Exploring the Use of Semantic Technologies for Cross-Search of Archaeological Grey Literature and Data

Presented by

Keith May

@keith_may

Based on the work of

Andreas Vlachidis,
Ceri Binding, Keith May, Douglas Tudhope
University of Glamorgan / South Wales

STAR

Semantic Technologies for Archaeological Resources

http://hypermedia.research.southwales.ac.uk/kos/star/

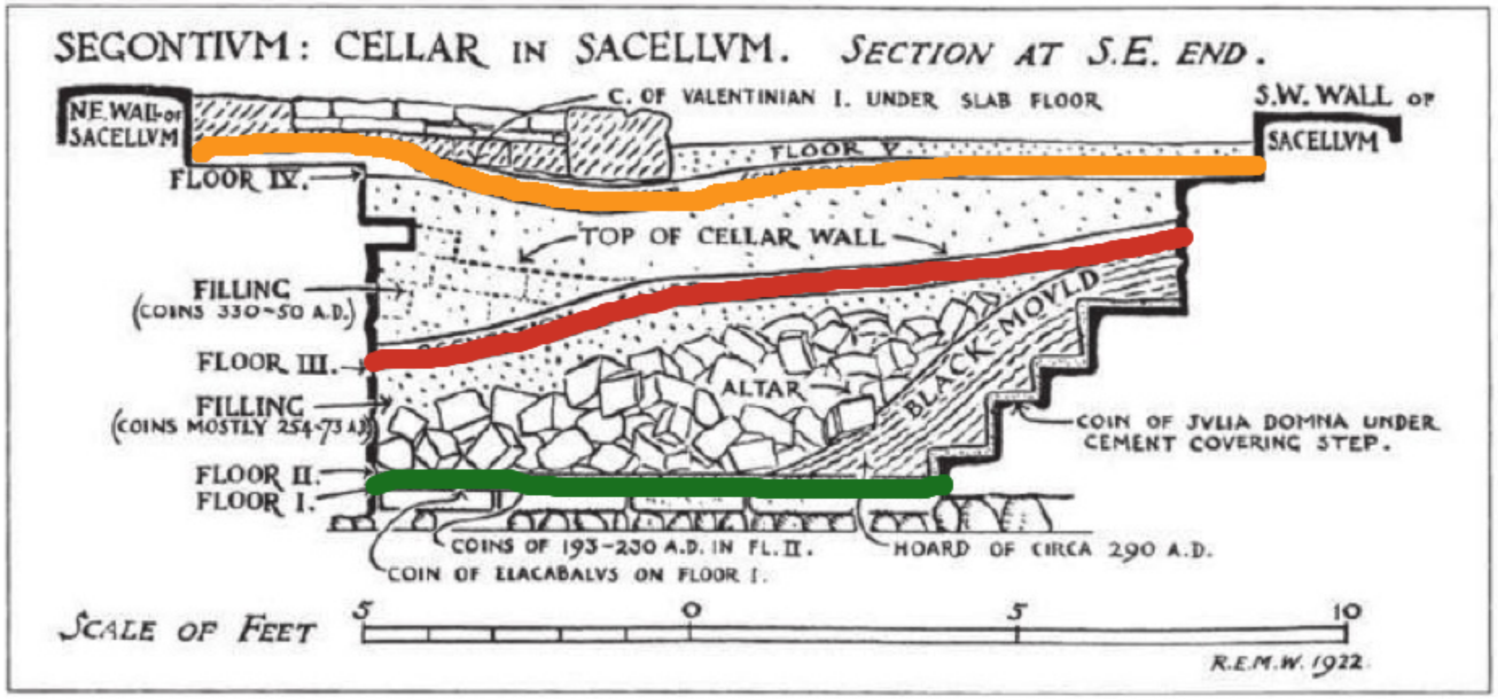
Arts & Humanities Research Council

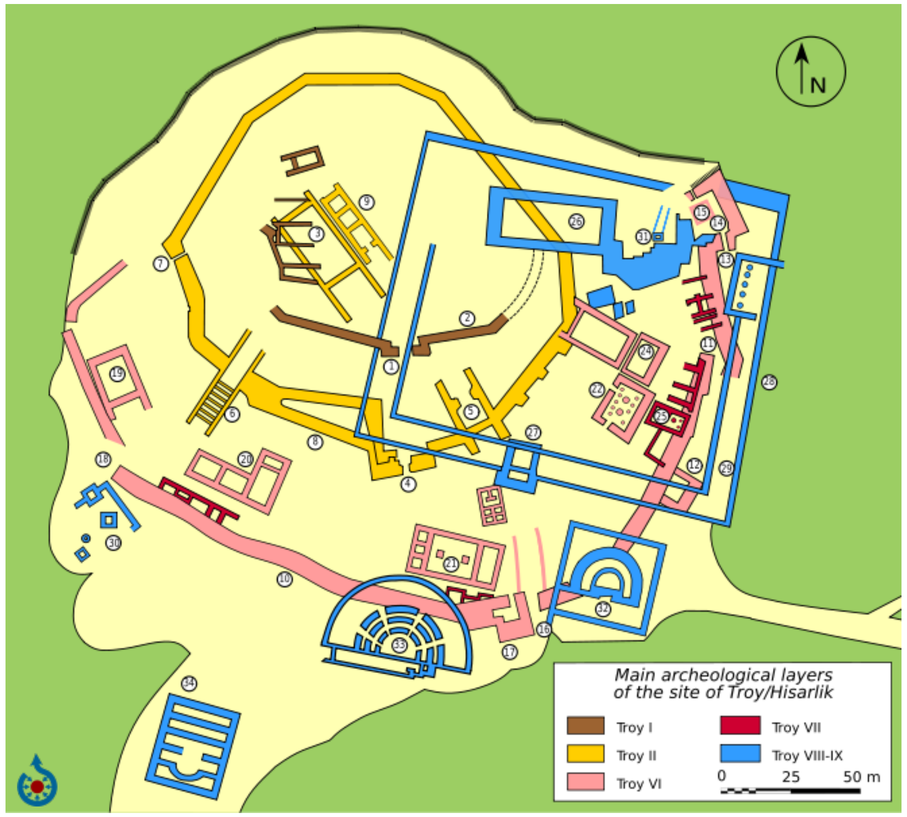
University of South Wales
Prifysgol De Cymru

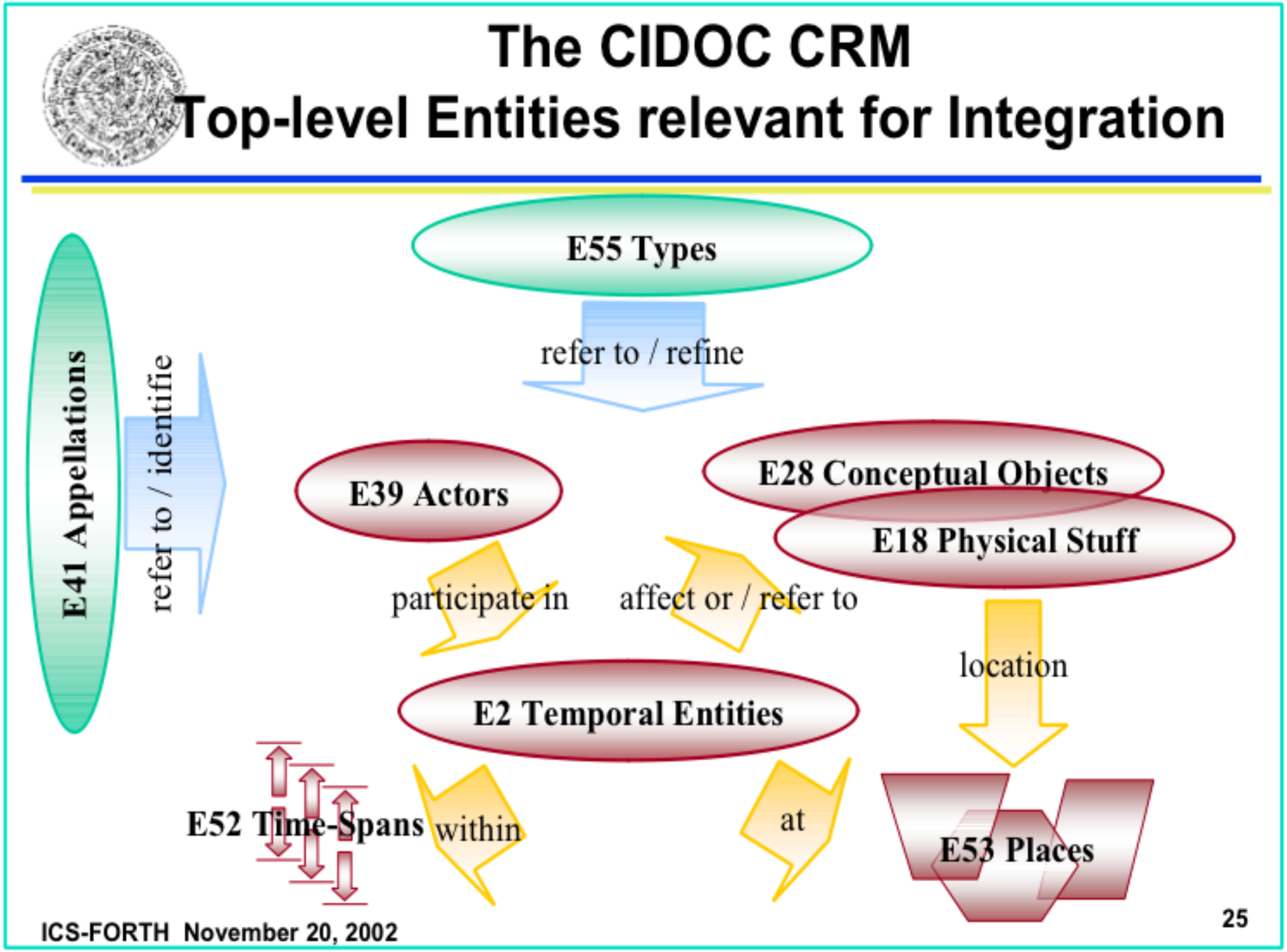
ENGLISH HERITAGE

Some Key Archaeological Places

- Investigation extents
- Contexts - positive & negative
- Finds spots - may record 3D spatial attributes.
- Sample locations/extents
- Groups of contexts e.g. Buildings - this will be more challenging as it gets its spatial information from several contexts
- Phases of Activity (Events) - Spatio-Temporal relationships between Group info


SEGONTIVM: CELLAR IN SACELLVM. SECTION AT S.E. END.
NE. WALL of SACELLVM
C. OF VALENTINIAN I. UNDER SLAB FLOOR
S.W. WALL of SACELLVM
FLOOR V
FLOOR IV.
TOP OF CELLAR WALL
FILLING (COINS 330-50 A.D.)
FLOOR III.
BLACK MOVLD
ALTAR
FILLING (COINS MOSTLY 254-73 A.D.)
COIN OF JVLIA DOMNA UNDER CEMENT COVERING STEP.
FLOOR II.
FLOOR I.
COINS OF 193-230 A.D. IN FL. II.
HOARD OF CIRCA 290 A.D.
COIN OF ELACABALVS ON FLOOR I.
SCALE OF FEET
5
0
5
10
R.E.M.W. 1922





The CIDOC CRM
Top-level Entities relevant for Integration

E55 Types

refer to / refine

E41 Appellations

refer to / identifie

E39 Actors

E28 Conceptual Objects

E18 Physical Stuff

participate in

affect or / refer to

location

E2 Temporal Entities

E52 Time-Spans

within

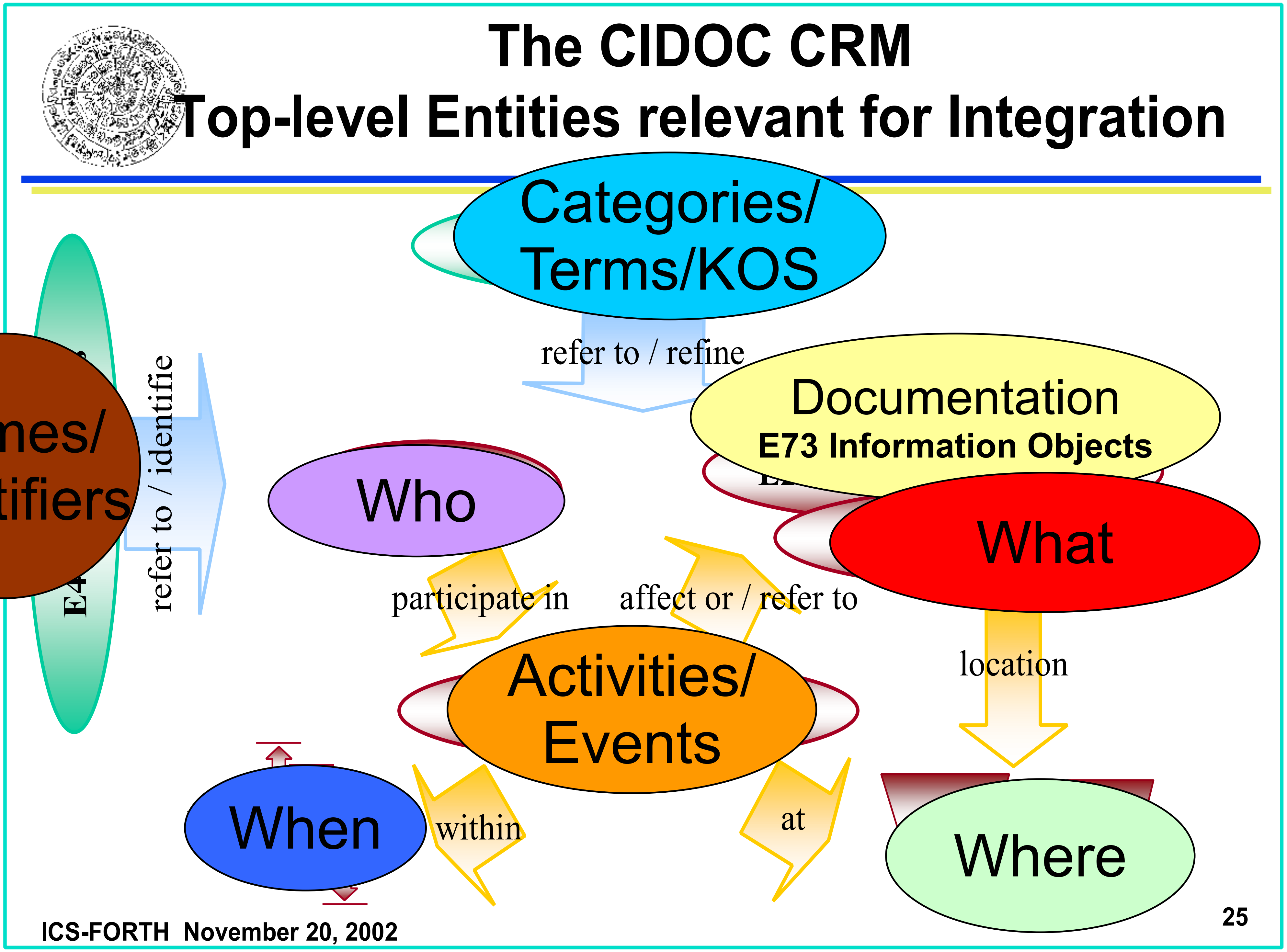
at

E53 Places

ICS-FORTH November 20, 2002

25

With thanks to M. Doerr et al



The CIDOC CRM
Top-level Entities relevant for Integration

Categories/ Terms/KOS

refer to / refine

Documentation

E73 Information Objects

Names/ Identifiers

E4

refer to / identifie

Who

What

participate in

affect or / refer to

Activities/ Events

location

When

within

at

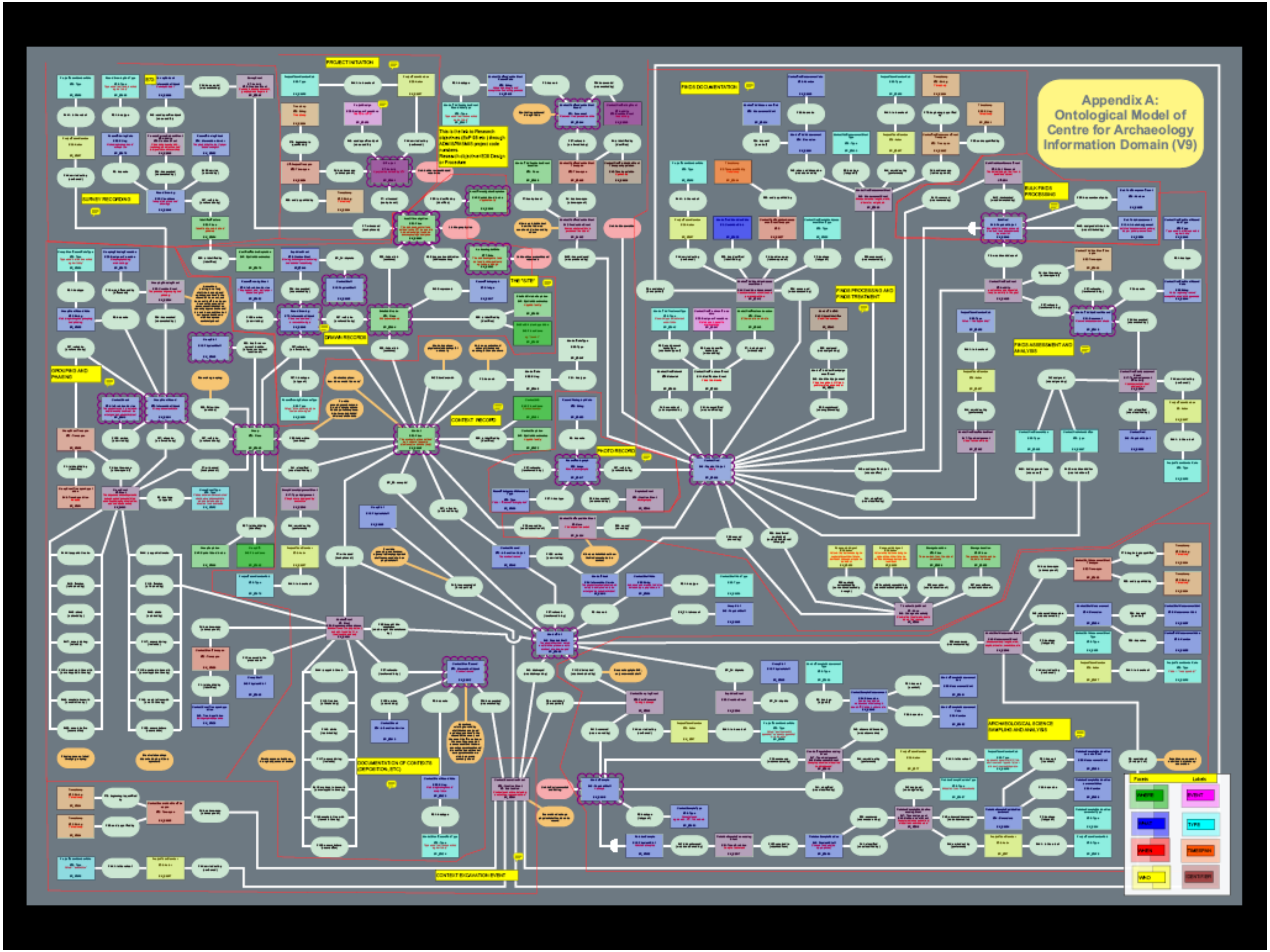
Where

ICS-FORTH November 20, 2002

25

With thanks to M. Doerr et al

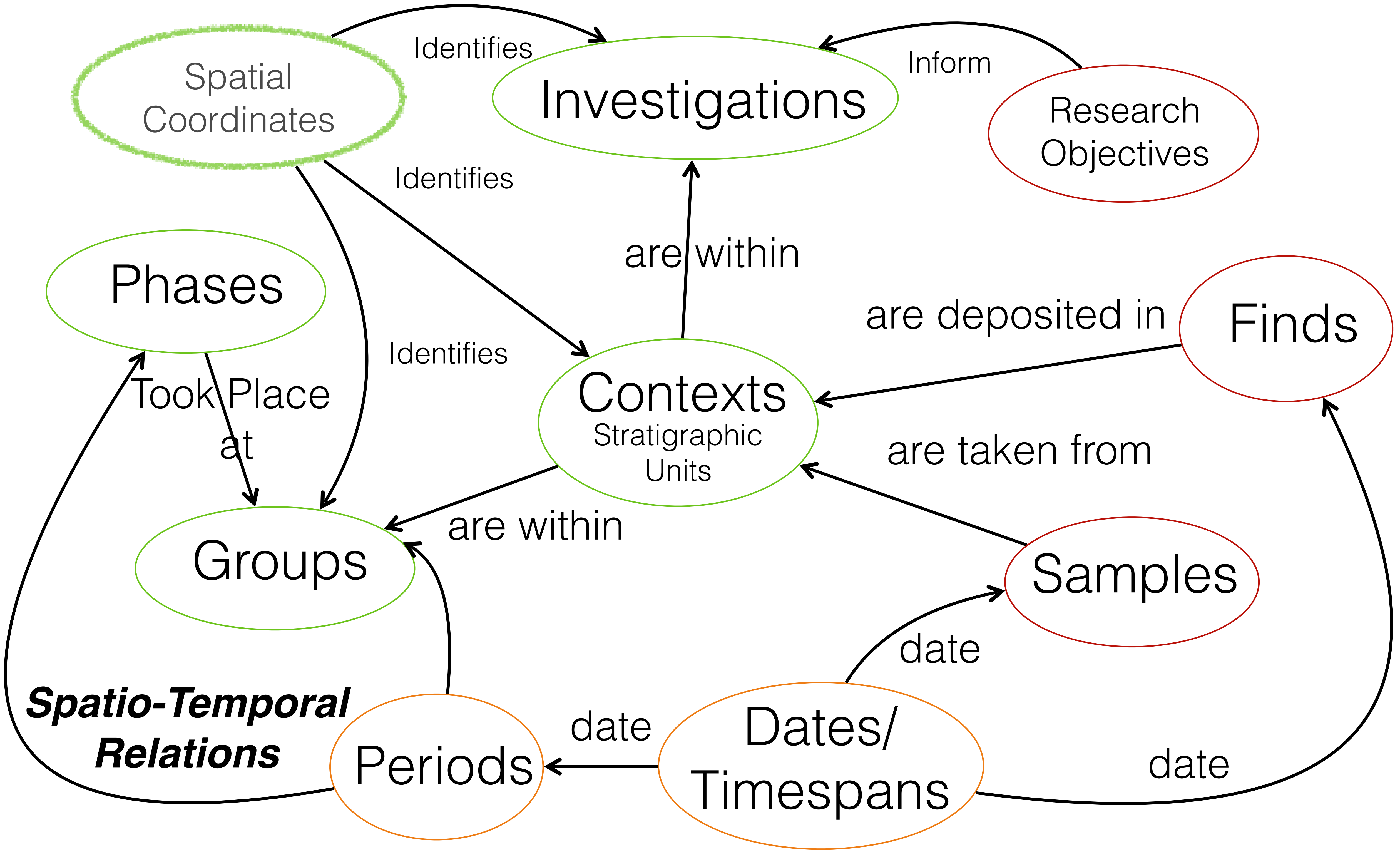
CRM-EH diagram of Archaeological Information Domain


Appendix A: Ontological Model of Centre for Archaeology Information Domain (V9)
PROJECT INITIATION
FINDS DOCUMENTATION
SURVEY RECORDING
BULK FINDS PROCESSING
THE 'SITE'
FINDS PROCESSING AND FINDS TREATMENT
DRAWN RECORDS
GROUPING AND PHASING
FINDS ASSESSMENT AND ANALYSIS
CONTEXT RECORD
PHOTO RECORD
ARCHAEOLOGICAL SCIENCE SAMPLING AND ANALYSIS
DOCUMENTATION OF CONTEXTS (DEPOSITION, ETC)
CONTEXT EXCAVATION EVENT

Archaeological extensions of CIDOC CRM

http:// purl.org/crmeh

Simpler interoperable CRM-EH Model



Archaeological Context/Stratigraphic Unit represented by 2 CRM entities - Spatial E53 - Physical E18

Context as a spatial entity - E53 Place

(e.g. pit cut)

- (E53.Place)
  - (Context_EHE0007)

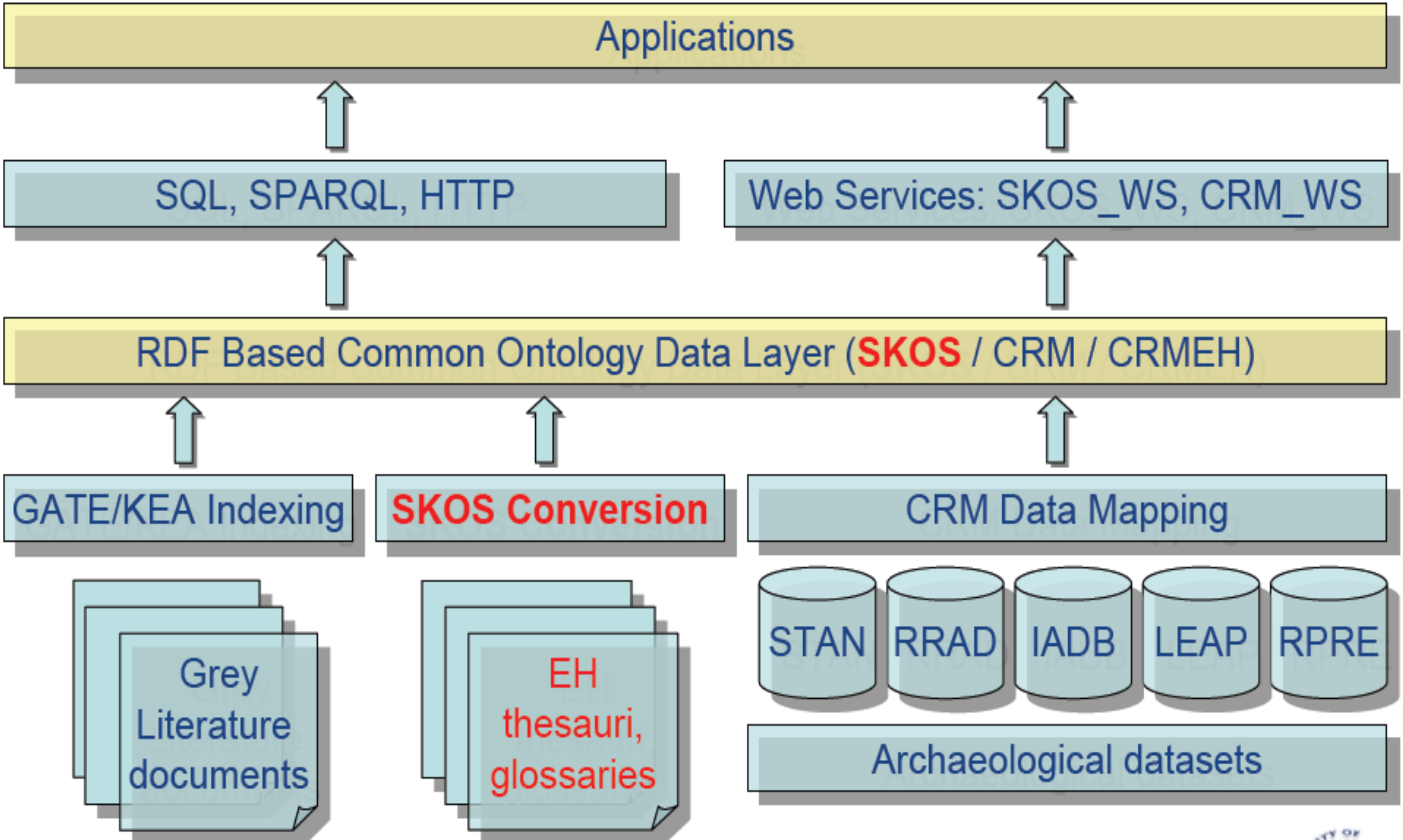
Context material a physical entity

- E18 Physical Thing

(e.g. pit fill)

- (E18.Physical_Thing)
  - (ContextStuff_EHE0008)

STAR - General Architecture



Applications

SQL, SPARQL, HTTP

Web Services: SKOS_WS, CRM_WS

RDF Based Common Ontology Data Layer (**SKOS** / CRM / CRMEH)

GATE/KEA Indexing

SKOS Conversion

CRM Data Mapping

Grey Literature documents

EH thesauri, glossaries

STAN

RRAD

IADB

LEAP

RPRE

Archaeological datasets

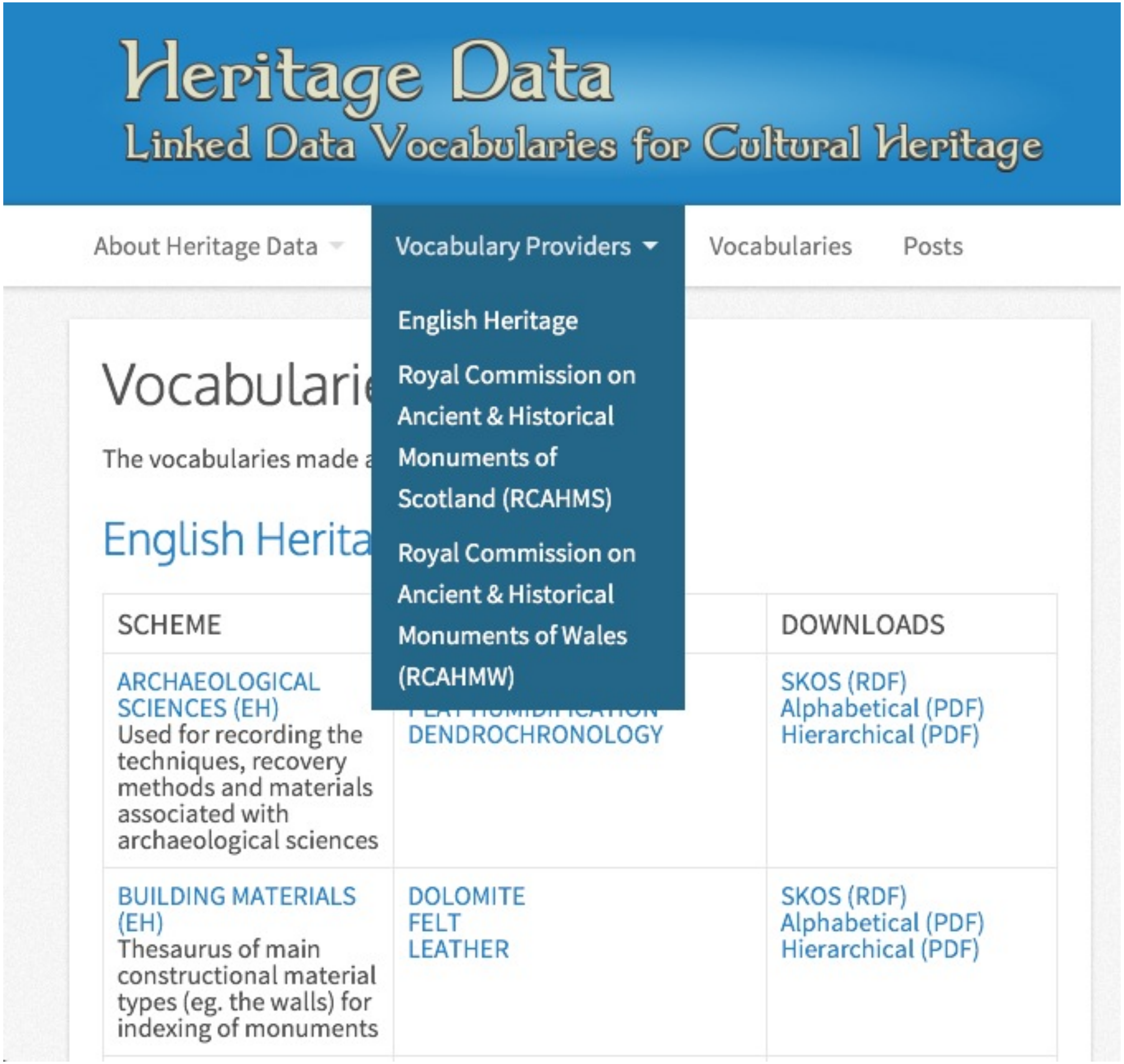
Conceptual Models and Knowledge Resources

- **CRM** [http://cidoc.ics.forth.gr/]
  - CIDOC Conceptual Reference Model
  - International standard ISO 21127:2006
- **CRMEH** [http://purl.org/crmeh]
  - 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.

English Heritage Thesauri

- **Monument types thesaurus**
  - classification of monument type records
- **Evidence thesaurus**
  - archaeological evidence
- **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

LOD Heritage Vocabularies: http://heritagedata.org



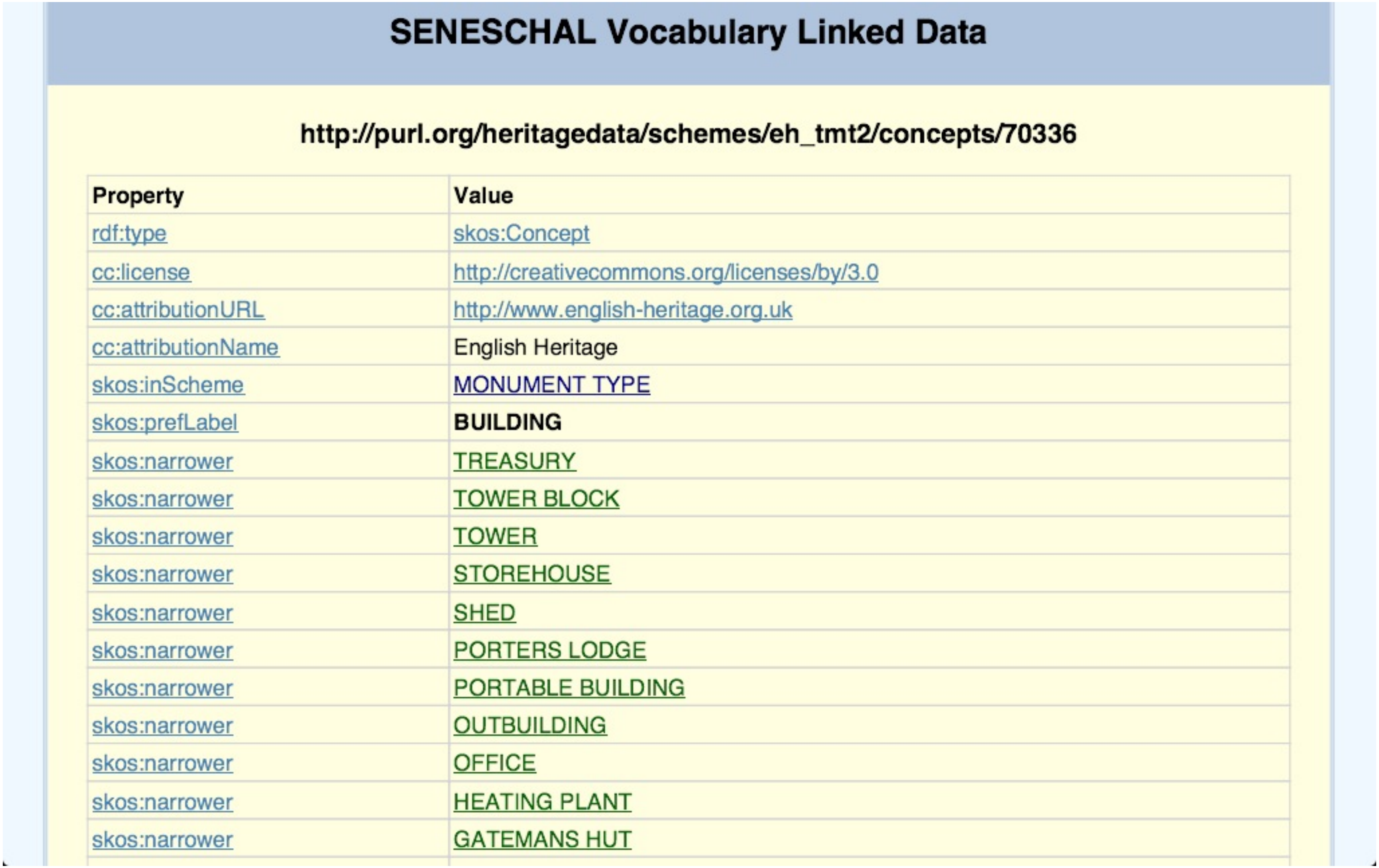
Heritage Data
Linked Data Vocabularies for Cultural Heritage
About Heritage Data
Vocabulary Providers
Vocabularies
Posts
English Heritage
Royal Commission on Ancient & Historical Monuments of Scotland (RCAHMS)
Royal Commission on Ancient & Historical Monuments of Wales (RCAHMW)

Vocabulari

The vocabularies made a

English Herita

SCHEME		DOWNLOADS
ARCHAEOLOGICAL SCIENCES (EH) Used for recording the techniques, recovery methods and materials associated with archaeological sciences	DENDROCHRONOLOGY	SKOS (RDF) Alphabetical (PDF) Hierarchical (PDF)
BUILDING MATERIALS (EH) Thesaurus of main constructional material types (eg. the walls) for indexing of monuments	DOLOMITE FELT LEATHER	SKOS (RDF) Alphabetical (PDF) Hierarchical (PDF)



SENESCHAL Vocabulary Linked Data

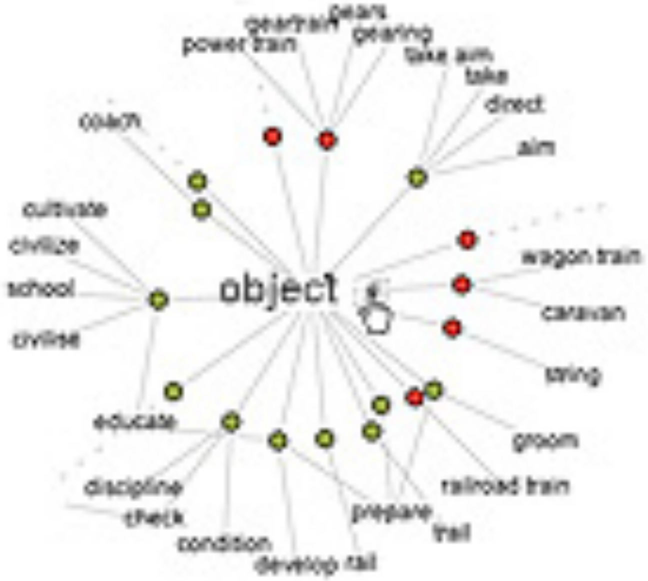
http://purl.org/heritagedata/schemes/eh_tmt2/concepts/70336

Property	Value
rdf:type	skos:Concept
cc:license	http://creativecommons.org/licenses/by/3.0
cc:attributionURL	http://www.english-heritage.org.uk
cc:attributionName	English Heritage
skos:inScheme	MONUMENT TYPE
skos:prefLabel	**BUILDING**
skos:narrower	TREASURY
skos:narrower	TOWER BLOCK
skos:narrower	TOWER
skos:narrower	STOREHOUSE
skos:narrower	SHED
skos:narrower	PORTERS LODGE
skos:narrower	PORTABLE BUILDING
skos:narrower	OUTBUILDING
skos:narrower	OFFICE
skos:narrower	HEATING PLANT
skos:narrower	GATEMANS HUT

Ontology Based Information Extraction

- Ontologies; a mediation language between concepts and their worded representations

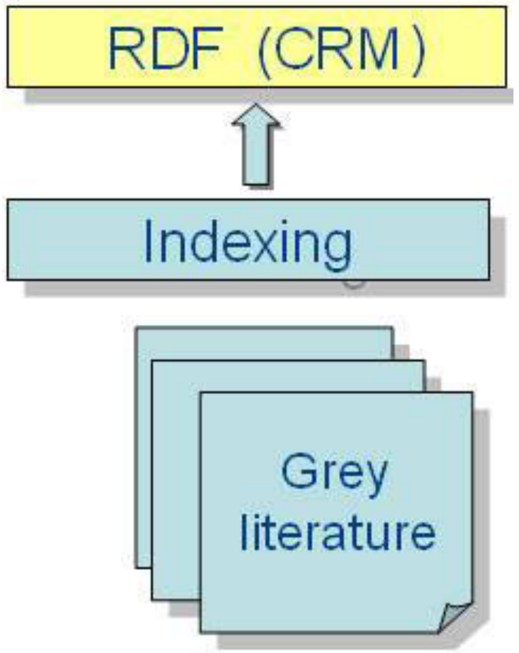
- Advance Information Retrieval
  - Beyond the limitations of (key)words to the level of concepts and semantic relationships
- Aid Information Retrieval
  - To make inferences from diverse data sources
- Information Extraction (IE)
  - A specific text analysis task aimed to extract specific information snippets from documents
  - Ontologies to drive/inform IE
  - To describe the conceptual arrangements of semantic annotations.



Excavating Grey Literature Documents

- **Grey Literature**; *source materials that can not be found through the conventional means of publication*

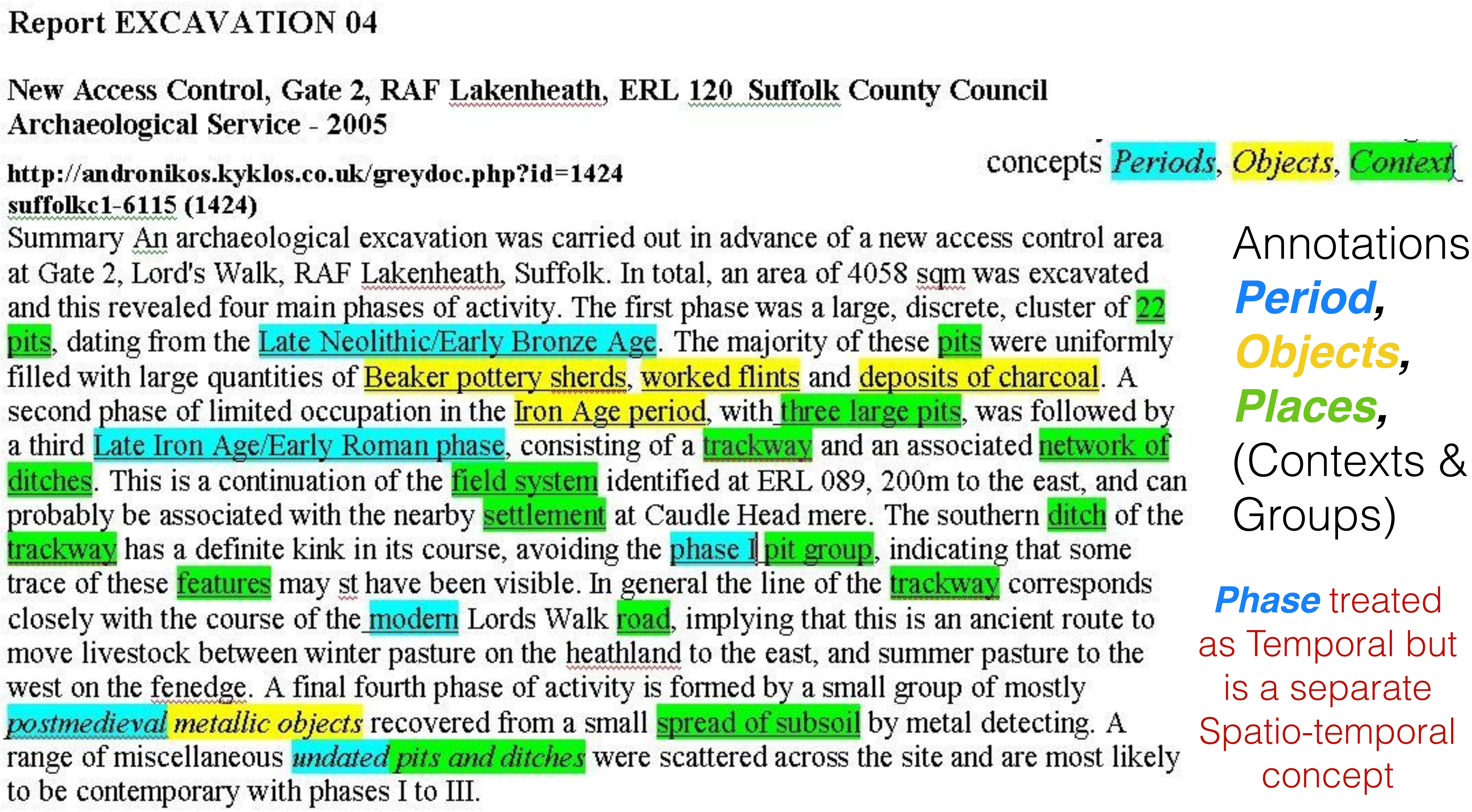
- Online AccesS to the Index of archaeological investgationS (OASIS) http://www.oasis.ac.uk
- Library of unpublished fieldwork reports on ADS now with DOIs
- Other publication reports e.g. Raunds
- Internet Archaeology LEAP article - Silchester
- Semantic Indexing
- Interoperable technologies W3C standards
- XML, RDF representation



Example of the Annotation Methodology

Rule-based method

Focused on Evaluations & Excavations Summaries from **OASIS**



Report EXCAVATION 04

New Access Control, Gate 2, RAF Lakenheath, ERL 120 Suffolk County Council Archaeological Service - 2005

concepts *Periods*, *Objects*, *Context*

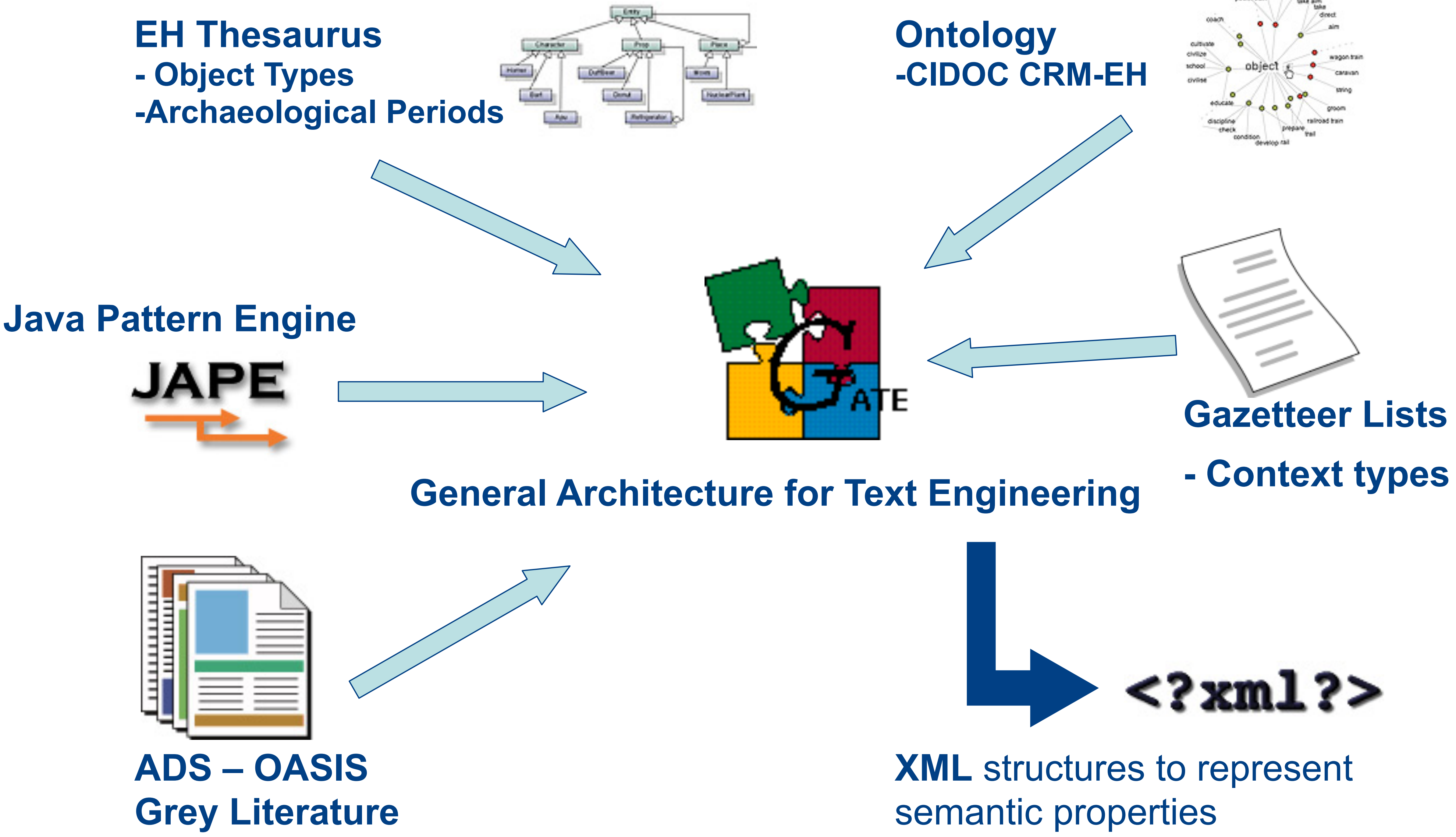
http://andronikos.kyklos.co.uk/greydoc.php?id=1424
suffolkc1-6115 (1424)

Summary An archaeological excavation was carried out in advance of a new access control area at Gate 2, Lord's Walk, RAF Lakenheath, Suffolk. In total, an area of 4058 sqm was excavated and this revealed four main phases of activity. The first phase was a large, discrete, cluster of 22 pits, dating from the Late Neolithic/Early Bronze Age. The majority of these pits were uniformly filled with large quantities of Beaker pottery sherds, worked flints and deposits of charcoal. A second phase of limited occupation in the Iron Age period, with three large pits, was followed by a third Late Iron Age/Early Roman phase, consisting of a trackway and an associated network of ditches. This is a continuation of the field system identified at ERL 089, 200m to the east, and can probably be associated with the nearby settlement at Caudle Head mere. The southern ditch of the trackway has a definite kink in its course, avoiding the phase I pit group, indicating that some trace of these features may st have been visible. In general the line of the trackway corresponds closely with the course of the modern Lords Walk road, implying that this is an ancient route to move livestock between winter pasture on the heathland to the east, and summer pasture to the west on the fenedge. A final fourth phase of activity is formed by a small group of mostly *postmedieval metallic objects* recovered from a small spread of subsoil by metal detecting. A range of miscellaneous *undated pits and ditches* were scattered across the site and are most likely to be contemporary with phases I to III.

Annotations
Period,
Objects,
Places,
(Contexts & Groups)

Phase treated as Temporal but is a separate Spatio-temporal concept

Information Extraction Framework



EH Thesaurus
- Object Types
-Archaeological Periods

Ontology
-CIDOC CRM-EH

Java Pattern Engine

Gazetteer Lists

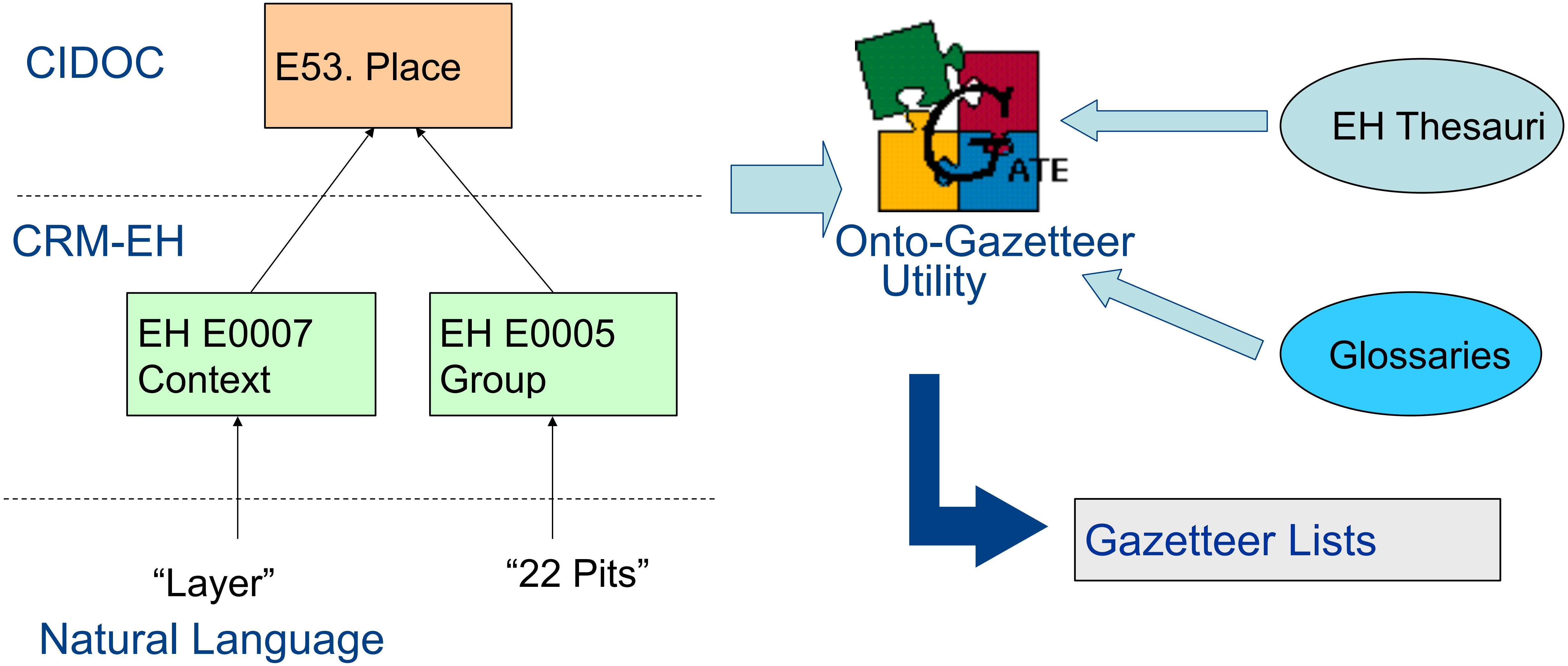
- Context types

General Architecture for Text Engineering

ADS – OASIS
Grey Literature

XML structures to represent semantic properties

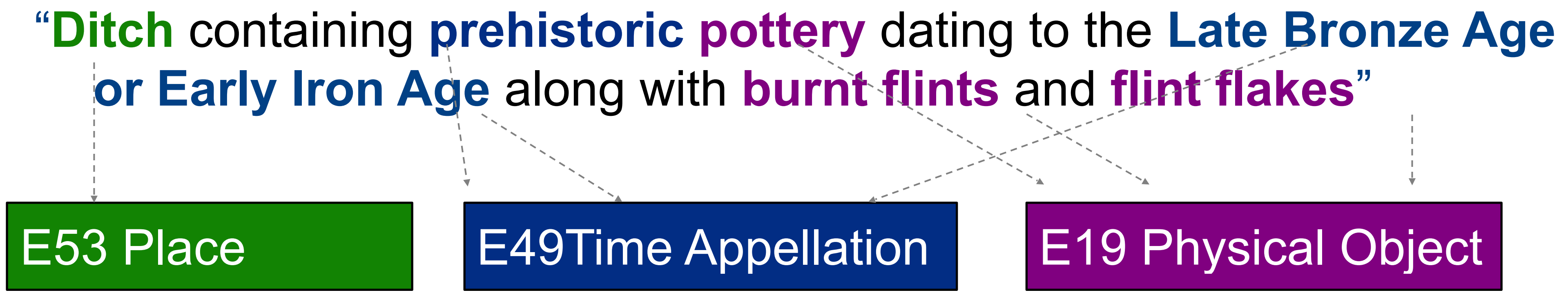
GATE Mapping of Knowledge Resources



Reference to SKOS mapped to the MinorType attribute of list entries

JAPE Pattern Matching Rules

Natural Language – Gazetteer Look-up



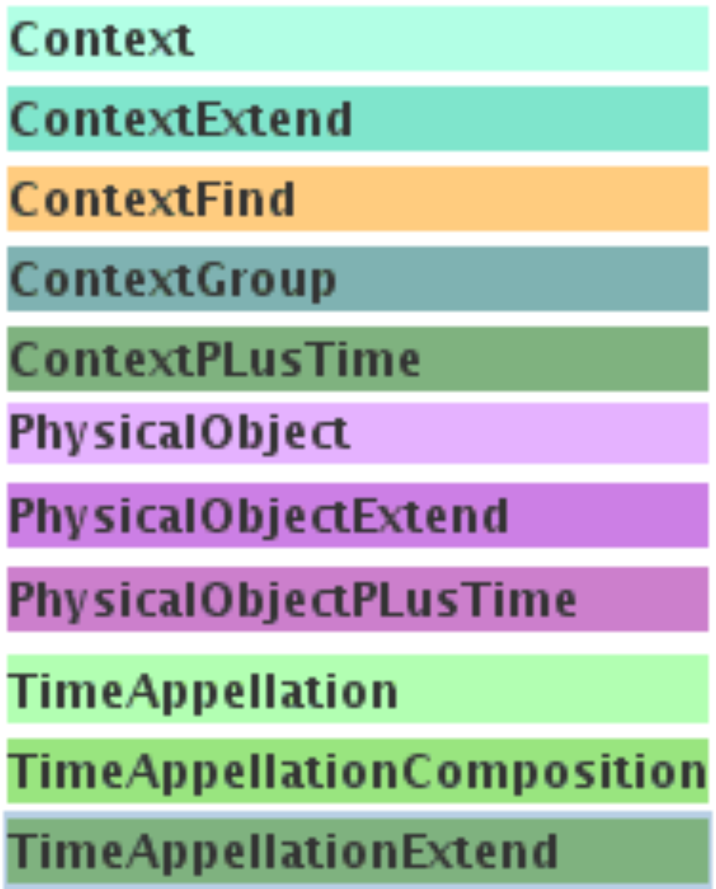
"**Ditch** containing **prehistoric pottery** dating to the **Late Bronze Age or Early Iron Age** along with **burnt flints** and **flint flakes**"

Pattern Matching Rules expanded beyond simple gazetteer look-up

<entity><*same*-entity>	E49 E49	***"Late Bronze Age or Early Iron Age"***
<entity><*other*-entity>	E49 E19	**"prehistoric pottery"**
<entity><verb>(<entity>/ <structure>)	E53	**"Ditch containing prehistoric pottery"**

Annotation Types exposed in XML

Annotation Types



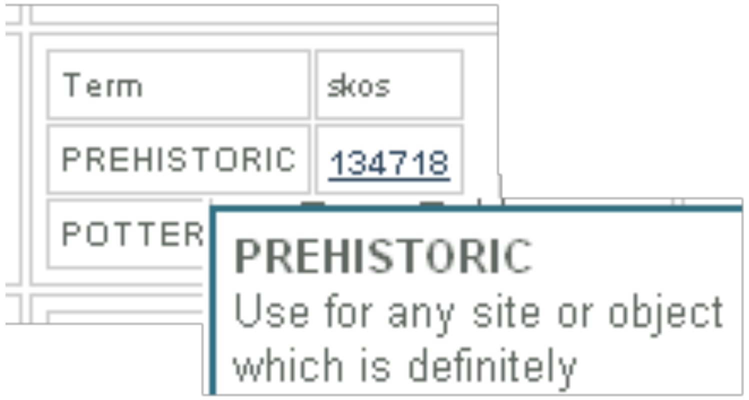
- Context
- ContextExtend
- ContextFind
- ContextGroup
- ContextPLusTime
- PhysicalObject
- PhysicalObjectExtend
- PhysicalObjectPLusTime
- TimeAppellation
- TimeAppellationComposition
- TimeAppellationExtend

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	

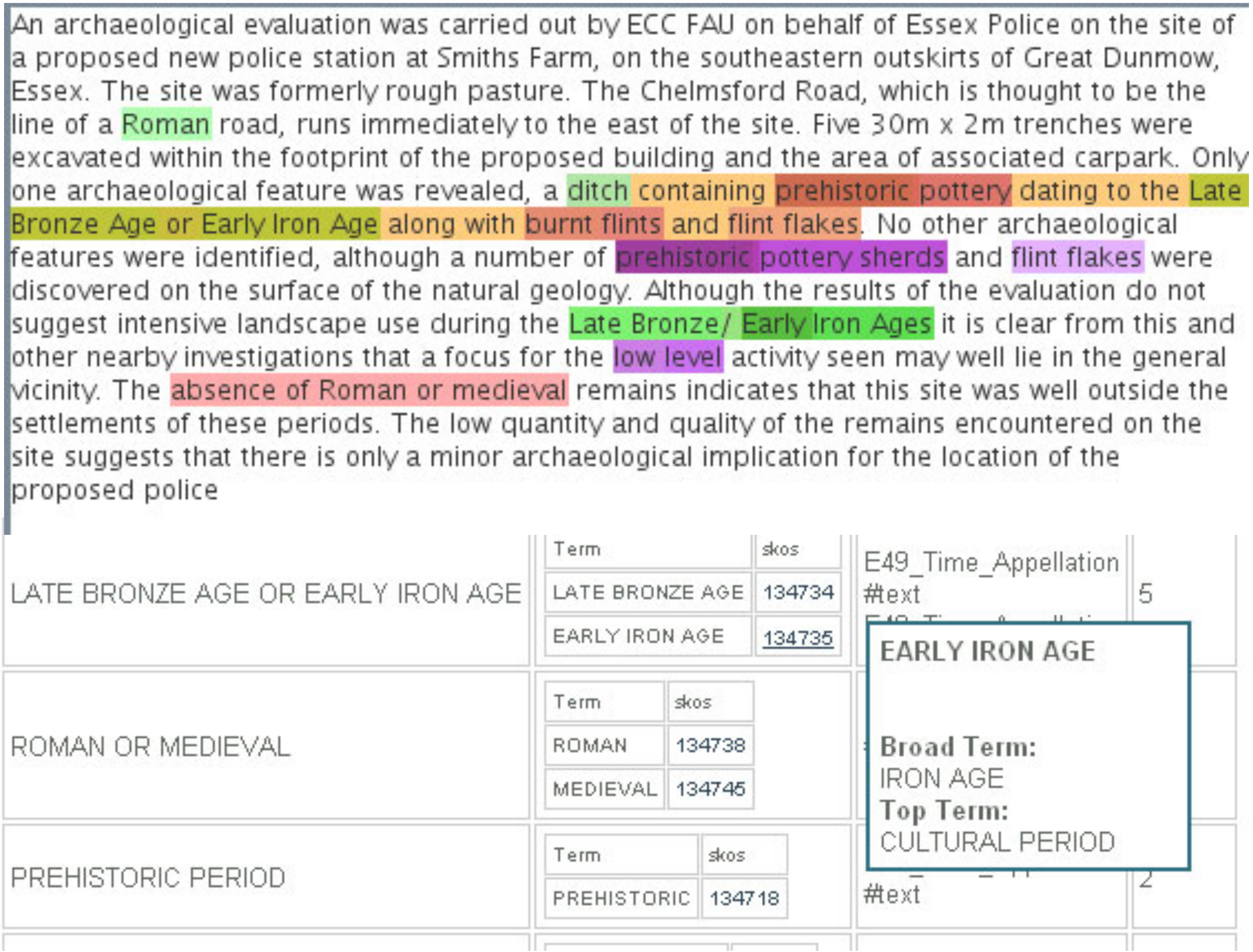
PREHISTORIC
Use for any site or object which is definitely

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/2008-04-01/
  CIDOC_v4.2_extensions_eh_.rdf"}
```

Andronikos Web Portal Interface



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	Term LATE BRONZE AGE EARLY IRON AGE	skos 134734 134735	E49_Time_Appellation #text	5
ROMAN OR MEDIEVAL	Term ROMAN MEDIEVAL	skos 134738 134745		
PREHISTORIC PERIOD	Term PREHISTORIC	skos 134718	#text	2

EARLY IRON AGE

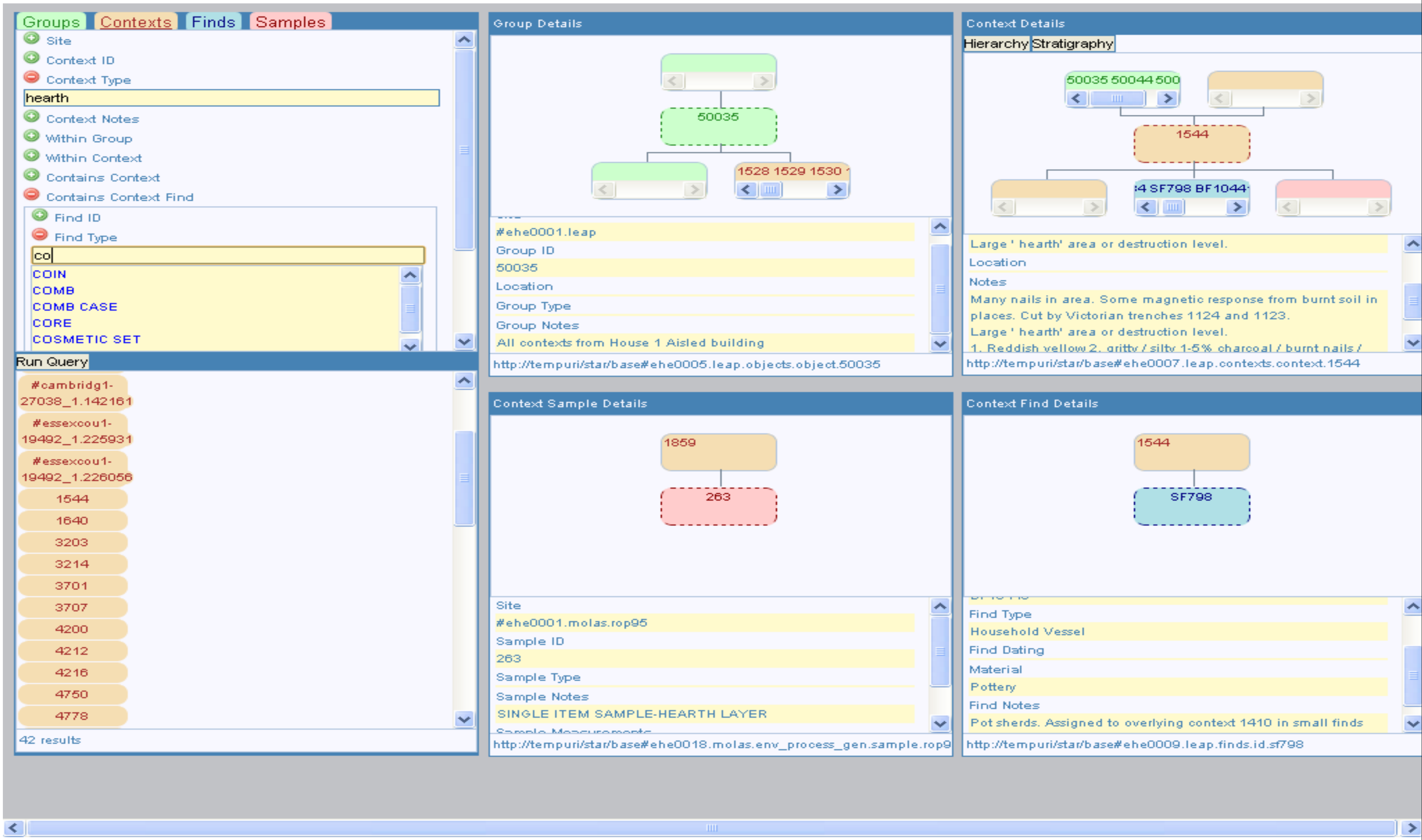
Broad Term:
IRON AGE
Top Term:
CULTURAL PERIOD

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

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')
- Distinguishing Materials from Objects hardest e.g. Pottery
- The potential of the method to extract complex phrases associated to two or more ontological entities
- Further work
  - Incorporation of additional Ontological Entities (Phases, Samples)
  - Gazetteer enhancement e.g. more terms for Places
  - Pattern matching rules expansion
  - Formal evaluation of the Extraction method and overall retrieval performance

STAR interface for cross-search of integrated data



Groups | Contexts | Finds | Samples

Site
Context ID
Context Type
hearth
Context Notes
Within Group
Within Context
Contains Context
Contains Context Find
Find ID
Find Type
co
COIN
COMB
COMB CASE
CORE
COSMETIC SET

Run Query

#cambridg1-27038_1.142161
#essexcou1-19492_1.225931
#essexcou1-19492_1.226056
1544
1640
3203
3214
3701
3707
4200
4212
4216
4750
4778

42 results

Group Details

50035
1528 1529 1530

#ehe0001.leap
Group ID
50035
Location
Group Type
Group Notes
All contexts from House 1 Aisled building

http://tempuri/star/base#ehe0005.leap.objects.object.50035

Context Details

Hierarchy | Stratigraphy

50035 50044 500
1544
4 SF798 BF1044

Large ' hearth' area or destruction level.
Location
Notes
Many nails in area. Some magnetic response from burnt soil in places. Cut by Victorian trenches 1124 and 1123.
Large ' hearth' area or destruction level.
1. Reddish yellow 2. gritty / silty 1-5% charcoal / burnt nails /

http://tempuri/star/base#ehe0007.leap.contexts.context.1544

Context Sample Details

1859
263

Site
#ehe0001.molas.rop95
Sample ID
263
Sample Type
Sample Notes
SINGLE ITEM SAMPLE-HEARTH LAYER
Sample Measurements

http://tempuri/star/base#ehe0018.molas.env_process_gen.sample.rop9

Context Find Details

1544
SF798

Find Type
Household Vessel
Find Dating
Material
Pottery
Find Notes
Pot sherds. Assigned to overlying context 1410 in small finds

http://tempuri/star/base#ehe0009.leap.finds.id.sf798

Internet Archaeology Vol 30 http://intarch.ac.uk/journal/issue30/tudhope_index.html

Bibliography

Reference papers

Andreas Vlachidis 2012. Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature. PhD Thesis, University of South Wales (USW) http://hypermedia.research.southwales.ac.uk/media/files/documents/2013-07-11/Andreas-Vlachidis_Thesis_print_ready.pdf

Vlachidis A, Tudhope D. 2012. A pilot investigation of information extraction in the semantic annotation of archaeological reports. International Journal of Metadata, Semantics and Ontologies, 7(3), 222-235. Inderscience.

Vlachidis A, Binding C, May K, Tudhope D. 2011. Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature. Proceedings CLA'11 Computational Linguistic Applications, Warsaw

Vlachidis A, Binding C, May K, Tudhope D. 2010 . Excavating Grey Literature: a case study on the rich indexing of archaeological documents via Natural Language Processing techniques and Knowledge Based resources. ASLIB Proceedings journal, 62 (4&5), 466 – 475.

Vlachidis A, Binding C, May K, Tudhope D. 2009. Semantic Annotations in the Archaeological Domain. Proceedings First biennial Conference of the British Chapter of the International Society for Knowledge Organization (ISKO UK), London

STAR
Semantic Technologies for Archaeological Resources

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

keith.may@english-heritage.org.uk
andreas.vlachidis@southwales.ac.uk
ceri.binding@southwales.ac.uk
douglas.tudhope@southwales.ac.uk

Arts & Humanities
Research Council

University of
South Wales
Prifysgol
De Cymru

ENGLISH HERITAGE