

KOS-based enrichment of archaeological fieldwork reports

Ceri Binding¹ and Douglas Tudhope¹

¹ *Hypermedia Research Group, University of South Wales, Pontypridd, CF371DL, UK*

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 Within the UK, the OASIS online index of fieldwork events and their unpublished reports represent a major initiative to make archaeological fieldwork available to a wider public.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 goes on to reflect 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.

Keywords

Automatic subject indexing, named entity recognition, knowledge organization systems.

1. 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 (e.g. [1]). The FAIR [2] Data Principles (Findable, Accessible, Interoperable, Reusable) entail re-use of data and key information in archaeological archives (see e.g. [3]). Semantic interoperability entails a key role in archaeological 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’ (see discussion in [4]), 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 ([5]). 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 ([6]) is encouraged.

ISKO UK conference, July 24–25, 2023, Glasgow, Scotland
EMAIL: ceri.binding@southwales.ac.uk; douglas.tudhope@southwales.ac.uk
ORCID: 0000-0002-6376-9613 (Binding); 0000-0002-5222-0430 (Tudhope)



© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

2. 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 FISH Archaeological Object Thesaurus and the FISH Thesaurus of Monument Types (SKOS versions [7]). The Historic England Periods PeriodO Authority [8] provided a source for identifying named periods, together with some regular expression patterns to identify temporal expressions, such as English century and year span expressions (for a report on work with a temporal expression normalization tool in different languages, see [9]). The STELETO tool [10] 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 on 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 (lists and HTML markup). Example results (HTML output) are illustrated in Figure 1 and Figure 2.

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**.

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

3. 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 ([11], ch7) 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.

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**.

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

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)*). Enhanced entry vocabularies may also have wider value for the FISH KOS. 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 [12] proposal of an ‘end-user thesaurus’ where a large entry vocabulary includes a wide variety 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. 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 [13]. This could form part of a post-processing set of filters that tackled negation expression and common problematic cases found in evaluation, where it was considered appropriate to remove a suggested subject indexing concept or reduce confidence values.

4. 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 [14]; 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 according to the indexing policy ‘the machine is right and the human (annotator) is wrong’ (see also [15], section 2.5). However, 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 in [16], [17]). 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 (see e.g. [18]; [19]; [16]) and differences in the overall policy guiding the indexing. Soergel [20] 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) policy we follow, following the FAIR principles and OASIS deposit guidelines, as discussed in the Introduction, 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 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” [21]. 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 [22] 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 [23] 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 (see e.g. [24]).

NER is considered a sub-task of information extraction within NLP (see e.g. [25]). 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 [26]. 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 [17] 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 [27]. 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 [28]. 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 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 [13]. ML approaches often work on a whole document basis (seen as an advantage in 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), see e.g. [21] and [15]. A detailed discussion of indexing strategies (exhaustivity and specificity are key characteristics) can be found in [16], 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. Our approach, in other words, is hybrid. 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 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 an option for intellectual subject indexing). Rather the automatic indexing outcomes are intended as *suggestions*, as part of a future interactive indexing utility, which would also 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 policy is dependent on the guidelines for writing abstracts.

Future work may benefit from extending the scope of indexing to the whole document. 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 wide variation in styles of report writing.

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 end-users understand the significant findings of the archaeological report [29]. The keywords will form part of the ADS Library metadata for 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 would be considered relevant.

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 [30]. 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 techniques. 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. 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 PeriodO Authority were employed as part of pattern based NER rules, together with NLP techniques such as lemmatisation and POS tagging, on abstracts from some 1600 OASIS metadata records. 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. Results could be fine-tuned by the incorporation of post-processing filters to prioritise subject indexing significant for the document in question and to reduce confidence attributes or rankings of common problematic cases as found in evaluation. Negative results are important in archaeology and negation detection capability would be an important component of recommendation tools.

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 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, 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.

6. 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.

7. References

- [1] A. Isaac et al., Report on enrichment and evaluation, Europeana Task Force on Enrichment and Evaluation, 2015. URL: <https://pro.europeana.eu/project/evaluation-and-enrichments>.

- [2] M. Wilkinson, M. Dumontier, I. Aalbersberg et al, The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data* 3, 160018. (2016). doi:10.1038/sdata.2016.18.
- [3] P. McKeague, A snapshot in time: a review of current approaches to archaeological archiving in Scotland, *Internet Archaeology* 58 (2021). doi:10.11141/ia.58.14.
- [4] T. Evans, A Reassessment of archaeological grey literature: semantics and paradoxes, *Internet Archaeology* 40 (2015). doi:10.11141/ia.40.6.
- [5] OASIS, Online system for reporting archaeological investigations, 2024. URL: <https://oasis.ac.uk/>.
- [6] FISH, Forum on Information Standards in Heritage, 2024. URL: <http://www.heritage-standards.org.uk>.
- [7] Heritage Data, Linked Data Vocabularies for Cultural Heritage, 2024. URL: <https://www.heritagedata.org/blog/vocabularies-provided/>.
- [8] HE Periods, Historic England Periods PeriodO Authority, 2014. URL: <http://n2t.net/ark:/99152/p0kh9ds>.
- [9] C. Binding, D. Tudhope, Automatic normalization of temporal expressions, *Journal of Computer Applications in Archaeology* 6(1) (2023) 24-39. doi:10.5334/jcaa.105.
- [10] C. Binding, D. Tudhope, A. Vlachidis, A study of semantic integration across archaeological data and reports in different languages, *Journal of Information Science* 45(3) (2019) 364-386. doi:10.1177/0165551518789874.
- [11] ISO 25964-1, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval 2013. URL: <http://www.niso.org/schemas/iso25964/>.
- [12] M. Bates, Subject access in online catalogs: a design model, *Journal of the American Society for Information Science* 37(6) (1986) 357-376. doi:10.1002/(SICI)1097-4571(1986)37:6<357::AID-ASII>3.0.CO;2-H.
- [13] A. Vlachidis, D. Tudhope, 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. doi:10.1002/asi.23485.
- [14] A. Vlachidis, D. Tudhope, A pilot investigation of information extraction in the semantic annotation of archaeological reports, *International Journal of Metadata, Semantics and Ontologies* 7(3), (2012) 222 - 235. doi:10.1504/IJMSO.2012.050183.
- [15] B. Hjørland, Subject (of documents), *Knowledge Organization* 44 (1) (2017) 55-64. doi:10.5771/0943-7444-2017-1-55.
- [16] K. Golub, D. Soergel, G. Buchanan, D. Tudhope, M. Lykke, D. Hiom, A framework for evaluating automatic indexing or classification in the context of retrieval, *Journal of the Association for Information Science and Technology* 67(1) (2016) 3-16.
- [17] K. Golub, Automatic subject indexing of text. *Knowledge Organization* 46 (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>.
- [18] F. Lancaster, *Indexing and abstracting in theory and practice*, 3rd ed., Champaign: University of Illinois, 2003.
- [19] K. Markey, Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials, *Library & Information Science Research* 6(2) (1984) 155-177.
- [20] D. Soergel, Indexing and retrieval performance: The logical evidence, *Journal of the American Society for Information Science* 45(8) (1994) 589-599. doi:10.1002/(SICI)1097-4571(199409)45:8<589::AID-ASII4>3.0.CO;2-E.
- [21] M. Zeng, M. Zumer, A. Salaba, (Eds.) *Functional Requirements for Subject Authority Data (FRSAD): a conceptual model*, IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR) 2010. URL: <https://www.ifla.org/wp-content/uploads/2019/05/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf>.
- [22] G. Crane, A. Jones, 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)*, ACM Press, Chapel Hill, 2006, pp 31-40. doi:10.1145/1141753.1141759.
- [23] VIAF, The Virtual International Authority File, 2023. URL: <https://viaf.org/>.

- [24] S. Wolf, Automating authority control processes, *Code4Lib Journal*, 47. 2020. URL: <https://journal.code4lib.org/articles/15014>.
- [25] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, *Linguisticae Investigationes* 30 (1) (2007) 3–26. doi: 10.1075/LI.30.1.03NAD.
- [26] J. Finkel, T. Grenager, C. Manning, 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, 2005, pp. 363-370. URL: <http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf>.
- [27] O. Suominen, Annif: DIY automated subject indexing using multiple algorithms, *LIBER Quarterly: The Journal of the Association of European Research Libraries* 29(1) (2019) 1–25. doi:10.18352/lq.10285.
- [28] J. Mork, A. Aronson, D. Demner-Fushman, 12 Years On - Is the NLM Medical Text Indexer still useful and relevant?, *Journal of Biomedical Semantics* 8 (8). (2017). doi:10.1186/s13326-017-0113-5.
- [29] T. Evans, J. Gilham, *OASIS V User Manual (V1.7)*, 2021. URL: https://docs.google.com/document/d/1vNnhFrgoURyOM5SLaKB_1JIHI6DjV4anykzVMKJPj_k/edit#heading=h.ck3klg4ce7v.
- [30] D. Tudhope, P. Beynon-Davies, H. Mackay, Prototyping Praxis: Constructing Computer Systems and Building Belief, *Human-Computer Interaction* 15 (2000) 353-383. doi: 10.1207/S15327051HCI1504_3