Reflections on Excavating Archaeological Grey Literature :

and on the challenges in information extraction

Overview

- Andreas Vlachidis
  - Lecturer in Information Science
  - UCL – Department of Information Studies
- Background and Motivation
- Information Extraction for Semantic Indexing
- Natural Language Processing (NLP) Pipelines
- Challenges and Solutions

All began in late 2007

- 2 Research Projects
- 8 Information Extraction Pipelines
- 3 Languages

Excavating Grey Literature: A case study on rich indexing of archaeological documents by the use of Natural Language Processing Techniques and Knowledge Based resources.

Andreas Vlachidis, Ceri Binding, Keith May, Douglas Tudhope

The paper describes the use of Information ... Language Processing (NLP) ... indexing of diverse ... unpublished online ... Literature', ... to integrate the data from various archaeolo... and their associated activities, and seeks ... potential of semantic technologies and na... processing techniques, for enabling ... semantically defined queries over archae... resources. [1][2]

... project has initially chosen ... base along with ... sampli...

Semantic **T**echnologies for **A**rchaeological **R**esources

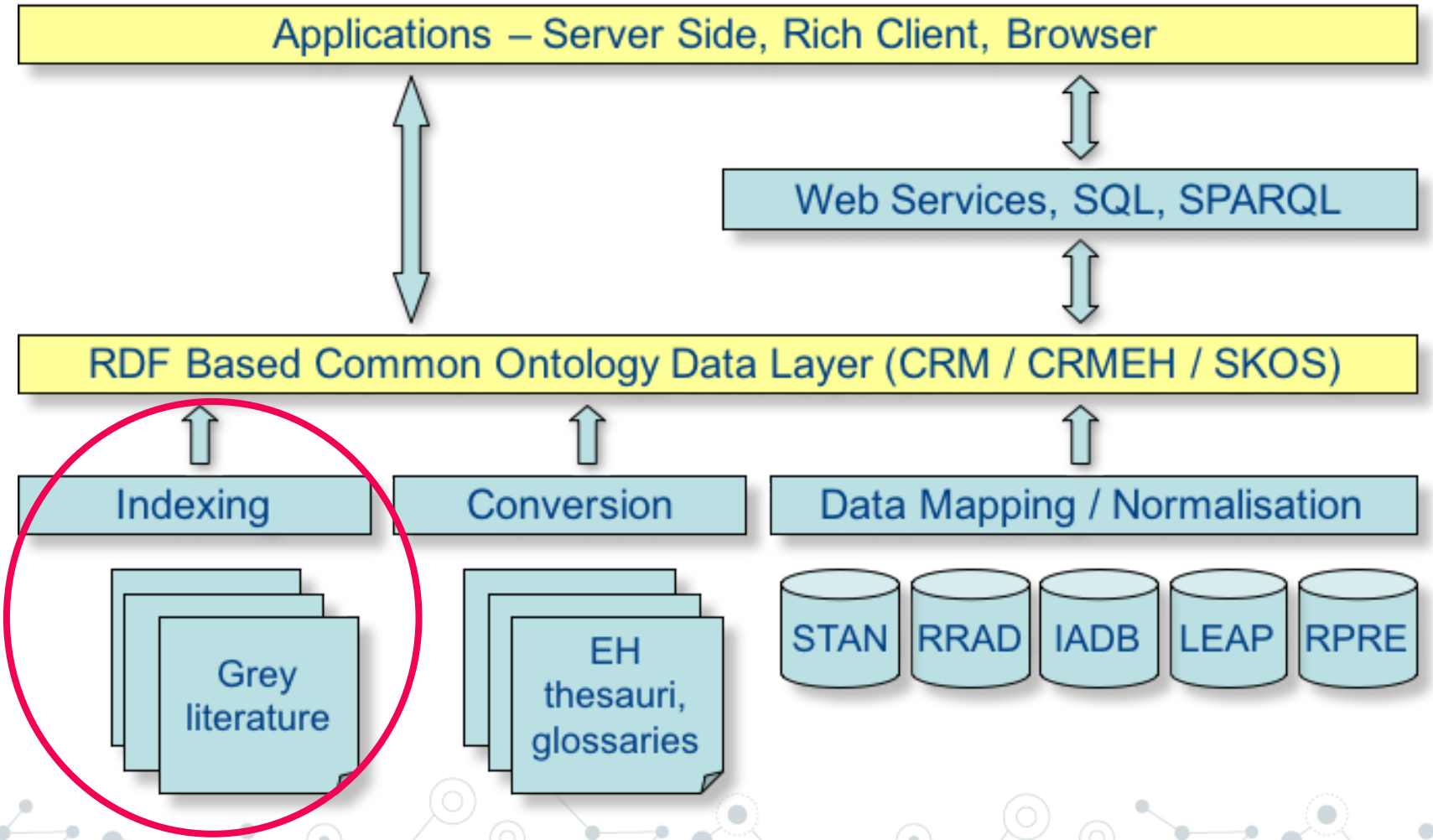
Develop new methods for linking digital archive databases, vocabularies and the associated grey literature, exploiting the potential of a high level, core ontology and natural language processing techniques

AHRC Funded Project; 2007-2010, £222,139
http://hypermedia.research.southwales.ac.uk/kos/star/

Semantic Indexing

Interoperable metadata generation of semantic attributes for supporting information retrieval and cross searching of archaeological Archaeology Grey Literature (fieldwork reports) with respect to a given ontology and terminology (CIDOC-CRM, SKOSified Vocabularies)

STAR Architecture



Archaeological Grey Literature

Reflect different stages of a fieldwork project worth recording and disseminating information about, such as watching briefs, excavation, evaluation, survey reports and related artefact and ecofact analysis

Unlocking the Potential of Grey Literature

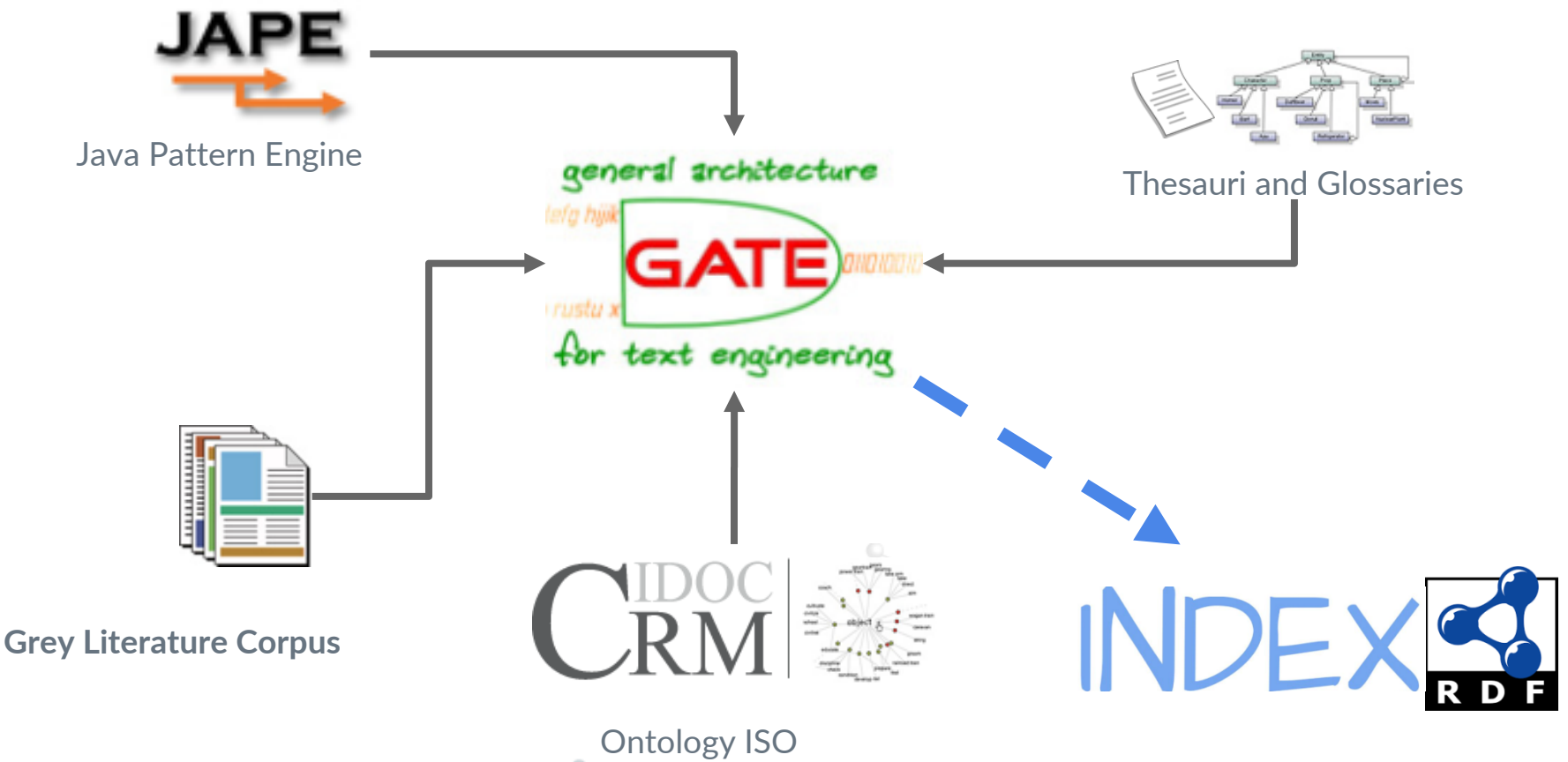
WWW offers the opportunity to improve archaeological practice, not only by enabling access to information but also by changing how information is structured and the way research is conducted (Falkingham, 2005)

Online AccesS to the Index of archaeological investigationS (OASIS)
http://oasis.ac.uk

Information Extraction for Semantic Indexing

A Rule–Based & KOS Driven Method

Semantic Indexing, NLP Framework



NLP for Semantic Indexing

- Rule-Based Information Extraction (IE)
  - IE a text analysis task
    - Extraction of targeted information from context
    - Turning unstructured text to structured data
  - Rules: Java Annotation Pattern Engine
    - Input from generic NLP (eg Tokenizer, POS)
    - Input from controlled vocabulary
    - Linguistic patterns
- Why not Machine Learning?
  - Lack of training corpus
  - Availability of domain vocabulary

NLP for Semantic Indexing

- Knowledge Organization Systems Driven
  - Domain Thesauri and Glossaries
    - Terminology and Coverage
    - SKOSified Resources
    - Well-defined Semantics
  - Conceptual Framework - Ontology
    - CIDOC-CRM (ISO standard ontology in CH)
    - Upper-level (Abstract Concepts)
    - Properties and Relationships
- Interoperable Output
  - **Conceptual** Level: Ontological Coherence (CIDOC-CRM, SKOS)
  - **Technology** Level: Standard Semantic Web (RDF, SPARQL)



NLP Pipelines

- STAR OPTIMA (EN)
- Ariadne NER (EN, NL, SE)
- Ariadne Dendrochronology (EN, NL, SE)
- Ariadne Numismatic (EN)

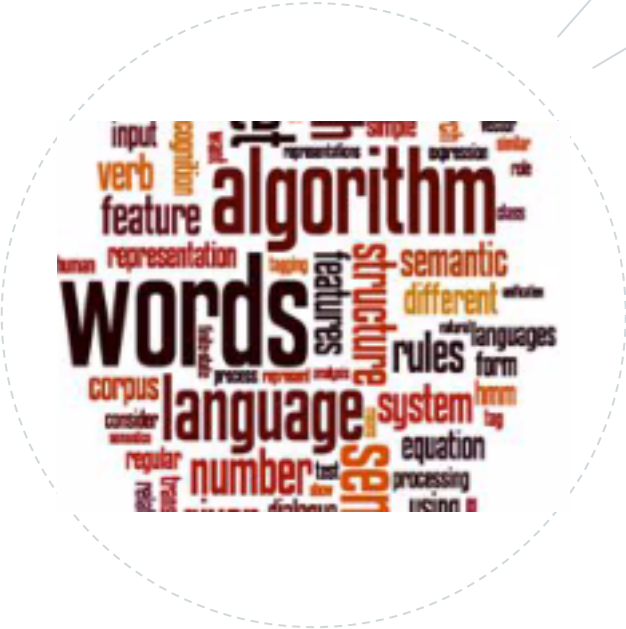
NLP – Pipeline Focus

Named Entity Recognition

Classify named entity mentions in unstructured text into predefined categories (NER)

Relation Extraction

Identifying the links between Named Entities and deciding which ones are meaningful (RE)

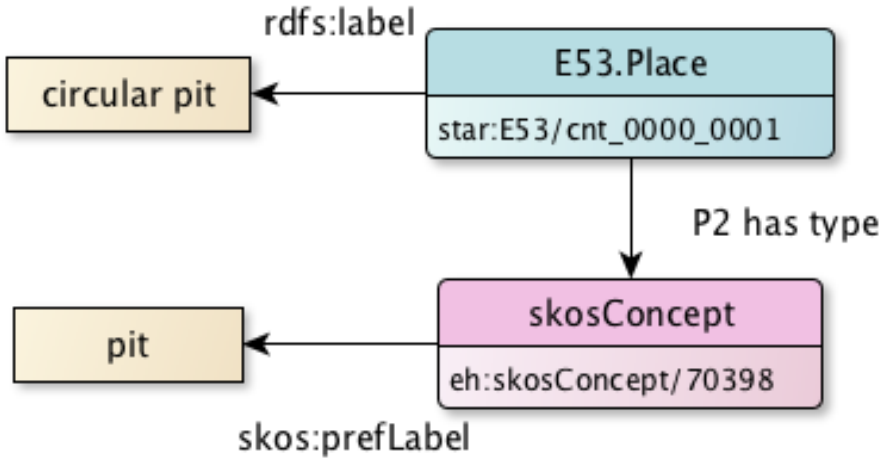


OPTIMA Pipeline

- Focus on NER and ER
- Contributed to the STAR Project
- English

E19 Physical Object	e.g. arrowhead, small stones, pottery, bowl, coin, flint flake
E49 Time Appellation	e.g. roman, medieval, 15th century, early to mid 18th century
E53 Place	e.g. deposit, pit, foundation, cut, wall
E57 Material	e.g. pottery, glass, clay, bowl, copper alloy, charcoal

OPTIMA Named Entity Recognition



E53.Place

TERM	SKOS(1)	SKOS(2)	Count
fill	ehg003.142		76
pit	ehg003.55	70398	45
ditch	ehg003.20	70351	42
pits	ehg003.55	70398	36

0009, was a mid brown/grey sand with a
/charcoal layer, 0014, contained some po
o contained some pottery. It was 50% exc
he site. It measured 1.8m wide and 0.6m
curved base. Its fill, 0031/0089, was a m
% excavated. 0074 was a large, circular p
ded layers of pale brown and silver/grey
yer of dark grey, charcoal-rich sand lying
h contained pottery sherds. Figure 7. Pha
ugh the majority of the ditches on the site
of a Late Iron Age/Early Roman field sy
ERL 089 and ERL 112). Many of the ditc
hes, 0057 and 0082, running parallel to ea
ing between the two. Ditch 0002, althou
the north-south ditches on the site, imply
man ditch system with Iron Age pottery
material and with 0209, 0211 and 0064 fo
t to that seen in phase IIIa. As 0064 cuts
ack of material recovered from these ditc
was possible, principally of ditch interse
ints. Only about 5% of the length of thes

OPTIMA Relation Extraction

EHE1001.ContextEvent

Relation between archaeological context (place) and time

pits which are predominantly Iron Age	
EHE0007.Context	EHE0026.TimeSpanAppellation
pits [ehg003.55]	Iron Age [134722]

EHE1002.ContextFindProductionEvent

Relation between find (object) and time

pottery from the Bronze Age, Iron Age and Medieval periods	
EHE0009.ContextFind	EHE0039.TimeSpanAppellation
pottery [ehg027.2]	Bronze Age [134723] Iron Age [134722] Medieval periods [134745]

OPTIMA Relation Extraction

EHE1004.ContextFindDepositionEvent

Relation between find (object) and archaeological context (place)

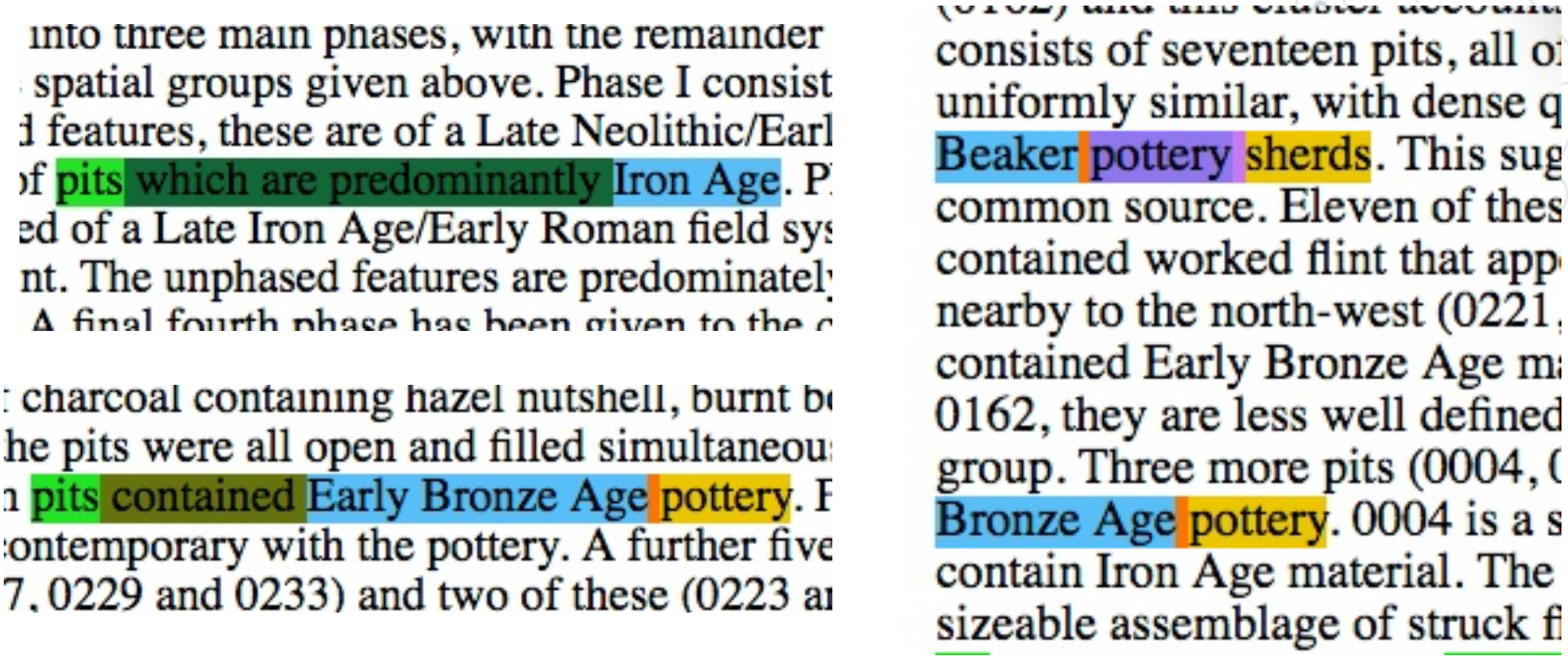
pits contained Early Bronze Age pottery	
EHE0007.Context	EHE0009.ContextFind
pits [ehg003.55]	pottery [ehg027.2]

P45.consists_of

Relation between find (object) and material

Copper alloy coin	
EHE0009.ContextFind	EHE0030.ContextFindMaterial
coin [95423]	Copper alloy [ehg019.6]

OPTIMA Relation Extraction



into three main phases, with the remainder
spatial groups given above. Phase I consist
l features, these are of a Late Neolithic/Earl
of pits which are predominantly Iron Age. P
ed of a Late Iron Age/Early Roman field sys
nt. The unphased features are predominatel
A final fourth phase has been given to the

charcoal containing hazel nutshell, burnt b
he pits were all open and filled simultaneou
n pits contained Early Bronze Age pottery. F
ontemporary with the pottery. A further five
7, 0229 and 0233) and two of these (0223 a

consists of seventeen pits, all o
uniformly similar, with dense q
Beaker pottery sherds. This sug
common source. Eleven of thes
contained worked flint that app
nearby to the north-west (0221
contained Early Bronze Age ma
0162, they are less well defined
group. Three more pits (0004,
Bronze Age pottery. 0004 is a s
contain Iron Age material. The
sizeable assemblage of struck f

Crooks Farm mansion. They were overlain by deposits of demolition material which almost certainly represent destruction of the mansion buildings. A possible pond was also discovered which appeared to be contemporary with the buildings. Remains of further buildings dating to the 19th century were found cutting into the earlier postmedieval structures. Medieval pottery remains were located within one of the evaluation trenches, but the nature of these were to some extent ambiguous due in part to the limited area

ARIADNE – NER Pipelines

Languages

- English, Dutch, Swedish

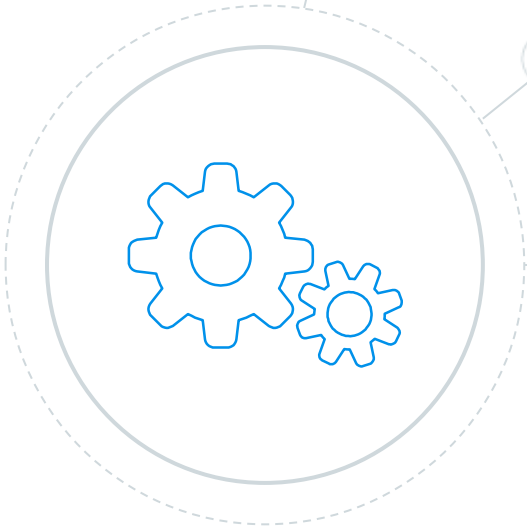
Focus

- Context (archaeological)
- Place (spatial)
- Object (find)
- Monument (building)
- Material
- Time

KOS

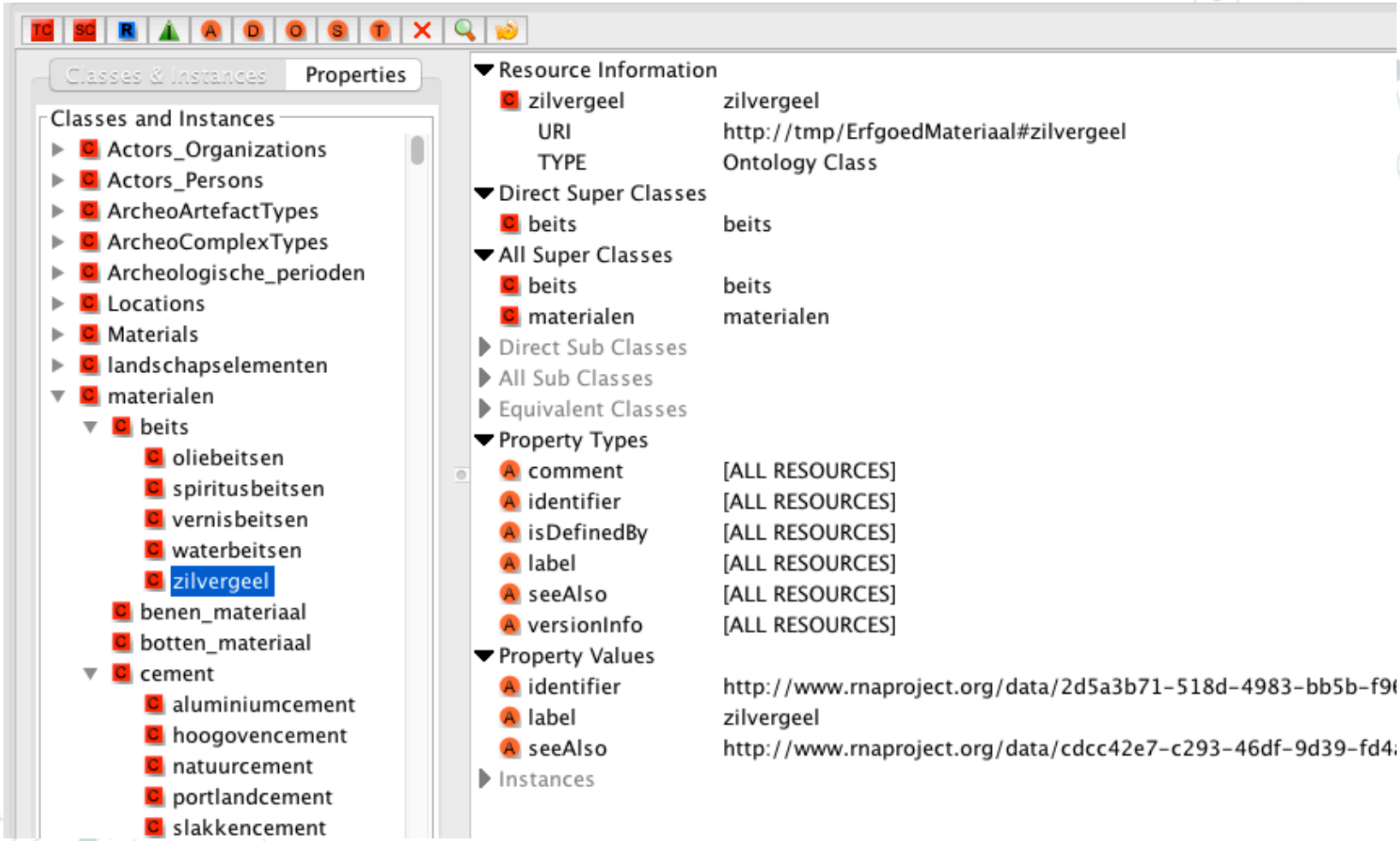
- Historic England (Heritage Data Thesauri)
- Cultural Heritage Agency of the Netherland (Erfgoed Thesaurus)
- Swedish National Data Service (Glossaries)

GATEfication of KOS Resources

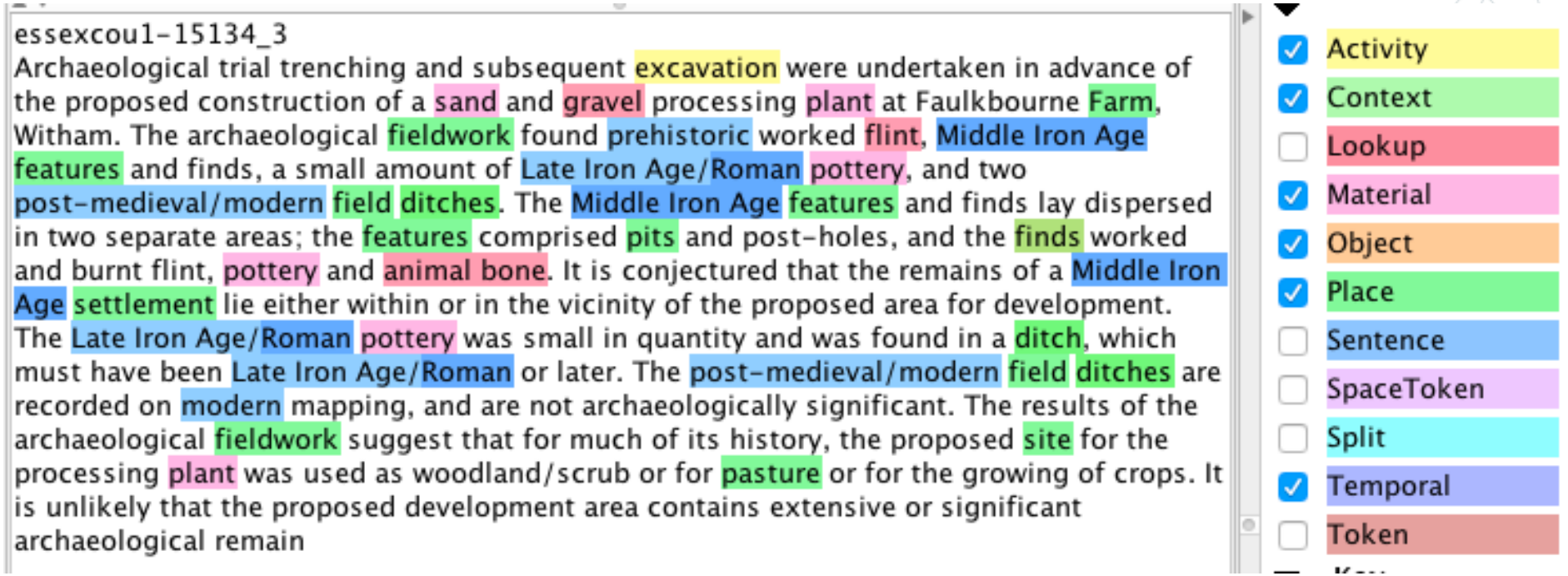


Design and transformation options for translating the original resources (SKOSified RDF files) into GATE friendly OWL--Lite structures capable of supporting the NER matching mechanism for the Information Extraction pipeline

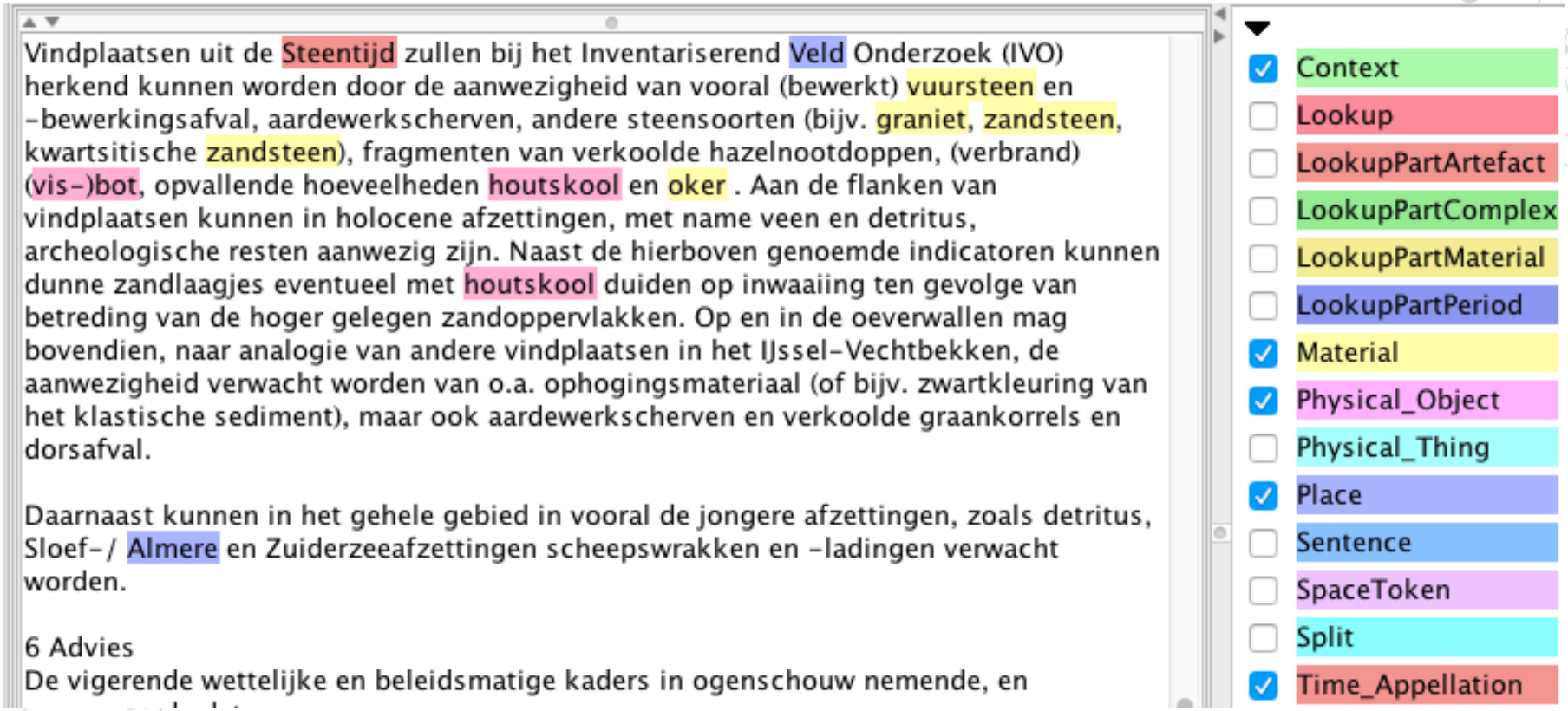
GATEfication – View of Ontology



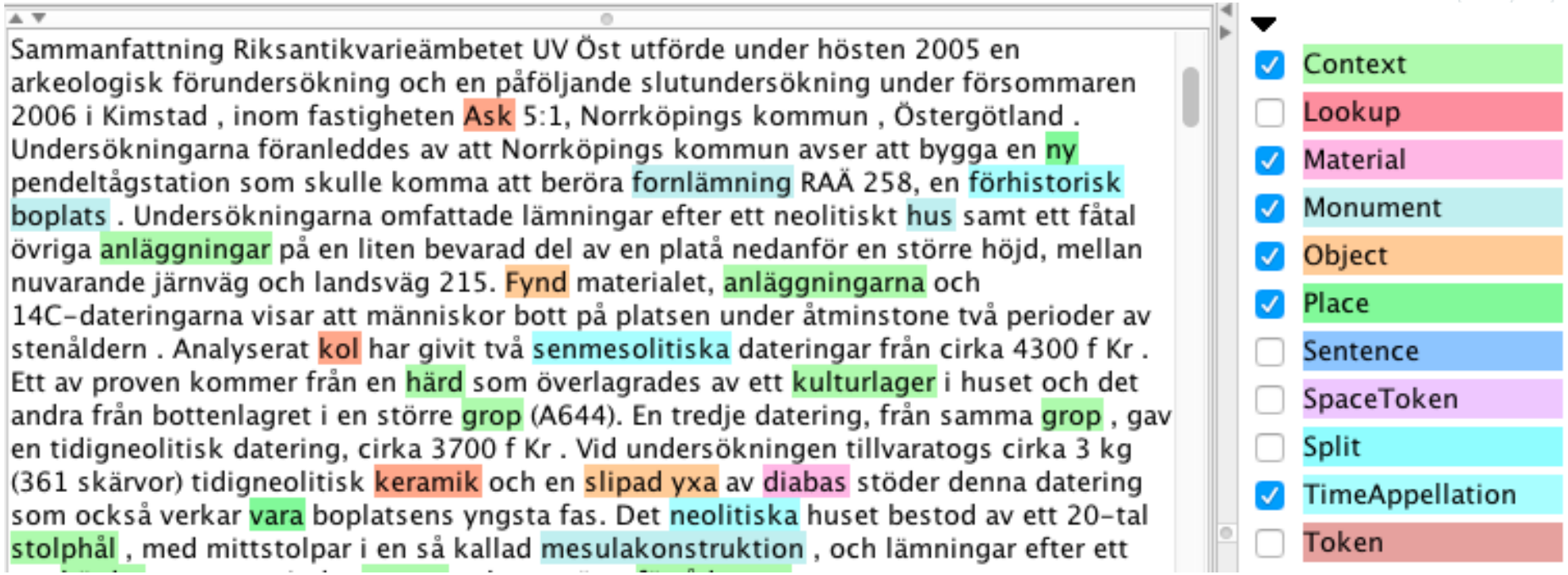
English– NER


essexcou1-15134_3
Archaeological trial trenching and subsequent excavation were undertaken in advance of the proposed construction of a sand and gravel processing plant at Faulkbourne Farm, Witham. The archaeological fieldwork found prehistoric worked flint, Middle Iron Age features and finds, a small amount of Late Iron Age/Roman pottery, and two post-medieval/modern field ditches. The Middle Iron Age features and finds lay dispersed in two separate areas; the features comprised pits and post-holes, and the finds worked and burnt flint, pottery and animal bone. It is conjectured that the remains of a Middle Iron Age settlement lie either within or in the vicinity of the proposed area for development. The Late Iron Age/Roman pottery was small in quantity and was found in a ditch, which must have been Late Iron Age/Roman or later. The post-medieval/modern field ditches are recorded on modern mapping, and are not archaeologically significant. The results of the archaeological fieldwork suggest that for much of its history, the proposed site for the processing plant was used as woodland/scrub or for pasture or for the growing of crops. It is unlikely that the proposed development area contains extensive or significant archaeological remain
Activity
Context
Lookup
Material
Object
Place
Sentence
SpaceToken
Split
Temporal
Token

Dutch – NER


Vindplaatsen uit de Steentijd zullen bij het Inventariserend Veld Onderzoek (IVO) herkend kunnen worden door de aanwezigheid van vooral (bewerkt) vuursteen en -bewerkingsafval, aardewerkscherven, andere steensoorten (bijv. graniet, zandsteen, kwartsitische zandsteen), fragmenten van verkoolde hazelnootdoppen, (verbrand) (vis-)bot, opvallende hoeveelheden houtskool en oker . Aan de flanken van vindplaatsen kunnen in holocene afzettingen, met name veen en detritus, archeologische resten aanwezig zijn. Naast de hierboven genoemde indicatoren kunnen dunne zandlaagjes eventueel met houtskool duiden op inwaaiing ten gevolge van betreding van de hoger gelegen zandoppervlakken. Op en in de oeverwallen mag bovendien, naar analogie van andere vindplaatsen in het IJssel-Vechtbekken, de aanwezigheid verwacht worden van o.a. ophogingsmateriaal (of bijv. zwartkleuring van het klastische sediment), maar ook aardewerkscherven en verkoolde graankorrels en dorsafval.
Daarnaast kunnen in het gehele gebied in vooral de jongere afzettingen, zoals detritus, Sloef-/ Almere en Zuiderzeeafzettingen scheepswrakken en -ladingen verwacht worden.
6 Advies
De vigerende wettelijke en beleidsmatige kaders in ogenschouw nemende, en
Context
Lookup
LookupPartArtefact
LookupPartComplex
LookupPartMaterial
LookupPartPeriod
Material
Physical_Object
Physical_Thing
Place
Sentence
SpaceToken
Split
Time_Appellation

Swedish NER


Sammanfattning Riksantikvarieämbetet UV Öst utförde under hösten 2005 en arkeologisk förundersökning och en påföljande slutundersökning under försommaren 2006 i Kimstad , inom fastigheten Ask 5:1, Norrköpings kommun , Östergötland . Undersökningarna föranleddes av att Norrköpings kommun avser att bygga en ny pendeltågstation som skulle komma att beröra fornlämning RAÄ 258, en förhistorisk boplats . Undersökningarna omfattade lämningar efter ett neolitiskt hus samt ett fåtal övriga anläggningar på en liten bevarad del av en platå nedanför en större höjd, mellan nuvarande järnväg och landsväg 215. Fynd materialet, anläggningarna och 14C-dateringarna visar att människor bott på platsen under åtminstone två perioder av stenåldern . Analyserat kol har givit två senmesolitiska dateringar från cirka 4300 f Kr . Ett av proven kommer från en härd som överlagrades av ett kulturlager i huset och det andra från bottenlagret i en större grop (A644). En tredje datering, från samma grop , gav en tidigneolitisk datering, cirka 3700 f Kr . Vid undersökningen tillvaratogs cirka 3 kg (361 skärvor) tidigneolitisk keramik och en slipad yxa av diabas stöder denna datering som också verkar vara boplatsens yngsta fas. Det neolitiska huset bestod av ett 20-tal stolphål , med mittstolpar i en så kallad mesulakonstruktion , och lämningar efter ett
Context
Lookup
Material
Monument
Object
Place
Sentence
SpaceToken
Split
TimeAppellation
Token

ARIADNE Dendrochronology Pipelines

Languages

- English, Dutch, Swedish

Aim

- Extract entities and rich sentences relating to dendrochronology analysis and discourse

Focus

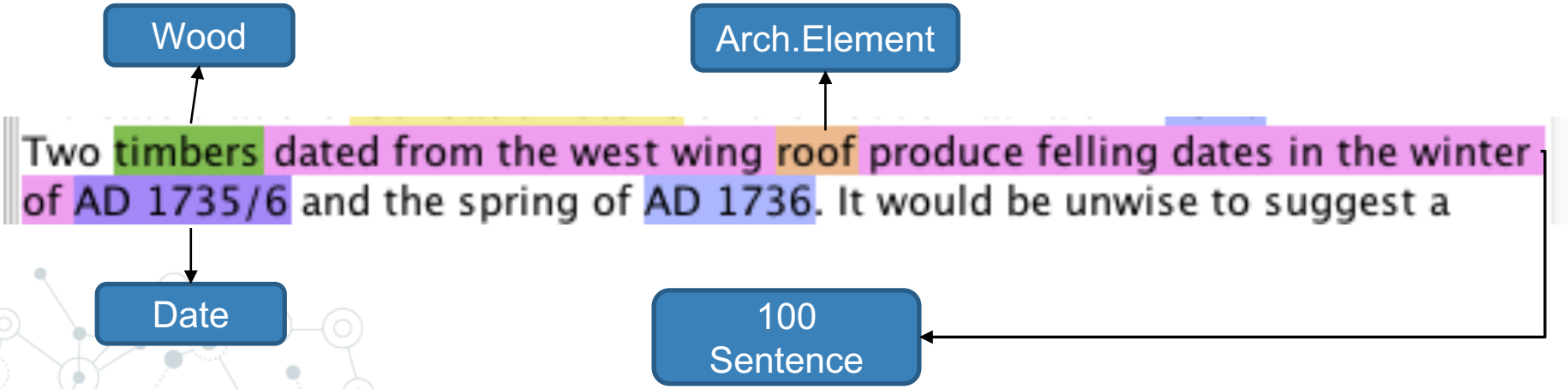
- Architectural Objects
- Date – Time Appellations
- Sample
- Wood Material
  - Types (oak, mahogany)
  - Products (lumber, plywood)

KOS

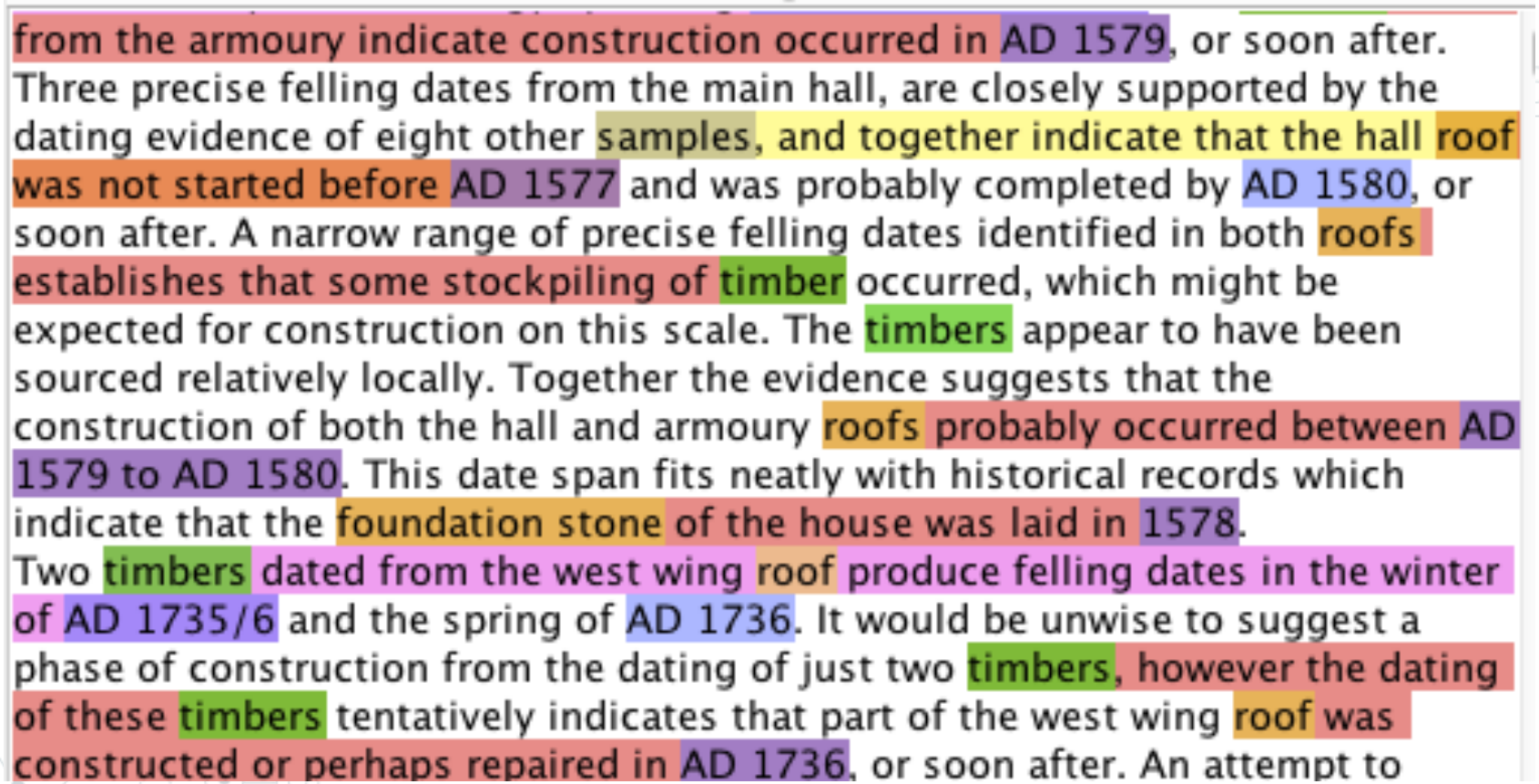
- Getty AAT
  - Architectural Elements
  - Wood and Wood Products

Dendrochronology Pipeline

- Sentence Weighting
  - 100 : (Architectural Object + Wood Material + Date)
  - 60 : (Any **two** of above entities)
  - 30 : (Any of the above entities)
  - Sample phrase (mentioning sampling activity)



Dendrochronology Pipeline



ARIADNE Numismatic Pipeline

Languages

- English

Aim

- Extract entities and relationships between coin, material and date

Focus

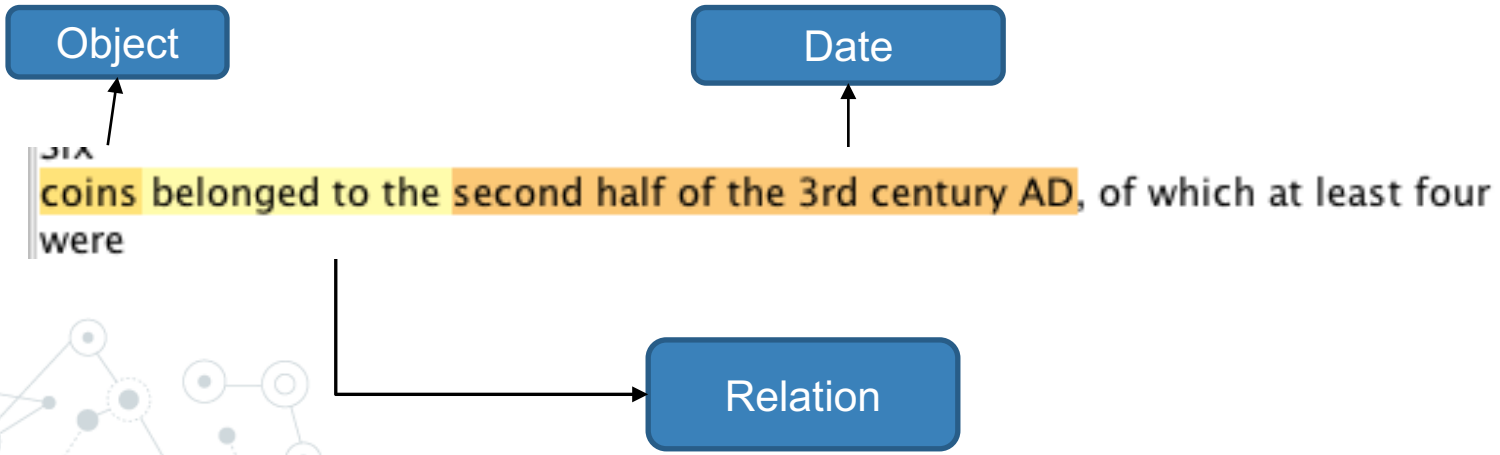
- Objects - Coins
- Date – Time Appellations
- Material

KOS

- Nomisma Ontology
  - Object Types
  - Denomination
  - Material
- Historic England
  - Period Thesaurus

Coin Pipeline

- Focused NER
  - Coin instances, Material, and Dates
- Focused Relation Extraction
  - OPTIMA – *Production-Event* Rules
  - OPTIMA – *Consists-of* Rules



Coin Pipeline

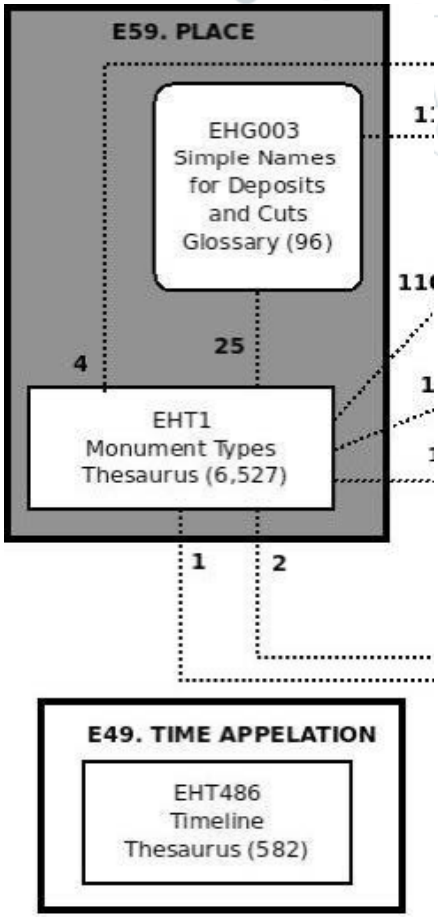
The assemblage is too small for detailed analysis, but the coins are fairly typical of lower status rural settlement of the Upper Thames Valley. There is only one certain early Roman coin (SF201) while a further piece (SF200) may be of this period. Six coins belonged to the second half of the 3rd century AD, of which at least four were probable or certain radiate copies, characteristic of the period from c AD 260/70–296. One of these (SF204) was extremely debased. This, plus two of the other radiate copies and a coin of Postumus came from the same (unstratified) location. The second quarter of the 4th century, always a period of high coin loss in this region, is well represented. Of the eight coins assigned to the House of Constantine all can probably be dated to the period AD 320–341, with all but SF312 dated AD 330 and later. Only SF174 might have been later, but thus is quite uncertain. The module of the corroded and unidentifiable coin SF189 is also consistent with a date in the second quarter of the 4th century, though earlier or later dates are possible. There were coins of the House of Valentinian – which is unusual for the region – and only one coin (SF335) which was probably later. Although its condition precluded certain identification the

Challenges and Solutions

- Vocabulary Use and Coverage
- Negation Detection
- Word Sense Disambiguation
- Multilingual

Vocabulary Use and Coverage

- Vocabulary is critical
  - Domain Oriented
    - Thesauri, Glossaries, Flat Lists
    - Entries of Non NLP relevance
      e.g. descriptive <Material By Form>
      coupled entries "*term/term2/term3*"
- Vocabulary Alignment
  - Align to entities of interest
  - Whole Structure or a Subset
- Vocabulary Coverage
  - Non included concepts
  - **How much to use of the available vocabulary ?**



Semantic Expansion

Controlled **synonym** and **hierarchical** relationships expansion

All Available: glossary and thesaurus terms

pit, hole, cavity, ash pit, fire pit, latrine pit, slag pit, hearth, posthole, site, buried landscape, impact crater, ..

Hypernym: broader terms and narrower terms and synonyms

pit, hole, cavity, ash pit, fire pit, latrine pit, slag pit

Hyponym: synonym and narrower term relationships

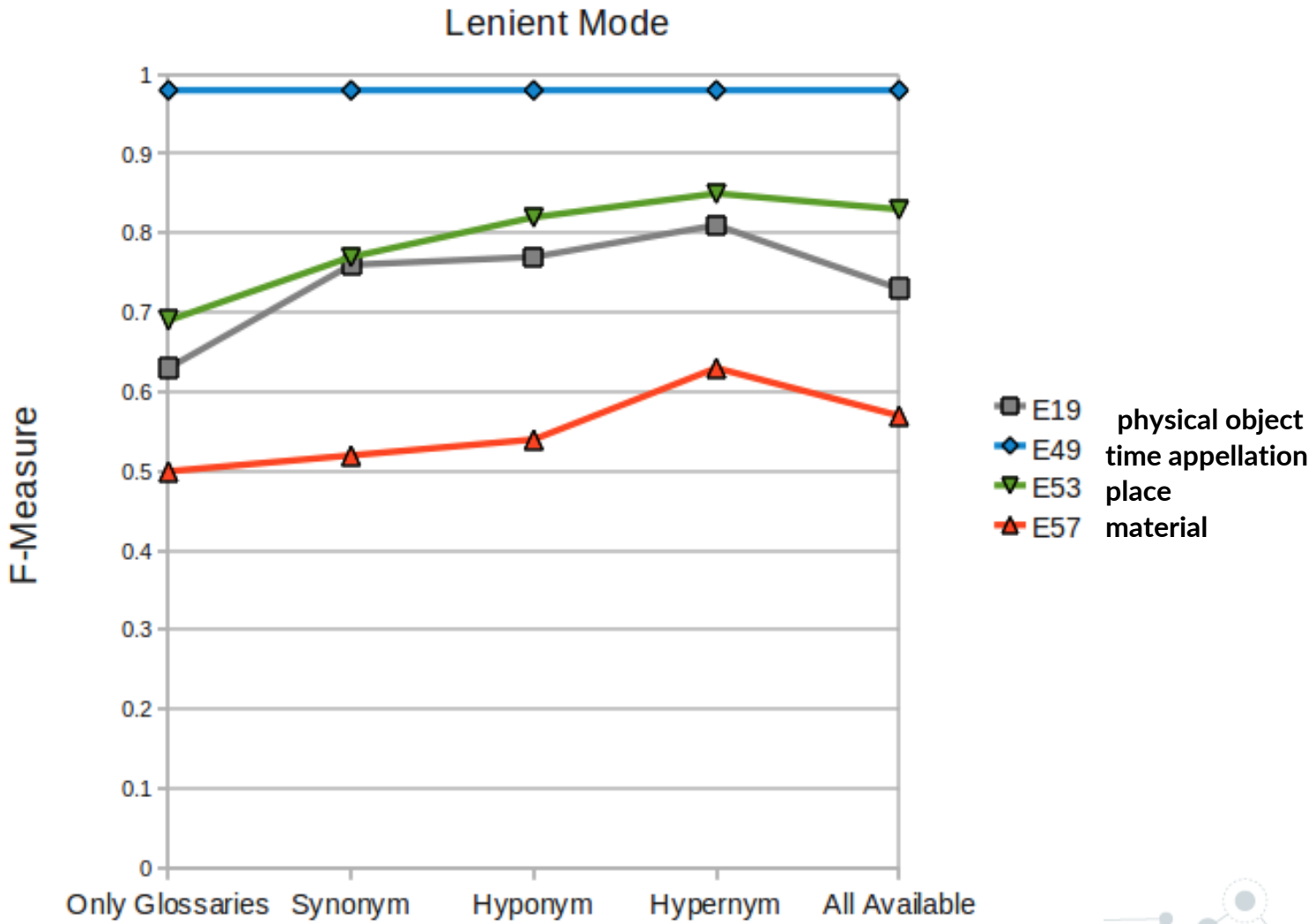
pit, hole, cavity

Synonym: of the glossary terms located in the thesauri

pit

No Expansion: does not make use of the semantic expansion mechanism

Semantic Expansion – Evaluation

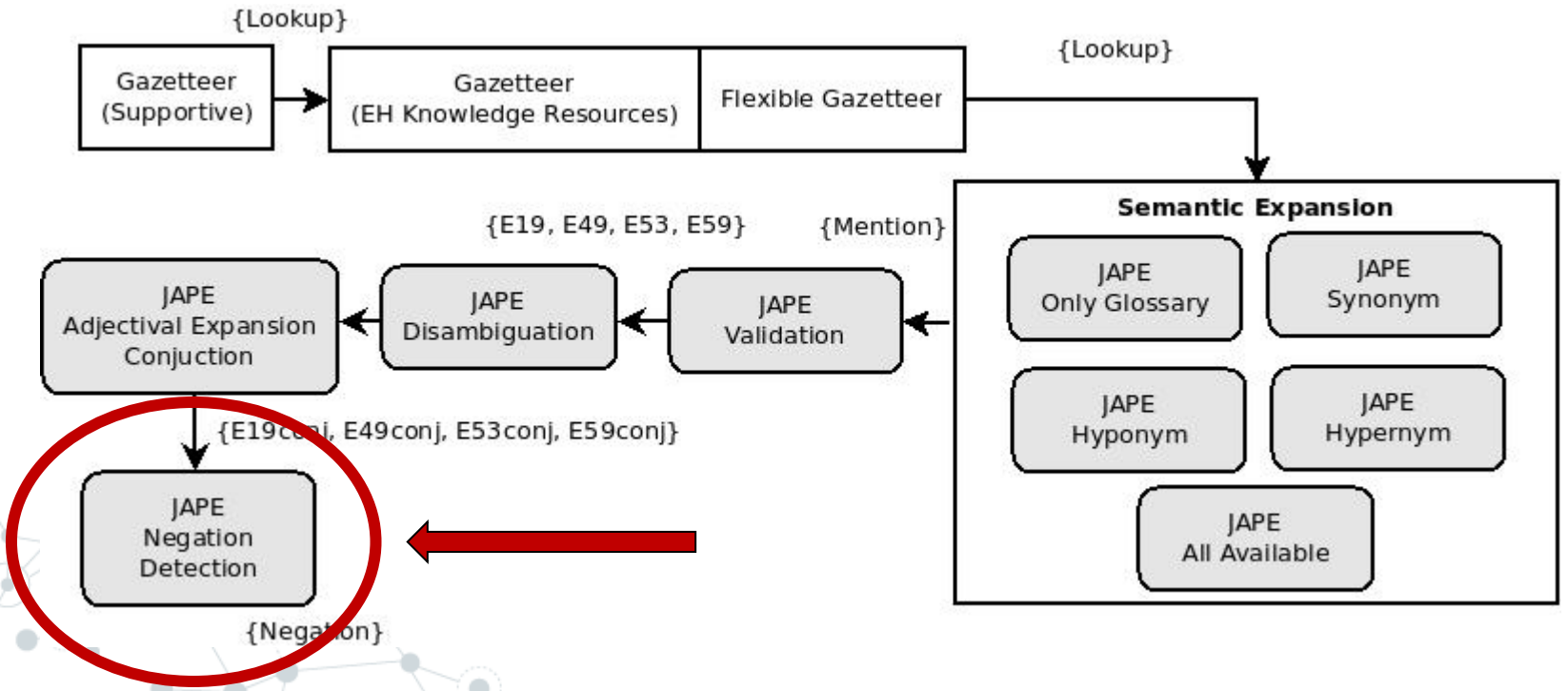


Negation Detection

- Negation: is an integral part of any natural language system that enables users to
  - communicate erroneous messages
  - the truth value of a proposition
  - contradictions
  - irony and sarcasm
- Being able to distinguish negative assertions in context is highly desirable
  - For research and analysis of facts, IR systems.
  - In archaeology appreciation and understanding of negated facts is equally **important** as the interpretation of positive findings.

Negation Detection

- *"No traces of a Roman settlement have been discovered in the area"*
- *"absence of any datable small finds or artefacts"*
- *"wares such as tea bowl are particularly unlikely to exist"*

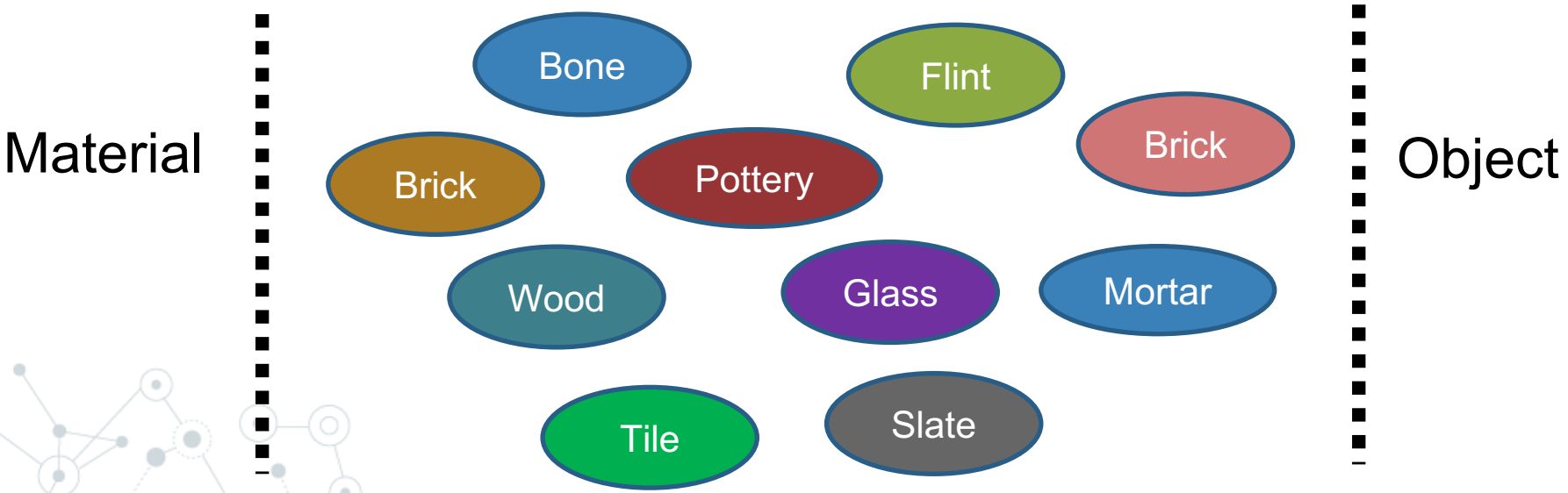


Negation Limitations

- Negated cases removed from indexing
- Need to be able to model negative assertions
  - Ontologies do not clearly address negation
- Potential Pathways
  - Integrated Argumentation Model
    - factual argumentation in a broader epistemological sense
    - may be too complex for IR purposes
  - Introduce a *has_sense* property (positive or negative)
    - Significant increase on the chain of triples
  - A Negative print (anti-matter) for all ontology classes
    - Will double the size of an ontology

Word Sense Disambiguation

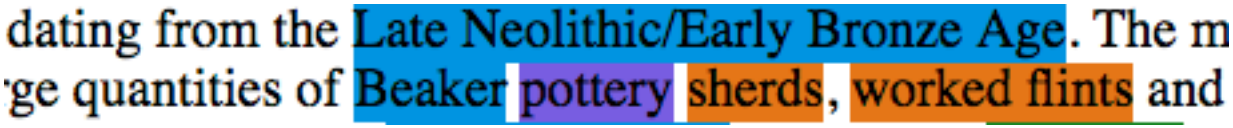
- Polysemy: same word can carry multiple meanings (senses) e.g “mouse”
  - CIDOC CRM ontology for driving the NER brought a specific form of polysemy, which is inflicted by the definition of ontology classes.

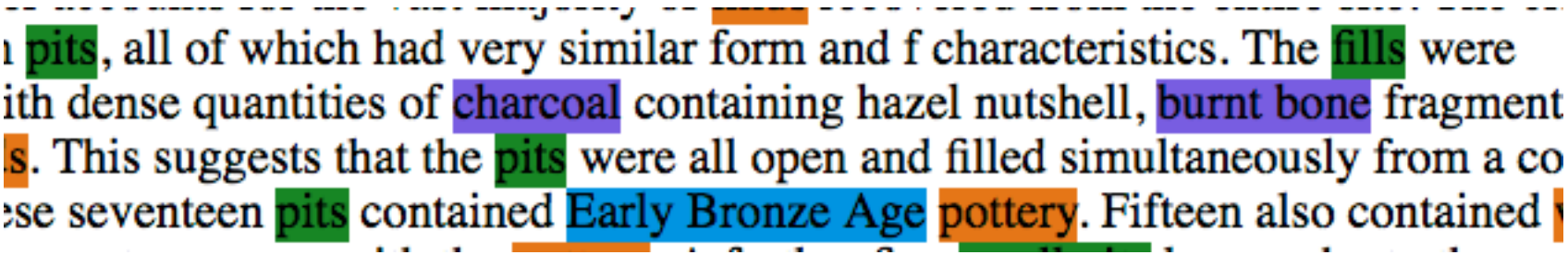


Word Sense Disambiguation

- Word pair cases
  - Left part of the pair as material and the right part as physical object based on the empirical use of English
    e.g. “pottery fragment”, “plaster tile”
- Concatenate pattern rules
  - assumption that co-ordinating concatenations join terms of the same kind
    e.g. “plaster and brick”
- Syntactical pattern rules
  - a determiner preceding an ambiguous term or use of the “made of” clause
    e.g. “artefacts made of wood” “the Iron Age pottery”

Word Sense Disambiguation





Multilingual – Dutch and Swedish

- **Compound** Nouns: A common linguistic behaviour where words are joined together to make a new word.
  - fornlämningsområdet (SE Archaeological Place)
  - aardewerkfragment (NL pottery fragment)
- Whole word matching limitation
- Part word matching prone to false positive and noise
- Several Annotation Span Options
  - a **single** span annotation (aardewerkfragment) associated with two SKOS
  - **two** separate annotations each associated with a SKOS reference
  - **three** annotations, two separate annotations (as above) and a third for the whole span annotated as “P45.consists_of” property

In Conclusion

Rule-based, KOS Driven Information Extraction is capable of delivering indexes of semantic attributes, carrying terminological and ontological qualities which can be expressed in interoperable formats for the purposes of information retrieval.

Thanks!

Any questions?

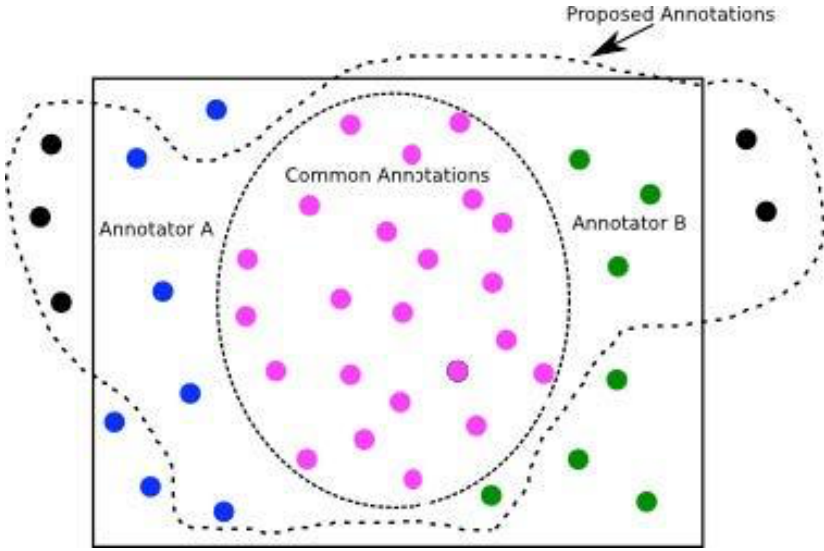
You can find me at:
a.vlachidis@ucl.ac.uk

Evaluation

- Gold Standard
- Inter Annotator Agreement
- Recall – Precision – F Measure

Building the Gold Standard

- ADS - 12 archaeology experts
  - Staff
  - Post-graduate students
- Each document Annotated by two different annotators
- Six composite documents
  - 10 individual summary passages per document
  - 250-300 words each summary
- Inter Annotator Agreement
  - 60% -74%
- Super Annotator - Reconciliation

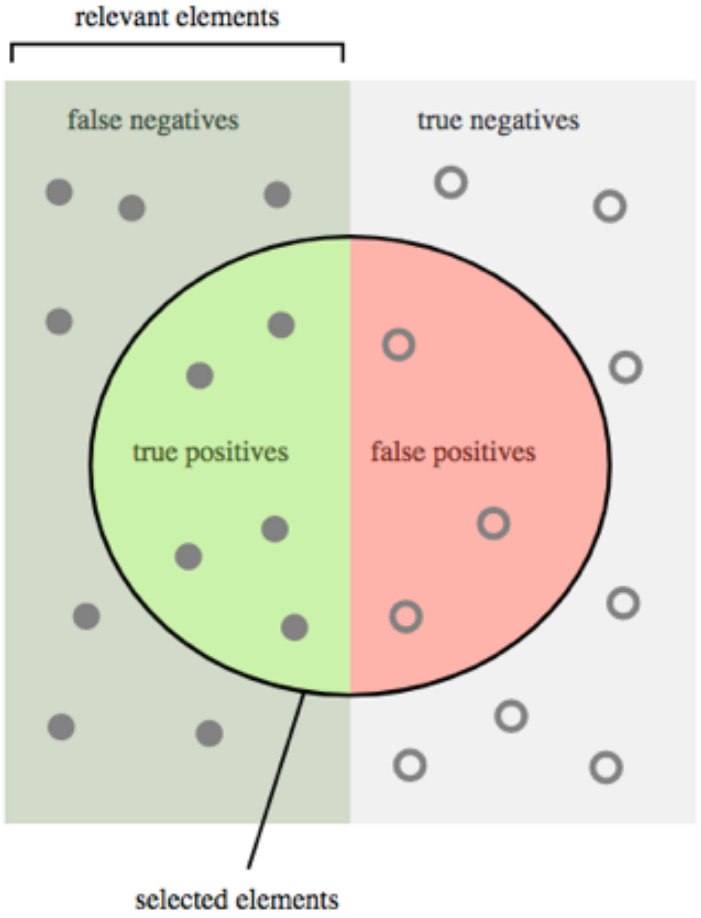


Precision and Recall

Precision = 

Recall = 

$$F_{\beta=1} = \frac{2PR}{P+R}$$



Evaluation Result

	Recall	Precision	F1
No Expansion	0.65	0.78	0.70
Synonym	0.72	0.80	0.76
Hyponym	0.76	0.80	0.78
Hypernym	0.87	0.78	0.82
All Available	0.88	0.73	0.79

Using the IE Output

- The STAR demonstrator
  - Making use of the decoupled RDF files
  - Cross searching between grey literature and datasets
  - A SPARQL engine supports the semantic search
- Semantic Search Examples
  - Context of type X containing Find of type Y“hearth” containing “coin”,
  - Context Find of type X within Context of type Y “Animal Remains” within “pit”.