KOS-based enrichment of archaeological fieldwork reports

Douglas Tudhope and Ceri Binding
Hypermedia Research Group
University of South Wales (USW)

ISKO UK 2023 CONFERENCE
Glasgow, 24-25 July 2023

ARIADNEplus is funded by the European Commission's Horizon 2020 Programme

KOS-based semi-automatic subject indexing

- OASIS online index of archaeological fieldwork and its unpublished reports (grey literature) hosted by the Archaeology Data Service (ADS) and supported by Historic England and Historic Environment Scotland (amongst others).
- OASIS reports contributed by variety of groups, including archaeological contractors (developer-funded) , community groups and academics.
- Existing subject indexing inconsistent and sometimes sparse
- Interactive search of standard KOS (thesauri etc.) from the Forum on Information Standards in Heritage available in latest OASIS version.

- Project aims to develop semi-automatic indexing tools generating (ranked) list of suggestions to assist intellectual judgment by OASIS data producers

- Presentation
  - report results from case study on selection of existing records
  - discuss findings
  - reflect on experience in order to inform future work in the project

Case Study

- Case study on an extract of some 1600 OASIS metadata records.

- Textual summaries/abstracts matched against preferred and alternate terms from Archaeological Object Thesaurus and Thesaurus of Monument Types (SKOS versions)

- Also matched named periods from Historic England Periods (via PeriodO linked data) with regular expression patterns identifying temporal expressions, such as English century and year span expressions

- Other vocabularies (eg Materials) would be possible but case study follows current OASIS cataloguing guidelines

FISH Archaeological Object Thesaurus (SKOS)

FISH Thesaurus of Monument Types (SKOS)

rdfs:label	FISH Thesaurus of Monument Types
skos:hasTopConcept	Agriculture And Subsistence
skos:hasTopConcept	Civil
skos:hasTopConcept	Education
skos:hasTopConcept	Health And Welfare
skos:hasTopConcept	Commemorative
skos:hasTopConcept	Commercial
skos:hasTopConcept	Defence
skos:hasTopConcept	Domestic
skos:hasTopConcept	Industrial
skos:hasTopConcept	Maritime
skos:hasTopConcept	Transport
skos:hasTopConcept	Unassigned
skos:hasTopConcept	Recreational
skos:hasTopConcept	Communications
skos:hasTopConcept	Water Supply And Drainage
skos:hasTopConcept	Gardens Parks And Urban Spaces
skos:hasTopConcept	Religious Ritual And Funerary
skos:hasTopConcept	Monument <By Form>
dct:description	Terminology relating to the built and buried heritage of the British Isles and used for recording sites, monuments, buildings and structures.
dct:publisher	http://www.historicengland.org.uk
dct:identifier	http://purl.org/heritagedata/schemes/eh_tmt2
dct:issued	2022-03-08

RDF downloads (N-Triples Turtle JSON XML)

Case Study

- Named Entity Recognition (NER) via spaCy NLP library "token pattern" rules augmenting the vocabulary term look up

  – pattern rules case-insensitive with whitespace normalisation
  – lemmatisation for object and monument entities
  – Part of speech (POS) tagging looking specifically for nouns
    to reduce false positives (e.g., *building* as a verb instead of a noun)

- Output in a variety of formats (TSV, JSON and HTML markup)
  including inline markup and list of suggested indexing concepts

Consider some (HTML) examples …

Example NER results (HTML)

The earliest feature MONUMENT was a large palaeochannel MONUMENT (probably a former branch of the Thames), which occupied much of the south-east half of the site MONUMENT. It was mainly filled with fine-grained sediments OBJECT, some of which were organic and dated by radiocarbon assay. The earliest deposits OBJECT were dated to 19480-19039 cal BP. Other sediments OBJECT were similarly dated to the Neolithic NAMEDPERIOD, Bronze Age NAMEDPERIOD to Middle Iron Age NAMEDPERIOD and Late Saxon. Most evidence for human activity was in the north-west half of the site MONUMENT. The earliest artefacts comprised 61 residual struck flints dated from the Mesolithic NAMEDPERIOD to the Bronze Age NAMEDPERIOD and part of a Late Bronze Age NAMEDPERIOD gold bracelet OBJECT. The excavations provided a transect across a Roman NAMEDPERIOD landscape. Two phases of the London-Silchester Roman NAMEDPERIOD road were revealed next to the modern London Road. The earlier road MONUMENT was flanked by a ditch MONUMENT and later by a fence MONUMENT. Evidence for Roman NAMEDPERIOD occupation on the SE side of the road MONUMENT, clearly represented part a linear settlement MONUMENT that, as previous excavations have shown, extended alongside the road MONUMENT into what is now the centre of Brentford. The evidence included the remains of two substantial timber buildings MONUMENT that had burnt down bread ovens MONUMENT, hearths MONUMENT, pits MONUMENT and gravel OBJECT surfaces. Successive Roman NAMEDPERIOD field systems MONUMENT defined by ditches MONUMENT lay between the settlement MONUMENT and the channel. The ditches MONUMENT also defined a track MONUMENT running down from the settlement MONUMENT to the channel. One ditch MONUMENT contained a human skeleton MONUMENT, and a crouched burial MONUMENT lay in a small grave MONUMENT next to another ditch MONUMENT. Roman NAMEDPERIOD artefacts included pottery, fragments building MONUMENT material, two shale armlets OBJECT, a sandstone palette, fragments of lava quernstone, pieces of glass vessels OBJECT, iron objects (a stylus OBJECT, a hipposandal OBJECT, cleavers OBJECT, the bowl OBJECT of a ladle OBJECT), a lead weight OBJECT and three bone pins OBJECT. Copper alloy objects included a large number of coins OBJECT, several cosmetic or medical implements OBJECT, fragments of brooches OBJECT, a bracelet OBJECT, a hair pin OBJECT and a fine circular seal box OBJECT lid. Later activity was represented by a few medieval NAMEDPERIOD and post-medieval NAMEDPERIOD pits MONUMENT.

Example NER results (HTML)

Wessex Archaeology was commissioned by PMSS to undertake a Stage 3 archaeological assessment of samples OBJECT taken from vibrocore VC7, recovered during a programme of geotechnical investigations on the proposed Project NEMO, UK-Belgium Electrical Interconnector. The vibrocore was located c.12km east of Ramsgate on the margins of a palaeochannel MONUMENT feature visible on geophysical data. The vibrocore was chosen for Stage 3 assessment as it contained probable prehistoric NAMEDPERIOD terrestrial deposits OBJECT with potential to provide information on the nature of past environments.The results show successive environments including an early Holocene freshwater channel and freshwater pool MONUMENT within a wooded river valley that became progressively choked with vegetation. This woodland comprised pine and hazel with a possible highly significant record of beech. The increasing amounts of vegetation lead to peat formation, with radiocarbon dates indicating that this terrestrial environment dates from c.10,000 years ago, equivalent to the early Mesolithic NAMEDPERIOD period. Potential evidence of human activity in the form of charcoal has been recovered from the sediments OBJECT . The well preserved remains of pollen OBJECT , ostracods, molluscs and foraminifera OBJECT are considered to be highly significant in the understanding of this early Mesolithic NAMEDPERIOD environment. This peat OBJECT deposit OBJECT has been truncated by sea level OBJECT rise with subsequent deposition evident of possibly late Mesolithic NAMEDPERIOD date within outer estuarine and shallow marine environments. It is recommended that further Stage 4 analysis work on the molluscs, ostracods and pollen OBJECT is undertaken. This should be supported by further radiocarbon dating of the sediments OBJECT to discover the timing of deposition of significant sediments OBJECT .

Case Study - Findings (from examples)

two shale armlets OBJECT , a sandstone palette, fragments of lava quernstone,

- NER not identify *quernstone* as not included in Object thesaurus
  - (*quern* is included, described as "a stone for grinding grain")
- NER not identify *post-medieval* (matching instead on *medieval*)
  - hyphenated form not in PeriodO authority (*post medieval* is included)

medieval NAMEDPERIOD and post- medieval NAMEDPERIOD pits MONUMENT

➔ need to extend entry vocabulary (and flexibility in the matching) for syntactical and synonym variants.

Stage 4 analysis work on the molluscs,

- NER not identify *ostracods* and *molluscs* as not included in Object thesaurus
  - (*ostracod remains* and *mollusca remains* are included)

➔ need to extend entry vocabulary and particular consideration of compound terms

Eg add constituent terms as BTs or ALTs to working entry vocabulary

Case Study - General findings

- Further entry vocabulary needed for NER purposes for
  - spelling alternatives (e.g. *palaeolithic/paleolithic*, *mediaeval/medieval)*
  - preferred terms with a context qualifier or not in natural language order (e.g., *hermitage (religious)*, *palette (artists)*
- Extended entry vocabulary offers wider value generally
  - cf 'end-user thesaurus' (Bates)
- Identify faceted combination of concepts

  more elaborate patterns needed for combinations important to OASIS, such as period-object and period-monument phrases

- Assign properties to suggestions reflecting confidence or priority
- Apply a post-processing filter
  - Problematic cases identified in evaluation
  - Patterns signifying negation (negative results important in archaeology)

Case Study – Negation detection examples

- No Roman, medieval or early post-medieval artefacts were recovered from the evaluation. The lack of any medieval or early post-medieval artefacts suggests that either there was no precursor to the existing early 18th-century farmhouse on the site, or that any such remains have been extensively disturbed by landscaping associated with the current buildings.
- No associated roadside ditches or structures were exposed. There was no evidence for a postulated second Roman road crossing the cable run, indicating either that this road does not exist, or it has been completely removed by later activity.
- *Probable negation?* No firm evidence was encountered however for actual settlement foci predating the Iron Age. … No Anglo-Saxon artefacts or features and few medieval finds or deposits were encountered.
- *Tentative?* There was little dating evidence for this phase, but it is suggested that these fields or enclosures were later Roman and they perhaps formed part of a nearby Roman villa estate.

➔ need to also consider appropriate metadata (model) for negation

Case Study - General findings

- Case study approach was - select all KOS concepts present in the *abstract* of a document that match the rule patterns
- Restriction of source text to the abstract yields reasonable results for some cases but is very dependent on the writing of the abstract and the cataloging guidelines

Consider extending the scope of indexing to whole document:

- Some past work in archaeology returned every occurrence of subject entities, using a frequency count to approximate relative importance
- However, this unreliable as multiple occurrences can derive from common objects or background sections discussing previous work on the site or nearby
- Pre-processing to identify common categories of sections in OASIS reports potentially helpful though challenging due to wide variation in writing styles

Reflections - Evaluation is complex

- Lack of any corpus of good practice indexing of the reports
- Wide variety of contributors and report styles

- Notion of definitive ‘gold standard’ for subject indexing might be considered problematic, in light of the wide variation in human subject indexing and indexing policies

- Designing instructions for annotators complex
  - Strict experimental protocols may hinder generalisation from the laboratory to the actual contexts of use in retrieval

- Assessing future utility of indexing tools should take account of intended retrieval system, range of user experience and the nature of queries and (re)search questions to be investigated

Reflections – consider broader context?

• In a previous project (ESRC study of commercial software prototyping practice) we drew on participatory design and sociology of technology.

• Here the broader context of software development was considered as combining technical and social components in 'messy networks' of evolving prototypes, user expectations, requirements, and working practices.

➔ Future (participatory design) work in current project could incorporate broader contextual elements, including the guidelines (and practice) for indexing and report writing, user experience/expectations, variations of search functionality seeking to take advantage of enriched subject metadata.

Reflections - Reviewing underlying approach

- Semi-automatic indexing (suggestions) based on standard KOS
  - following FAIR principles and OASIS guidelines

Is this blurring the boundaries between

- traditional subject indexing
- named entity recognition (NER)
- named entity authority control (name authorities) ?

Reflections - Reviewing underlying approach

All three approaches associate entities (with names and PIDs) to a document or segment of a document, either automatically or semi-automatically, sometimes using vocabularies, and thus the approaches share some family resemblance.

Key features of the approaches for this case study include:

- scope of the methods and balance between intellectual and automatic activity
- source document scope and the output format
- scope and extent of any vocabularies
- scope and the extent of the indexed entities
- ultimate purpose of the exercise

Touching on a few of these issues …

Reflections - Reviewing underlying approach

- Name authorities are always based and subject indexing is often based on vocabularies, while NER is often not.

- Name authorities and NER work with the specific set of entities given names in the domain, while subject indexing extends to more general concepts
- Subject indexing vocabularies can be large and deep (with a small set of top-level concepts/facets.

- It might be argued that subject indexing concepts can be more abstract entities, depending on the subject domain and thus may pose more difficulties for identification and offer wider scope for differing judgments.

- Name authorities and NER entities may (arguably) be more more straight forward to distinguish as regards homonyms and different senses?

Reflections - Reviewing underlying approach

Perhaps the most distinguishing feature is the purpose for which the approach is applied to the document and relationship of the concept with the text string

- NER focuses on the immediate identification of a name; the relationship between text string and name may be *instanceOf*, leaving determination of further purpose to the end-application.
- Name authorities can involve various relationships connecting works with persons and places but can also include the subject(s) a document is *about*, which is the key focus for subject indexing.

- The *aboutness* relationship is a thorny topic within subject indexing
- *Aboutness* sometimes distinguished from *isness* (similar to the *instance* relationship) and *ofness* (e.g. for picture indexing). Closely related to subject indexing strategy (*exhaustivity* and *specificity*)

Reflections – hybrid approach?

Case study approach is hybrid?

- Entities involved are more typical of subject indexing and are arguably less clear cut in identification than *some* NER applications.
- NER would typically aim to identify every occurrence of an instance
- Subject indexing traditionally provides (vocabulary concept) subject metadata that best represents the *aboutness* of the document.
  - may be a subset of the terms mentioned explicitly in document
  - may include terms not present in the document

However

- OASIS cataloguing guidelines more complicated …

Reflections – going beyond aboutness?

Ultimate purpose of OASIS indexing goes beyond overall aboutness to address FAIR principles?

- Cataloguers asked not to record all the different individual finds but to help end-users understand the *significant* findings of the archaeological report

 - asked to add keywords on interesting or relevant objects found (or flag that report has no significant findings)

- Indexing strategy encompasses *significant* elements of a particular investigation, intended to reflect the (re)search needs for which the report would be considered important and might be reused.

→ Investigate contextual patterns reflecting significance and incorporate those patterns in the post-processing prioritisation filters

Conclusions

- Results demonstrate that overall approach is feasible - the NER patterns could be extended to accommodate other vocabularies

- Findings include need for some pre-processing to extend the entry vocabulary of the KOS employed for NLP purposes. Compound terms merit particular attention as does the faceted combination of separate concepts.

- Post-processing filters could prioritise subject indexing considered significant and reduce rankings of common problematic cases. Negation detection could be an important component.

Conclusions ctd.

- Techniques can be characterised as a hybrid approach.

- The purpose or indexing policy for OASIS goes beyond overall *aboutness* to request indexers to include *significant* objects or artefacts found.

- Utility assessment is complex as user behaviour and subject indexing practice and guidelines all change over time in an evolving complex network.

- Ideally (co)design of future best practice indexing policy and guidelines for abstract writing could operate in tandem with participatory design of an automatic indexing recommendation system and corresponding search services.

References

- ARIADNEplus. https://ariadne-infrastructure.eu
- Hypermedia research group. http://hypermedia.research.southwales.ac.uk/kos/
- HeritageData LOD vocabularies https://www.heritagedata.org/

- Binding C, Tudhope D. forthcoming. KOS-based enrichment of archaeological fieldwork reports. ISKO-UK 2023 Proceedings.
- Binding C, Tudhope D, Vlachidis A. (2018) A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science. 45(3), 364-386.
- Tudhope D, Beynon-Davies P, Mackay H. 2000. Prototyping Praxis: Constructing Computer Systems and Building Belief. Human-Computer Interaction, 15(4), 353–383.
- Vlachidis A, Tudhope D. A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology, 67(5), (2016). 1138-1152.

Open Access versions available from https://bit.ly/2ocaHC6

THANK YOU!

ARIADNEplus is a project funded by the European Commission under the H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914.

The views and opinions expressed in this presentation are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

Thanks are due to the OASIS team, including Tim Evans, Jo Gilham and Holly Wright.

Contact:
douglas.tudhope@southwales.ac.uk
ceri.binding@southwales.ac.uk

http://www.ariadne-infrastructure.eu/