KOS-based Enrichment of Archaeological Fieldwork Reports

Ceri Binding* and Douglas Tudhope**

* **Hypermedia Research Group, University of South Wales, Pontypridd, CF371DL, UK
* ceri.binding@southwales.ac.uk, ** douglas.tudhope@southwales.ac.uk

Ceri Binding has been a researcher in the Hypermedia Research Group at University of South Wales since 2007, having previously worked in civil engineering and then software development. During that time he has jointly published several research papers focussed on the subjects of controlled vocabularies, data integration & interoperability, Linked Open Data and the semantic web. He developed the 'heritage data' platform making national cultural heritage controlled vocabularies available as Linked Open Data. He produced a SKOS RDF conversion for the Integrative Levels Classification (2nd Edition, ILC2). He created and maintains an open archive of Networked Knowledge Organisation Systems (NKOS) workshop proceedings. Recent research projects include ARIADNEplus H2020 project (and its predecessor ARIADNE), and the Historic England MATRIX project. His research interests include knowledge organisation, controlled vocabularies and semantic web technologies. http://hypermedia.research.southwales.ac.uk/kos/ https://orcid.org/0000-0002-6376-9613

Douglas Tudhope is Professor in the Faculty of Computing, Engineering and Science, University of South Wales and leads the Hypermedia Research Group. His main research interests are in Knowledge Organization Systems (KOS) and Services and their potential for assisting interactive and automatic indexing and retrieval. He has participated in the FP7 ARIADNE and H2020 ARIADNEplus Archaeological Infrastructures Projects and was PI on the AHRC funded STAR, STELLAR and SENESCHAL projects and the preceding EPSRC funded FACET project, investigating thesaurus-based query expansion. He is a member of the Networked Knowledge Organisation Systems/Services (NKOS) network and has (co)organised various NKOS workshops. He was a member of the ISO TC46/SC9/SC8 (and NISO) working group developing the thesaurus standard (ISO 25964). http://hypermedia.research.southwales.ac.uk/kos/ https://orcid.org/0000-0002-5222-0430

Binding, Ceri, and Douglas Tudhope. 2024. "KOS-based Enrichment of Archaeological Fieldwork Reports". *Knowledge Organization* 51, no. 5: 292-299. 32 references. DOI:10.5771/0943-7444-2024-5-292.

Abstract: Semantic enrichment techniques and tools based on knowledge organization systems (KOS) have an important role to play in supporting information discovery. This paper reports on work investigating and developing automatic indexing techniques (for final intellectual judgment) based on KOS. Within the UK, the OASIS online index of fieldwork events and their unpublished reports represent a major initiative to make archaeological fieldwork available to a wider public. OASIS is hosted by the Archaeology Data Service and is funded by Historic England and Historic Environment Scotland. A wide variety of organisations provide OASIS reports. Subject indexing is inconsistent and sometimes sparse, although use of standard KOS from the Forum on Information Standards in Heritage is encouraged. Results from a case study for an automatic (KOS-based) subject indexing recommendation system are reported. Findings include the need to extend the KOS entry vocabularies and the need for post-processing filters to prioritise subject indexing significant for the document in question. The paper reflects on the experience with future work in mind, including discussion of evaluation issues and positioning the approach within the context of previous work on subject indexing, automatic indexing for Name Authorities and Named Entity Recognition (NER). The techniques followed in the case study can be characterised as a hybrid approach. The purpose for which the indexing is applied is a key distinguishing feature. In this case, the purpose or indexing policy for OASIS goes beyond overall aboutness to request indexers to include significant objects or artefacts found during the project. Future work will investigate contextual patterns reflecting significance and incorporate those patterns in post-processing prioritisation measures.

Received: 20 March 2024; Revised 24 March 2024; Accepted 25 April 2024.

Keywords: automatic subject indexing; named entity recognition; knowledge organization systems.

† This article was selected as one of the best papers at the Seventh ISKO UK biennial Conference, July 24-25, 2023, held at the University of Strathclyde, Glasgow, in partnership with the School of Computing, Engineering & the Built Environment at Edinburgh Napier University.

1.0 Introduction

Semantic enrichment techniques and tools based on knowledge organization systems (KOS), such as thesauri, classifications and other vocabularies, have an important role to play in supporting information discovery, as seen for example in Europeana experience (Isaac 2015). The FAIR (Wilkinson et al. 2016) Data Principles (Findable, Accessible, Interoperable, Reusable) support re-use of data and key information in archaeological archives, with machine readable vocabularies playing an important role. In particular, as discussed by McKeague (2021), semantic interoperability entails a key role in archaeological data and metadata for standard KOS containing persistent identifiers, which allow concepts to be unambiguously referenced.

This paper reports on work investigating KOS-based automatic subject indexing recommendation techniques (for final intellectual judgment) in the archaeology domain. Results from the case study are reported and findings are discussed, together with some reflections from the experience intended to inform future development.

Within the UK, the OASIS online index of fieldwork events and their unpublished reports, sometimes termed 'grey literature' (Evans 2015), has been a major initiative to make archaeological fieldwork available to a wider public. OASIS is hosted by the Archaeology Data Service (ADS) and is supported by Historic England and Historic Environment Scotland (OASIS 2024). A wide variety of organisations provide OASIS reports, including archaeological contractors (developer-funded), community groups and academics. Subject indexing is inconsistent and sometimes sparse, although use of standard KOS from the Forum on Information Standards in Heritage (FISH 2024) is encouraged.

2.0 Case study

We conducted a (semi-automatic) indexing case study on an extract of some 1600 OASIS report metadata records. The textual summaries/abstracts were matched against preferred and alternate terms extracted from the SKOS versions of the FISH Archaeological Object Thesaurus and the FISH Thesaurus of Monument Types (Heritage Data 2024). The Historic England Periods list (HE Periods 2024) provided a source for identifying named periods via PeriodO linked data, together with some regular expression patterns to identify temporal expressions, such as English century and year span expressions. Binding and Tudhope (2023) discuss work on archaeological temporal expressions in more detail and report on work with a temporal expression normalization tool in different languages. The STELETO tool (Binding et al. 2019) was used to render the KOS resources to a form suitable for Natural Language Processing (NLP) techniques, which augmented the term look up. This Named Entity Recognition (NER) processing employed lemmatisation for object and monument types, enabling variant forms of the words to be detected (e.g., singular/plural forms). The pattern rules were made case-insensitive and the input text had whitespace normalisation applied to avoid non-matching due to trivial differences in whitespace. Part of speech (POS) tagging looked specifically for nouns, in order to reduce 'false positive' matches (e.g., building as a verb instead of a noun). The automatic indexing suggestions were output in a variety of formats (TSV, JSON and HTML markup). Example results (HTML output) are illustrated in Figure 1 and Figure 2.

3.0 Findings from the case study

The case study is intended to inform planning of future work. A selection of prominent issues encountered is discussed in this section. The term post-medieval was not identified by the NER process (see Figure 1, matching instead on the term medieval). This is because it does not exist in the specified PeriodO authority in hyphenated form (the term post medieval is present). This suggests the need to extend the entry vocabulary (and flexibility in the matching) to account for syntactical and synonym variants. As another example, the term quernstone was not identified (Figure 1) because it does not exist in the Object Types thesaurus (the term quern is present in the thesaurus, described as 'a stone for grinding grain'). The terms ostracods and molluscs are potentially interesting but are not matched (Figure 2) because the specific Archaeological Objects thesaurus terms are ostracod remains and mollusca remains – again pointing to the need for extension of the entry vocabulary and particular consideration of compound terms. The ISO 25964 international thesaurus standard (ISO 25964-1 2011) outlines different options for compound terms and their consequences for retrieval. In this case, the consequence is for NER but the principles are similar. One option is to retain a compound preferred term and additionally include the constituents as preferred (possibly broader) terms or possibly non-preferred alternate terms.

Looking at the results overall, further entry vocabulary is needed for NLP purposes, for example where spelling alternatives exist (e.g. palaeolithic/paleolithic, mediaeval/medieval) and where the preferred term contains a context qualifier or is not in natural language order (e.g., hermitage (religious), palette (artists). A few spelling mistakes were observed; common mistakes can be addressed in the entry vocabulary or a term distance component could be included in the matching function. Enhanced entry vocabularies may have wider value for the FISH KOS in the longer term. While in this case, the immediate users are NLP indexing agents, adding a wider range of alternate terms can be seen as reminiscent of Bates (1986) proposal of an 'end-user thesaurus' where a large entry vocabulary includes a wide vari-

Figure 1. Example NER results (HTML) for objects, monuments and periods.

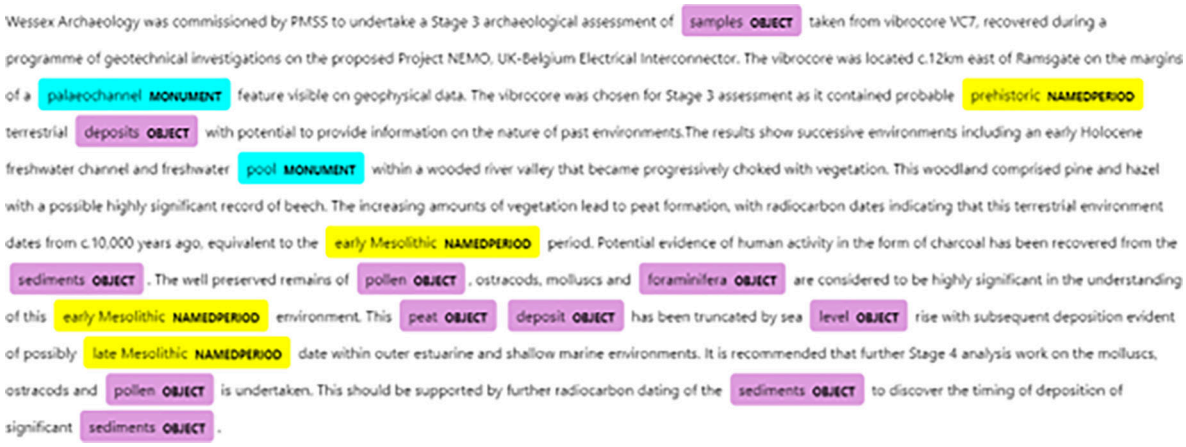

Figure 2. Example NER results (HTML) for objects, monuments and periods.

ety of terms to support search by non-expert users. Faceted combination of concepts requires particular consideration; more elaborate rules need to be developed for combinations important to OASIS, such as period-object and period-monument phrases. Many developer based interventions will not result in any findings (e.g. if there is no particular archaeological significance to a development area) and this is sometimes explicitly reported. Negative results are important in archaeology and work is required on patterns signifying negation, building on previous work that adapted a method used in biomedicine (Vlachidis and Tudhope 2016). This could form part of a post-processing set of filters that tackled negation expression and any common problematic cases found in the evaluation of results, where it was considered appropriate to remove a suggested subject indexing concept or reduce its confidence value. For example, case study analysis uncovered various unhelpful matches on terms, such as site, trench, road, etc. referring to the archaeological intervention, rather than its findings. In fact, the term 'site' was the most frequently matched monument type in the case study, followed by the term 'road'. Since such terms sometimes occur in fairly stylised patterns they may be capable of being addressed by specific filters to remove the index term or reduce confidence values.

4.0 Reflections

Reflections on the case study include the complexity of evaluating (and thus improving) the results of the work, particularly in light of the lack of a corpus of good practice indexing of the reports and the wide variety of contributors and report styles. In previous work with archaeological NER, we have encountered complexities in designing the instructions for human annotators (indexers) of documents when creating a 'gold standard' for evaluation (Vlachidis and Tudhope 2012); our instructions were intended to be relevant to the intended future retrieval use case and annotators were asked to exercise judgment when encountering instances of textual expressions from the vocabularies. Some NER practices involve highly detailed instructions, to the point of specifying all vocabulary terms that must (always) be annotated. Without this specificity, we may encounter situations in an evaluation where it might be considered that 'the machine is right and the human (annotator) is wrong' – see also the discussion by Hjørland (2017, section 2.5). However, such highly specific instructions appear to undermine the ecological validity of the evaluation and its relevance to future use cases involving intellectual judgments and differing information needs. Comparisons of evaluation outcomes with alternative methods and techniques can be complicated in practice. Due to typical length restrictions in publication, evaluation details are often not provided in full, hindering the comparison of evaluation results, an exception being those from research communities with an established experimental framework. Strict experimental protocols may hinder the generalisation from the laboratory to the actual contexts of use in retrieval - see discussion on evaluation by Golub et al. (2016) and see also Golub (2019). The notion of any definitive 'gold standard' for subject indexing might be considered problematic in light of the wide variation in human subject indexing (inter-indexing consistency) revealed by previous studies (for example, Markey 1894; Lancaster 2003; Golub et al. 2016) and differences in the overall policy guiding the indexing. Soergel (1994) draws attention to the complexity of assessing the future utility of indexing tools. Evaluation should take account of the intended retrieval system and the nature of queries and (re)search questions that will be investigated; if a concept-based search system is envisaged then basing evaluation on literal string matches may be unhelpful.

4.1 Reviewing the underlying approach

Taking stock and considering options for possible next steps following the case study, it would seem an appropriate time to review the underlying approach. The automatic indexing (recommendation) strategy we adopted, following the FAIR principles and OASIS deposit guidelines, as discussed in Section 1, is based on standard KOS in the archaeology domain. Is this blurring the boundaries between subject indexing, NER and named entity authority control? With that in mind, we first give a very brief overview of these three approaches and then attempt to compare them for purposes of this discussion. This section is not intended as an exhaustive review or definition, rather a comparison of some key features, as relevant to our case study.

In information science, named entity authority control seeks to enrich documents (traditionally library catalogues of different kinds) by identifying and correctly using the named entities, the names of people and works (also extended to places, organisations, dates) with the assistance of Name Authorities (NA), sometimes called authority files. Different name variants are joined together and treated similarly while identical name forms referring to different entities are disambiguated. The Functional Requirements for Subject Authority Data considers authority control for names of persons/places the same process as for the names (terms) of subject concepts and thus would tend to group it with subject indexing: "The purpose of authority control is to ensure consistency in representing a value - a name of a person, a place name, or a term or code representing a subject - in the elements used as access points in information retrieval" (Zeng et al. 2010). While some names (e.g. of events, places) are used for subject indexing, it can be useful to make a practical distinction with subject indexing particularly when considering application software. It might be argued that need is intensified with NA due to the frequency and importance of name occurrence. In the Perseus project (on the Greco-Roman world), which draws on text mining and NER techniques with authority files, for the Perseus Digital Library, (Crane and Jones 2006) estimated that 6-7% of words in texts are named entities of some kind. Authoritative NA are provided by institutions such as National Libraries and Archives, Library of Congress and aggregation services, such as OCLC's VIAF (2024) which merges authority data for names, locations, works, expressions while preserving regional terminology. There are also standard domain and subject specific authority files. Each recognised instance of a name in a source document is enriched by the authoritative version, usually employing markup of some kind, e.g. in MARC format or TEI (Text Encoding Initiative). These authorities hold name variants, definitive versions, unique identifiers (PIDs) and also contextual information which can help to disambiguate the same or similar names denoting different entities. The enrichment process was traditionally intellectual but is now usually at least partly automated, with a variety of tools available, also due to the need for updating catalogues with operational authority files and services constantly growing and refining (Wolf 2020).

NER is considered a sub-task of information extraction within NLP (Nadeau and Sekine 2007). NER locates and identifies entities within a body of text that have predefined

categories (names). The original focus was on the news, business and defence domains but this has broadened to a range of areas, with some current focus on the medical and bioinformatics domains. Entities are often names of persons, organisations, places, events but can be domain specific. For example, the largest model in the Stanford NER system identifies entities as persons, organizations, locations, money, time, percentages, and dates (Finkel et al. 2005) NER tends to be fully automated (with iterative evaluation). It may or may not employ a dictionary. Methods can be rule-based or machine learning (ML) with the latter being more common in recent years and requiring training sets rather than being vocabulary-based. However, this may not yield the authoritative entities (with PIDs) that support semantic data integration and reuse. Output may be in one of several serialisation formats e.g., XML, delimited (TSV), slash tagging or HTML markup.

Subject indexing is a key application of knowledge organization. It was traditionally intellectual but is now often automated to different degrees. Golub (2019) gives an overview of automated subject indexing and distinguishes 'text categorization' (machine learning approaches), 'document clustering' (automatic clustering and derivation of names, via unsupervised learning) and 'document classification' which takes advantage of existing KOS vocabularies and is the method we employ in the work with OASIS. ML approaches generally (and for NER) tend not to use KOS vocabularies, relying instead on training sets, but some work is vocabulary-based. For example, Annif combines different ML tools and KOS (Suominen 2019). The National Library of Medicine's Medical Text Indexer (MTI) is an operational, vocabulary-based machine aided indexing system for MEDLINE with recommendations based on the Medical Subject Headings (MeSH®). The MTI pipeline combines different pathways for recommendations, including ML for some types of indexing (Mork et al. 2017). It is reported recently as moving towards greater reliance on automated suggestions with less input from cataloguers. Our work with OASIS relies on a variety of archaeological KOS.

All three approaches associate entities (with names and possibly IDs) to a document or segment of a document, either automatically or semi-automatically and sometimes using vocabularies, and thus the approaches share some family resemblance. It is useful to attempt a (very broad brush) comparison of key features for this paper's purposes. These include:

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

Taking the features in turn, all three approaches employ automatic methods with some intellectual input (if we count annotation for training sets under intellectual). Subject indexing has been the most reliant on intellectual cataloguing though with the volume of material available today, some form of automated recommendation is common. However, the balance of responsibility (intellectual vs automatic) for final indexing decision lies on a continuum that can vary and that can be adapted over time. All methods work on text documents, with NA traditionally applied to catalogues of different kinds. All tend to work on whole documents though can be applied to segments, with subject indexing sometimes prioritising abstracts as representing key aspects of a document and this was the approach followed in our previous archaeological work (Vlachidis and Tudhope 2016). ML approaches often work on a whole document basis (seen as an advantage in the MTI work). Output format can be inline markup and/or a metadata set (both formats were generated in our case study) usually employing PIDs if the output is vocabulary based. On the whole, subject indexing has tended to produce a set of subject metadata. NA is always based on vocabularies and subject indexing is often based on vocabularies (but can be keyword based), while NER is often not. ML methods tend not to use vocabularies, though see counter examples in the approaches review above. Regarding the entities targeted, NA and NER work with the specific set of entities that are given names in the domain, as with the examples above. Arguably, NA and NER entities may be more clearly distinguished from homonyms and different senses and perhaps more straight forward than subject indexing in some cases. Subject indexing vocabularies can be large and deep but tend to have a small set of top-level concepts (or facets in a faceted system). 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.

Perhaps the most distinguishing feature is the purpose for which the approach is applied to the document and this necessarily involves the relationship between the named entity (or concept) and the document and the ultimate end-use. Arguably, NER focuses on the immediate identification of a name; the relationship between text string and name is instanceOf, leaving determination of further purpose to the end-application. NA involves various relationships connecting works with persons and places but can also include the subject a document is about, which is the key focus for subject indexing. The aboutness relationship is a thorny topic within information science and certainly capable of being treated differently by different indexers or cataloguers. Aboutness has been discussed and distinguished from isness (similar to the instance relationship frequently used with ontologies) and ofness (e.g. for picture indexing),

for more information see Zeng et al. (2010) and Hjørland (2017). A detailed discussion of indexing strategies (exhaustivity and specificity are key characteristics) is provided by Golub et al. (2016), including issues of aboutness (with its contribution to relevance judgements) and the difficulty of evaluation.

4.2 Reflecting on the approach

Following the review of the three approaches in Section 4.1, in the case study we have employed an NLP NER pipeline, taking advantage of pre-processing elements such as tokenisation, lemmatisation, POS, albeit with a pattern/rule vocabulary-based strategy rather than the ML approach more common today. This can be considered a hybrid approach. The entities involved are more typical of subject indexing (and NA depending on the definition) and are arguably less clear cut in identification than some NER applications. NER would typically aim to identify (and annotate via markup) every occurrence of an entity instance in the document in question. Subject indexing (vocabulary based) would traditionally provide (vocabulary concept) subject metadata that best represents the aboutness of the document. These may be a subset of the concepts mentioned explicitly in the document and may include terms not even present in the document at all. In the case study, our strategy has been to select all vocabulary concepts present in the abstract of a document that match the rule patterns designed (applying entry vocabulary, using POS to help disambiguate homonyms, etc.). We do not have rules capable of assigning concepts based on a high-level understanding of the overall text (as would be the goal for intellectual subject indexing). Rather the automatic indexing outcomes are intended as suggestions, as part of a future interactive indexing utility, which would also aim to facilitate intellectual choice of vocabulary concepts not necessarily present in the text. Additionally, in future work we have the option to assign properties to the automatic annotation (suggestion), reflecting confidence or priority, and also to change priorities (or remove annotations) via post-processing rule-based filters (as seen in the MTI pipeline). The narrowing down of the source text to the abstract might yield reasonable results for some subject indexing cases but is very dependent on the writing of the abstract; the indexing outcomes are dependent on the guidelines for writing abstracts and the extent to which the guidelines are followed.

Future work may benefit from extending the scope of indexing to include the whole document and also the title. One strategy might be to identify every occurrence of the subject entities involved, using a frequency count to approximate relative importance. However, this does not reliably give an indication of significance as multiple occurrences can derive from listings or tables of common objects or from background sections that discuss previous work on the site or even nearby sites at other locations. In future work, the automatic identification of common categories of archaeological report sections would help to guide the focus of automatic indexing, although the task is made difficult by the fairly wide variation in style of report writing. Confidence values could be associated with the source text, in order to distinguish whether a subject entity derived from the abstract, the title, the whole document (or possibly particular sections of the document).

Another complication is the ultimate purpose of OASIS indexing. In fact, the purpose or indexing goal for OASIS goes beyond overall aboutness to address some of the FAIR principles discussed in the Introduction. The OASIS manual asks cataloguers to add keyword information (from the KOS listed in Section 2) on interesting or relevant objects and artefacts found during the project. Cataloguers are asked not to record all the different individual finds but to help the end-users understand the significant findings of the archaeological report (Evans and Gilham 2021). The keywords will form part of the ADS Library metadata to assist resource discovery. There is an option to flag a record as having no significant findings. Thus, the indexing strategy is specialised beyond general aboutness to encompass significance, intended to reflect the archaeological information needs of future searches for which the report might be considered relevant. Significance of course is a complex concept that will vary to some extent with particular archaeological research interests and priorities. Any overall significance of an archaeological intervention or excavation (for a given set of objectives) differs from the significance that may be attached to the individual finds or discoveries of a particular excavation and its report. For the purposes of this case study of automatic indexing, significance can be considered in terms of a combination of methods that will result in the prioritisation of particular findings of an archaeological report in the subject indexing metadata of an OASIS record.

Sociologically informed studies of technology have represented technical and social components as a seamless web rather than an orderly working out of user requirements. Developer strategies can be considered as attempts to stabilize a network of evolving prototypes, user expectations, requirements, and working practices (Tudhope et al. 2000). With such 'messy networks' in mind, future development planning and study of the recommendation system in trial use should incorporate broader contextual elements, including the guidelines provided and user practice both for indexing and for abstract/report writing, together with prototyped variations of search functionality in the retrieval system that seek to take advantage of the resulting subject metadata. The current OASIS indexing guidelines will tend to result in a fairly time-consuming task, one of the motivations for developing automatic indexing recommendation tools. In future work,

we intend to investigate contextual patterns reflecting significance and incorporate those patterns in the post-processing prioritisation filters discussed in Section 3.

5.0 Conclusions

Results have been reported from a case study of KOS-based automatic indexing recommendation techniques intended to support the entry of subject metadata to the OASIS archaeological archive. The FISH Archaeological Object Thesaurus, the FISH Thesaurus of Monument Types File and the Historic England Periods list were employed as part of pattern based NER rules, together with NLP techniques such as lemmatisation and POS tagging, on the abstracts from some 1600 OASIS metadata records. Other vocabularies (eg Materials) would also have been possible if considered useful for OASIS. The KOS-based NER functionality was adopted to this use-case and the results demonstrate that the approach is feasible. The NER patterns could be extended to accommodate other vocabularies, and local specialisation patterns could be created to handle idiosyncrasies of natural language.

Findings from the case study include the 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, such as Object - Period. Results could be fine-tuned by the incorporation of post-processing filters with confidence attributes to prioritise subject indexing significant for the document in question and to reduce the confidence attributes or rankings of common problematic cases as uncovered in evaluation. Negative results are important in archaeology and a negation detection capability should be a component of future recommendation tools. Consideration of the patterns in a document representing significance and also lack of significance are important.

Reflections on the case study experience touch on the complexity of evaluation in real life settings. The overall KOS-based NER techniques are discussed within the context of work on subject indexing, automatic indexing for Name Authorities and NER generally. The techniques followed in the case study can be characterised as a hybrid approach. The purpose for which the indexing is applied is a key distinguishing feature. In this case, the purpose or indexing policy for OASIS goes beyond overall aboutness; indexers are requested to include the significant objects or artefacts found in the archaeological fieldwork, thus reflecting FAIR principles for reuse. As discussed, the assessment of the future utility of new indexing tools is inherently complex in a world where user behaviour and subject indexing practice and guidelines all change over time and can be seen as an evolving complex network. Ideally, the (co)design of future best practice indexing policy and guidelines for writing abstracts can operate in tandem with the design of an automatic indexing recommendation system and corresponding search services.

References

Bates, Marcia. 1986. "Subject Access In Online Catalogs: A Design Model". *Journal of the American Society for Information Science* 37 no 6:357-76. https://doi.org/10.1002/(SICI)1097-4571(198611)37:6<357::AID-ASI1>3.0.CO;2-H

Binding, Ceri, Douglas Tudhope and Andreas Vlachidis. 2019. "A Study of Semantic Integration Across Archaeological Data And Reports in Different Languages". *Journal of Information Science* 45 no 3: 364-386. https//doi:10.1177/0165551518789874.

Binding, Ceri and Douglas Tudhope. "Automatic Normalization of Temporal Expressions". *Journal of Computer Applications in Archaeology* 6 no. 1: 24-39. https//doi: 10.5334/jcaa.105.

Crane, Gregory, and Alison Jones. 2006, "The Challenge of Virginia Banks: An Evaluation Of Named Entity Analysis In A 19th-Century Newspaper Collection". In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL'06),* Chapel Hill: ACM Press, 31-40. https//doi:10.1145/1141753.1141759.

Tudhope, Douglas, Paul Beynon-Davies and Hugh Mackay. 2000. "Prototyping Praxis: Constructing Computer Systems and Building Belief". *Human-Computer Interaction* 15: 353-383. https//doi: 10.1207/S15327051HCI1504_3.

Evans, Tim. 2015. "Reassessment of archaeological Grey Literature: Semantics and Paradoxes". *Internet Archaeology* 40. https//doi:10.11141/ia.40.6.

Evans, Tim and Jo Gilham. 2021. "OASIS V User Manual (V1.7)". https://docs.google.com/document/d/1vNnhFrgoURyOM5SLaKB_1JIHI6DjV4anykzVMKJPj_k/edit#heading=h.ck3klg4ce7v

Finkel, Jenny R., Trond Grenager and Christopher Manning. 2005. "Incorporating non-local information into information extraction systems by Gibbs sampling". In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05),* Ann Arbor, Michigan, 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf.

FISH. 2024. Forum on Information Standards in Heritage. http://www.heritage-standards.org.uk.

Golub, Koraljka, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Lykke and Debra Hiom. 2016. "A Framework for Evaluating Automatic Indexing Or Classification In The Context Of Retrieval". *Journal of the Association for Information Science and Technology* 67 no1:3-16. https://doi.org/10.1002/asi.23600

Golub, Koraljka. 2019. "Automatic Subject Indexing of text." *Knowledge Organization* 46 no.2: 104-121. Also available in ISKO Encyclopedia of Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli, 2019. URL: https://www.isko.org/cyclo/automatic.

HE Periods. 2024. Historic England Periods. https://historicengland.org.uk/listing/the-list/historic-periods/.

Heritage Data. 2024. Linked Data Vocabularies for Cultural Heritage. https://www.heritagedata.org/blog/vocabularies-provided/.

Hjørland, Birger. 2017. "Subject (of documents)". *Knowledge Organization* 44 no1: 55-64. https//doi:10.5771/0943-7444-2017-1-55.

Isaac, Antoine. ed. 2015. "Report on Enrichment and Evaluation". *Europeana Task Force on Enrichment and Evaluation*. https://pro.europeana.eu/project/evaluation-and-enrichments.

ISO 25964-1. 2011. "Thesauri and Interoperability with Other Vocabularies. Part 1: Thesauri for Information Retrieval". http://www.niso.org/schemas/iso25964/

Lancaster, Frederick W. 2003. *Indexing and abstracting in theory and practice,* 3rd ed., Champaign: University of Illinois.

Markey, Karen. 1984. "Interindexer Consistency Tests: A Literature Review and Report of a Test of Consistency in Indexing Visual Materials". *Library & Information Science Research* 6 no. 2:155-177.

McKeague, Peter. 2021. "A Snapshot in Time: A Review of Current Approaches to Archaeological Archiving in Scotland". *Internet Archaeology* 58. https//doi:10.11141/ia.58.14.

Mork, James, Alan Aronson and Dina Demner-Fushman. 2017. "12 Years On - Is the NLM Medical Text Indexer Still Useful and Relevant?". *Journal of Biomedical Semantics* 8 no. 8 https//doi:10.1186/s13326-017-0113-5.

Nadeau, David and Satochi Sekine. 2007. "A Survey of Named Entity Recognition And Classification". *Lingvisticae Investigationes* 30 no.1: 3–26. https//doi: 10.1075/LI.30.1.03NAD.

OASIS. 2024. Online System for Reporting Archaeological Investigations. https://oasis.ac.uk/.

Soergel, Dagobert. 1994. "Indexing and Retrieval Performance: The logical Evidence". *Journal of the American Society for Information Science* 45 no. 8: 589-599. https//doi:10.1002/(SICI)1097-4571(199409)45:8<589:AID-ASI14>3.0.CO;2-E.

Suominen, Osma. 2019. "Annif: DIY Automated Subject Indexing Using Multiple Algorithms." *LIBER Quarterly: The Journal of the Association of European Research Libraries* 29 no.1:1–25. https//doi:10.18352/lq.10285.

VIAF. 2024. The Virtual International Authority File. https://viaf.org/.

Vlachidis, Andreas and Douglas Tudhope. 2016. "A Knowledge-Based Approach To Information Extraction For Semantic Interoperability In The Archaeology Domain". *Journal of the Association for Information Science and Technology* 67, no.5: 1138-1152. doi:10.1002/asi.23485.

Vlachidis, Andreas and Douglas Tudhope. 2012. "A Pilot Investigation of Information Extraction In The Semantic Annotation Of Archaeological Reports". *International Journal of Metadata, Semantics and Ontologies 7* no.3: 222-235. https//doi:10.1504/IJMSO.2012.050183.

Wilkinson, Mark, Michel. Dumontier, Jan Aalbersberg et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3, 160018. https//doi:10.1038/sdata.2016.18.

Wolf, Stacey. 2020. "Automating Authority Control Processes." *Code4Lib Journal*, 47. https://journal.code4lib.org/articles/15014.

Zeng, Marcia, Maja Zumer and Athena Salaba, eds. 2010. "Functional Requirements for Subject Authority Data (FRSAD): a Conceptual Model." IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR). https://www.ifla.org/wp-content/uploads/2019/05/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf.

Acknowledgements

Parts of this work were supported by the ARIADNEplus project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823914. Thanks are also due to the OASIS team, including Tim Evans, Jo Gilham and Holly Wright. The views and opinions expressed in this article are the sole responsibility of the authors.