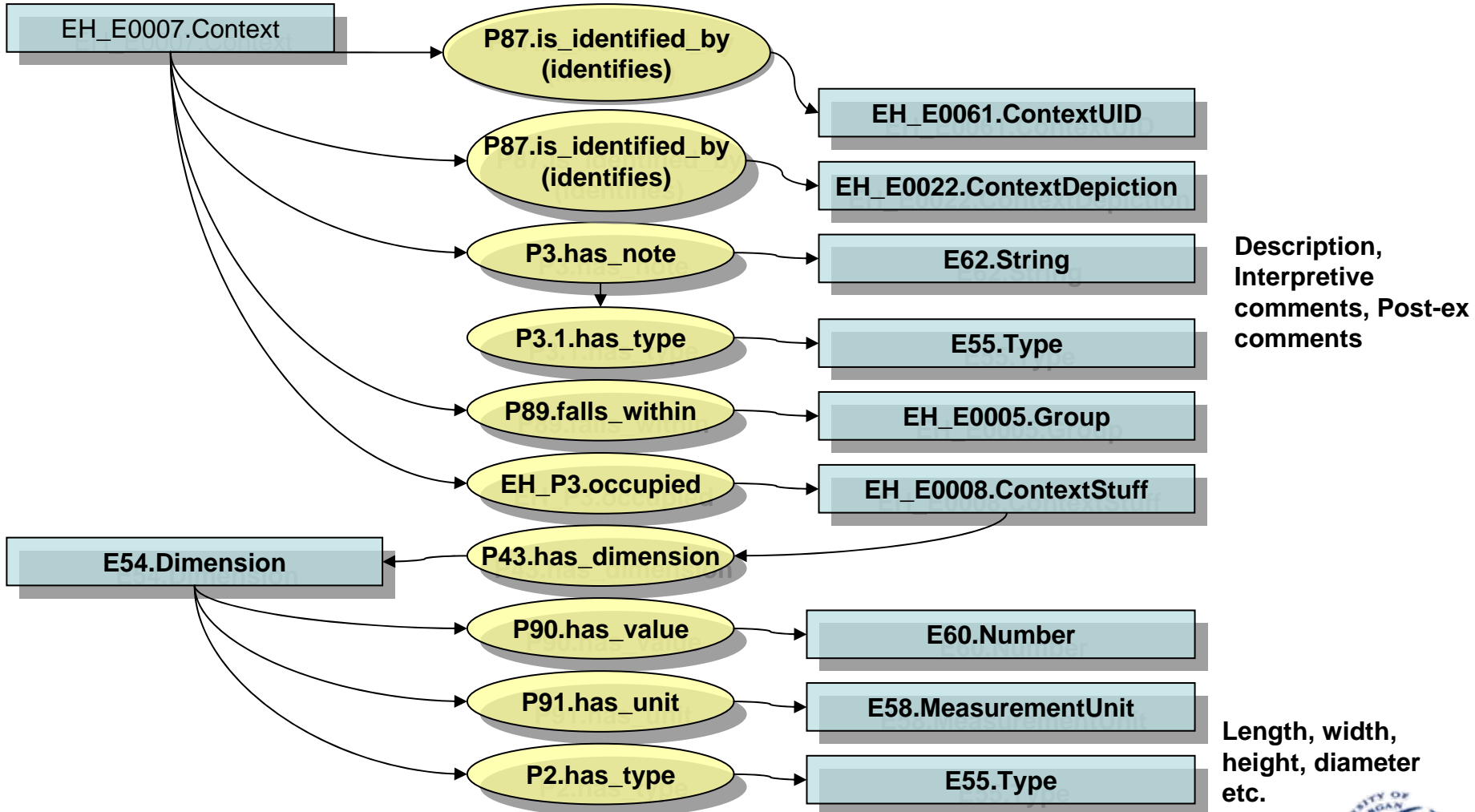
## **CRMEH- English Heritage Ontological Model**

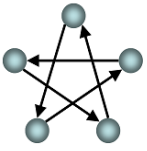
- Adopting and **extending** CRM for complete picture of on-site and off-site processes.
- **Entities and relationships** relating to Stratigraphic relations and phasing information, finds recording and environmental sampling.
- The extended CRM model CRM-EH, comprises **125** extension sub-classes and **4** extension sub-properties.
- Multiple disconnected **databases** and **legacy data**: CRM as ‘semantic glue’ to pull the data together





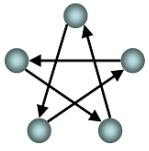
# CRMEH A Closer Look





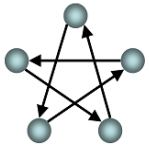
# Simple Knowledge Organisation System

- Standard set for representation
  - Thesauri, Taxonomies, Classification Schemes
- Publication of controlled structured vocabularies
  - Intended for the Semantic Web
  - Built upon standard RDF(S)/XML W3C technologies
- Looser semantics than e.g. OWL



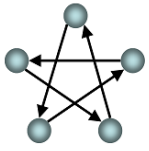
# English Heritage Thesauri

- Monument types thesaurus
  - Classification of monument type records
- Evidence thesaurus
  - Archaeological evidence
- MDA object types thesaurus
  - Archaeological objects
- Building materials thesaurus
  - Construction materials
- Archaeological sciences thesaurus
  - Sampling and processing methods and materials
- Timelines thesaurus
  - Periods, and time-based entities



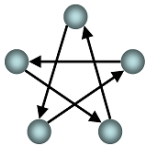
# Data Mapping and Extraction

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



## Applications and Utilities

- **Data Mapping and Extraction Utility**
  - Bespoke mapping/extraction utility
  - Extract archaeological data conforming to mapping
  - Semi-automated manner
- **Prototype CRM Browser**
  - Prototype CRM browser
  - Query entry of free-text search terms
  - Option to navigate the results of returned queries.



# STAR Data Mapping and Extraction Utility

Database: RRAD

Subject: <http://tempuri/star/cmeh#EHE0007.Context>

Predicate: [http://cidoc.ics.forth.gr/rdfs/cidoc\\_y4.2.rdfs#P3F.has\\_note](http://cidoc.ics.forth.gr/rdfs/cidoc_y4.2.rdfs#P3F.has_note)

Object: <http://tempuri/star/cmeh#EHE0046.ContextNote>

Prefix: context.contextno.

Column: ContextNo

Literal value: Description

FROM clause: Context

WHERE clause: trim(Description) <> ""

Generated SQL:

```
SELECT DISTINCT
'http://tempuri/star/cmeh#EHE0007.Context' AS [SUBJECTTYPE],
'http://tempuri/star/base#EHE0007.rad.context.contextno.' & ContextNo AS [SUBJECT],
'http://cidoc.ics.forth.gr/rdfs/cidoc_y4.2.rdfs#P3F.has_note' AS [PREDICATE],
'http://tempuri/star/cmeh#EHE0046.ContextNote' AS [OBJECTTYPE],
'http://tempuri/star/base#EHE0046.rad.context.description.' & ContextNo AS [OBJECT],
Description AS [LITERAL]
FROM Context
WHERE 1 = 1
AND trim(Description) <> ""
```

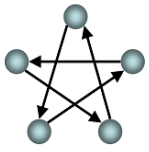
Resultant Data:

SUBJECTTYPE	SUBJECT	PREDICATE	OBJECTTYPE	OBJECT	LITERAL
http://tempuri/...	http://tempuri/...	http://cidoc.ic...	http://tempuri/...	http://tempuri/...	Upper ploughsoil over...
http://tempuri/...	http://tempuri/...	http://cidoc.ic...	http://tempuri/...	http://tempuri/...	Original recorded cob...
http://tempuri/...	http://tempuri/...	http://cidoc.ic...	http://tempuri/...	http://tempuri/...	Alluvial clay deposit; ...
http://tempuri/...	http://tempuri/...	http://cidoc.ic...	http://tempuri/...	http://tempuri/...	Natural subsoil
http://tempuri/...	http://tempuri/...	http://cidoc.ic...	http://tempuri/...	http://tempuri/...	This context comprise...
http://tempuri/...	http://tempuri/...	http://cidoc.ic...	http://tempuri/...	http://tempuri/...	Post-excavation cont...

- Entry boxes corresponding to Entity-Relationship-Entity elements of the CRM-EH statement.
- SQL query building up: SQL query incorporating selectable consistent URIs (CRM, CRM-EH, SKOS, Dublin Core and others).
- Query execution against the selected database
- Tabular data export to RDF format file







## Prototype CRM Browser

STAR.CRM.WSClient v1.4

Suggest terms: brooch

Suggested terms: AESICA BROOCH, ANIMAL BROOCH, ANNULAR BROOCH, Ansate Brooch, AUCISSA BROOCH, BOW AND FANTAIL BROOCH, BOW BROOCH, BROOCH, BROOCH PIN, Brooch Spring, CATERPILLAR BROOCH, COLCHESTER BROOCH, CROSSBOW BROOCH, CRUCIFORM BROOCH, DISC BROOCH, DOLPHIN BROOCH, DRAGONESQUE BROOCH, HEADSTUD BROOCH, HOD HILL BROOCH, Hook Norton Brooch, KNEE BROOCH, Lambertton Moor Brooch, LANGTON DOWN BROOCH, LONG BROOCH, NAUHEIM DERIVATE BROOCH, PENANNULAR BROOCH

Expanded query for selected term: +([NAUHEIM]) ("NAUHEIM DERIVATE BROOCH" "Simple One-Piece Brooch" "BOW BROOCH" )

Search: +([NAUHEIM]) ("NAUHEIM DERIVATE BROOCH" "Simple One-Piece Brooch" "BOW BROOCH" )

Legend: Raunds Roman, Raunds Prehistoric, LEAP Silchester, Show Colours

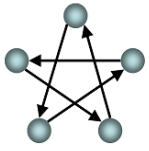
Silchester Town Life Project - Project SILCH: I3736 - Microsoft Internet Explorer

Address: http://www.silchester.reading.ac.uk

Property	Text
value	http://www.silchester.reading.ac.uk/i3/ite

http://tempuri/star/base#ehe0083.leap.photos.caption.3736

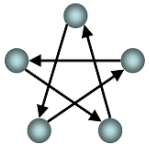
- Test and demonstrate interoperability between datasets.
- Incorporated the SKOS based thesauri browsing interface
- Distinguish between results, colour coding
- Search for “Nauheim Brooch”, Browse results and ‘drill’ deeper
- Link to live data, via returned URL hyperlinks



# Semantic Annotations Pathway

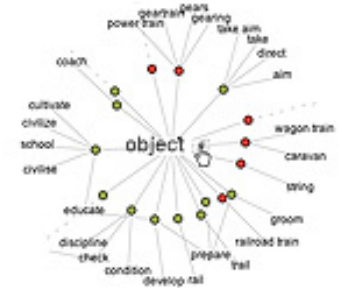
- Semantic Annotations
  - specific **metadata generation** and usage schema
  - aimed to automate identification of concepts and their relationships in documents
- Research effort
  - Directed towards the generation of **rich document indices** carrying semantic and interoperable properties for the purposes of semantic interoperability .

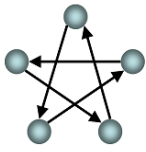




# Ontology Based Information Extraction

- Ontologies; a mediator technology between concepts and their worded representations
- Advance Information Retrieval
  - Beyond the limitations of words to the level of concepts
- Aid Information Retrieval
  - To make inferences from heterogeneous data sources
- Information Extraction
  - A specific text analysis task aimed to extract specific information snippets from documents
  - Ontologies to drive/inform IE
  - To describe the conceptual arrangements of semantic annotations.





# Archaeology Domain & Upper Level Ontologies

**Gnosis**  
Powered by ClearForest a Thomson Reuters company

Facility

- Chelmsford Road (1)
- Smiths Farm (1)

Organization

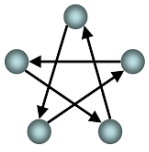
- Essex Police (2)

- Thompson Reuters - Gnosis Plug-in
- Limitations of Upper level and Lightweight Ontologies in specialised domains
- e.g. Archaeology Grey Literature Document

An archaeological evaluation was carried out by ECC FAU on behalf of Essex Police on the site of a proposed new police station at Smiths Farm, on the southeastern outskirts of Great Dunmow, Essex. The site was formerly rough pasture. The Chelmsford Road, which is thought to be the line of a Roman road, runs immediately to the east of the site. Five 30m x 2m trenches were excavated within the footprint of the proposed building and the area of associated carpark. Only one archaeological feature was revealed, a ditch containing prehistoric pottery dating to the Late Bronze Age or Early Iron Age along with burnt flints and flint flakes. No other archaeological features were identified, although a number of prehistoric pottery sherds and flint flakes were discovered on the surface of the natural geology. Although the results of the evaluation do not suggest intensive landscape use during the Late Bronze/ Early Iron Ages it is clear from this and other nearby investigations that a focus for the low level activity seen may well lie in the general vicinity. The absence of Roman or medieval remains indicates that this site was well outside the settlements of these periods. The low quantity and quality of the remains encountered on the site suggests that there is only a minor archaeological implication for the location of the proposed police

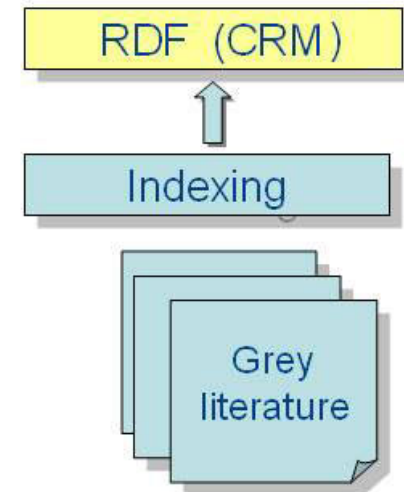
Highlighting powered by ClearForest ©

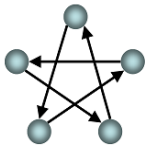




# Excavating Grey Literature Documents

- **Grey Literature**; *source materials that can not be found through the conventional means of publication*
  - Raunds reports
  - Online Access to the Index of archaeological excavationS (OASIS)  
[<http://ads.ahds.ac.uk/project/oasis/>]
  - Library of unpublished fieldwork reports
  - English Heritage listed Buildings System (LBS)
- **Semantic Indexing**
  - Interoperable technologies W3C standards
    - XML, RDF representation
  - TEI adoption



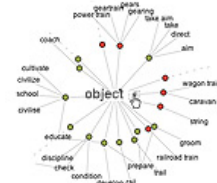


# Information Extraction Framework

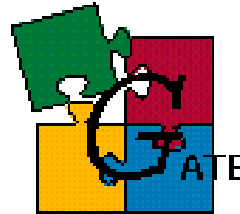
**EH Thesaurus**  
- Object Types  
- Archaeological Periods



**Ontology**  
-CIDOC CRM-EH



**Java Pattern Engine**



**Gazetteer Lists**

**General Architecture for Text Engineering**



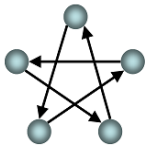
**ADS - OASIS  
Grey Literature**



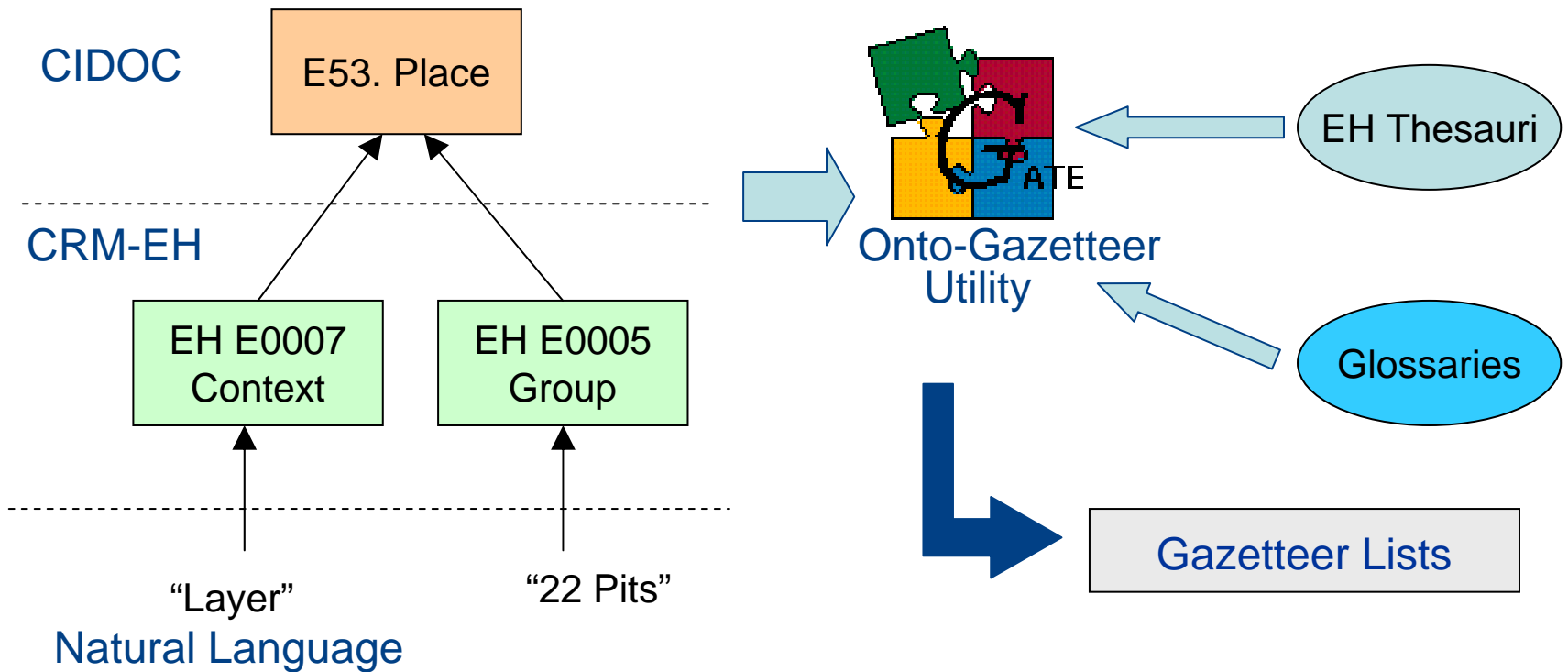
**<?xml?>**

**XML structures to represent  
semantic properties**



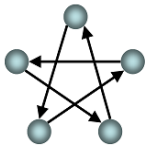


# GATE Mapping of Knowledge Resources



Reference to SKOS mapped to the MinorType attribute of list entries





# JAPE Pattern Matching Rules



Natural Language – Gazetteer Look-up

“**Ditch** containing **prehistoric pottery** dating to the **Late Bronze Age or Early Iron Age** along with **burnt flints** and **flint flakes**”

E53 Place

E49Time Appellation

E19 Physical Object

Pattern Matching Rules expanded beyond simple gazetteer look-up

<entity><same-entity>



“*Late Bronze Age or Early Iron Age*”

<entity><other-entity>



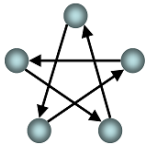
“prehistoric pottery”

<entity><verb>(<entity>  
/<structure>)



“Ditch containing prehistoric pottery”





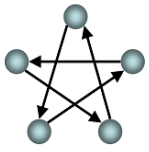
# A Cascading Extraction Process

- Processing Resources
  - JAPE Context\_Find
  - JAPE Context\_Extend
  - JAPE Context\_Group
  - JAPE Contexts\_Lookup
  - JAPE Physical\_Object\_Exte
  - JAPE Physical\_Object\_Plus
  - JAPE Physical\_Object\_Look
  - JAPE Periods\_Compositio
  - JAPE Periods\_Extend
  - JAPE Periods\_Lookup
  - OntoGazetteer
  - Hash Gazetteer\_0001
  - VerbChunker
  - POS
  - SentenceSplitter
  - Tokenizer

- A cascading order of natural language processes over text
- Expanding from simple gazetteer *Look-Up* matching rules to complex JAPE transducers
- Build up from previously defined annotations to express annotation structures (templates) of ontological concepts







# Annotation Types exposed in XML

## Annotation Types

- Context
- ContextExtend
- ContextFind
- ContextGroup
- ContextPlusTime
- PhysicalObject
- PhysicalObjectExtend
- PhysicalObjectPlusTime
- TimeAppellation
- TimeAppellationComposition
- TimeAppellationExtend

## XML Annotation Structures

("Ditch containing prehistoric pottery")

```

<ContextFind>
  <Context>Ditch</Context>
  <VG>containing</VG>
  <PhysicalObjectPlusTime>
    <Time_Appellation>
      prehistoric
    </Time_Appellation>
    <PhysicalObject>
      pottery
    </PhysicalObject>
  </PhysicalObjectPlusTime>
</ContextFind>

```

## DOM – XML Applications

Term	skos
PREHISTORIC	134718
POTTER	PREHISTORIC Use for any site or object which is definitely

**Andronikos\***  
 Uses PHP-MySQL to  
 display semantic indices  
 values in HTML format

## Semantic Attributes for Annotation Types

```

<PhysicalObject gateId="8749" SKOS-EH="134718" thesaurus ="EH-Object  

Types" class="EHE0009.ContextFind"  

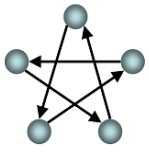
ontology="http://hypermedia.research.glam.ac.uk/media/files/documents/2  

008-04-01/CIDOC_v4.2_extensions_eh_.rdf" }

```

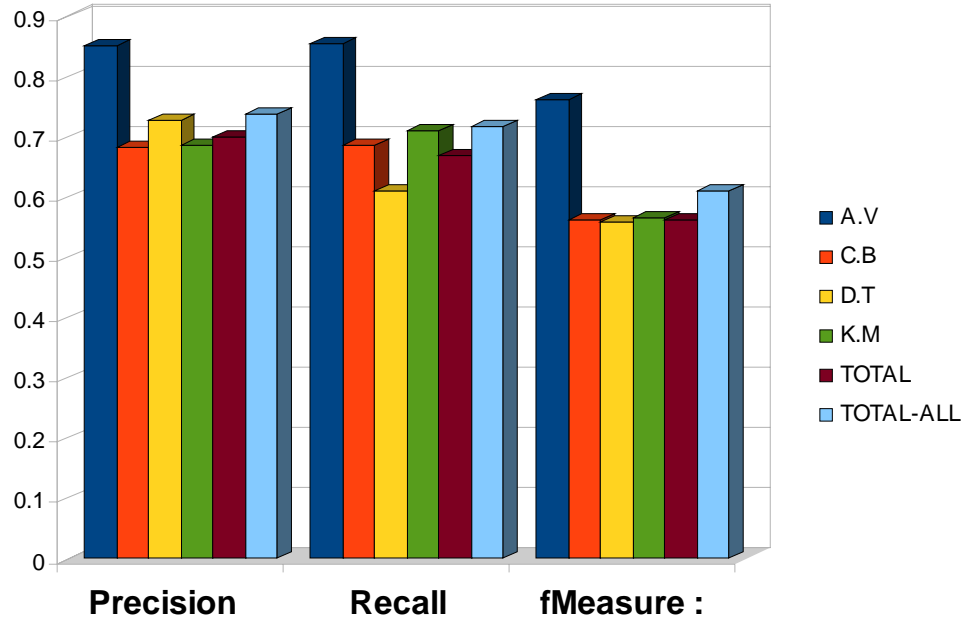






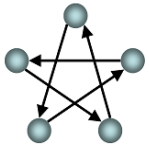
# 'Gold Standard' PILOT Evaluation

- 'Gold standard'; a collective effort of human annotators
- Manual annotation of GS with respect to the Annotation Types (aimed to suggest expansion)
- Pilot study (formative assessment).
- Aimed to benchmark the performance of the extraction mechanism
- **Inter-Annotators Scores**



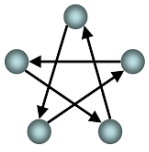
	AV	CB	DT	KM	TOTAL	TOTAL-ALL
<b>Precision</b>	0.85	0.68	0.72	0.68	0.69	0.73
<b>Recall</b>	0.85	0.68	0.61	0.71	0.66	0.71
<b>fMeasure :</b>	0.76	0.56	0.56	0.56	0.56	0.61





# Pilot Evaluation Results - Discussion

- Encouraging Recall and Precision rates over 70% for *Time Appellation* concepts
- The limited amount of glossary terms (*Places*) has influenced the performance
- Agreement for *Place* and *Physical Objects* was not always clear cut (i.e 'burnt tree throws')
- The potential of the method to extract complex phrases associated to two or more ontological entities
- Future work
  - Incorporation of additional Ontological Entities (Material, Samples)
  - Gazetteer enhancement
  - Pattern matching rules expansion
  - Formal evaluation of the Extraction method and overall retrieval performance



# Model Adaptation Issues

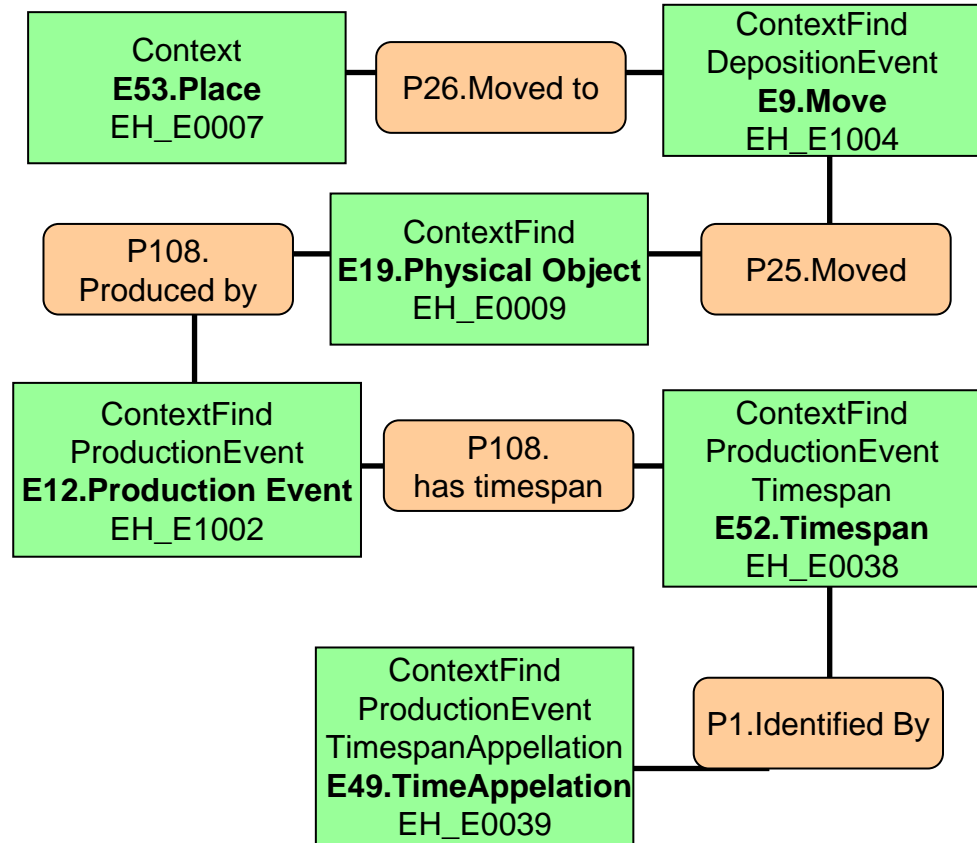
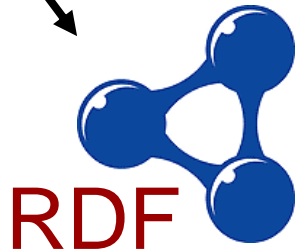
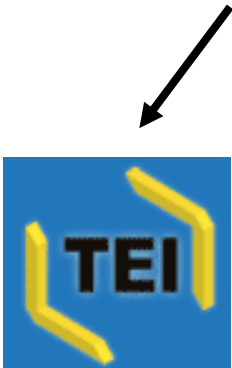
- CRM-EH is a detailed event driven model. Natural Language can be abstract. Mapping with entities/properties can by-pass model verbosity

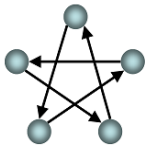
```

<ContextFind>
  <Context>Ditch<Context>
  <VG>containing</VG>
  <PhysicalObjectPlusTime>
  <Time_Appellation>prehistoric
  <Time_Appellation>
    <PhysicalObject>pottery
  </ PhysicalObject>
  <PhysicalObjectPlusTime>
</ContextFind>

```

## Interoperable Indices Formats





# Prototype Query Builder

- Inter-relationships of the CRM-EH modeled data.
- Short-cuts for traversing the commonly followed relationships between key entities

Query Builder

Query type:  Group  Context  Find  Sample

Find

Show all fields

ID:  (any)

Type:  Brooch

Note:  Nauheim

Dated:  (any)

Material:  (any)

Within Context:  (any)

(JSON)

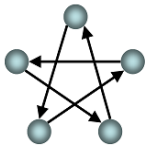
```
{"contextFind":[{"has_type":"Brooch","has_note":"Nauheim"}]}
```

(SPARQL)

```
#STAR SPARQL query: [Wed May 27 17:14:45 2009]
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX crm: <http://oidc.ics.forth.gr/rdfs/oidc_v4.2.rdfs#>
PREFIX crmeh: <http://tempuri/star/crmeh#>
PREFIX ...
```

Run query

- Archaeological *Context* associated key relationships:
  - Find
  - Sample
  - Stratigraphic, Spatial, Temporal
  - Group



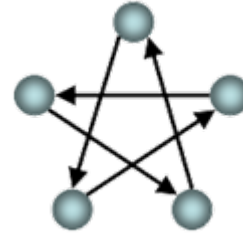
# Prototype Indices Deployment

An archaeological evaluation was carried out by ECC FAU on behalf of Essex Police on the site of a proposed new police station at Smiths Farm, on the southeastern outskirts of Great Dunmow, Essex. The site was formerly rough pasture. The Chelmsford Road, which is thought to be the line of a Roman road, runs immediately to the east of the site. Five 30m x 2m trenches were excavated within the footprint of the proposed building and the area of associated carpark. Only one archaeological feature was revealed, a ditch containing prehistoric pottery dating to the Late Bronze Age or Early Iron Age along with burnt flints and flint flakes. No other archaeological features were identified, although a number of prehistoric pottery sherds and flint flakes were discovered on the surface of the natural geology. Although the results of the evaluation do not suggest intensive landscape use during the Late Bronze/ Early Iron Ages it is clear from this and other nearby investigations that a focus for the low level activity seen may well lie in the general vicinity. The absence of Roman or medieval remains indicates that this site was well outside the settlements of these periods. The low quantity and quality of the remains encountered on the site suggests that there is only a minor archaeological implication for the location of the proposed police

LATE BRONZE AGE OR EARLY IRON AGE	<table border="1"><tr><td>Term</td><td>skos</td></tr><tr><td>LATE BRONZE AGE</td><td>134734</td></tr><tr><td>EARLY IRON AGE</td><td>134735</td></tr></table>	Term	skos	LATE BRONZE AGE	134734	EARLY IRON AGE	134735	E49_Time_Appellation #text 5			
Term	skos										
LATE BRONZE AGE	134734										
EARLY IRON AGE	134735										
ROMAN OR MEDIEVAL	<table border="1"><tr><td>Term</td><td>skos</td></tr><tr><td>ROMAN</td><td>134738</td></tr><tr><td>MEDIEVAL</td><td>134745</td></tr></table>	Term	skos	ROMAN	134738	MEDIEVAL	134745	<table border="1"><tr><td>EARLY IRON AGE</td></tr><tr><td>Broad Term: IRON AGE</td></tr><tr><td>Top Term: CULTURAL PERIOD</td></tr></table>	EARLY IRON AGE	Broad Term: IRON AGE	Top Term: CULTURAL PERIOD
Term	skos										
ROMAN	134738										
MEDIEVAL	134745										
EARLY IRON AGE											
Broad Term: IRON AGE											
Top Term: CULTURAL PERIOD											
PREHISTORIC PERIOD	<table border="1"><tr><td>Term</td><td>skos</td></tr><tr><td>PREHISTORIC</td><td>134718</td></tr></table>	Term	skos	PREHISTORIC	134718	#text 2					
Term	skos										
PREHISTORIC	134718										

- Andronikos web-portal development
- Utilise semantic annotation XML files
- The server side technology PHP DOM XML
- MySQL database server to store relevant thesauri structures.





## **STAR** Semantic Technologies for Archaeological Resources

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

avlachid@glam.ac.uk  
cbinding@glam.ac.uk  
keith.may@english-heritage.org.uk  
dstudhope@glam.ac.uk



Arts & Humanities  
Research Council

DANMARKS  
BIBLIOTEKSSKOLE



ENGLISH HERITAGE

University of Glamorgan

you live, you learn

