A SEMANTIC SPACE FOR ARCHAEOLOGY:

TOWARDS DESCRIPTIVE ONTOLOGIES FOR ARCHAEOLOGY

by Georgia Burnett

THE UNIVERSITY OF
NEW SOUTH WALES

SYDNEY • AUSTRALIA

Submitted to the School of Humanities, the University of New South Wales, in partial fulfilment of the requirements of the Bachelor of Arts with Honours in History

ACKNOWLEDGEMENTS

To Corinne,
my thesis buddy.

Many thanks to Dr. Brian Ballsun-Stanton,
without whom, this thesis would have remained theory.

Also my supervisors,
Drs. Shawn Ross and Adela Sobotkova,
for their contributions and support throughout the year.

And my family and friends,
for putting up with me.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 2

TABLE OF CONTENTS 3

TABLE OF FIGURES 6

ABSTRACT 7

CHAPTER ONE: ARCHAEOLOGY AND THE DIGITAL HUMANITIES 15

1.1 THE CRISIS OF DATA ACCESSIBILITY 15

1.2 SYNTHESIS IN ARCHAEOLOGY AND DATA SHARING 18

1.3 STANDARDS IN ARCHAEOLOGICAL DATA MANAGEMENT 30

1.3.1 DATA STANDARDS IN DIGITAL ARCHAEOLOGY 32

1.3.1.1 POSSIBILITIES AS A RESULT OF DATA STANDARDS 34

1.3.1.2 PROBLEMS AND "COSTS" RELATED TO DATA STANDARDS 37

1.3.2 ANALYSIS OF MULTIPLE: ARCHAEOLOGICAL DATABASES 41

1.3.2.1 THE ONLINE ACCESS TO THE INDEX OF ARCHAEOLOGICAL INVESTIGATIONS (OASIS) PROJECT 41

1.3.2.2 THE DIGITAL ARCHAEOLOGICAL RECORD (TDAR) 44

1.3.2.3 THE VICTORIAN HERITAGE DATABASE 46

1.4 CONCLUSIONS 47

CHAPTER TWO: CONTENT STANDARDS AND APPLIED TEXT ANALYSIS IN ARCHAEOLOGY 49

2.1 CONTENT STANDARDS IN ARCHAEOLOGY 50

2.1.1 CONTROLLED LISTS 51

2.1.1.1 CASE STUDY: THE IFRAO GLOSSARY 52

2.1.2 THESAURI 56

2.1.2.1 CASE STUDY: THE ENGLISH HERITAGE THESAURI 59

2.1.3 ONTOLOGIES 61

2.1.3.1 TDAR'S ONTOLOGY FOR METADATA 64

2.1.3.2 THE CIDOC-CRM ONTOLOGY 65

2.3 (COMPUTER-ASSISTED) TEXT ANALYSIS 67

2.2.1 LATENT SEMANTIC ANALYSIS 72

2.3 APPLIED TEXT ANALYSIS IN ARCHAEOLOGY 76

2.3.1 THE ARCHAEOTOOLS PROJECT 77

2.3.2 THE ARMADILLO PROJECT 79

2.4 CONCLUSIONS 80

CHAPTER THREE: PRELIMINARY ANALYSIS OF A ROCK ART CORPUS 82

3.1 THE CORPUS 82

3.1.1 CORPUS GENERATION AND DOMAIN 83

3.1.2 CORPUS CONVERSION AND CLEANSING 84

3.2 CONTEXTUAL PROCESSING 86

3.2.1 CORPUS SIZE RELATIVE TO OTHER CORPORA 86

3.2.2 GOOGLE NGRAM VIEWER 2.0 87

3.2.3 WORD CLOUD GENERATION 93

3.3. SOFTWARE INFRASTRUCTURE 96

3.3.1 TARGET TERM LISTS 96

3.3.2 SEMANTIC VECTORS 97

3.3.2.1 RANDOM PROJECTION AND POSITIONAL INDEXING 99

3.4 RESULTS 103

3.4.1 COMPARING "KNOWN" SYNONYMS 104

3.4.2 BATCH TERM VECTOR COMPARISONS 106

3.4.3 SEMANTIC LINKS REGARDING BIGRAM TERMS 108

CONCLUSION 110

APPENDIX I: WORD CLOUD 114

APPENDIX II: BIGRAM GENERATION 115

APPENDIX III: BIGRAMS 120

APPENDIX IV: COMPARING RELATED TERMS 129

APPENDIX V: BATCH VECTOR COMPARISONS 131

APPENDIX V: BIGRAM VECTORS 137

REFERENCES 147

TABLE OF FIGURES

Figure A: An ontological representation of the television show, The Simpsons.[1] *62*

Figure B: A possible representation of the class/slot feature of ontologies. *63*

Figure C: An ontological representation of "artefact scatter." *64*

Figure D: An example how the SVD algorithm sums vectors. *75*

Figure E: An illustration of how faceted classification works.[2] *78*

Figure F: Ngram graph representing the use of "rock art", "rock marking" and "rock image" as a percentage in the Google corpus over time.[3] *90*

Figure G: Ngram graph representing the use of "rock marking" and "rock image" as a percentage in the Google corpus over time.[4] *90*

Figure H: Ngram graph representing the use of "pictogram", "pictograph" and "rock painting" as a percentage in the Google corpus over time.[5] *91*

Figure I: Ngram graph representing the use of "Petroglyph", "rock carving" and "rock engraving" as a percentage in the Google corpus over time.[6] *92*

Table A: Significant cosine similarities and their respective terms. *108*

[1] Reproduced from: "Protege," <http://www.upriss.org.uk/awt/s12.html> (Accessed: 4 October, 2013).

[2] "Archaeotools: Data mining, facetted classification and E-archaeology," Archaeology Data Service, <http://archaeologydataservice.ac.uk/research/archaeotools> (Accessed: 4 October, 2013).

[3] Created via Google Ngram Viewer: Jean-Baptiste Michel et al., "Quantitative analysis of culture using millions of digitized books," *Science* 331, no. 6014 (2011).

[4] Ibid.

[5] Ibid.

[6] Ibid.

ABSTRACT

Archaeology suffers from a disconnect between datasets. Without interoperable data, projects remain local, relevant only to a pocket of discussion. Interoperable data is only achievable through standardisation. Metadata has been favoured in archaeology as a means of capturing datasets; however, metadata cannot facilitate the required level of interoperability needed when performing large-scale automated comparisons. Content standards—specifically, standardised vocabularies to ensure semantic interoperability—are needed. Ontologies are the best representation for standardised vocabularies in archaeology. Ontologies not only quantify the term in a hierarchy, but also allow for the term to exist over several layers of the network.

Not only does archaeology require content standards, however; the standards that are designed must reflect the natural language of the discipline. They must reflect the nuanced vocabularies of archaeologists in practice. Text analysis provides a viable way of producing a natural language ontology. This thesis makes the first steps towards a natural language ontology by determining relationships between related terms. Positional indexing, a derivative of latent semantic analysis, provides a method of determining relationships between terms in a domain-specific corpus of literature. By doing so, it ensures an ontology reflective of the domain. Synonym extraction for ontology development is considered a key feature of future research.

INTRODUCTION

"It is only in recent years…that an awareness of the need for ontologies, controlled vocabularies and agreed data structures has emerged. There is little doubt that such goals will not be easily accomplished, but they are an aim worth aspiring to."[7]

Enhancing research data for interoperability and synthetic research has a huge impact on the wider research community; meaningful comparisons across multiple and distinct data sets can improve research productivity on a phenomenal scale.[8] Aside from data that are dependent on superlatives for their legitimacy (for example, first, oldest, or largest), data relating to individual projects alone can only rarely be related to questions outside of local or restricted interest.[9] Obstacles inhibiting efforts by scholars engaged in synthetic research block interdisciplinary efforts to recognise phenomena on large spatio-temporal scales, or conduct meaningful comparative studies.[10] In terms of long-term social and socio-ecological dynamics in world history and prehistory,

[7] Julian D. Richards and Catherine Hardman, "Stepping Back from the Trench Edge: An Archaeological Perspective on the Development of Standards for Recording and Publication," in *The Virtual Representation of the Past*, ed. M. Greengrass and L. Hughes (Farnham, Surrey: Ashgate Publishing Company, 2008): 188.

[8] Keith W. Kintigh, Francis P. McManamon, and Katherine Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)" (paper presented at the Towards a Data Standard for Paleolithic Archaeology, 78th Annual Meeting of the Society for American Archaeology, Honolulu, Hawaii, 3-7 April 2013), 1.

[9] Ibid.

[10] Ibid., 2.

archaeological data has fantastic potential to contribute to long-term studies. However, due to the complex nature of the data itself and the alarming lack of data interoperability—by any universal standard—the data remains unused and, at worst, unusable.

The obstacles facing data sharing are numerous. From concerns about quantity and quality, to practical questions of how data should be shared—how it should be digitally formatted, even how contextual information (metadata, when it exists) should be structured—ineffective data sharing has stunted the success of scholars involved in synthetic research. Inadequately defined, or non-existent, data standards are a primary cause of data absence.[11] "The all-too-frequent result [of ineffective or non-existent data standardisation]," writes Nelson, "is fragmented and often mutually incomprehensible scientific information."[12]

In order for archaeological data to have relevance in future academic endeavours, data produced in the field must be commensurate; in other words, it must be consistent, interoperable, and easily retrieved and reconciled with other data.[13] At present, a large portion of modern archaeological research runs the risk of falling quickly into obscurity or obsolescence due to inadequate measures to ensure data preservation.[14] An exception can be made only for those whose

[11] Bryn Nelson, "Empty Archives," *Nature* 46 (2009): 161.

[12] Ibid.

[13] For an outline of ontologies see chapter two, section 2.1.3.

[14] For a discussion on adequate measures for data preservation, see chapter one, section 1.2.

work have them labelled as a *cause célèbre*, in that their conclusions must be revisited in the light of more recent evidence.[15] Data that is interoperable allows archaeological data to exist in the context of the whole past, rather than an isolated on the spectrum.[16] It is the proposal of this paper that an archaeological ontology integrated into a field application is the most viable way to ensure the continuity of archaeological data.

The material cultures archaeologists engage with are not easily categorised or defined; they do not fit easily into objective taxonomic labels. Indeed, one of the primary tasks of archaeology is to categorise material cultures in a way that is easily approachable, without running the risk of indiscriminate labelling.[17] Therefore, archaeological data not only reflects the objective realities of the objects but the subjective realities of the investigators.[18] However, discarding interpretation from the data refutes the core of archaeological practice; to reconcile subjective interpretation and objective fact the relational nature of archaeology must be recognised and made distinct. Archaeological data

[15] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 2.

[16] A. Baines and K. Brophy, "What's Another Word for Thesaurus?: Data standards and classifying the past," in *Digital Archaeology: Bridging Method and Theory*, ed. T. L. Evans and P. Daly (London: Routledge, 2006): 238.

[17] Richards and Hardman, "Stepping Back from the Trench Edge," 169.

[18] A. Baines and K. Brophy, "Archaeology without -isms," *Archaeological Dialogues* 13 (2006): 69.

must exist both with and without subjective interpretation for the aspirations of data preservation to be fully realised.[19]

Ontologies provide a viable way for these two realities to coexist; using text analysis software, this paper will show that one can retrieve a descriptive, reclaimed list of terms associated from existing works, and map this to a core ontology. An ontology, for the purpose of this paper, is "the representation of meaning of terms in vocabularies and the relationships between those terms."[20] Supplementary to this, if personal idiosyncrasies regarding terminology are mapped to a core standard at data creation, this method allows for a more dynamic, versatile bank of natural language terms. This method has the intention of promoting vocabulary growth and academic freedom of expression.[21] The primary aim of this paper is to prove text analysis as the best device for developing archaeological ontologies; ontologies allow for datasets to retain their individual schemas and terms, while simultaneously being able to conform multiple datasets to a interoperable standard.

Controlled vocabularies, as a means of standardising content, is a contentious topic in archaeology. Primarily amongst these, is the lack of understanding regarding problems with the absence of agreed data standards;

[19] Baines and Brophy, "What's Another Word for Thesaurus?," 240.

[20] Deborah L. McGuinness and Frank van Harmelen, "OWL Web Ontology Language Overview," W3C, <http://www.w3.org/TR/owl-features/> (Accessed: 9 September, 2013).

[21] The nature of unique sites and "low hanging fruit" syndrome are two major concerns for data standards in academic archaeology: see chapter one, section 1.3.1.2.

the problem is everybody's, therefore it is nobody's.[22] While there is an awareness of the issue, it is so pervasive—almost ingrained in the very fabric of the discipline—that it is difficult to know where to start. Some have argued that is the very nature of archaeological research—a drive for reinterpretation, for reinvestigation—that allows for multiplicities in vocabularies.[23] If a controlled vocabulary is to be created, it needs to reflect the evolution of the discipline in real time.

Text analysis allows for the ontology to be developed in a way that reflects the natural language of the discipline. Natural language is a key variable when considering controlled vocabularies and data structures; several projects in the past have failed because of their lack of understanding of how vocabulary is deployed in reality. These failings rest on ascribing a bank of *prescriptive* vocabulary standards, rather than *descriptive* standards; that is, standards that are prescribed to the user independently, rather than descriptively retrieved from existing literature or research material, thus reflecting the natural language norms of the discipline.[24]

Using text analysis to detect synonyms and concept mapping, it is possible to construct a set of terms that reflect the reality of the vocabulary of practitioners. The primary aim of this thesis is to present a method for the use of

[22] Nelson, "Empty Archives," 162.

[23] W. Fredrick Limp, "Web 2.0 and Beyond, or On the Web Nobody Knows You're an Archaeologist," in *Archaeology 2.0: New Tools For Communication and Collaboration*, ed. Eric C. Kansa (Cotsen Institute of Archaeology, 2011): 277.

[24] For a discussion on these projects, see chapter one, section 1.3.1.

text analysis in the archaeological domain to aid in the construction of ontologies.[1] As this is a preliminary experiment, the data is not considered complete enough to attempt a full ontological representation of the chosen specialisation.

[1] See chapter three for an detailed outline of this method.

CHAPTER ONE: ARCHAEOLOGY AND THE DIGITAL HUMANITIES

"The practice of archaeology is not as objective as fieldworkers would like to believe; nor is it as subjective as theorists often suppose."[1]

This chapter will outline the place of archaeology in the realm of the digital humanities. Data availability and discoverability is the core problem in regards to archaeological data management, and its roots lay in the theory and practice of archaeology as a discipline. Data availability and discoverability is the primary challenge faced by scholars engaged in research synthesis. Efforts have been underway for some time in archaeology to digitise data and standardise metadata, in an attempt to address the problem of data availability and discoverability. Even when primary datasets are available and discoverable, however, a lack of data content standards limits their compatibility. This chapter presents the problem of terminology in archaeology, and introduces computational text analysis as a solution to producing frameworks improving the interoperability of archaeological data.

1.1 THE CRISIS OF DATA ACCESSIBILITY

When it comes to data synthesis, the first obstacle is the lack of access to primary data. The core of this data problem lies in practices surrounding of archaeological data management and the theory behind them. The statement

[1] Richard Bradley, *The Significance of Monuments: On the Shaping of Human Experience in Neolithic and Bronze Age Europe* (London: Routledge, 1998): 2.

presented above by Richard Bradley, a professor of prehistoric archaeology at the University of Reading, encompasses the apparent schism that is appearing in the discipline of archaeology.

The "gulf between practice and theory," as Baines and Brophy refer to it, has appeared to many archaeologists to be a result of the discontinuity in archaeological theory.[1] In spite of Bradley's comment, theory emphasises a need for subjective assessment of archaeological data, while the publications resulting from fieldwork often accentuate conclusions of the investigators as objective fact. However, the conflict between objective and subjective approaches illustrates one basis of the problem of data accessibility in archaeology. Archaeological approaches vary widely, and individual projects are often driven by specific research questions or hypotheses. These research questions then determine the methods used during fieldwork, including data management.

Traditional venues for the dissemination of archaeological information – hard-copy publications – also limit the availability of data. Due to space and cost constraints, comprehensive datasets are rarely published. Instead, data is often "cherry picked" for its intrinsic interest (for example, only the high-quality decorated ceramics are fully published, while plain wares are ignored), or its relevance to a project's research agenda or hypothesis. Thus, data entry into databases becomes a subjective exercise. While scholars consider all relevant data as their complete dataset, this is not always the case.[2] The data is given

[1] Baines and Brophy, "Archaeology without -isms," 69.

[2] Jon Holmen, Christian-Emil Ore, and Øyvind Eide, "Documenting two histories at once: Digging into archaeology" (paper presented at the Beyond the artefact: Digital

significance based on the research requirements of the investigator, and not on objective merit; data that could be considered important for other studies of similar phenomena might be shelved in favour of more relevant analysis. This selective process creates a feedback loop where the "correct" answers are published with only the data necessary to support them, regardless of the problematic nature through which the data was collected or selected for presentation, or the existence of any opposing evidence.[3]

If comprehensive datasets were available, the misconceptions born of published interpretations would not be a significant problem. The availability of primary datasets would provide researchers with the means to evaluate interpretations, and otherwise reuse, repurpose or reinterpret the data. Although archaeology has a long tradition of active fieldwork and publication, copy datasets are rarely comprehensive due to selective publication of the data. These publications often do not appear for many years after the completion of fieldwork. Furthermore, the majority of works from the past two decades are "published via short-run journals or not published at all beyond a typescript report lodged at with the local planning authority."[4] Often, these publications are

---

Interpretation of the Past, Proceedings of CAA 2004, Budapest, Hungary, 13-17 April 2004), 1.

[3] Linda E. Patrik, "Is There an Archaeological Record?," *Advances in Archaeological Method and Theory* 8 (1985): 27-28.

[4] Stuart Jeffrey et al., "The Archaeotools Project: faceted classification and natural language processing in an archaeological context," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367, no. 1897 (2009): 2508.

not widely available. While comprehensive datasets are preferable, the lack of digital data availability is even more concerning.

Online dissemination of digital data has great potential to redress these problems. Despite the IT revolution and the introduction of digital archives for the storage and management of data, only a small minority of archaeological data is made widely available digitally. In most cases the problem lies with individuals or projects. They do not prepare digital datasets, describe them adequately with metadata, or submit them to appropriate online repositories.[5] Within archaeology the problem is reaching a crisis point; archaeological data, as rightly maintained by a majority of archaeologists, is irreplaceable. "Once we have excavated a context," writes Keith Kintigh, an advocate of data sharing in archaeology, "it is gone."[6] If primary data is not successfully shared and described for future use, there is only the conclusion drawn by the excavators; missing data does not bode well for those engaged in projects that require data synthesis or reinterpretation.

1.2 Synthesis in Archaeology and Data Sharing

The search for patterning in archaeological data is key in archaeological research.[7] Pattering, as Julian Richards defines it, is "the repeated and consistent

[5] Nelson, "Empty Archives," 160.

[6] Keith W. Kintigh, "The Challenge of Archaeological Data Integration," in *Technology and Methodology for Archaeological Practice: Practical applications for the past reconstruction* (British Archaeological Reports, 2009): 3.

[7] Julian D. Richards, "From anarchy to good practice: the evolution of standards in archaeological computing," *Archeologia e Calcolatori* 20 (2009): 27.

re-occurrence of a range of attributes" that are essential to categorise and describe artefact types, typologies and classifications.[8] Archaeology relies on recoverable patterns in the archaeological record. Patterns allow for researchers to recognise where societies have adapted, "including the internal organisation of that system whereby goods, energy or information are stored and circulated."[9] Archaeology has moved away from identifying unique social systems in terms of pre-existing social institutions in recent decades, and patterning is not without its critics.[10] However, the acknowledgement of broader social and material patterns remains important to the discipline.

For archaeology to reach its fullest potential as a scientific endeavour, and to be able to make substantive predictions about the unknown, the identification of patterning and other facets of archaeological research must be shared on a broad scale, rather than just locally. The identification of patterning over multiple datasets requires that the primary data in those datasets be interoperable. Interoperability is the ability to exchange and make use of multiple datasets, usually by conforming data to a shared technical and content standard. There are three types of interoperability: technical, syntactical and semantic.

[8] Ibid., 29.

[9] John C. Barrett, "Fields of Discourse: Reconstituting a Social Archaeology," *Critique of Anthropology* 7, no. 5 (1988): 6.

[10] See: Barrett, "Fields of Discourse: Reconstituting a Social Archaeology," 7-8.

Technical interoperability exists on the level of hardware. It refers to the degree of compatibility between computer systems.[11] Syntactic interoperability refers to the use of compatible data formats (audio or video, for example), so the information systems holding the data can exchange it. [12] Semantic interoperability concerns the exchange of shared, unambiguous information about data content between computer systems.[13]

Data interoperability offers a number of prospects for archaeology, and is not limited to new or recently created datasets. Given the nature of archaeological finds, data must be collected and stored appropriately. As a result, many technical platforms have been used, and subsequently fallen out of favour over time; an abundance of legacy data is the result. In IT, legacy data refers to any data stored on hardware or software that has become obsolete or replaced with newer technology.[14] In archaeology, the term is more nuanced—it refers to any pre-existing data used by a project that that project itself did not produce. If data interoperability is encouraged and data is given adequate metadata and mapping, legacy data has a chance of long-term preservation. Legacy data interoperability could stem the ongoing loss of irreplaceable archaeological data to time. Data integration would also expand international access to legacy data

[11] Richards, "From anarchy to good practice," 28.

[12] Herbert Kubicek, Ralf Cimander, and HansJochen Scholl, "Layers of Interoperability," in *Organizational Interoperability in E-Government* (Springer Berlin Heidelberg, 2011): 87.

[13] Ibid., 89.

[14] "Legacy data," <http://www.businessdictionary.com/definition/legacy-data.html> (Accessed: 31 August, 2013).

on a local scale, and allow for integration and synthesis on larger spatio-temporal scale. Enhanced access will give the data larger significance in interdisciplinary studies.[15]

For the purposes of archaeological data sharing, syntactic and technical interoperability has mostly been achieved by the broader IT marketplace.[16] Semantic interoperability is hindered by a number of factors, however. As James Boyle, a founding member of Creative Commons,[17] commented:

> Researchers generally create their own formats because they believe that they know how their users want to use the data…[But there are roughly a billion people with Internet access] and at least one of them has a smarter idea about what to do with your content than you do.[18]

The biggest problem faced by semantic interoperability arises from individualised research agendas with tailored field practices producing idiosyncratic datasets. Larger patterns of variation also exist, however. As an example, archaeological datasets are often specific (in terms of terminology, chronological periods, typical field practices, etc.), to individual jurisdictions or geographic regions.[19]

---

[15] Kintigh, "The Challenge of Archaeological Data Integration," 7-8.

[16] Richards, "From anarchy to good practice," 28.

[17] Creative Commons is a non-profit organisation for the promotion of creative content sharing.

[18] Nelson, "Empty Archives," 161.

[19] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 1.

The variables chosen by a project, such as unique artefact taxonomies, and inconsistent collection intensities (due to a number of factors, such as project constraints and variable durability of artefacts), can all affect the consistency of data. In turn, variables effect the ability to integrate and compare disparate data. For example, any long-term socio-environmental study "requires long-term plant, animal, and human subsistence data at spatial scales not available in any single study."[20] An assumption of demographic trends on a regional scale underlie many arguments about this type of change, but "assembling reliable population estimates requires consistent integration of data that may span a century of archaeological research."[21] If data is tailored to suit the needs of each project, interoperability cannot be achieved and access to the data is blocked by semantic incompatibility.[22]

Datasets can also come to varying conclusions—drawn from identical primary data—not because of error, but because they employ interpretive processes differently. For example, in Southwest American archaeology, more than one procedure is used to defend estimated population numbers as they are predicted from the number of residential rooms or total floor area of a dwelling. When analysts apply different criteria (such as opposing models for estimating population sizes) to certain data, as David Doyle argues, results can lead to

[20] Ibid., 2.

[21] Ibid.

[22] Ibid., 1.

widely varying population estimates.[23] Kintigh suggests that this deviation means that there are "even more subtle difficulties [to data interoperability] than those deriving from out-dated or erroneous conclusions [as made by primary investigators]."[24] Thus, even when primary data is available, interpretative process can deviate results. If these semantics are not explicitly outlined, or the process through which the conclusion was made is not available, conclusions that are arguable can appear as facts.

The problem of data interoperability, however, works on the presumption that comprehensive primary datasets are widely available to those engaged in synthetic research. Yet they are not:

> …scholars engaged in synthetic research rarely compare their data-driven interpretations with data recovered by other archaeological projects, but rather with the conclusions drawn by the investigators of these projects. The pernicious effect is that erroneous conclusions based on inconsistent premises become entrenched in the literature as "facts" that continue to serve as faulty premises in scientific arguments.[25]

Without access to comprehensive, primary datasets, researchers involved in synthesis rely on published interpretations and the selective datasets underlying

[23] David E. Doyle, *Lake Hohokam prehistory in Southern Arizona* (Scottsdale: Gila Press, 1981); In Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 2.

[24] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 2.

[25] Ibid.

them. Such data may not be representative and interpretations can be difficult to re-evaluate: often "published" eventually equates to "fact", given enough time. Without complete datasets, even the "objective" facts are problematic; their validity cannot be confirmed.

The absence of primary datasets presents a problem for research communities. Daniel Garner notes that asking researchers for primary data is problematic because, "…it's a lot of work and because in many cases the databases don't exist for it. So there is kind of a chicken and egg problem here."[26] That is, until the infrastructure exists, it is difficult to produce sharable, comprehensive datasets; but without the existence of such datasets and a culture of sharing, it is difficult to justify the production of the infrastructure. "Infrastructure is the thing that we always fail to fund," writes Boyle, "because it's kind of everybody's problem, and therefore it's nobody's problem."[27]

Multidisciplinary researchers put the responsibility for data sharing on the shoulders of three groups: "the funding agencies, which can demand data sharing in return for support; the [relevant disciplinary] societies, which can establish it as a precedent; and the journals, which can make sharing a condition of publication." [28] Without requirements to share primary datasets, archaeologists cannot view the possibilities of interoperable data outside their own domain of research; thus, data incompatibility compounds as a disciplinary problem.

[26] Nelson, "Empty Archives," 162.

[27] Ibid.

[28] Ibid., 161.

While the absence of data sharing as the final step of the data life cycle is in itself problematic, it must be limited in in the scope of this paper. The lack of primary data is an obstacle; however, synthesis can still be achieved, but will have disputable foundations.[29]

However, issues with combining heterogeneous data sources are not without consideration; semantic interoperability is equally important, and is the focus of this thesis. It is important to note first that archaeology as a discipline offers incentives for those who deviate from traditional or established conventions. As Frederick Limp notes,

> …archaeological scholarship provides a powerful *dis*incentive for participation in the development of semantic interoperability and, instead, privileges the individual to develop and defend individual terms/structures and categories.[30]

Disincentives for standardised research are a major problem for synthesis, interoperability and data standardisation. Strong incentives exist, especially increased prestige, for those who offer a unique perspective on a issue once considered closed. For example, revisions may come in the form of new evidence suggesting alternative interpretation, new fieldwork informed by better methods or techniques, or indeed new vocabularies that "better suit" the problem. This attitude of awarding individuality makes for a powerful disincentive for

[29] Kintigh, "The Challenge of Archaeological Data Integration," 1.

[30] Limp, "Web 2.0 and Beyond," 277.

standardised datasets.[31] Subjective interpretation is considered by some to be more important than the raw data itself. Baines and Brophy, for example, argue that "subjectivity, experience, agency and empathy are presented as the preferred subjects of archaeological enquiry."[32] With subjective conclusions based on individual research goals being made paramount, publications often reflect only one way of thinking about the dataset. However, without the primary data being made available, this way of thinking becomes the *only* way of thinking about the dataset. When it comes to synthesis in archaeological research, scholars usually have to rely on the *conclusions* and *interpretations* made by investigators, rather than primary data itself.

The lack of primary datasets comparison can lead to "faulty inferences" being entrenched as facts. Considering that these errors are presented as objective facts, they are often difficult to detect.[33] These facts are entrenched until another researcher is willing to invest in revisiting the archaeology, whether it be the physical site itself or the often sparse and incomprehensible primary datasets (when they are available at all). When it comes to revisiting the physical archaeology itself, another problem is encountered, and that is the nature of archaeological data.

Archaeological data is highly contextual, and once the site is excavated—a destructive, if necessary, process—context is left only in the field records.[34]

---

[31] Ibid.

[32] Baines and Brophy, "Archaeology without -isms," 70.

[33] Kintigh, "The Challenge of Archaeological Data Integration," 3.

[34] Ibid.

Methods of data collection by fieldworkers often suffer from inconsistency and frequent missing values—often not the fault of the fieldworker, but due to the sensitive and unpredictable nature of archaeological data collection.[35] Even if the data is of a high quality—complete and undisturbed—the way of archaeological data collected is naturally fragmented; archaeologists create spreadsheets, narrative texts such as journals, photos, drawings and geospatial data, to name a few. The very structure of this data is complex and diverse, not to mention the content itself. In addition, with considerations of the research question at hand, the nature of the data collected is also often highly subjective in nature. Steps are often taken during a research project to separate objective and factual observations from subjective variables of interest.[36]

This problem of idiosyncratic and variable datasets is encompassed in Eric Kansa's 'small science' dilemma.[37] 'Small science' refers to decentralised institutions, usually working on a case-specific basis and with specific research questions to be answered. Customised methods of data collection are to be expected, and unique recording forms and independent databases are also common.[38] Archaeology is a classic example of a small science. Due to the diverse nature of archaeological data in context, archaeologists find themselves with a

[35] Patrik, "Is There an Archaeological Record?," 28.

[36] Kintigh, "The Challenge of Archaeological Data Integration," 7-8.

[37] Eric C. Kansa and Ahrash Bissell, "Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains," *Data Science Journal* 9 (2010).

[38] Christine L. Borgman, Jillian C. Wallis, and Noel Enyedy, "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries," *International Journal on Digital Libraries* 7, no. 1-2 (2007).

diverse collection of methods of data collection. In turn, they find themselves with data in a diverse set of structures.

However, it can be argued that archaeologists are not fully equip to deal with their data to begin with. Tool training and data management capabilities of archaeologists are central to this issue. The conclusion made by Kenneth Aitchison in his 2004 assessment of professional practice in archaeology stated that;

> …universities are *not* the producers of archaeologists, but of the education and training that leads to academic qualifications. The employers are not the consumers of what the universities produce; that is the role of the individual graduates, who consume that education and training. The universities then see the responsibility falling to those individual graduates or their employers to find ways to gain sufficient experience and expertise to be able to enter practice as archaeologists.[39]

Archaeologists, he argued, are not trained for archaeology but for "tertiary learning and generalised skill acquisition…[for] wherever the job market takes them."[40] Aitchison argues that, due to a large number of students enrolling in archaeology degrees with no intention of making archaeology a career, it would

[39] Kenneth Aitchison, "Supply, demand and a failure of understanding: addressing the culture clash between archaeologists' expectations for training and employment in 'academia' versus 'practice'," *World Archaeology* 36, no. 2 (2004): 205.

[40] David Austin, "Archaeology, funding and the responsibilities of the university," in *The Responsibilities of Archaeologists: Archaeology and Ethics*, ed. Mark Pluciennik, *BAR International Series 981* (Oxford: Archaeopress, 2001): 35.

be “inappropriate to try and shoehorn these students into purely vocational courses designed to meet the requirements of the archaeological workplace.”[1] Archaeologists, then, are not trained for their vocation; they are trained for generalised entry into the workforce. The lack of specialist training, as described here, is one of the most problematic aspects of archaeology as a discipline, especially when it comes to maintaining data.

Archaeology is a data intensive discipline; data that is material culture, which in itself is an extremely subjective and diverse, and often of poor quality.[2] Geographic Information Systems (GIS) and spreadsheets are popular ways of representing archaeological data; however, these formats are not the limits of data representation. “Poor system models,” as Brian Ballsun-Stanton comments, “beget worse system models.”[3] That is, archaeologists submitting data may not have considered other methods that would best represent their data; poor data communication leads to poor data results. This problem is exacerbated when it comes to processes of database use.

Normal business extract, transform, load (ETL) procedures require vast amounts of funding and technical expertise to function optimally. Data that is poorly modelled or structured, incomplete, unreliable or biased require ETL procedures of vast magnitude to give data even a semblance of order. Unfortunately, these factors are all ones of which describes the archaeological

[1] Aitchison, "Supply, demand and a failure of understanding," 205.

[2] Patrik, "Is There an Archaeological Record?," 27-28.

[3] Brian Ballsun-Stanton, "Asking About Data: Exploring Different Realities of Data via the Social Data Flow Network Methodology" (University of New South Wales, 2012), 349.

record itself. Archaeologists largely lack the programming skills that allow for the mass manipulation of diverse, unstructured data. It is not to say they cannot manipulate data; rather, they are unable to creatively assess the best way to represent data. Due to the poor quality of archaeological data, and its representation notwithstanding, most archaeologists do not think about their data in terms of reuse. Results pertaining to their specific research questions are paramount, and data can rarely be considered outside of their own focus.

Even without the problem of structuring data, datasets must be first be made available. If they are not, and researchers wish to revaluate sites or conclusions, they face a number of problems. If researchers revisit a site that has already been excavated, the chance of retrieving any conclusive data is low; archaeology is often a destructive process, especially excavation.[4] In conjunction, if they are revaluating conclusions, the primary dataset often contains competing taxonomies and definitions of terms, usually defined at the project leaders' discretion.

1.3 Standards in Archaeological Data Management

"Archaeologists," Kintigh writes, "are stewards of the archaeological record."[5] The seventh principle on records and preservation, as stated by the Society for American Archaeology, states: "Archaeologists should work actively for the preservation of, and long-term access to, archaeological collections,

[4] Kintigh, "The Challenge of Archaeological Data Integration," 3.

[5] Keith W. Kintigh, "The Promise and Challenge of Archaeological Data Integration," *American Antiquity* 71, no. 3 (2006): 572.

records, and reports."[6] Archaeologists are ethically bound—and often legally bound by government or cultural heritage management agencies—to deposit all data and documentation collected or created during fieldwork in a data repository. These requirements ensure the long-term preservation and access to irreplaceable data. They also intend to maintain existing data in a way that prevents loss of data due to incompatible systems, software, or the physical deterioration of the storage media.[7] For example, the state of Victoria, Australia, requires archaeological data collected by heritage consultants to be deposited in the Victorian Heritage Database, or another approved database, by the overlaying government body, Heritage Victoria.[8]

Practice, however, often falls short of the desired outcomes, especially in the area of academic research as compared to statutory or commercial projects.[9] "Authority only exists in as far as it elicits a response,"[10] as John Barrett writes, and the response to online data archiving from the archaeological research community has been lukewarm at best.

When the data is submitted adequately at all, often repositories do not prepare the information in a way that allows the data to be commensurate.

---

[6] "Society for American Archaeology Principles of Archaeological Ethics," *American Antiquity* 61, no. 3 (1996).

[7] Kintigh, "The Promise and Challenge of Archaeological Data Integration," 571-72.

[8] Department of Planning and Community Development, "Guidelines for Investigating Historical Archaeological Artefacts and Sites," State of Victoria, 2012.

[9] Kintigh, "The Promise and Challenge of Archaeological Data Integration," 572.

[10] Barrett, "Fields of Discourse," 12.

Rather, repositories typically treat media in a way that transforms them into artefacts themselves; by "storing them in boxes on shelves."[1] Incommensurate artefacts become stored symbols without a common meaning: trivial to find if you know what you are looking for, impossible to find if you do not. As a result, the standardisation of data submitted to these repositories is becoming of vital importance to the long-term preservation and accessibility of data.

1.3.1 Data Standards in Digital Archaeology

"It is the sharing of data between projects," writes Julian Richards and Catherine Hardman, "that requires the development of shared standards."[2] Standards, in terms of archaeological databases and data, usually refer to three areas: technical (software and hardware) standards, content standards, and metadata standards.[3] It should be noted that when referring to data standards, it is a reference not to data quality, but data structures.[4] Meaningful comparisons across multiple datasets created independently require two main components: one, adequate documentation of data semantics (metadata, or how the data has been created), to allow for other researchers to understand the research structure; and two, the capacity to "represent the datasets within a common

[1] Francis P. McManamon and Keith W. Kintigh, "Digital Antiquity: Transforming Archaeological Data Into Knowledge," *The SAA Archaeolgical Record* 10, no. 2 (2010): 37.

[2] Richards and Hardman, "Stepping Back from the Trench Edge," 170.

[3] Richards, "From anarchy to good practice," 28.

[4] Jakob Voß, "Revealing digital documents: Concealed structures in data" (Cornell University, 2011).

schema," to allow for meaningful comparisons between them.[5] Most initial archaeological applications have been designed for the recording of data during fieldwork; however, semantic comparison is not a central part of the discipline.

Archaeological databases, in some ways, are considered a solution. Databases somewhat resolve the problem of handpicked conclusions; all data can be submitted—at the investigators discretion—regardless of relevance. It does not solve the problem of selective investigation, but data is available in its primary form, reducing the reliance on in literary conclusions.

Databases, however, are not without their obstacles. As a discipline straddling the humanities, social sciences and environmental studies, complex data collection methods and structures are to be expected; it makes for a difficult solution when it comes to standardisation.[6] Archaeology is contextual and any system of standardisation must reflect this diversity. During the 1970s and 1980s, there were several attempts to develop universal archaeological databases.[7] One such example, the ArchéoDATA system, commented that large scale centralised projects usually failed because they have been imposed "from

[5] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 1.

[6] Kansa and Bissell, "Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains," 42.

[7] For examples see: D. Arroyo-Bishop, "The ArchéoDATA Project," in *Computer Applications and Quantitative Methods in Archaeology 1989*, ed. S. P. Q. Rahtz and J. D. Richards 548 (Oxford: BAR International Series 1989); R. G. Chenhall, "The Archaeological Data Bank: a progress report," *Computers and the Humanities* 5, no. 3 (1971).

the top down"; that is, from a centralised body to individual researchers. This method resulted in limited consultation from practitioners. ArchéoDATA used specialised personnel to develop the system and standards, resulting in "unreasonable demands" being made on the archaeologist.[1] The outcome of these ambitious projects was a newly found appreciation of realism; centralised standards and systems may function, but would never be adopted in a discipline where "there are diverse practitioners, and geographical, historical and political drivers behind diversity."[2]

As Kintigh notes, however, "the cost in lost data of not moving forward is incalculable." There is a pressing need in archaeology for an "information infrastructure that will allow us to archive, access, integrate, and mine disparate datasets."[3] Without a minimum standard, data exists in isolation.

1.3.1.1 Possibilities as a Result of Data Standards

Data standardisation solves a number of the key problems relating to data comparability, interoperability, and synthesis in archaeological research.[4] Standards enforced at the beginning of a data life cycle enhance data

[1] John Wilcock, "On the importance of high-level communication formats in world archaeology," in *Archaeology and the Information Age: a global perspective*, ed. Sebastian Rahtz and Paul Reilly (London: Routledge, 2004): 72. It should be noted this project, the ArchéoDATA system, is now defunct.

[2] Richards, "From anarchy to good practice," 30.

[3] Kintigh, "The Promise and Challenge of Archaeological Data Integration," 567.

[4] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 1-3.

preservation in the long-term, because "a standard schema is more likely to be adequately documented than a one-off recording system."[5]

Standards also reduce time spent mapping an individual dataset to an archival standard, which is an important ethical responsibility. "At least," comments Kintigh et. al., "for those investigators who take seriously their ethical responsibilities to make their data available to others."[6] It is not to be said that all relationships and functions have to be mapped on the onset of a project. However, it is essential that control of relevant terminologies and project semantics is defined early on, and that a dynamic process is enforced to encourage the expansion of these concepts.[7]

The control of variables is essential when it comes to the interoperability of data. Standards allow for data interoperability to be achieved; some have argued the success and value of a dataset is based on its interoperability. "There is a power law at work here," notes Limp. "The value of any single data element grows along with our ability to relate it to other elements."[8] As discussed earlier, standards allow for multiple datasets to produce, and make understandable, long-term multi-disciplinary studies of phenomena outside the scope a single project.[9] It also allows for data to be repurposed and reused by others in ways

[5] Ibid., 1.

[6] Ibid.

[7] Limp, "Web 2.0 and Beyond," 278.

[8] Ibid., 277.

[9] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 2.

the original investigator may have overlooked. The possibilities of data that follows some form of standard cannot be overstated; however, it is neither without its problems nor objections.

Artefacts that are no longer present to be analysed are problematic without data standardisation. Over time, archaeological artefacts have been lost to reparation legislation: one such example is the Aboriginal Heritage Act (2006),[10] which gives Aboriginal communities rights to indigenous human remains. Artefacts have also been destroyed; by natural degradation or human fault, or even excavation itself. In theory, with the introduction of standards, this primary data can be accessible even after the artefact in question (or its context) has been lost, destroyed or rightfully returned to its community.

Legacy data, too, also benefits. Mapping legacy data to a standard allows for the integration of this irreplaceable data into databases. It increases the accessibility of legacy data to researchers, while also sustaining the utility of this data, stemming the loss of data to media degradation, software obsolescence, or inadequate metadata.[11]

Standards also allow for data to be integrated into synthetic, interdisciplinary research, by allowing for data to be represented within a common schema. As a result, researchers can "extract a sensibly integrated and appropriately scaled database of analytically comparable observations from

[10] *Aboriginal Hertiage Act 2006* (Cth), <http://www.austlii.edu.au/au/legis/vic/consol_act/aha2006164/> (17 October 2013).

[11] Kintigh, "The Promise and Challenge of Archaeological Data Integration," 567.

multiple data sets employing different recording protocols."[12] This data, then, can be compared automatically and cost-effectively.

Standards, and the benefit of data interoperability, promises investigation into long-term and large-scale questions, with a level of empirical support that cannot be otherwise achieved. It allows for patterning on samples larger and more distributed than any single project could collect.[13] As archaeological data starts subscribing to a comparable standard, the more useful it will be in interdisciplinary studies, particularly in terms of long-term socio-ecological development.[14]

1.3.1.2 Problems and "Costs" related to Data Standards

The challenges to data standards, by the inherent complexities of archaeological data and the nature of archaeology as a discipline, are not insignificant. Small sciences are inherently problematic when it comes to applying "universal" standards.[15] The problems regarding the implementation of data standards are theoretical, systematic and social in nature.[16]

With regards to legacy data, much data and metadata has been lost to the researchers own inadequate storage and documentation of their research. It is

[12] Ibid., 570.

[13] Ibid.

[14] Ibid., 571.

[15] Kansa and Bissell, "Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains," 43-44.

[16] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 1.

coupled with an unwillingness to spend the time required to complete such endeavours due to little reward for the original researchers themselves.[17] Even when dealing with the conclusions drawn by others, and the data that can be gleaned from such endeavours, research often suffers because such conclusions have inevitably "smoothed" the detail-rich primary data.[18]

There is also a certain distrust of standards within research communities, springing from a "general resistance to the imposition of controls or restraints that restrict freedom of archaeological research and interpretation."[19] Kintigh et al. argue that social factors are the main obstruction to data standards; there is a lack of leadership when it comes to adopting standards, and a sense of inertia suffered by the archaeological community as a whole.[20] Resistance to overarching controls perhaps originates from the search for novelty in archaeology, as discussed previously. Some have also argued that data standards stifle intellectual enquiry by imposing a rigid means of data collection, thereby restricting the research question being investigated and the circumstances in which the research is taking place.[21]

[17] Nelson, "Empty Archives," 160.

[18] Kintigh, "The Promise and Challenge of Archaeological Data Integration," 570.

[19] Baines and Brophy, "What's Another Word for Thesaurus?," 237.

[20] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 3.

[21] Richards and Hardman, "Stepping Back from the Trench Edge," 168.

When dealing with the integration of legacy data into databases, it is unlikely that any legacy data would meet a proposed contemporary standard.[22] If data were required to follow a standard format at submission, and the standard was high, much would be discarded rather than deposited. If the standard represented the lowest common denominator, high precision data would be lost to the standard.[23] A compromise requiring higher precision would result in the same loss of data, albeit legacy data. As Kintigh writes: "The history of our field shows that data standards do, and should, continue to evolve."[24]

Similarly, there are hundreds of competing standards in fields related to archaeology, particularly in the cultural departments. Many of these core systems lack either the authority to make change, or are too small to capture the nuances of complex, advanced data.[25] There is a possible advantage, however; as standards explicitly dicate a structure, a well-defined standard is preferable (however comprehensive) to no standard at all.

Even in places such as the United Kingdom, where data standards have been integrated into commonplace heritage management successfully, a problem is clear. With the introduction of projects such as the Monument Inventory Data Standard (MIDAS) Heritage, some strides have been taken towards

[22] Kintigh, "The Challenge of Archaeological Data Integration," 6.

[23] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 2.

[24] Kintigh, "The Challenge of Archaeological Data Integration," 6.

[25] Martin Doerr, "The CIDOC Conceptual Reference Model: an ontological approach to semantic interoperability of metadata," *AI Magazine* 24, no. 3 (2003): 74.

standardisation but they lack interoperable qualities outside a local spectrum. Take periodisation as an example; as soon as one leaves England, it becomes clear that "archaeological periodicity is, in reality, culturally determined, and relative date ranges are spatially dependent."[1] For example, even with the standard as it exists, most of Scotland lacks a 'Roman period', and where 'Viking period' is the predominant term in England, it is often replaced with 'Norse' in Scottish literature; it is only compounded in the European stage, where new periods are encountered relative to spatial boarders.[2] Some progress is being made in this cultural conflict, with the introduction of the Advanced Research Infrastructure for Archaeological Dataset Networking in Europe (ARIADNE) database, launched in February 2013.[3] ARIADNE is aimed at overcoming the fractures caused by disparate archaeological data repositories in European archaeology.

Most would argue that the benefits of having data that is semantically interoperable outweighs the costs and dangers. Moreover, they argue that a level of data commensuration should be aimed at where possible.[4] Several database projects have been implemented worldwide, with varying degrees of success, each with varying protocols on data standardisation.

[1] Richards and Hardman, "Stepping Back from the Trench Edge," 171.

[2] Ibid.

[3] Kate Fernie, "Launch of FP7 infrastructure project - Ariadne," <http://www.ariadne-infrastructure.eu/News/Press-release> (Accessed: 11 October, 2013).

[4] Nelson, "Empty Archives," 162.

1.3.2 Analysis of Multiple: Archaeological Databases

In the last decade, archaeological databases have become popular modes of communicating primary datasets. Standards, primarily concerning standards for metadata, have proliferated as a result of these databases and data repositories. At present, there is a trend to regard multiple standards as preferable to no standards. Logic dictates that 10,000 documents subscribing to 30 competing standards are easier to integrate and compare than 10,000 documents using individual data schema. Standards by their nature define data and texts; drawing from the example, comparing 30 defined standards as opposed to 10,000 individual schemas—that are not necessarily explicitly described—is much less costly and time consuming. The following summaries of existing databases outline several key issues in data management in archaeology.

1.3.2.1 The Online Access to the Index of archaeological investigations (OASIS) Project

The OASIS project, a United Kingdom based scheme, operates under the University of York and brings together a number of strategic partners: the Archaeology Data Service (ADS), the Archaeological Investigations Project (AIP) of Bournemouth University, the National Heritage Protection Commissions Programme, and the National Monuments Record of English Heritage. It has been continuously developing since 2002, and its primary concern is grey literature as it relates to archaeology.[1]

[1] Grey literature is generally loosely defined as documents not readily available via conventional channels, with the intention of a restricted audience, and usually found only in hardcopy form. (As defined by: Marie-Claire Debachere, "Problems in obtaining grey literature," *IFLA Journal* 21, no. 2 (1996): 94.)

Within archaeology, grey literature can refer to anything from survey or object recording forms to field reports or summaries disseminated within a restrictive channel; in most cases, these documents form a substantial part of the most recent, up-to-date archaeological data. The aim of the OASIS project is to "provide an online index to the mass of archaeological grey literature that has been produced as a result of the advent of large-scale developer funded fieldwork and a similar increase in fieldwork undertaken by volunteers."[2] The primary purpose of this project is "to bridge the research gap [by providing] archaeological scholars with up-to-date information about the latest archaeological results."[3] The data is uploaded via an OASIS Data Collection Form, which then forwards the information to the Library of Unpublished Fieldwork Reports, managed by the ADS.

While the establishment of such an archive is certainly a set in the right direction when it comes to stemming archaeological data loss (namely, to software obsolescence, as OASIS provides a data format standard), there are two main problems: first, concerning the enforcement of OASIS procedures, and second, concerning the retrieval of literature from the database. Concerning enforcement, the primary issue is that there is little to no legal requirement to use the scheme. Larger than that, even when data is uploaded, there is no base content standard. While the OASIS form provides an outline of what is required,

[2] Kieron Niven, "OASIS: Online AccesS to the Index of archaeological investigationS," <http://oasis.ac.uk/england/> (Accessed: 31 July, 2013).

[3] Julian D. Richards and Catherine Hardman, "OASIS: dealing with the digital revolution" (paper presented at the Digital Heritage of Archaeology, Proceedings of CAA 2002, Heraklion, Crete, 2-6 April 2002), 327.

there is no agreed universal recording system for archaeological fieldwork. Therefore, the knowledge representation is unique from project to project.

The ADS provides a "Guide to Good Practice", but only suggests what metadata is needed for data preservation and reuse.[4] Julian Richards notes that, "[while] there is a common denominator to the fields used on most [field] forms, and also overlapping vocabularies, most excavators have stuck to their own systems."[5] This disconnect brings forth the second issue; lacking a universal content standard means search queries do not always bring up all relevant results, especially when terms have more than one meaning or when concepts can be defined by more than one term. Conflicting terminology is an issue for other projects within the UK, especially when dealing with data outside their region; for example, when dealing with chronological periods, the ARENA project mapped national terminology to an absolute timescale.[6] However, the absolute timescale reduces each data set to the lowest common denominator; ergo, searching 'Iron Age', for example, recovers "sites dated to anywhere between 1000 BC and AD 1000, depending upon the country in question."[7]

OASIS provides a solid example for preliminary data standardisation. However, without legal requirements for neither submission nor standardised

[4] Richards and Hardman, "Stepping Back from the Trench Edge," 176; "Archaeology Data Service/Digital Antiquity: Guides to Good Practice," Archaeology Data Service & Digital Antiquity, <http://guides.archaeologydataservice.ac.uk/> (Accessed: 31 July, 2013).

[5] Richards and Hardman, "Stepping Back from the Trench Edge," 177.

[6] Archaeological Records of Europe - Networked Access: a portal intended to cross-search six European national or regional sites and monuments databases.

[7] Richards and Hardman, "Stepping Back from the Trench Edge," 171.

query terms, OASIS can only provide access on a limited scope. Access to documents is not inclusive to access to information. Without content standardisation, metadata can only go so far in adequately describing data.

1.3.2.2 THE DIGITAL ARCHAEOLOGICAL RECORD (TDAR)

tDAR is a sub-project of Digital Antiquity, a non-profit organization concerned with the preservation of archaeological data and records, and is "an international digital repository for the digital records of archaeological investigations." Like OASIS, the ADS and a number of individuals from a number of universities support tDAR.[8] tDAR's project statement places key importance on the following system capabilities:

> ….[the system must] integrate data collected at different scales, at different times, by different investigators using variable data recovery strategies and inconsistent typologies; adequately encode complex typologies, data recording schemes, archaeological contexts, and recovery techniques; and, most importantly that it is neither necessary nor advisable to reduce data to a single standard at registration time. Instead, the semantics of new and legacy data must be preserved.[9]

To achieve this, tDAR offers a complex ontological system regarding metadata but little in terms of content mapping. Their ontology allows for the articulation of relationships between items in a controlled vocabulary, but does allows little deviation or editing of these relationships. That being said, metadata can go

[8] "tDAR: The Digital Archaeological Record," Digital Antiquity, <http://www.tdar.org/> (Accessed: 31 July, 2013).

[9] Kintigh, "The Challenge of Archaeological Data Integration," 5.

some ways in describing content—though not to depth needed for cross-searching mass databases automatically. Metadata allows for data preservation and open access—but when it comes to data comparison across multiple projects, the researcher must take a considerable amount of time ontology mapping with regards to content. Ontology mapping is the process of reconstructing data semantics of a dataset to understand the structural aspects of the data.

tDAR considers any data standards in regard to content unwise; it is their option that "any centralized attempt to impose standards would fail utterly."[10] It also poses some problems with regards to legacy data; it is often noted that essential metadata is lost when a primary investigator or specialist responsible for a project is unavailable.[11] Even when this is not the case, legacy data itself is difficult to locate—the chance of choosing adequate metadata to describe the semantics of the dataset is slim.[12]

However, tDAR remains a well-conceived repository that grants discoverability, and at least some degree of interoperability, to primary archaeological data, for both archaeologists and specialists from other fields. Unlike OASIS, a standard is enforced that can allow personal idiosyncrasies while also ensuring a standardised form. Using hierarchical concept mapping in the form of ontologies, queries submitted to the database return results are scaled

[10] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 4.

[11] Kintigh, "The Challenge of Archaeological Data Integration," 3.

[12] Ibid.

and reconciled to match the scope outline of the researcher.[13] However, like OASIS, tDAR lacks enforcement—while researchers may feel ethically obligated to submit their data, there is no official requirement for them to do so.

1.3.2.3 The Victorian Heritage Database

In Australian archaeology, digital archaeology is just finding its place in the cultural management sector. The Victorian Heritage Database (VHD) is one of few archaeology databases, and is supported by the Victorian state government and the Department of Planning and Community Development. The database advertises itself as "a fully searchable online database containing information about Victorian Heritage Places and Precincts, including statements of significance, physical descriptions, historical information, builder, architectural style, photographs and heritage overlay number."[14]

As a cultural resource management (CRM) database, there is a legal obligation for consultants and heritage advisors to submit their primary data to the database.[15] This statutory requirement gives VHD an advantage over both tDAR and OASIS. However, as seen with OASIS, there is a significant issue of meaningful data accessibility when it comes to CRM databases. While the data is present and accounted for, meaningful metadata and content standards are very rarely adequate for data interoperability. The data submitted to the archive does not meet any unified terminology standard, for either metadata or content.

[13] Kintigh, "The Promise and Challenge of Archaeological Data Integration," 571.

[14] "Heritage Victoria," Department of Planning and Community Development, <http://www.dpcd.vic.gov.au/heritage> (Accessed: 31 July 2013, 2013).

[15] "Guidelines for Investigating Historical Archaeological Artefacts and Sites," 6.

Keyword searches cannot function adequately, as vocabularies of researchers and users may not meet semantically. Furthermore, it is limited to the CRM sector; academic research, while supported, is neither mandatory nor enforced. While submission to VHD (or another approved database) is required by law, VHD does not overcome many of the problems that are somewhat solved by OASIS and tDAR. Like many database-by-legislation efforts, VHD lacks the organisation enabled by OASIS' standardised submission form, and also lacks the degree of interoperability enabled by tDAR

1.4 Conclusions

This chapter has given a brief glimpse into the problems facing archaeology as a digital domain. It has been established that metadata standards are the preferred solution for the crisis of data accessibility in archaeology. Metadata certainly grants access to data; however, this is often limited to access on the superficial "discovery" level. Conflicting content standards remain problematic when performing large-scale cross searching across databases and datasets; without some standardised content, researchers can spend hundreds of hours mapping datasets to a united schema. Federal agencies in the United States, for example, report approximately 50,000 field projects that contained some element of archaeological resources in the past 12 months.[16] While heritage management usually has a better record of standardising information—usually at data creation with universal data entry forms—research projects can rarely compare data at a meaningful level without tedious and time-consuming

[16] McManamon and Kintigh, "Digital Antiquity," 38.

data mapping. With projects numbering like this for just one year, the process of mapping concepts over multiple datasets would be near on impossible.

Terminology standardisation is the most obvious solution to this problem. However, it is far easier said than done, and any attempt at terminology standardisation must reconcile the relational and diverse nature of archaeological data.

OASIS, tDAR and VHD both solve and create problems concerning metadata and content standards. It is clear that metadata—the key words and tags that are associated with documents for query purposes—are the most popular form of standardisation in archaeological databases. While OASIS, in particular, provides a technical standard by applying data formatting, none of these standards delve deeply into the problem of content standardisation. Content standardisation, if it appears at all, is secondary and less developed. It is the obvious solution for data incompatibility that data standards are conformed to at data creation; by applying data standardisation this early, data is commensurate on the content level. Metadata allows access to datasets, and is the first step to stem data loss; however, access to the data itself will continue to be problematic if data content is not adequately defined.

CHAPTER TWO: CONTENT STANDARDS AND APPLIED TEXT ANALYSIS IN ARCHAEOLOGY

This chapter will discuss the relationship between standardised terminology, ontologies and computer-assisted text analysis. Metadata enhances the discoverability and interpretation of archaeological data, but lacks the ability to define the semantic nature of the data. With metadata standardisation alone, interoperability remains limited. Semantic interoperability requires data content standards. A standardised terminology across a domain is the best route for content standards. Ontologies provide the most powerful system for standardising terminology. Ontologies, as it is used here, refer to networks of concepts, often represented by a relationally organised vocabulary.[1] Automated text analysis offers a method for producing a descriptive, rather than prescriptive ontology. By extracting synonymous terms and concepts, a natural language ontology can be developed. Natural language ontologies reflect the reality of the discipline, and encourage participation by the wider community.

This chapter will offer an outline of the problems of terminology in archaeology, and explore applications of automated text analysis for one stage of ontology production: relationship mapping. Chapter three will present a case study of retrieving synonyms from a select corpus of texts pertaining to the sub-discipline of rock art.

[1] Natalya F. Noy and Deborah L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," (Stanford University: Stanford, 2001), 1.

2.1 Content Standards in Archaeology

From the earliest days in the field, archaeologists have sought to create archaeological taxonomies, classifications, systematics and semantics. A central goal of the archaeologist is to properly classify materials, and thereby "[process] a multidimensional matrix of variables into [a] proper single classificatory category."[1] Creating vocabulary structures has been a primary goal, so placing controls on the use of terminology is the next logical step. Placing controls on vocabularies used in a discipline ensures meaningful interoperability. Archaeologists primarily work from representations of material culture, namely the datasets collected by other archaeologists, so it is imperative that meaning is conveyed effectively and consistently.

Control of the terminology used in datasets plays a large role in the level of semantic interoperability exhibited by a dataset. For data comparability to improve, standardised terminology, taxonomies and vocabularies must be adopted to minimise problems associated with semantic interoperability.[2] Terminology plays a significant role in how data is accessed, consumed and reused. Unified vocabularies across a discipline allow the past to be brought into relation with larger socio-ecological processes, rather than exist in insolation to one another.[3] By defining shared terminologies, where the investigator outlines their own semantic process, terminology becomes the precondition for content

[1] Limp, "Web 2.0 and Beyond," 279.

[2] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)."

[3] Baines and Brophy, "What's Another Word for Thesaurus?," 238.

accessibility for data comparison.[4] Terminology standards facilitate and improve the discovery of relevant digital resources in search queries.

Researchers have a choice when it comes to how they achieve semantic interoperability through their use of terminology: controlled lists of terms, thesauri or ontologies. Each method has its own structure and complexity level; a flat list for a controlled list, a tree for taxonomies, multiple trees for thesauri, and a network for ontologies. Controlled lists and thesauri of standard and alternative terms are the two most common ways of approaching content standards. However, simply naming items is not always the most effective way to demonstrate the relational variables of an archaeological object or concept. Hierarchies are more effective for displaying the relational properties of a term.[5] Related terms in a hierarchy can modify or expand the meaning of the original term; for example, terms that provide properties that define how the primary term can be used in context.

Ontologies have only recently begun to be developed for archaeology and cultural resource management. The relational properties of ontologies, however, make ontologies perfect for archaeological concepts, as concepts and items in archaeology often overlap.

2.1.1 Controlled Lists

Flat lists of terms, often in the form of glossaries, are the simplest form of terminology control. In a well-conceived list, terms do not overlap in meaning or

[4] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 3.

[5] Baines and Brophy, "What's Another Word for Thesaurus?," 246.

occurrence (i.e. each term is unique), all terms are equal in level of specificity, and all "have the same level of rank in a classification system."[6] These terms are also known as "flat term list."

Controlled lists have a limited utility. In a database context, they are "best employed in certain fields of a database where a short list of values is appropriate and where terms are unlikely to have synonyms or ancillary information." [7] Outside databases, however, controlled lists often do not adequately define all the possible variables of a concept, leading to the use of new terms that do not occur within the flat term list. In a relatively short amount of time, the list of terms does not effectively reflect the reality of the domain's vocabulary; it is quickly out-dated.

Archaeologists work with a highly variable material culture. While they require a controlled vocabulary to meaningfully share concepts, the controlled vocabulary also needs to account and make room for what is unknown. Furthermore, these updates to others need to be effectively communicated. As a static entity, controlled lists are an unreliable and ineffective means of vocabulary control in archaeology.

2.1.1.1 Case Study: The IFRAO Glossary

The International Federation of Rock Art Organisations (IFRAO) glossary of terms is a typical example of a flat vocabulary. The IFRAO was established in 1988, and is "a federation of national and regional [rock art] Organisations

[6] Patricia Harping, *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, Cultural Works* (New Hampshire: Odyssey Press, Inc., 2010): 19.

[7] Ibid., 20.

promoting the study of palaeoart and cognitive archaeology globally."[8] The glossary itself was produced as a draft in 1999, with the first edition published in 2001 and the second, in 2010.[9] The IFRAO glossary illustrates the problematic nature of flat vocabulary lists. Lists do not adequately compile terms for all the possible variables found in the field, and practitioners commonly require custom terms to describe their dataset.

One of the most controversial terms of the rock art discipline is, indeed, "rock art" itself.[10] As defined by the IFRAO glossary of terms, "rock art" is a term used to describe "non-utilitarian anthropic markings on rock surfaces, made either by an additive process or by a reductive process."[11] Many rock art specialists would consider this definition to be exactly the sort of subjective imposition that is problematic. Controversy surrounds the term "art" as it is used in "rock art". David Whitely, a renowned rock art specialist, notes that indigenous peoples often lack literal terms for "art" in the sense that Westerners mean it. Western connotation associates art with aesthetics, thereby attributing

[8] Robert G. Bednarik, "IFRAO," <http://home.vicnet.net.au/~auranet/ifrao/web/> (Accessed: 2 October, 2013).

[9] Robert G. Bednarik, "IFRAO 1988-2000: the first dozen years," <http://home.vicnet.net.au/~auranet/ifrao/web/hist.html> (Accessed: 2 October, 2013).

[10] For the purpose of this example, I will be using the conventional term "rock art" as listed by the glossary of the International Federation of Rock Art Organisation (IFRAO) to describe the study of rock art within archaeology. All formal terminology regarding rock art in this paper will also draw from the IFRAO glossary.

[11] Robert G. Bednarik, "IFRAO Rock Art Glossary," <http://home.vicnet.net.au/~auranet/glossar/web/glossary.html> (Accessed: 8 September, 2013).

indigenous peoples with values they did not necessarily intend when creating the work.[12] "What is wrong in this instance is not the use of the term 'art' for the graphic imagery of non-Western cultures," argues Whitely, "so much as a reductionist and intellectualised view of Western art that this argument promotes."[13] Several alternative terms have been proposed, including "rock markings" and the hyphenated "rock-art", though little progress has been made towards the wide-spread adoption of either alternative.[14] In conjunction, much of the study of rock art concerns markings functioning as maps, trail markers, property lines, warnings, or identification purposes. That is, the marks are indeed utilitarian and are, some argue, almost exclusively so.[15]

A similar conflict is apparent in the use of other terms. "Petroglyph" and "pictogram" are two important terms denoting categories of rock art. In terms of the International Federation of Rock Art Organisation (IFRAO) definition given above, "pictogram" refers to the additive (painting) process of creating rock art, and "petroglyph" refers to the reductive process. On the other hand, David

---

[12] David S. Whitely, ed. *Handbook of Rock Art Research* (Walnut Creek: AltaMira Press, 2001): 22.

[13] Ibid., 22-23.

[14] Christopher Chippindale and Paul S. C. Taçon, "What's in a Word, What's in a Hyphen? A Modest Proposal That We Abandon the Words 'Petroglyph' and 'Pictograph', and Hyphenate 'Rock-Painting', 'Rock-Engraving', 'Rock-Art' among the Words We Use," *Rock Art Research* 23, no. 2 (2006); B. K. Shwartz, "Further comments on Christopher Chippindale and Paul S. C. Taçon's 'What's in a word, what's in a hyphen?': On 'Rock Art' History and Terminology," *Rock Art Research* 24, no. 1 (2007).

[15] Shwartz, "Further comments on Christopher Chippindale and Paul S. C. Taçon's 'What's in a word, what's in a hyphen?': On 'Rock Art' History and Terminology," 125.

Whitely uses “pictograph” to refer to both rock carvings and drawings.[16] In spite of the conflict at the fundamental level, “rock painting” has also been suggested as a substitute term for “pictograph.”[17] “Rock marking” and “rock engraving” have been identified as substitutes for “petroglyph.”[18]

This debate highlights some of the problems associated with flat vocabularies. Without a relational, hierarchical, or networked structure, terms exist in isolation. Flat vocabularies provide no suggestion of alternative meanings or synonymous terms. The IFRAO does not advertise its glossary as a means of controlling terminology. However, the very suggestion that the glossary was produced to provide a “basis of determining a uniform terminology” suggests some attempt towards standardised terminology.[19] Its failings highlight the problems associated with controlled lists. Controlled lists do not adequately provide description for the full list of variables found in the field—and cannot, considering the unknowns to be had in archaeology—nor do they provide a means for the reconciliation of conflicting terms.

---

[16] David S. Whitley, *Introduction to Rock Art Research* (Walnut Creek: Left Coast Press, 2011): 23.

[17] See Chippindale and Taçon, "What's in a Word, What's in a Hyphen?."; Shwartz, "Further comments on Christopher Chippindale and Paul S. C. Taçon's 'What's in a word, what's in a hyphen?': On 'Rock Art' History and Terminology."

[18] Whitley, *Introduction to Rock Art Research*: 24; Shwartz, "Further comments on Christopher Chippindale and Paul S. C. Taçon's 'What's in a word, what's in a hyphen?': On 'Rock Art' History and Terminology," 125.

[19] Bednarik, "IFRAO 1988-2000".

2.1.2 THESAURI

Thesauri, in particular thesauri with relational properties, produce a more robust and nuanced approach to terminology standardisation than do controlled lists. No universal thesaurus for archaeology as a discipline has yet emerged, though English Heritage has developed thesauri for use in the UK.[1]

Thesauri allow for competing terms to be reconciled without detracting from the original semantic structure of the system.[2] They accommodate representation of competing or conflicting terms, as thesauri by their nature "tend to discuss ambiguous or multiple interpretations."[3] Conflicting terms refer to synonyms and homonyms; synonyms being identical meanings with multiple terms, homonyms being identical terms with multiple meanings. Unlike controlled lists, ambiguous terms can be grouped together in a thesaurus, often with the favoured term identified as a primary term for use.

However, thesauri only allow for one relation between terms in their structure: synonymy. Ontologies, as a network of relationships, allow for more complex relationships to be made clear; ontologies not only quantify the term, but also allow for the term to exist over several layers of the network. By doing so, the impact of conflicting use of a term in the literature is lessened. They also allow for the specification of the type of relationship between terms; for example, one term may be associated with one concept, but can also be associated with its other homonyms.

---

[1] Richards and Hardman, "Stepping Back from the Trench Edge," 172-73.

[2] Ibid., 169.

[3] Baines and Brophy, "What's Another Word for Thesaurus?," 240.

The two primary problems of thesauri stem from the fact that they have, at most, a two-tier hierarchy. That is, the preferred term plus synonyms; thesauri may also be "flat", with no preferred term amongst the synonyms. A flat thesaurus, or even a two-tier hierarchy, can create ambiguities; it limits the amount of information that can be inherited from one term to another. Terms with more than one use or classification (for example, an "artefact scatter" can be classed as a location itself or a feature within a site) must be quantified by defining the application of the term.[4] The term "artefact scatter" should be quantified into (at least) two distinctions, "artefact scatter (feature)" and "artefact scatter (location)." This solution, however, presents its own problems. Without formal induction, the distinction between the two may not be clear to the user. The problem of classification itself is more serious: when terms depend on context for their meaning, it is difficult to predict their primary use in literature. For example, if artefact scatters do appear in equal instances in the field within sites and as singular locations, there is no way of predicting the primary use of "artefact scatter": "artefact scatter (location)" and "artefact scatter (feature)" would be used equally. Therefore, if the classification in a thesaurus were to favour "artefact scatter (location)", an uninitiated user may interpret all uses of the term "artefact scatter" in as a location of a site.

Thus, comes the problem of classifying the term in terms of its use. The term could be classed under multiple categories, but it may not be apparent to

[4] B. J. Wielinga et al., "From thesaurus to ontology" (paper presented at the Formal Ontology in Information Systems, Trento, Italy, 6-8 June 1998), 196.

the user that this is the case.[5] For example, should "artefact scatter (location)" be the primary term, and "artefact scatter (feature)" be classed under another term (say, "assemblage"), a user may expect an artefact scatter to *always* be the primary location, rather than, *sometimes* a feature of a site. As an ontology specifies relationships between terms (for example, "sits within"/"is a part of"), the term "artefact scatter" can be defined appropriately according to its context.

The second problem stemming from the one-/two-tier hierarchy of a thesaurus also concerns term ambiguity. Thesauri usually require terms to be linked to a particular group of synonyms, with a single preferred term. Using the example above, the two distinctions of "artefact scatter" encounter problems when attempting to conform. "Artefact scatter (location)" could be a synonym of "site", while "artefact scatter (feature)" could be a synonym of "assemblage." Working in theory, the two terms are limited by the weak structure of thesauri; these terms are independent of one another. The user must be aware of this conflict—otherwise assumptions could be made, for example, that "artefact scatter" *always* refers to a feature.[6] Within an ontology, the problem is solved though not only multiple levels of hierarchy—"artefact scatter" could be its own hierarchy with its second tier denoting the two sub groups—but by its relational properties as a network of concepts. "Artefact scatter (location)" can be associated with "artefact scatter (feature)" while also being listed as a synonym of "site."

[5] Ibid.

[6] Ibid.

Thesauri can be used to reconcile archaeological datasets.[7] Indeed, some argue that searching by keyword (for example, through metadata) should be the most common approach to the integration of archaeological datasets.[8] As discussed above, however, ambiguities limit their effectiveness. Thesauri lack the network aspect of ontologies; while they relate synonyms to one another, they lack the ability to define various types of relationships between concepts and the multi-level hierarchy supported by ontologies. The relationship between synonyms and homonyms must be clear in whatever means of controlled vocabulary. Thesauri, while communicating the relationship between synonyms, can risk isolation of terms by restricting its structure to a static two-tier system.

2.1.2.1 Case Study: The English Heritage Thesauri

English Heritage has developed various archaeological thesauri related to sub-domains in an effort to create a shared language for British archaeologists. These thesauri include terminology for monument types, archaeological objects and building materials, amongst others.[9] The thesauri created by English Heritage aim to promote usage of a master term, rather than open usage to a variety of terms, creating a two-tiered hierarchy for each term. The English Heritage Thesauri state a primary term, and then list the synonyms. Synonyms

[7] Richards and Hardman, "Stepping Back from the Trench Edge," 173.

[8] Baines and Brophy, "What's Another Word for Thesaurus?," 236.

[9] "English Heritage Thesauri," <http://thesaurus.english-heritage.org.uk/frequentuser.htm> (Accessed: 13 October, 2013).

are provided to describe the intended usage of the master term in various contexts.[10]

However, outside the United Kingdom these standards have not been widely adopted. They neither correlate to international terminology regarding the phenomena they describe, nor to archaeological concepts that vary from one region to another (for example, chronological periods). Even within the United Kingdom, these standards are promoted as "Good Practice" for conscientious archaeologists, rather than "Required Practice." The utility of the thesauri are perceived as limited enough to preclude their imposition even within the regulated environment of UK archaeology. The standards, therefore, remain limited to the UK, and optional even there.[11]

Problems of the two-tier hierarchy system are apparent within the English Heritage thesauri. For example, "artefact scatter" is listed under the Monuments thesaurus as a "monument <by form>". In addition, the listing for "assemblage" does not include "artefact scatter" as a related term. As a result, two problems present themselves. One, by having "artefact scatter" solely under "monument <by class>", the only meaning conveyed is an artefact scatter as a primary site; its function as a secondary element of a larger site is lost. The second problem is one faced by the user and highlights the problem of arbitrary, prescribed vocabulary control. If the archaeologist does not consider their data to describe an artefact scatter as a primary site, the listing under "monument <by form>" would not fit their dataset. They may be forced to use the term "artefact

[10] Baines and Brophy, "What's Another Word for Thesaurus?," 239.

[11] Richards and Hardman, "Stepping Back from the Trench Edge," 172-73.

scatter" anyway, but the standard meaning enforced by English Heritage and the investigators own meaning are fractured.

English Heritage has made the first steps toward vocabulary control in the British archaeology sector. However, as thesauri, their controls fail as a result of the inherent structure of thesauri as a two-tier hierarchy. Without an adequate network of relationships outlined between terms, a user faces the problem of both term ambiguity and term classification.

2.1.3 Ontologies

Ontologies overcome many of the limitations associated with thesauri. Ontologies, in the context of this paper, refer to a network of related concepts and terms, providing a semantic framework for interoperable data (see Figure A for a simple example of an ontology).[12] In modern implementations, ontologies are machine-interpretable and often exist as part of a database. If an ontology is part of a database structure, the database can identify each dataset by their basic concepts and the relations among them.[13] Even further, if the data is mapped to an ontology when the data is being created, the researchers own idiosyncratic vocabulary can be linked to the ontology, allowing for semantic individuality while ensuring interoperable data. Linking ontologies to information input fields in an application could achieve this task.

[12] Irena Spasić et al., "Text mining and ontologies in biomedicine: Making sense of raw text," *Briefings in Bioinformatics* 6, no. 3 (2005): 239.

[13] Noy and McGuinness, "Ontology Development 101," 1.

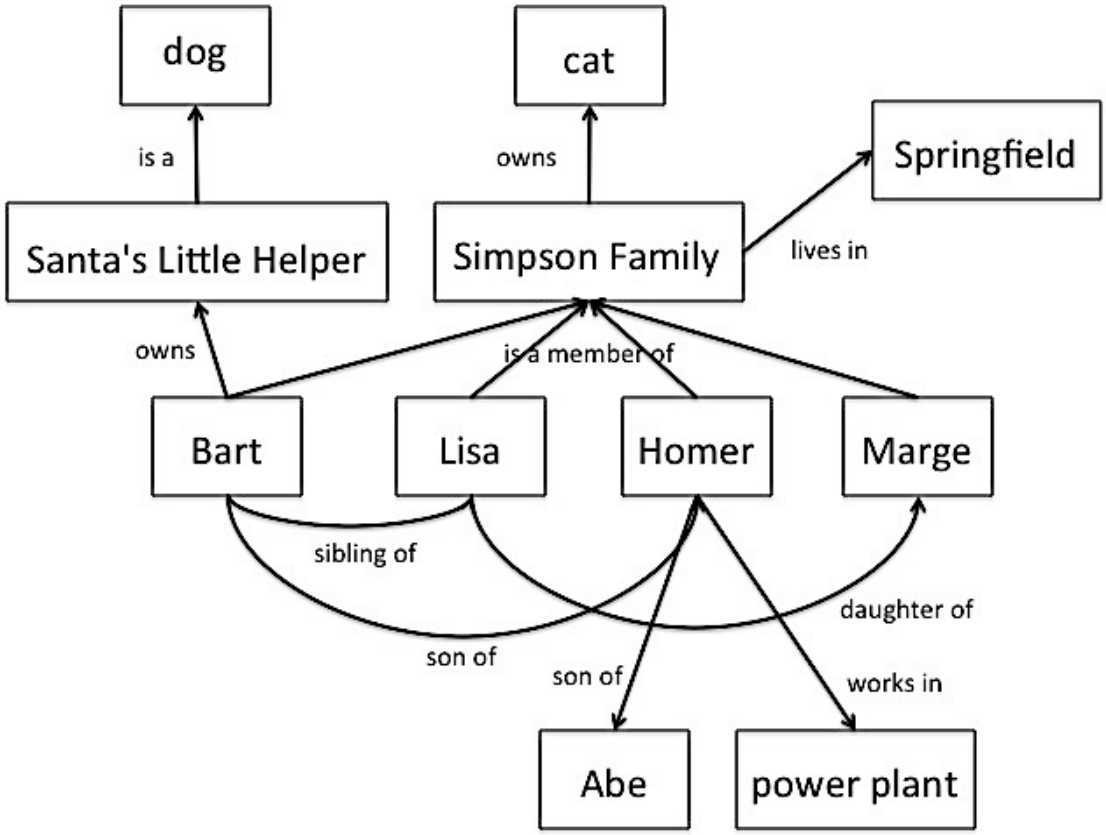


FIGURE A. AN ONTOLOGICAL REPRESENTATION OF THE TELEVISION SHOW, *THE SIMPSONS*.

Ontologies, unlike thesauri, allow for more than one relationship between terms to exist. It does so through a system of slots and classes; see Figure B for a simple diagram of this feature. Slots, also known as "properties," define the attributes or properties of a class, effectively describing its context. A class represents a word or concept.

The most useful feature of a controlled vocabulary for archaeology is for concepts to relate to each other. As Baines and Brophy write, "[Archaeologists] attempt to use language as a tool to *redescribe*, and in so doing *contextualise*, the mute objects we encounter in such a way as to further our goal as

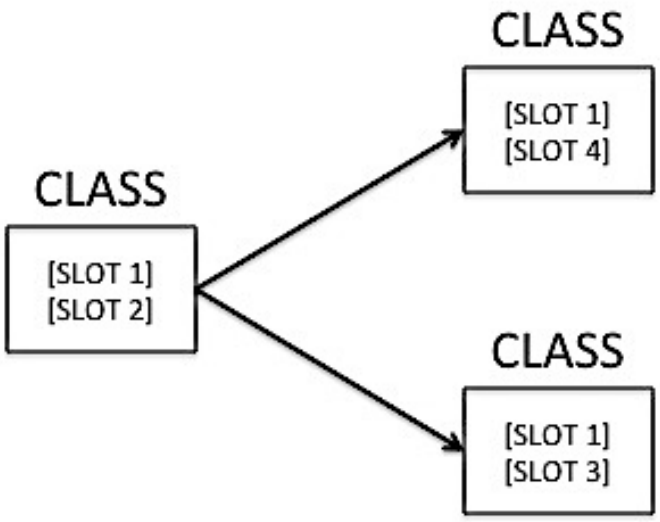


FIGURE B. A POSSIBLE REPRESENTATION OF THE CLASS/SLOT FEATURE OF ONTOLOGIES.

archaeologists." [14] In archaeology, concepts overlap and terms can have legitimate use in multiple contexts (thus, the context changing their meaning).

Controlled lists and thesauri do not adequately describe these relationships between terms. Controlled lists lack relationships to begin with, and the weak structure of thesauri struggles with term ambiguities that are complex. Ontologies facilitate a hierarchical structure, allowing for conceptual links to be grouped, but slots allow for a network of concepts rather than a pure hierarchy. In an ontology, ambiguous terms are easily dealt with; slots provide properties of the term (class) that define its context (see Figure C for an example that solves the problem of "artefact scatter" presented in section 2.1.2).

Ontologies have been developed for several purposes in archaeology. tDAR, as touched on in section 1.3.2.2, uses ontologies to regulate the metadata associated with datasets. By doing so, they attempt to make the data more discoverable; this implementation stems from the idea that keyword searches

[14] Baines and Brophy, "What's Another Word for Thesaurus?," 237.

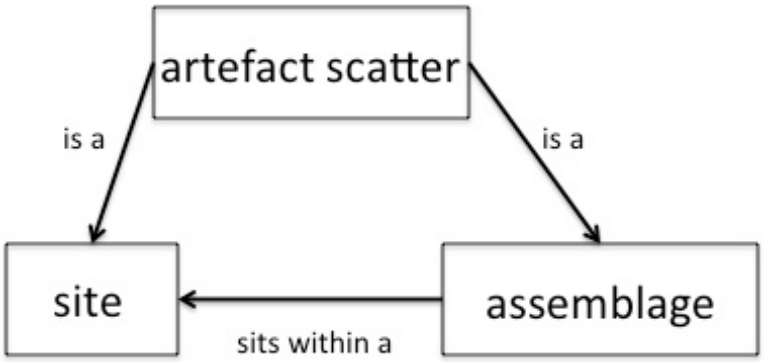


FIGURE C. AN ONTOLOGICAL REPRESENTATION OF "ARTEFACT SCATTER."

best facilitate data integration.[15] The CIDOC-CRM ontology was designed for museum use; it is used to describe cultural heritage information, and is intended to standardise library and archival information.[16]

2.1.3.1 tDAR's Ontology for Metadata

Ontologies are used by tDAR as a system of normalising vocabularies used in metadata. That is, the ontology controls the metadata associated with each dataset submitted to the repository, allowing for a standardised set of keyword searches. Search queries using standardised terms retrieve relevant datasets for the researcher, making datasets discoverable. Data that is discoverable, however, is not data that is interoperable. Metadata can make an effort towards describing data content. However, this feature limited even if the metadata describes the data schema; if a researcher wishes to compile data from several datasets over several databases automatically, they must go through a lengthy process of data mapping to achieve the level of interoperability required.

The tDAR ontology functions as a hierarchy with a first order, or listing, of presence or absence of a feature. That is, if an object is identified as being burnt,

[15] Ibid., 236.

[16] Doerr, "The CIDOC Conceptual Reference Model."

this is the first class entered before identifying the level of burning in the next class. Subsequent levels of the tDAR ontology deal with greater specificity regarding presence; in the context of the example, that would be details such as intensity of burn, and so on.[17]

Varying levels of presence allows for inconsistent or deteriorated data to be found in search queries, as any level of specificity can be searched for. As previously mentioned, a high precision standard would result in the loss of data, particularly legacy data.[18] Legacy data contains many out-dated terms and incomplete datasets; however, legacy data remains vital to the discipline. With a method of varying levels of presence, legacy data can remain incorporated into archaeological databases.

The tDAR ontology provides a basis for data accessibility. However, they continue the trend of metadata standards in archaeology. Metadata does not facilitate true data interoperability. Furthermore, access to the full extent of features allowed by an integrated content ontology is lost through metadata standards. The CIDOC-CRM ontology—a content standard for cultural heritage information—shows the full potential of an ontology when applied to content.

2.1.3.2 THE CIDOC-CRM ONTOLOGY

The CIDOC Conceptual Reference Model (CIDOC-CRM), its development starting in around 1996, is a "high-level ontology to enable information

[17] Kintigh, McManamon, and Spielmann, "Enhancing Data Comparability and Enabling Synthesis With tDAR (the Digital Archaeological Record)," 7.

[18] Ibid., 2.

integration for cultural heritage data and their correlation with library and archive information."[1] CIDOC is a subset of the International Council of Museums (ICOM), known as the Committee for Documentation of the International Council of Museums (CIDOC). CIDOC provide the museum community with advice on good practice and developments in museum documentation. The CIDOC-CRM was built to facilitate the exchange of information between cultural heritage institutions. In short, the CIDOC-CRM can be defined as the "curated knowledge of museums."[2]

The CIDOC-CRM exemplifies the successful implementation of an ontology in archaeology. The CIDOC CRM is "an object-oriented approach [designed to deal] with the necessary diversity and complexity of data structures in the domain."[3] Its aim is to provide a reference model and information standard that can be used by museums, and other institutions of cultural heritage, to describe and annotate their collections. By doing so, the CIDOC-CRM provides a minimum standard for information annotation for museum collections. The ontology has

[1] Doerr, "The CIDOC Conceptual Reference Model."

[2] "ISO 21127:2006," <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=34424> (Accessed: September 15, 2013).

[3] Martin Doerr, "The CIDOC Conceptual Reference Model: Who We Are," <http://www.cidoc-crm.org/who_we_are.html> (Accessed: 13 September, 2013).

been implemented in various applications, including the object database of the German Arachne Project and the MIDAS Heritage XML schema.[4]

The CIDOC-CRM is a content standard—it describes artefacts in a museum collection in relation to how it relates to other events, actors, etc. It does so through the slots associated with each term. For example, using the example in Figure C (section 2.1.3), an artefact scatter "is a" assemblage that "sits within a" site; in addition, an artefact scatter "is a" site. The item is *defined* by its relationship to other terms. It is the same process for the CIDOC-CRM. It provides standardisation at the base level of the data: the terms used to describe it. Each item is a "fixed point," and the ontology constrains the interpretation of terms to how they are used in context.[5]

The CIDOC-CRM shows the potential of ontologies when applied to content. Unfortunately, the CIDOC-CRM is limited to the cultural heritage sector; any attempt to use its standard in field data collection would fail. The relationships defined within the ontology are domain-specific, and require more information than is known at the time of data collection.

2.3 (Computer-Assisted) Text Analysis

Automated text analysis can be used to extract synonymous terms from existing domain literature, and provide evidence for semantic links between

[4] For a complete list of applications of the CIDOC-CRM, see: "The CIDOC Conceptual Reference Model: Applications," <http://www.cidoc-crm.org/uses_applications.html> (Accessed: September 24, 2013).

[5] Doerr, "The CIDOC Conceptual Reference Model," 7.

terms. By doing so, what is being reclaimed is the natural language of the discipline, and the terms discovered can be translated into an ontology.

Text analysis has a long history of applications to content analysis and standardisation, including synonym and homonym extraction, which together allow taxonomies to be reclaimed from domain literature. Synonyms are different words that share meanings; homonyms are similar words with multiple meanings. The terms discovered using this method reflect the real-time language of the domain, without reliance of out-dated lists or static thesauri. By describing vocabularies from actual usage, rather than *prescribing* a standard set of terms, users are more likely to respond favorably.

Text analysis emerged as a prominent field of study in the 1980s. Early attempts were a labour-intensive, manual processes; with the advent of the IT age, text analytics became much more efficient. Text analysis also became a multidisciplinary effort, drawing on the fields of information retrieval, data mining, machine learning, statistics, and computational linguistics. For these fields, text analysis was used to model, structure, and analyse textual information.[6]

In the 1990s, text analytics refocused from algorithm *development* to algorithm *application*. Researchers went from using large corpora as resources to produce better text analysis algorithms, to deploying these algorithms for

[6] Mark N. Brock, "Computerised Text Analysis: roots and research," *Computer Assisted Language Learning* 8, no. 2-3 (1995): 231.

practical extraction.[7] Today, modern text analysis systems are moving from "data processing" to "concept processing." That is, the basic unit of processing has moved from being an atomic piece of data to a semantic concept that may contain many pieces. A semantic concept carries an element of interpretation and exists in context with other concepts.[8] This shift to concept processing is important for homonym retrieval. For example, terms exist as relations of their context, and different meanings can be derived from different contexts. Context-aware, concept-level analysis can distinguish between the various meanings of homonyms.

Text analysis usually works with a corpus, which is a domain-based or general collection of literature. Corpus-based analysis allows for association patterning, "quantitative relations, measuring the extent to which features and variants are associated with contextual factors."[9] The method is empirical, allowing the analysis of patterns in natural language of the corpus. Corpus-based analysis can be used both quantitatively and qualitatively. The results are more reliable because natural language has many nuances, and is best observed

[7] Marti A. Hearst, "Untangling text data mining" (paper presented at the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, 1999), 8.

[8] Janez Brank, Marko Grobelnik, and Dunja Mladenić, "Automatic Evaluation of Ontologies," in *Natural Language Processing and Text Mining*, ed. Anne Kao and Stephen R. Poteet (London: Springer-Verlag, 2007): 193.

[9] Douglas Biber, Susan Conrad, and Randi Reppen, *Corpus Linguistics: investigating language structure and use* (Cambridge: Cambridge University Press, 1998): 4.

through multiple examples. Corpora allow for a scope and reliability of research not otherwise possible.[10]

In the recent past, terminology extraction has become an area of innovation in text analytics. Text analysis has been used for terminology and thesaurus extraction, as well as the generation of ontologies. The biomedical domain, for example, has developed controlled vocabularies and ontologies using text analysis.[11] The degree of similarity between words, in terms of synonyms and homonyms, is important for developing archaeological ontologies. Word similarity has been traditionally captured with thesauri.[12] However, by retrieving synonyms and homonyms from a corpus of domain specific literature, the list of words can be developed into a natural language ontology. By doing so, the ontology can reflect all possible values for a concept as it is actually used in the literature.

Similarity between words can be computed under two relations: grammatical and non-grammatical. A grammatical relation occurs when words

---

[10] Ibid.

[11] For more information on ontology development using text analysis in the biomed domain, see: Irena Spasić, Sophia Ananiadou, John McNaught, and Anand Kumar. "Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text." *Briefings in Bioinformatics* 6, no. 3 (2005): 239-51; Irena Spasić, Daniel Schober, Susanna-Assunta Sansone, Dietrich Rebholz-Schuhmann, Douglas B. Kell, and Norman W. Paton. "Facilitating the Development of Controlled Vocabularies for Metabolomics Technologies with Text Mining." *BMC Bioinformatics* 9, no. 5 (2008): 1-16.

[12] Ido Dagan, "Contextual Word Similarity," in *Handbook of Natural Language Processing*, ed. Robert Dale, Herman Moisl, and Harold Somers (New York: Marcel Dekker, Inc., 2000): 459.

co-occur within known syntactic constraints. That is, the order or arrangement of words dictates where words can conceivably appear. The context of a word restricts the words that can occur with it, constrained by specific syntactic relations (for example, grammatical relations can assess the likelihood of the adjective "edible" and the verb "cook" appear together more often than "edible" and "drive").[13] Non-grammatical relations are easier to analyse when mining for synonyms because they do not require grammatical tagging of documents. Grammatical tagging can be partially automated, but the results are unreliable and manual correction is time-consuming. The analysis below relies on non-grammatical analysis.

Non-grammatical relations can be broken down into three main areas: n-grams, and large and small co-occurrence windows.[14] In the area of computational linguistics, an *n*-gram is a "sequence of *n* words that appear consecutively in [a] text."[15] Usually forming bigrams and trigrams (sequences of two or three words), *n*-grams can be applied as vectors, allowing for comparison. The vector space model, also known as vectorial semantics, is an algebraic model for representing text and/or documents as vectors.[16] Vectors represent both direction and magnitude, giving the vector space model a multi-dimensional aspect. The magnitude of a vector is set by the frequency with which it appears;

[13] Ibid., 462.

[14] Ibid.

[15] Ibid., 462-63.

[16] "What is Vector Space Model," <http://www.igi-global.com/dictionary/vector-space-model/31436> (Accessed: 14 September, 2013).

the direction is relative to other terms.[17] If two *n*-grams appear "close" in their vector representation, then it is likely they are similar in meaning.[18] Co-occurrence windows work on a similar, if simplified, principle. Co-occurrence of in a large window simply suggests that the words are related to the general topic; a "large window" is relative, and can be anything from a paragraph to a whole document.[19] Co-occurrence in a small window refers to sentences or phrases, and is used to by some models to index grammatical relations between words for automatic processing.[20]

2.2.1 Latent Semantic Analysis

Latent semantic analysis (LSA) is a popular method of computing word similarity using corpora. Susan Dumais et al. first introduced LSA in 1988 as a method of information retrieval.[21] LSA uses the non-grammatical relations outlined above to compute similar words using vector space models; it is often referred to as a "bag of words" (BOW) model, as it treats the corpus as a selection of individual, unrelated terms.[22] LSA does not require grammatical tagging to work; while it incorporates context, word context is computed as a large window

[17] Peter D. Turney and Patrick Pantel, "From Frequency to Meaning:Vector Space Models of Semantics," *Journal of Articial Intelligence Research* 37 (2010): 158.

[18] Dagan, "Contextual Word Similarity," 463.

[19] Ibid.

[20] Ibid., 463-64.

[21] Susan T. Dumais et al., "Using latent semantic analysis to improve access to textual information" (paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Washinton, DC, 1988).

[22] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham, "Introduction to Latent Semantic Analysis," *Discourse Processes* 25 (1998): 2.

(a paragraph or whole document). It operates on the idea that the collective contexts of a word, across a corpus, reveal relationships that determine the similarity between words.[23] This core concept behind LSA lies in Zellig Harris' 1968 distributional hypothesis: "the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities."[24] That is, words that occur in similar contexts tend to have similar meanings.

To create a "semantic space," LSA uses vectorial semantics to represent a corpus as a series of term vectors and document vectors. As vectors represent direction and magnitude, each vector exists in its own dimension, allowing for similarity to be computed in multiple "directions." For example, it allows for "artefact scatter" to be similar to both "site" and "assemblage" without implying "site" and "assemblage" are similar (they are, but that similarity would be inferred from their own relationship to one another). By transforming words and documents into vectors, term similarity can be calculated. Term weighting, traditionally term frequency-inverse document frequency (tf-idf) in LSA, ensures rare terms are given extra "weight" to reflect their relative importance in constructing overall concepts and meaning (as apposed to "stop words" like "the", "is", "a", etc. which appear with high frequency).[25]

---

[23] Ibid.

[24] Z. S. Harris, *Mathematical Structures of Language* (New York: Wiley, 1968): 12.

[25] Thomas K. Landauer and Susan T. Dumais, "Latent semantic analysis," <http://www.scholarpedia.org/article/Latent_semantic_analysis> (Accessed: 8 October, 2013).

At this point, the true term-document relationships are obscured by variability in word choice; that is, synonyms and homonyms.[26] To extract these terms, a singular value decomposition (SVD) algorithm is applied to the corpus. The SVD algorithm acts on the vectors in such a way that "every passage is represented as a vector whose value is the sum of vectors standing for its component words."[27] That is, a set of word vectors is transformed into a phrase vector, the phrase vector being the sum of the term vectors (shown in Figure D). The term vectors for "the", "quick", "brown", "fox" are summed into the phrase vector "the quick brown fox." Phrase vectors allow associations to be made between words and documents based on phrases rather than individual words, which brings a degree of conceptual similarity. As a result of SVD, it is similar concepts, too, that bring the vectors closer in the semantic space.[28] Unfortunately, visualising the processes used by LSA is impossible; LSA can project thousands of dimensions and manipulate them in ways that the brain cannot imagine.

[26] Ludovic Lebart and Martin Rajman, "Computing Similarity," in *Handbook of Natural Language Processing*, ed. Robert Dale, Herman Moisl, and Harold Somers (New York: Marcel Dekker, Inc., 2000): 487.

[27] Landauer and Dumais, "Latent semantic analysis".

[28] Lebart and Rajman, "Computing Similarity," 487.

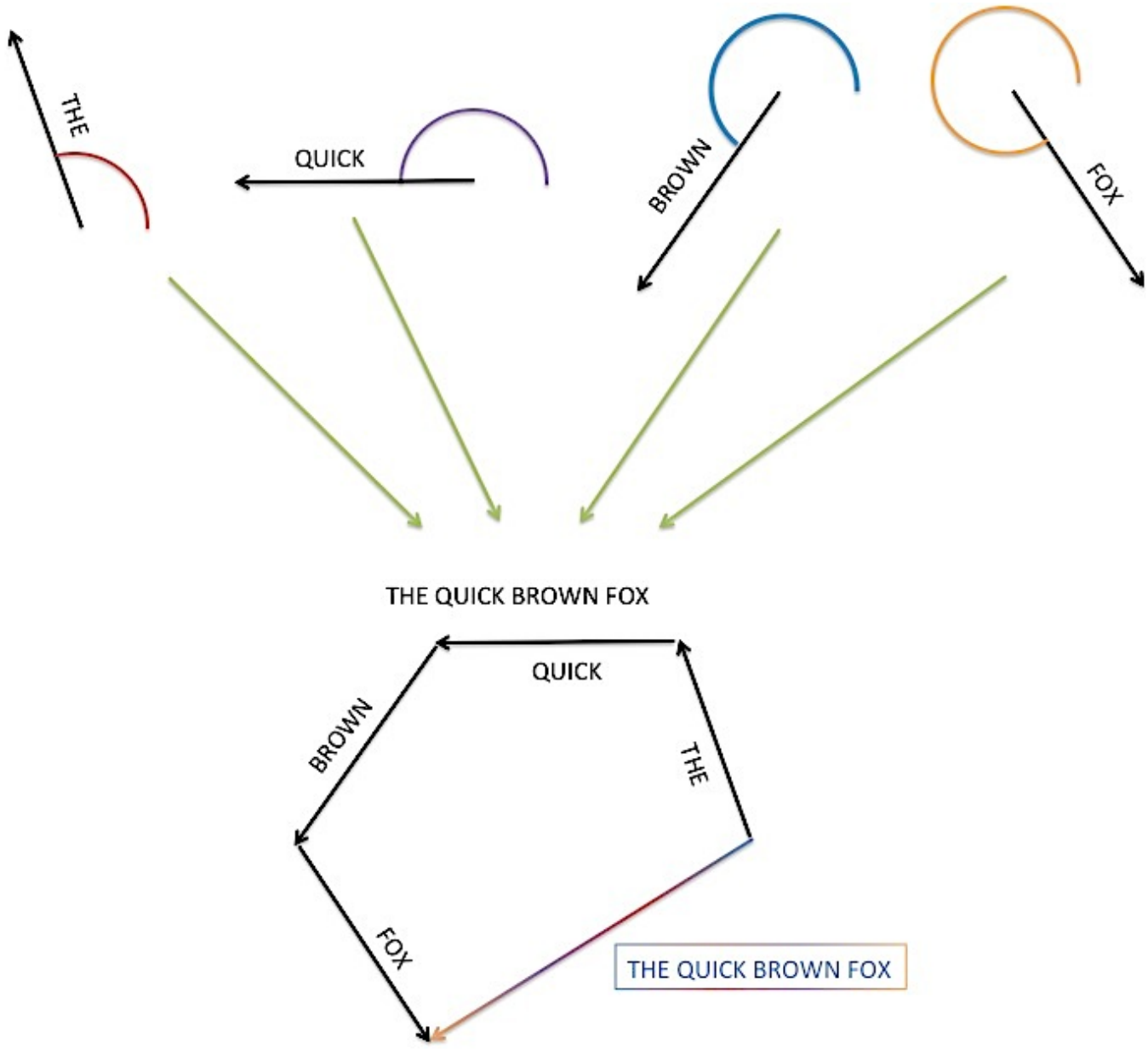


FIGURE D. AN EXAMPLE HOW THE SVD ALGORITHM SUMS VECTORS.

LSA has many practical uses, particularly in web search engines. Google, for example, makes use of LSA principles for ranking pages in terms of relevance and expanding search terms to include synonymous terms.[29] As a result, Google can return "themed content" rather than strictly adhering to search terms. LSA is the method of choice for this pilot, and is discussed further in chapter three.

2.3 Applied Text Analysis in Archaeology

Several projects in archaeology have used computer-assisted text analysis. Among these are the Archaeotools Project and the Armadillo Project. These two project focus on populating instances of ontologies as form of metadata. That is, using text analysis, they identify "buzz words" (such as names, dates, or places) in order to link them to an ontology. As a result, these projects focused on information extraction using, e.g., named-entity recognition and fact extraction (see 2.3.1. below), not terminology extraction, as is being attempted here. The focus on linking ontologies to documents likely results from the attention given to metadata standards in archaeological databases; content standards are secondary, if addressed at all. These projects focused on extracting metadata from document through text analysis. This project differs by extracting terminology and the relationships between terms from archaeological literature to inform data content standards

[29] The Google search function is in constant development and, due to the inconceivable size of the web as a corpus, is not traditional LSA as described here. However, it provides a relatable example of the process and is useful for illustration purposes.

2.3.1 The Archaeotools Project

The Archaeotools Project is an example of a successful application of text analysis in archaeology. The project lasted from 2007-2009, funded by the UK's e-Science Research Grants Scheme, a collaboration between three major funding bodies: the Arts and Humanities Research Council (AHRC), the Engineering and Physical Sciences Research Committee (EPSRC) and the Joint Information Systems Committee (JISC). Based in the United Kingdom, the project was developed in partnership with the ADS and the Natural Language Processing Research Group at the University of Sheffield. In 2009, this project was integrated into the ADS ArcheoSearch.[1] The aim of the project was to integrate published records referring to UK archaeological sites and monuments with information from "grey literature". This information was then made searchable on a single, faceted browser interface.[2]

The key goal of the project was to develop a faceted classification system for grey literature on the OASIS database. Faceted classification is the assignment of constructed attributes, and allows for multiple attributes to be assigned to one object. Constructed attributes essentially act like metadata. They provide a way to discover literature by the features (attributes) associated with it. Such a classification system allows various documents to be associated with multiple attributes (facets), improving discoverability (see Figure E for an illustration). Each facet created by the system covers aspects such as what,

[1] http://archaeologydataservice.ac.uk/archsearch

[2] Julian D. Richards et al., "The Archaeology Data Service and the Archaeotools Project: Faceted Classification and Natural Language Processing," in *Archaeology 2.0*, ed. Eric C. Kansa (Cotsen Institute of Archaeology, 2011).

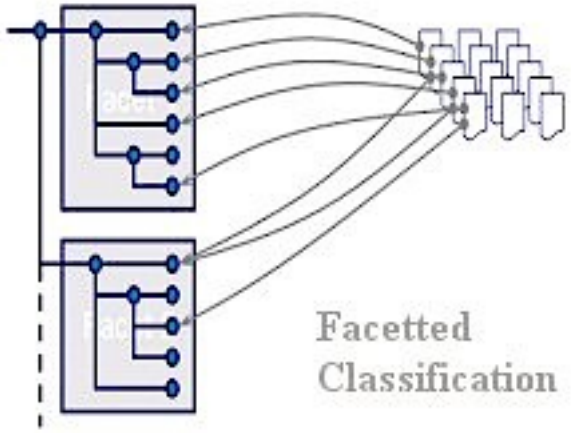


FIGURE E. AN ILLUSTRATION OF HOW FACETED CLASSIFICATION WORKS.

where and when.[3] That is, a researcher can hypothetically query something like "pottery," "Londinium," "2nd C AD" to retrieve articles related to any or all of those metadata elements (artefact class, place, date), since each relevant document has been previously associated with the appropriate metadata.

In order to create these facets of classification, the Archaetools project employed two main methods of information extraction: fact extraction and named-entity recognition (NER). Fact extraction is explicit in the title; it is the "identification of facts."[4] For example, fact extraction could be used to link archaeological artefacts to locations, creating a relationship of the form "artefact-found-at-place."[5] Thus, the fact of "location" is extracted from a document describing one or more artefacts, linking those artefacts to that place. NER aims to recognise named entities within a text, such as an archaeological period or the name of an investigator. For example, a facet based on an investigator could

[3] Jeffrey et al., "The Archaeotools Project:," 2512.

[4] Ibid., 2515.

[5] Ibid.

bring up all relevant projects that an archaeologist is associated with.

This type of faceted classification created by fact extraction and NER is useful for archaeological literature, as archaeological data often crosses several fields of enquiry and contains multiple findings. Thus, fact extraction and NER contribute the automated generation of facets, associating documents with metadata and enhancing discoverability.

The Archaetools project used content analysis techniques to extract standardised metadata from documents, facilitating the discovery of datasets using keyword searches. The discovery of datasets, however, does not necessarily mean that the content will be fully accessible without some degree of semantic reconciliation.

2.3.2 The Armadillo Project

The Armadillo Project (2005-2007) retrieved information according to a predetermined ontology, and used this information to populate the ontology with instances.[1] The University of Sheffield, in conjunction with the Natural Language Processing Group, were the primary developers. The project was funded by the AHRC, like Archaeotools. The pilot was completed in June 2007, and the Armadillo system has seen little public development since.

The project was similar to the Archaeotools project in that it used information extraction to extract facts and entities. However, rather than reclaiming instances to create a way to search a database, Armadillo sought to

[1] "About Armadillo," <http://www.hrionline.ac.uk/armadillo/armadillo.html> (Accessed: 27 September, 2013).

populate ontologies with instances of its occurrence.[2] The ontology used in the Historical Data Mining Project, the pilot project of Armadillo, referred to "dates, names and places centred around 18th Century London." [3] Thus, Armadillo was used to mine a corpus of documents for instances of these dates, names and places, and link them to their corresponding class in the ontology. Again, the focus is aimed at the discovery of datasets by metadata, rather than data interoperability by content.

Armadillo relies on content analysis for ontology annotation; it extracts information to link a document to a concept. No attempt, however, was made to construct ontologies descriptively from the documents or to clarify the use of terms (e.g., through identification of synonyms or homonyms). Like the Archaeotools project, prescriptive ontologies were used, which likely overlooked some metadata associations due to word variation, change over time, and other problems. However, like Archaeotools, the project was deemed successful regarding the metadata improvement goals it set out to achieve.

2.4 Conclusions

Controlled vocabularies facilitate a higher degree of data interoperability than can be achieved by metadata standards alone. Controlled lists represent the simplest method of controlling vocabularies. However, their application is limited in archaeology. They do not allow for synonyms and homonyms to be represented or reconciled. Thesauri make an attempt to reconcile vocabularies,

[2] Ibid.

[3] Ibid.

however their weak structure does not capture the multi-level hierarchies or networks that can clarify relationships between ambiguous terms. Ontologies present the best option for describing content standards in archaeology. As a network of related concepts and terms, ontologies allow for a variety of relationships to be illustrated. These relationships define the ways in which a term can be used. Despite the problematic application of the CIDOC-CRM to fieldwork, the standard illustrates how ontologies have been developed in the cultural heritage domain. Ontologies have been effectively collected using automated text analysis methods in other domains. These descriptively produced ontologies are preferable, as they reflect the actual use of terms and concepts within the respective domain as it changes over time.

Latent semantic analysis presents the most viable option for computer-assisted descriptive generation of ontologies, as it not only collects similar words, but also concepts where the words themselves are ambiguous (e.g., synonymous or homonymous). Text analysis has been employed previously in archaeology; however, in keeping with larger trends, the focus has been on metadata generation and annotation. Chapter three will outline how a derivative of LSA, positional indexing, can be used to derive concept similarity using a domain-specific corpus.

CHAPTER THREE: PRELIMINARY ANALYSIS OF A ROCK ART CORPUS

This chapter will present the preliminary analysis of a rock art corpus processed using positional indexing, a derivative of latent semantic analysis (LSA). This chapter will cover four key aspects of investigation: the construction of the corpus, the methods of contextual processing, the software infrastructure, and the preliminary results.

Preliminary results are not sufficient for ontology creation on their own. Instead, they serve as a proof of concept for applications of automated text analysis in the construction of descriptive archaeological ontologies. While refinement and expansion of the corpus, as well as the application of complementary techniques, will be required for actual ontology generation, this study demonstrates the power of text analysis techniques as a method of conceptually linking terminology in archaeological literature.

3.1 THE CORPUS

The corpus assembled for this experiment totalled ~2.3 million words before processing. It consisted primarily of journal articles from one particular domain of study, Australian indigenous rock art. No archaeological domain corpora currently exist that are suitable for ontology generation, requiring the generation of one for this analysis. A well-constructed domain corpus ensures that a wide range of relevant terms appear within the corpus, so that a representative sample of concepts and meanings can be analysed.

3.1.1 Corpus Generation and Domain

The corpus for this experiment was manually generated via keyword searches of journal databases. The databases included JSTOR, Anthropology Plus, ProQuest Research Library, and Web of Science. Particularly relevant journals, such as *Rock Art Research* and *Australian Archaeology*, were also searched volume by volume for relevant articles.

Search queries were performed using keyword searches of "rock art", paired with geographic locations often associated with Australian indigenous rock art.[1] The method is somewhat problematic; ambiguous terms are ill suited for keyword searches. This problem is mitigated by the fact that "rock art" is, by far, the most popular descriptor for the domain (see section 3.2.2).

Specific articles were selected based on article title. This method is inclusive, and it produced some outliers only passingly or tangentially related to the core concept of rock art. The analysis software, however, uses a target word list to determine document relevance, so outliers have a limited impact.

A selection of both technical reports and reflective sources were collected. Technical reports include articles that address the process, progress or results of field investigations. Reflective sources refer to more theoretical or synthetic articles. The intended audiences for reflective sources may include non-specialists, making it broader than technical reports, which are aimed towards

[1] These included the general term "Australia", and the more specific "Arnhem Land," "Kimberley," "Mount Grenfell," "Kakadu," the "Olary region," "Pilbara," the "Cape York Peninsula," the "Bradshaws," "Carnarvon Gorge," "Murujuga," "Sydney," "Uluru," and "Borroloola."

domain experts. Specialist terms should be more frequently used in the latter. Together, these two resource types produce a rich pool of data to assess for a variety of domain-specific terms.

JSTOR returned a diverse selection of both types of sources, and contributed the majority of documents to the corpus. Unfortunately, it is the database that also produced the most documents (PDFs) of inferior quality, yielding a more difficult cleansing and conversion process. Anthropology Plus, on the other hand, returned a good selection of both technical and reflective sources. The database was also advantageous in the fact that the plain text was saved within the downloaded document for easy and consistent extraction (see section 3.1.2 for further information). ProQuest Research Library and Web of Science returned a limited selection of relevant sources, most of which were technical reports, but the documents were of high quality. Technical reports formed the majority of the corpus.

The corpus content dates from 1990 to 2013. The quality of the documents created before 1990 proved too unreliable for processing in the time available. Moreover, by limiting the search to the past two decades, the terms included in the corpus reflect the current domain vocabulary.

3.1.2 Corpus Conversion and Cleansing

The software used for analysis requires all documents to be in plain text format (i.e. text without mark up or formatting, commonly TXT files). The original corpus was made up of PDF files, generated from a variety of software packages and in a variety of PDF versions. PDFs can be formatted in three ways: as a bundle of images (for example, a scanned document), as text bounded by

paragraphs (best retaining the original formatting of the document), or simply as text characters placed on a page. As such, PDFs are unpredictable when it comes to text extraction, particularly when no indication is given of how they were generated. Conversion to plain text standardises all documents.

To convert the files from PDFs to TXTs, a workflow using Adobe Acrobat XI was used. Acrobat XI was chosen because Adobe is the creator of the PDF file format and their software is best positioned to convert PDFs to other formats. Acrobat XI also has powerful optical character recognition (OCR) capabilities for those documents stored only as images (such conversions still required visual inspection to ensure their quality).[2] Overall, the qualities of PDFs for this corpus were relatively high.

Document cleansing was a manual process, requiring two steps. First, every document was given a top and end cull, removing things like the journal listings page and reference lists. Titles like “abstract”, “acknowledgements” and “references” are good placeholders for this process. During this stage, the second task was performed: documents were visually inspected for quality control. If the PDF failed on visual inspection—if the text condensed itself (e.g., “thisisanexample”) or characters did not convert correctly (e.g., “thïš îś æń ęxåmpłè”)—the document was discarded.

[2] OCR refers to the conversion of files showing text—including scanned or printed text—into a machine-readable format. For more on Acrobat OCR capabilities, see: dstromfe, "Acrobat OCR: Make your scanned documents searchable," <http://blogs.adobe.com/acrobat/acrobat_ocr_make_your_scanned/> (Accessed: 14 October, 2013).

The corpus used here is relatively small, so manual cleansing was a practical step. For larger corpora, automated cleansing is required. Several sophisticated analytics tools are available for such a task, but even then some visual inspection of all documents is still recommended to control quality.

3.2 Contextual Processing

Contextual processing allows the corpus to be positioned within larger corpora studies. First, the size and structure of the domain corpus under study can be compared with general corpora. Second, results from the Google Ngram Viewer 2.0 highlights the popularity of domain terms in a general corpus setting. Third, a word cloud provides a useful representation of bigrams in a BOW model, and an explanation of relative term frequency in the corpus.

3.2.1 Corpus Size Relative to Other Corpora

After document cleansing, the corpus generated was ~1.6 million words. This corpus is miniscule when compared to general corpora. For example, the Google corpus (a general corpus) contains ~155 billion American English words and ~34 billion British English words.[3] Another example is the Corpus of Contemporary American English, which contains 450 million words dated between 1990 and 2012. The Oxford English Corpus, likewise, contains ~2 billion words.[4] These large general corpora are best suited for analysis of

[3] Mark Davies, "Google Books," <http://googlebooks.byu.edu/> (Accessed: 1 October, 2013).

[4] Mark Davies, "The Corpus of Contemporary American English," <http://corpus.byu.edu/coca/> (Accessed: 1 October 2013, 2013); Oxford Dictionaries, "About the Oxford English Corpus,"

language development over time. They are created using literature of all genres; for example, the Google corpus has been estimated to represent 6% of all literature ever published in English.[5]

General corpora have limited use when it comes to thesaurus generation, requiring that a domain-specific corpus be used. Analysis based upon a domain corpus ensures word usage is specific to the domain under study. A word such as "pictogram", for example, can occur over multiple domains. By ensuring the corpus is reflective of, e.g., archaeology, the context of "pictogram" is limited to how it is used in archaeology. Domain specific corpora also ensure that relevant terms occur frequently enough for meaningful analysis.

3.2.2 Google Ngram Viewer 2.0

In order to contextualise terms appearing in the domain corpus, however, it is important to understand how they are represented in general corpora. The Google Ngram Viewer, which operates on the general Google corpus, provides a useful tool that visualises the frequency with which *n*-grams (a set of *n* words) appear over any given period of time.[6]

Google Ngram Viewer was used to analyse three sets of three bigrams each. The chosen terms represent some common synonyms used within rock art

---

<http://www.oxforddictionaries.com/words/about-the-oxford-english-corpus> (Accessed: 1 October, 2013).

[5] Stephanie E. Vasko, "Examing Trends via Google Ngram: taking a closer look," <https://sites.psu.edu/stephanievasko/2013/09/29/examining-trends-via-google-ngram-taking-a-closer-look/> (Accessed: 14 October, 2013).

[6] Michel et al., "Quantitative analysis of culture using millions of digitized books."

studies (introduced in section 2.1.1.1). The synonyms were drawn primarily from a debate in *Rock Art Research* volume 23, supplemented by other controversies identified in section 2.1.1.1.[7] They are as follows:

Set 1	Set 2	Set 3
1. Rock art	1. Pictogram	1. Petroglyph
2. Rock image	2. Pictograph	2. Rock engraving
3. Rock marking	3. Rock painting	3. Rock carving

The graphs were produced using Google Ngram Viewer 2.0 in September 2013; they have been smoothed to best represent the trends of the data.[8] The x-axis represents the date. The y-axis of the graph represents the percentage of all occurrences of each bigram as it occurs within the Google corpus.[9]

Although the corpus for this study is limited to 1990-2013, the time frame for general corpus contextualisation is 1860-2009. The start date was determined by the first term to come into common use ('pictograph' begins to appear frequently just after 1860 and increases thereafter). The Google corpus

[7] Chippindale and Taçon, "What's in a Word, What's in a Hyphen?."; Shwartz, "Further comments on Christopher Chippindale and Paul S. C. Taçon's 'What's in a word, what's in a hyphen?': On 'Rock Art' History and Terminology."

[8] Smoothing ensures the data represented shows a moving average. For example, a smoothing of 1 (as is used in these graphs) when given the year 1900 means that the data for 1900 is the raw count for 1900 plus one value either side (i.e. the average of the raw data for 1899, 1990, and 1901). For more information, see: Google, "About Google Ngram Viewer," <https://books.google.com/ngrams/info> (Accessed: 1 October, 2013).

[9] Ibid.

does not extend reliably past 2009.[10] Contextualisation within a longer-duration general corpus shows how the terms in my domain corpus have changed over time.

Several limitations regarding Google Ngram Viewer have been raised. Most importantly, concepts can be expressed by any number of words or phrases at any given time, all of which adapt and change as time passes.[11] For example, a word such as "conscience" declined in use during the twentieth century, while the phrase "eat at you" increased. The decline of "conscience" does not mean there is less discussion about the underlying concept of morality; it has simply been replaced by another way (or several) of referring to it.[12] Word choice is important; I can only compare specific word frequencies over time, as compared to known synonyms. It is not an exhaustive analysis of word choice.

Figure F deals with the first set of terms. It shows the frequency of "rock art" and its synonyms, "rock image" and "rocking marking", as they occur between 1860 and 2009. "Rock art" is clearly the preferred term; lines representing "rock marking" and "rock image" barely register on the graph due to low representation in the corpus. ("Rock art", too, does not appear significantly before 1930; again, a result of low representation compared to later dates rather than no representation at all.) For this reason, Figure G shows a

[10] Vasko, "Examing Trends via Google Ngram".

[11] John McWhorter, "David Brooks' Favorite New Theory of Language Is Wrong," <http://www.newrepublic.com/article/113274/david-brooks-language-our-words-dont-reveal-our-worldview#> (Accessed: 4 October, 2013).

[12] Ibid.

detail of the graph, showing only “rock image and “rock marking”. “Rock image” has made significant progress in the past few decades. “Rock marking” may be considered synonymous with the technique of rock engraving for some users. Likewise, the “art” in “rock art” complicates the intended idea with modern, Western notions of art as a non-utilitarian activity. “Rock image” may be gaining in popularity because it is considered to be a more neutral term.

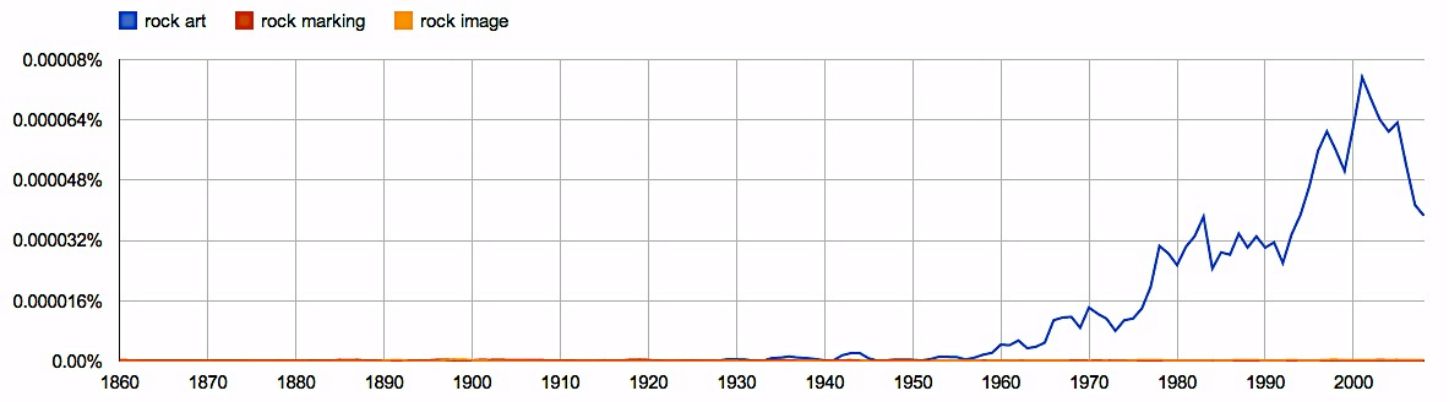


FIGURE F. NGRAM GRAPH REPRESENTING THE USE OF "ROCK ART", "ROCK MARKING" AND "ROCK IMAGE" AS A PERCENTAGE IN THE GOOGLE CORPUS OVER TIME.

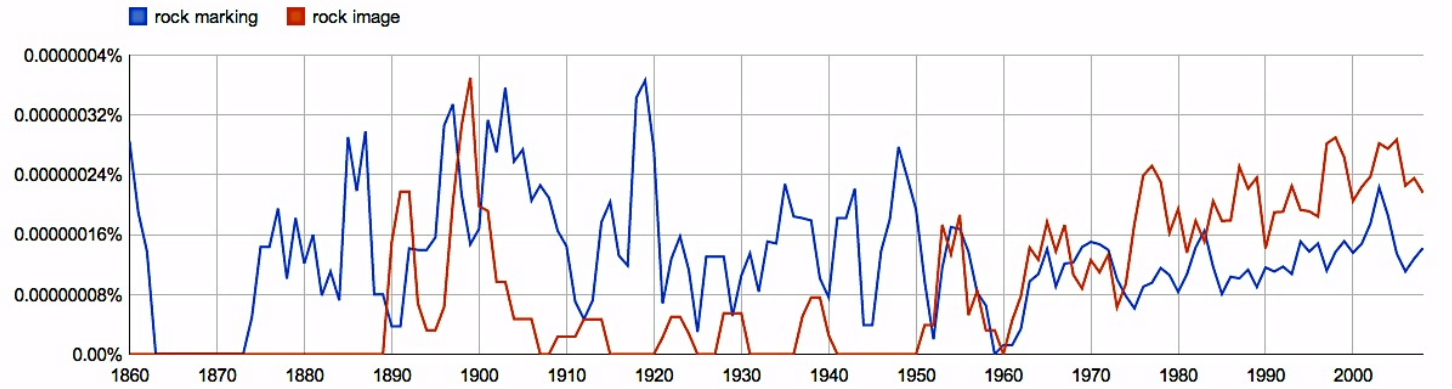


FIGURE G. NGRAM GRAPH REPRESENTING THE USE OF "ROCK MARKING" AND "ROCK IMAGE" AS A PERCENTAGE IN THE GOOGLE CORPUS OVER TIME.

Figure H deals with the second set of terms. It shows the frequency of three other terms, “pictogram”, “pictograph” and “rocking painting”, as they occur between 1860 and 2009. Use of all terms has increased since 1860. Unlike Figure F, the problem is word choice. The IFRAO glossary lists “pictogram” as a

painted "art motif" and "pictograph" as a "writing character of figurative appearance."[1] However, David Whitely, in his instructive *Introduction to Rock Art Research* (2010), gives the name "pictograph" to both rock carvings and drawings.[2] As a result, there is a tension between terms as they compete for use in related, but distinct, domains (for example, archaeology vs. palaeography). Even within archaeology, it would be wrong to say that "pictogram" is the most popular word used to refer to painted motifs as per the IFRAO definition, since it could just as easily refer to rock carvings as per Whitely's definition.

As a result, Figure H is an unreliable representation of word frequency due to conflicting word meanings and crossover from other domains. The intertwined meaning and use in literature belies a larger problem of synonyms and homonyms in archaeological vocabularies. Such ambiguities limit the analysis that can be performed on general corpora.

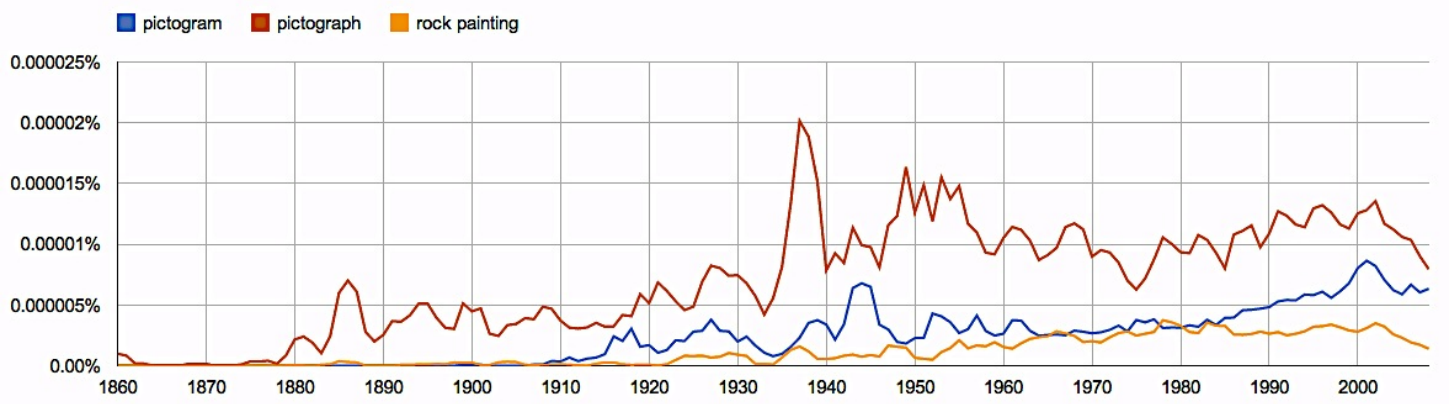


FIGURE H. NGRAM GRAPH REPRESENTING THE USE OF "PICTOGRAM", "PICTOGRAPH" AND "ROCK PAINTING" AS A PERCENTAGE IN THE GOOGLE CORPUS OVER TIME.

[1] Bednarik, "IFRAO Rock Art Glossary".

[2] Whitley, *Introduction to Rock Art Research*: 23.

Figure I deals with the third set of terms. It shows the frequency of yet another three terms: “petroglyph”, “rock engraving” and “rock carving”, as they occur between 1860-2009. The specialist term “petroglyph” is clearly more frequently used in comparison to its synonyms. Again, the important note the underlying problems that occur with managing corpora. When “rock carving_NOUN” was entered (“_NOUN” as denoting the use of “rock carving” where it occurs as a noun in the corpus), it returned no results. “Rock carving” appears, albeit with low frequency, in the bigrams generated from my own corpus (see Appendix II); while not conclusive (no parts of speech (POS) tagging was used to generate the bigrams), it is certainly suggestive of a problem. It also may suggest an unrepresentative sample of rock art literature in the Google corpus.

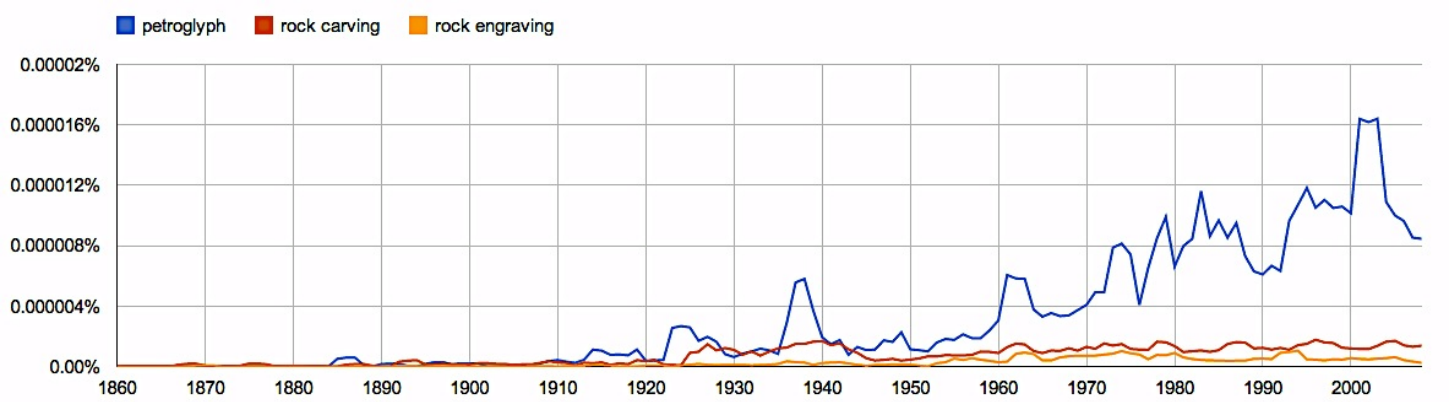


FIGURE I. NGRAM GRAPH REPRESENTING THE USE OF "PETROGLYPH", “ROCK CARVING” AND "ROCK ENGRAVING" AS A PERCENTAGE IN THE GOOGLE CORPUS OVER TIME.

This exercise also shows the problem of general corpora; as seen in Figure H, underrepresentation of domain-specific literature is likely, and can cause problems with results. Although analysis of a general corpus provides helpful context, it is important to use a domain corpus for synonym generation,

as it ensures that the most relevant and appropriate instances of a term are analysed.

3.2.3 Word Cloud Generation

Word clouds can produce intuitive representations of bigram frequency in a bag of words (BOW) analysis. Bigrams were considered to be the best representation of the corpus, considering that "rock art" itself is a bigram. The resulting word cloud can be viewed on the cover of this text, and in Appendix I. To produce a word cloud, a word (or, in this case, bigram) list must first be produced (see Appendix II for the code used to perform this task). The bigrams were created using the following conditions:

1. Words had to be at least three characters. This step was used to eliminate words with repeating characters, which were prominent in the corpus.
2. Stop words were ignored, and not included in the bigrams.
3. Output was lowercase, ignoring differences between, e.g., "Rock art" and "rock art."
4. Words starting or ending with a dash were ignored. This condition can result in some bias: "rock-art," for example, may be underrepresented because it can appear over two lines (thus ending or starting with a dash: "rock-" or "-art").
5. If the word contained a dash, it is a bigram in itself (for example, "rock-art" is treated like "rock art"). This is a personal preference, and is not a required feature.

6. Pairing each word with each of their neighbours created the bigrams. For the phrase “little painted rock”, for example, “little painted” and “painted rock” are two bigrams generated.

The list of bigrams produced appears in Appendix III. The open-source package WordCram was then used to produce a word cloud from the bigrams.[3]

While word clouds serve as a useful visualisation, analysis will focus on the bigram list found in Appendix II. It is unsurprising that rock art is the most frequently used bigram in the corpus. The top twenty terms represent the domain of the corpus particularly well; bigrams like “aboriginal people” and “central australia” matched with the top term “rock art” describe the corpus as a Australian rock art corpus. These relationships demonstrate that the collection method worked, producing a corpus relevant to the domain of choice.

Just from the selection of bigrams generated here, related terms were identified that would not appear in a typical thesaurus. From my own knowledge of the domain, I selected the following ten terms which I consider to be related to the primary term “rock art”:

1. Rock art
2. Rock painting
3. Cave art
4. Rock shelter

[3] Daniel Bernier, "WordCram," <http://wordcram.org/> (Accessed: 24 September, 2013).

5. Paleolithic art (unexpectedly, the US spelling featured predominantly in the corpus)
6. Aboriginal art
7. Art image
8. Rock marking
9. Painted rock
10. Parietal art

These terms highlight the need for ontologies in archaeological research; the terms listed here are undoubtedly related to "rock art", but are not necessarily synonyms. For example, "cave art" is a derivative of "rock art", reflecting a specific location of that art. In the context of archaeology, "aboriginal art"—itself related to, but not synonymous with, "paleolithic art"—can refer to "rock art" as well as other forms of pre-modern Australian art. "Rock art" (or, indeed, a "rock painting") can be found within a "rock shelter." A "rock marking" can appear with a "rock painting" forming "rock art". "Rock art" is a domain specific term for "parietal art." Relationships amongst these terms are more nuanced than synonymy, requiring a network to define them adequately. Using a network of slots (the properties that link concepts), ontologies can bring these terms into appropriate relation. The relationship between these terms will be investigated further in sections 3.4.2 and 3.4.3.

3.3. Software Infrastructure

For the present analysis, the software package Semantic Vectors was used to run this experiment.[1] The process also required a target term list from which to derive related terms.

3.3.1 Target Term Lists

Four short target term lists have been constructed for this experiment. The first three lists build upon the argument found in *Rock Art Research*, volume 23 (used for the *n*-grams in section 3.2.2, and first introduced in the discussion in section 2.1.1.1):

1. Rock art	1. Pictogram	1. Petroglyph
2. Rock image	2. Pictograph	2. Rock engraving
3. Rock marking	3. Rock painting	3. Rock carving

Identified as substitute terms in the literature, these bigram lists form the basis of comparison for the first run of the software. The fourth list consists of the 10 terms found in bigrams related to rock art that were generated for section 3.2.3.

1. Rock art
2. Rock painting

[1] Instructions for installation of Semantic Vectors can be found on the Semantic Vectors wiki. For this experiment, Semantic Vectors 3.8 was installed, using Lucene 3.6.2. While the instructions are fairly straight forward, it is recommended that any archaeology researcher without IT skills, who are intending to produce their own results, consult an IT specialist. Dr. Brian Ballsun-Stanton, data architect of the Federated Archaeological Information Management Systems (FAIMS) Project, is the IT specialist I give full credit to for this experiment, and its successful run.

3. Cave art
4. Rock shelter
5. Paleolithic art
6. Aboriginal art
7. Art image
8. Rock marking
9. Painted rock
10. Parietal art

This final list serves two purposes. First, it forms a basis of semantic comparison between terms that are considered related. Second, using these terms, I attempt to identify similarity between contexts comparing results between individual analyses through the software.

3.3.2 Semantic Vectors

Semantic Vectors is an open-source package that "creates word space models from free natural language text."[2] As Magnus Sahlgren explains,

> The word space model is a computational model of word meaning that utilises the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity.[3]

[2] Dominic Widdows, "Semantic Vectors," <https://code.google.com/p/semanticvectors/> (Accessed: 14 October, 2013).

[3] Magnus Sahlgren, "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional spaces" (Stockholm University, 2006), 5.

That is, terms and documents are transformed into vectors and these vectors are measured to approximate a value indicating "closeness" in meaning. As vectors represent direction and magnitude, the vectors give LSA its dimensional aspect. Each vector is its own dimension, allowing for the "semantic space."

Apache Lucene is the portion of Semantic Vectors that handles documents. [4] Lucene is a software library that provides ready-to-use infrastructure for document searching and indexing. Developed by the Apache Software Foundation, Lucene is an open source, Java application programming interface (API). It is best described as a library that stores the texts being analysed by Semantic Vectors; Lucene specifies how routines are run in regards to data structures, object classes and other variables. It can be incorporated into almost any software application. In terms of real-world applications, Twitter adapted Lucene as a means of indexing tweets and hashtags for the twitter search function.[5]

Semantic Vectors, in keeping with LSA principles, transforms each term into a vector. These vectors can be summed to represent phrases or document Semantic Vectors works on the assumption that, when observed in vector space (i.e. as vectors in *n* dimensional space, where *n* is the number of vectors), similarity between terms and documents is represented by proximity between

[4] The Apache Software Foundation, "Apache Lucene," <http://lucene.apache.org/> (Accessed: 14 October, 2013).

[5] M. G. Siegler, "Twitter Quietly Launched A New Search Backend Weeks Ago," <http://techcrunch.com/2010/10/06/new-twitter-search/> (Accessed: 14 October, 2013).

vectors. Using oversimplified simplified metaphor, SV creates a multidimensional "map" where similar terms show up next to each other because they are "closer."[1] This method produces both synonyms and related terms (i.e. terms related by concept).

Concept identification does not present just synonyms; for example, concept identification for "pictogram" brings up words like "colour", "paint" or "drawing". These are not synonymous words, they are words that would appear often when describing pictograms, and are therefore conceptually related. They are important to draw out from the text, as they lay down the necessary foundation for constructing associations between concepts. Associations can form the basis of slots (properties) as the terms relate between classes. By identifying common concepts, the construction of classes and slots become a reflection of the concepts considered important in the literature.

3.3.2.1 Random Projection and Positional Indexing

LSA has three primary limitations. First, it does not work well for small corpora (though the threshold of "small" is never well defined). As the corpus size decreases, the number of errors increases. Second, with a small corpus, LSA can present a result in a relatively small number of dimensions, resulting in

[1] Dominic Widdows and Trevor Cohen, "The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics" (paper presented at the Fourth IEEE International Conference on Semantic Computing, Pittsburgh, Pennsylvania, 22-24 September 2010).

problems when the dimensions are compressed using SVD.[2] That is, since the dimensions created using LSA cannot exceed the number of documents in the corpus, the number of dimensions may not be sufficient to execute the SVD algorithm. This problem often results in NaN (not a number) coordinates when the SVD is applied—that is, the dimensions are too few and no results can be returned. Third, dimensional reduction using SVD is a one-time process; a sample of *all* data must be produced in the preliminary steps, and this step cannot be skipped. If changes to any data need to be made, the entire corpora must be re-indexed. These restrictions make LSA a somewhat cumbersome and computationally heavy process.[3]

Because of these shortcomings, Semantic Vectors also supports other methodologies. Random projection (RP) and positional indexing (PI) methodology are both complements to LSA. They are "lighter" alternatives that work on the same principles, though run more effectively on small corpora. Rather than varying dimensionality on the basis of the number of terms or documents, using RP dimensionality is a fixed parameter set by the user. As such, dimensionality does not have to change when elements change. New data can be added without having to re-index the corpus. RP does not require the corpus to

---

[2] Dominic Widdows, "Latent Semantic Analysis," <https://code.google.com/p/semanticvectors/wiki/LatentSemanticAnalysis> (Accessed: 14 October, 2013).

[3] Magnus Sahlgren, "An introduction to random indexing," in *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering* (2005), 4.

be finalised when processing begins.[4] Dimensionality can also be as large or as small as desired, reducing the likelihood of a null result due to an insufficient number of dimensions. Unfortunately, RP does not support term weighting (that is, where terms are "weighted" proportionally in regards to their relevance to the corpus; discussed in section 2.3.1). Because terms are weighted equally without regard to domain terms (that is, in the context of this corpus, a word such as a "the" is given equal weighting to a word like "petroglyph"), RP is less suitable for thesaurus generation.

PI implements the basics of RP and LSA, with additional metadata associating words with their neighbours (thus, "positional indexing"). [5] This method uses a "sliding context window."[6] That is, whole sentences are not necessarily the "window" that is indexed; the tool moves along the text, indexing words in predetermined (for example, 10-word) groups.

PI can index words using permutation indexing or directional indexing. Permutation indexing encodes each word relative to each other word in the sliding window. For example, using the 4 word sliding window "the quick brown fox", "brown" is indexed as appearing with "the", "quick", and "fox" without

[4] Ibid., 6-7.

[5] This method of indexing was first introduced in Magnus Sahlgren, Anders Holst, and Pentti Kanerva, "Permutations as a means to encode order in word space" (paper presented at the The 30th Annual Meeting of the Cognitive Science Society, Washington D.C., USA, 23-26 July 2008).

[6] Dominic Widdows, "Semantic Vectors: Positional Indexes," <https://code.google.com/p/semanticvectors/wiki/PositionalIndexes> (Accessed: 4 October, 2013).

indicating where these terms are in relation to “brown.” Directional indexing classifies terms according to each term’s relative position to other words in the sliding window.[7] For example, in the window “the quick brown fox”, “brown” is indexed as appearing after “quick” and before “fox.” Context is based on the position of ”quick” and “fox” in relation to “brown.” Permutation indexing simply records “brown” as existing with the other three words in the window.

Unfortunately, traditional LSA could not produce results for this corpus. Though no definitive reason was observed, it was most likely a NaN failure: the corpus was too small, and LSA produced no results. The corpus simply could not produce enough dimensions for the method to work. Further work will be needed to determine the minimum size of the corpus needed for LSA to function. In regards to RP, it does not use term weighting, and therefore was unsuitable; the results using RP were polluted by too many stop words (“this”, “the”, etc.).

PI was, ultimately, the method used for this experiment. While it does not use term weighting, PI compensates by using word context (and therefore is not a BOW model). Unlike RP, results were more likely to reflect more specific terms. Although all words are treated equally, word context allows relevance to be computed, corresponding to how the word is used in context. Its relational nature explains some of the anomalies in the data, and highlights the problems of an untidy corpus (see Appendix V, ID #04, Term: “european” for an example). While BOW models are preferred for synonym identification, the results show

[7] Dominic Widdows, "Semantic Vectors: Permutation Indexing and Search," <https://code.google.com/p/semanticvectors/wiki/PermutationSearch> (Accessed: 4 October, 2013).

that concept identification is handled effectively when word contexts are incorporated. Related terms are important for ontology generation, reflecting a positive result for the corpus.

3.4 RESULTS

Three tests were run using two types of analysis, batch comparison and vector sums. A batch comparison simply means the comparisons were compared in one run, with one output where all the results were returned. For example, rather than analysing "rock art" vs. "rock marking", "rock art" vs. "rock image" etc. individually, a batch comparison processes all three at once, producing 6 results at once. The second analysis was a SUM. That is, the sum of two terms (see Figure D, section 2.2.1, for an example of how vectors are summed), and how they compare to the rest of the words in the corpus. The three tests included:

1. A three-batch comparison between the terms:
    a. Rock art vs. rock marking vs. rock image (results: Appendix IV, ID #01)
    b. Pictogram vs. pictograph vs. rock painting (results: Appendix IV, ID #02)
    c. Petroglyph vs. rock carving vs. rock engraving (results: Appendix IV, ID #03)
2. Comparison between the list of 10 terms identified in section 3.2.3 (results: Appendix V, ID #01-ID #10)
3. Individual analysis (SUM) of the same 10 word list, which was used to compute conceptual similarities (results: Appendix VI ID #01-ID #10)

When reading the results of each test, the similarity "score" is represented as a cosine similarity. A cosine similarity is used to measure

similarity between vectors. A cosine similarity of 0º (i.e. no distance between vectors: the terms are identical) is represented as a score of 1.[8] Therefore, a pair of terms scoring 0.9 or above was deemed to be closely related semantically. This threshold was informed by the similarity score of 0.98 between "rock art" and "parietal art" (see Appendix IV, ID #01). Future research will review and refine this threshold.

3.4.1 Comparing "Known" Synonyms

In section 3.2.2, "known" synonyms from the *Rock Art* debate were identified and represented using Google Ngram Viewer (for results, see Appendix III). When compared within this corpus, similarity scores were low compared to assertions made in the literature. The score between "rock image" and "rock marking" was the only score above 0.9 in the comparison (see Appendix III, ID #01). With all other results scoring below 0.9, there appears to be little similarity shared between the terms.

The first result—"rock art" vs. "rock marking" vs. "rock image" (Appendix IV, ID #01)—produced the highest score: a 0.93 between "rock marking" and "rock image." The first comparison was between two substitutes for "rock art" proposed in the literature. The lower score between "rock art" and "rock image" (0.85) and "rock art" and "rock marking" (0.86)—and their similar results—might suggest these substitute terms are being used in different ways to "rock art." However, it is likely that "rock art" occurs more regularly in the corpus; therefore, it appears in a wider selection of contexts from which to choose. In

[8] "Cosine Similarity," <http://en.wikipedia.org/wiki/Cosine_similarity> (Accessed: 2 October, 2013).

other words, “rock art” is more of a master term; the relationships between these terms are more complex than simply synonymy—they require the nuance of ontology to describe their relationship.

The over-representation suggests a comparatively wider meaning for “rock art.” The bigrams generated for section 3.2.3 support this judgement (see Appendix II): “rock art” appears over 7000 times, while “rock image” occurs 3 times and “rock marking” occurs 7 times. “Rock art” appears in thousands of different contexts—“rock image” and “rock marking” appear in 10. A wider selection of contexts gives “rock art” greater selection in the context of PI.

The second comparison—“pictogram” vs. “pictograph” vs. “rock painting”—generated the lowest scores of the test (see Appendix IV, ID #02). The low score between “pictogram” and “pictograph” (0.44) was expected considering the variation between the two definitions identified in section 2.1.1.1. Given the IFRAO definition of “pictogram” as an additive process on rock, the low score when compared to “rock painting” (0.47) contradicts use in the literature, where a “rock painting” is clearly seen as an “additive process.” In fact, the only relatively high score (0.67)—still well under the 0.9 threshold—was between “rock painting” and “pictograph.” The results (that is, “rock painting” as it compares to “pictogram” [0.47] and “pictograph” [0.67]) suggest that the preferred term is, at least in part, Whitely’s use of pictograph as a means of describing paintings on rock. The IFRAO glossary was a collaboration between rock art scholars, yet their preferred term (pictogram) for a painting on rock is not supported by the analysis. The results suggest that even authorities in sub-

disciplines have limited influence over vocabularies, which bodes poorly for any prescriptive process of vocabulary standardisation.

The third comparison—"petroglyph" vs. "rock carving" vs. "rock engraving"—generated results that were both lower than initially expected and confusing (see Appendix IV, ID #03). The comparison between non-specialist terms "rock carving" and "rock engraving" generated the highest similarity (0.88). Neither "rock carving" nor "rock engraving" compares well to "petroglyph" (0.69 and 0.8, respectively), even though they are simply non-specialist derivatives of "petroglyph." Considering its specialist nature, "petroglyph" could be being used in a more technical context than its non-specialist counterparts.

These comparisons emphasise the potential of PI for ontology generation. While ineffective for synonym generation, context allows established "synonyms" to be evaluated quantitatively. In regards to ontologies, this feature can be used to assess the levels of hierarchy or the nature of relationships within a network of terms. For example, using the results here, "pictogram" and "pictograph" would not be on the same level of hierarchy; further investigation would be needed to show whether or not they could even related in the network.

3.4.2 Batch Term Vector Comparisons

A list of related terms, generated from the bigram list (generated for section 3.2.3, see Appendix III), were then compared in a batch comparison (for results see Appendix V). See section 3.3.1 for the full list of 10 terms.

The process of comparison was the same as the analysis used in 3.4.1. Of 100 comparisons, 43 resulted in scores of 0.9 or higher and of these, 10 resulted in scores of 0.95 and higher. Only 6 comparisons resulted in a score lower than 0.8. As these terms are related, this high result is not unexpected. The 10 terms that scored 0.95 and above indicate extremely close semantic similarity (see Table A for the list). From the terms in Table A, complex relationships between terms can be derived. “Rock art” is a type of “paleolithic art”, and is also a type of “parietal art.” If “rock art” is a type of “parietal art”, then “parietal art” can be a type of “paleolithic art.” “Painted rock” explains how a “rock painting” is made. A “rock painting” often appears in a “rock shelter.” These types of relationships cannot be defined by thesauri, but are important to the way the words relate to each other. Ontologies can compute these relationships, though a network of slots.

The results shown in Table A show contextual similarity; the words are not simply related in the sense of their meaning (and some, indeed, are not related in meaning), but share similar contexts. Results such as “rock painting” and “rock shelter” which presented a score of 0.96 support this (Table A; Appendix V, ID #02). A “rock painting” is a feature of a “rock shelter”—the two are not similar in meaning, but they appear in the same contexts. The ability to predict relationships between concepts is an important feature in regards to ontology building. When building an ontology, these terms could be related by an explicitly defined slot (“sits within” or “is a feature of”). By being able to predict similar concepts, not only can one determine and link similar terms, but also define the structure of the ontology by specifying relationships between, e.g.,

hierarchies and classes. An ontology so informed would embody concept use in the domain literature.

TABLE A. SIGNIFICANT COSINE SIMILARITIES AND THEIR RESPECTIVE TERMS.

Cosine Similarity	Term 1	Term 2
0.979605	ROCK ART	PARIETAL ART
0.961564	ROCK ART	PALEOLITHIC ART
0.9596411	PALEOLITHIC ART	PARIETAL ART
0.957439	ROCK PAINTING	ROCK SHELTER
0.952342	ROCK PAINTING	PAINTED ROCK

3.4.3 Semantic Links Regarding Bigram Terms

The third analysis produced a sum of the terms, and compared them to the all the terms in corpus (see Appendix V). The target list used was the same as the list shown in section 3.4.2. The results show the sum of the terms (see Figure D for how vectors are summed), and their respective cosine similarity to other terms in the corpus. It is important to note while the primary term is a bigram, the results – related terms generated by the software – were limited to single words. Bigrams are not easily supported by Semantic Vectors, and are an area of further research outside the scope of this paper.

Concept similarity was the most significant result of this run, and show promise for ontology generation. “cave art”, for example, shows a high concept similarity with “varnish”, “paintings”, “shelters”, and “shelter” (0.91, 0.91, 0.9, and 0.9 respectively: see Appendix V, ID #03). “Cave art” is not synonymous with any of these terms, but the terms reflect common features of cave art. If “cave

art" was a class in an ontology, "varnish", "paintings", "shelters", and "shelter" could be prospective subordinate terms or attributes of the class.

* * *

Synonym generation is currently considered a primary goal of future research into descriptive ontologies. This experiment shows that PI can return results that could be used to predict and construct the relationships between concepts in ontologies (i.e the slots). Any future ontology must be reflective of the discipline in all aspects: synonym generation provides natural language classes (the terms), but the key feature of an ontology is its relationships between classes (the slots). By descriptively generating slots from the literature, the ontology is becomes a true natural language ontology.

CONCLUSION

"Researchers generally create their own formats because they believe that they know how their users want to use the data…[But there are roughly a billion people with Internet access] and at least one of them has a smarter idea about what to do with your content than you do."[9]

There is a crisis of data accessibility in archaeology. For archaeology to move forward as a discipline, one that is equipped to contribute synthetic studies, steps towards standardisation in digital spaces must be made. Without standards, researchers create their own semantic space through which they produce their work. The result is conflicting datasets, in regards to terminology and representation of meaning in the datasets. As a result, researchers involved in large-scale, comparative, or synthetic studies must spend time and resources mapping datasets to a common standard; assuming a complete primary dataset is available to begin with. It is imperative that archaeologists understand the importance of making their data available to be shared, in a way that allows it to be combined with other datasets. Archaeological data is irreplaceable, and there is usually just one chance to produce a complete dataset; it is essential that these datasets be made available for future study.

Metadata standards are a popular route of development for standards in archaeological databases. Metadata standards provide datasets with the means to be discovered and, to an extent, assessed. However, metadata provides very little information that gives access to the content of the dataset. That is, while the

[9] Nelson, "Empty Archives," 161.

dataset may be discoverable, the content is still semantically isolated. If a researcher is comparing multiple datasets from multiple projects with multiple semantic schemas, an unreasonable amount of time will be spent mapping these datasets to a common standard so they can be combined.

Data content standards must be established for meaningful comparison between archaeological datasets. This thesis provides proof of concept for one approach to content standardisation: descriptive ontology generation. By reclaiming terminology from existing literature, any steps towards content standardisation will be a reflection of the field itself. By providing a *descriptive* content standard, rather than *prescribing* an arbitrary standard, archaeologists are far more likely to not only use the standard, but also contribute to its development. Any successful standard requires active engagement from the relevant community. Using this method, suggestions for development would not be limited to contributions through user suggestions (like the CIDOC-CRM), but from continued analysis of publications for further extraction.

Ontologies, unlike flat vocabularies or thesauri, allow the evolution of concepts; they are relational and machine-readable, and thus, the user can map relations between classes. Ontologies are not a perfect solution. While ever a human element remains in research, however, there is little chance of a "perfect" solution; only solutions that can endeavour to provide the best coverage in regards to the problem. Because of their relational nature, ontologies provide the most suitable route for content standards in archaeology.

To implement content standards, synonymous terms must be identified and reconciled. This thesis provides evidence that automated text analysis as a

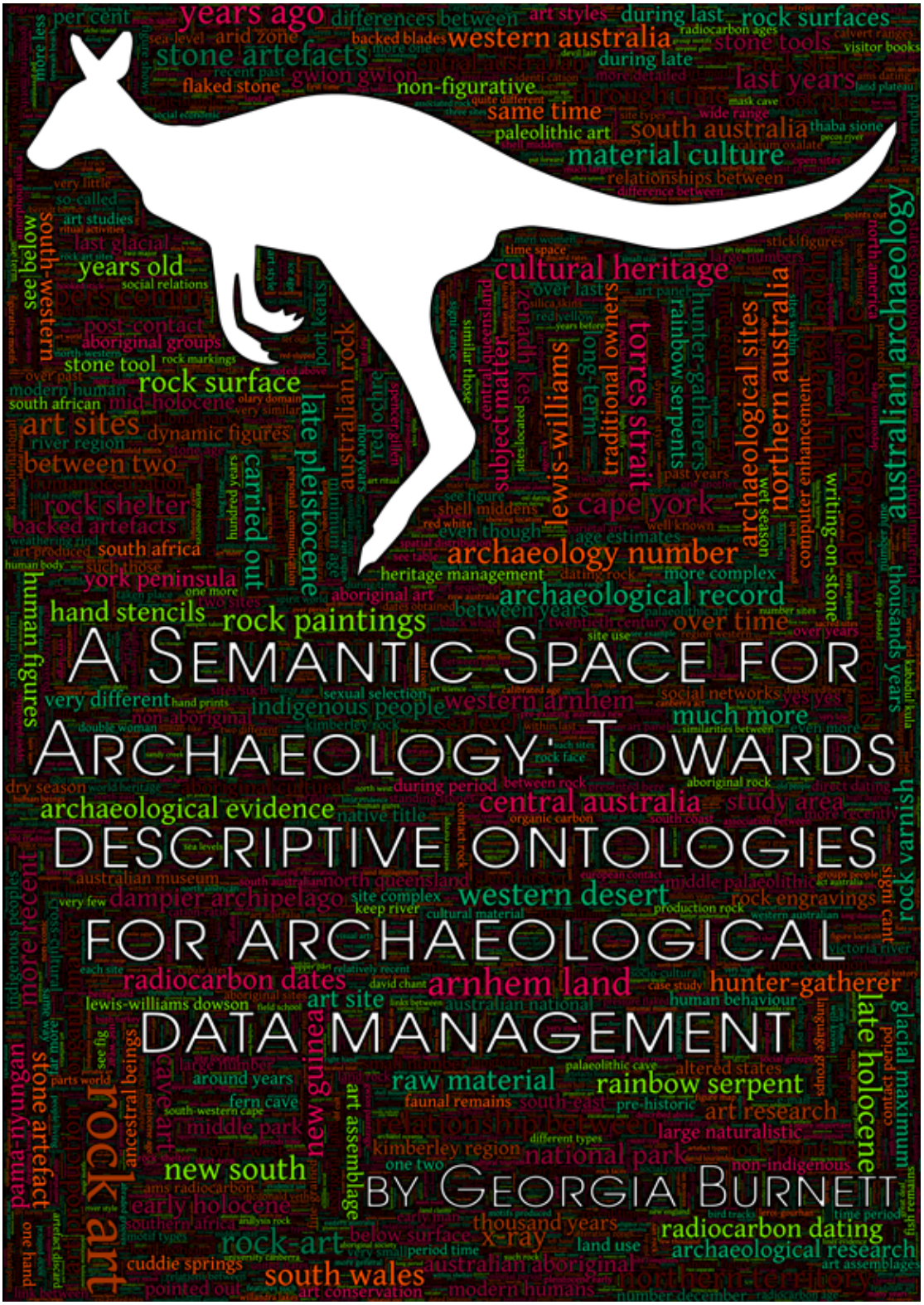
viable method for synonym extraction from archaeological literature. The approach taken, positional indexing (PI), provided an excellent check against established synonyms and identified conflicts between synonyms identified by previous literature. PI is also capable of concept recognition, which provides an excellent basis for ontology creation. That is, results regarding conceptual relations between words can be used to make relational links in ontologies. Relations between words are an essential part of ontology building. PI can be used to predict these relationships with greater efficiency. This thesis should only be considered a demonstration of PIs application in archaeological literature; further investigation is expected to produce significant conceptual relations between words. Context clearly plays a role in word use, and PI has significant possibilities are valuable on their own, and in combination with LSA and other approaches.

The production of a useful, descriptive ontology would require additional approaches. Latent semantic analysis (LSA) is also proposed as a potentially valuable approach. The bag of words (BOW) model used by LSA is expected to result in significantly fewer problems regarding context; PI produced outliers in some results that were clearly related to word context. Further research will be necessary to determine the corpus size required for LSA.

The research presented here presents a proof of concept for text analysis as a suitable means of taking the steps towards vocabulary standardisation and ontology creation. It also provides a methodology for vocabulary retrieval in the domain of humanities research. This method described did not exhaust the potential of automated text analysis in archaeology, but instead serves as a

springboard to further research. The power and possibility of descriptively generated ontologies promises a future for archaeological data outside the limited scope of local project reports.

APPENDIX I: WORD CLOUD



APPENDIX II: BIGRAM GENERATION

The word cloud featured on the cover of this thesis was made with the open-source package, WordCram.[10] The instructions for that package can be found on the website. This appendix gives instructions on how to generate bigrams from a corpus: putting these bigrams in a word cloud is optional, but provides a good visual. The process explained here does not deal with plurality: that is, words like “rock” and “rocks” are treated as two words.

To render the corpus for processing, Dr. Brian Ballsun-Stanton (FAIMS) created the code below. The following requirements were applied:

1. The bigram has to be at least 3 characters in length.
2. Stop words were ignored, excluding them from the process.
3. The corpus was forced into lowercase.
4. Ignore all words that start (-XXX) or end (YYY-) with a dash.
5. If the word features a dash (XXX-YYY), it is its own bigram.
6. Outside the code below, a regular expression editor was used to match repeating characters (for example, “aaa” or “aaaaaaaa”): `.*([a-zA-Z-])\1{2,}.*\n`
7. `Old` is a variable used in the code loop.

[10] Bernier, Daniel. "WordCram." http://wordcram.org/. Accessed: 24 September, 2013. Last Updated: 24 August, 2013.

```
      BEGIN {

   RS="[^A-Za-z-]";
   old = "";
}

/[a-zA-Z-]{3,}/ {
   if ( tolower($0) !~
/^(a|able|about|across|after|all|almost|also|am|among|an
|and|any|are|as|at|be|because|been|but|by|can|cannot|cou
ld|dear|did|do|does|either|else|ever|every|for|from|get|
got|had|has|have|he|her|hers|him|his|how|however|i|if|in
|into|is|it|its|just|least|let|like|likely|may|me|might|
most|must|my|neither|no|nor|not|of|off|often|on|only|or|
other|our|own|rather|said|say|says|she|should|since|so|s
ome|than|that|the|their|them|then|there|these|they|this|
tis|to|too|twas|us|wants|was|we|were|what|when|where|whi
ch|while|who|whom|why|will|with|would|yet|you|your)$/ &&
$0 !~ /^-/ && $0 !~ /-$/ && $0 !~/^[a-zA-Z-]-[a-zA-Z-
]$/) {
              print old " " tolower($0);
              if ($0 ~ /-/)
                    print tolower($0);
              old = tolower($0);

   }
}
```

THE PROCESS: PART ONE

For the example, “A quick brown fox jumps over to the lazy rock-art dog”:

STEP ONE: after `RS="[^A-Za-z-]"`

A

quick

brown

fox

jumps

over

to

the

lazy

rock-art

dog

STEP TWO: after `/[a-zA-Z-]{3,}/`

quick

brown

fox

jumps

over

the

lazy

rock-art

dog

STEP THREE: after `/^(a|able|...|you|your)$/`

quick

brown

fox

jumps

lazy

rock-art

dog

STEP FOUR: the loop then pairs each term,

quick

old is empty, so nothing happens

print [nothing] `space` quick

put “quick” into the variable old

quick brown

old is “quick”

print “quick” `space` “brown”

put “brown” into the variable old

brown fox

old is “brown”...etc.

STEP FIVE: the following bigram list is produced,

quick brown

brown fox

fox jumps

jumps lazy

lazy rock-art

rock-art

THE PROCESS: PART TWO

GAWK, on Linux, was then used to count word frequency and produce a .txt output. The following code was applied to the list of words generated in part one:

```
gawk -f ../biGramcleaner.awk *.txt | sort| uniq -c
| sort -nr > good/2output.txt
```

STEP ONE: `*.txt`

Operate on all files in the directory.

STEP TWO: `| sort`

Order all lines according to alphabetic order.

STEP THREE: `| uniq -c`

Find all unique lines, and count how many times each line appears.

STEP FOUR: `| sort -nr`

Sort the unique list by number, so that the largest number is at top. This is a visual step.

STEP FIVE: `> good/2output.txt`

Write the output of the above commands to the textfile in the directory good, to “2output.txt”

APPENDIX III: BIGRAMS

Over 200,000 bigrams were produced from the process shown in Appendix I. The following is the first 500 bigrams in descending order. The remaining 199,500 can be found at: http://hdl.handle.net/102.100.100/13673.

7840	rock art	353	archaeological record
1329	arnhem land	351	new guinea
1042	years ago	349	late pleistocene
762	rock-art	344	lewis-williams
716	aboriginal people	339	western australia
695	art sites	326	raw material
693	australian archaeology	325	more recent
529	stone artefacts	325	last years
507	western desert	312	upper palaeolithic
485	torres strait	304	rainbow serpent
485	rock paintings	296	archaeological sites
484	archaeology number	294	cave art
469	cape york	292	x-ray
445	cultural heritage	290	radiocarbon dates
409	new south	286	carried out
406	northern territory	285	archaeological evidence
404	south wales	275	rock shelter
382	pers comm	272	national park
378	late holocene	268	over time
377	central australia	257	between two
374	relationship between	254	much more
366	material culture	253	south australia
364	northern australia	246	hand stencils
354	rock surface	244	through time

243	raw materials	168	central australian
243	hunter-gatherer	167	backed artefacts
236	years old	166	gwion gwion
236	western arnhem	165	rock shelters
235	radiocarbon dating	164	see below
231	art production	164	rainbow serpents
228	rock varnish	162	glacial maximum
226	stone artefact	161	rock painting
224	dampier archipelago	160	post-contact
222	study area	160	middle park
222	rock surfaces	158	non-figurative
221	same time	156	thousand years
218	australian rock	156	during last
214	pama-nyungan	153	per cent
213	art site	152	sea level
211	art research	152	between years
205	early holocene	150	differences between
203	human figures	149	stone tool
202	subject matter	149	nineteenth century
202	stone tools	148	red ochre
193	indigenous people	148	mid-holocene
190	long-term	148	art assemblage
189	organic matter	146	very different
182	zenadh kes	146	thousands years
181	traditional owners	145	rock engravings
177	south-western	144	last glacial
177	hunter-gatherers	143	modern humans
176	archaeological research	143	arid zone
175	york peninsula	143	aboriginal cultural
169	australian aboriginal	142	dynamic figures

141	writing-on-stone
141	below surface
140	took place
139	south africa
138	kimberley region
137	during late
137	ancestral beings
134	human occupation
134	art styles
133	yes yes
133	pointed out
133	large naturalistic
132	south-east
131	port keats
131	north queensland
129	paleolithic art
129	over last
129	even though
128	native title
127	non-aboriginal
125	signi cant
125	flaked stone
125	computer enhancement
124	relationships between
124	cuddie springs
123	altered states
123	aboriginal groups
122	river region
122	more complex
122	middle palaeolithic
122	australian national
121	age estimates
120	land use
119	social networks
119	small number
119	north-west
119	around years
118	pigment art
118	non-indigenous
118	modern human
118	dry season
117	fern cave
117	during period
117	art conservation
117	aboriginal art
116	heritage management
116	australian museum
115	wet season
115	central queensland
114	one two
114	minimum age
113	lewis-williams dowson
113	indigenous peoples
113	glen thirsty
112	southern africa
112	shell middens
112	human behaviour
111	two sites
111	site complex
111	number december

111	faunal remains	100	large numbers
110	such those	100	cultural materials
110	more less	100	aboriginal communities
109	one hand	99	recent past
109	north america	99	land rights
109	large number	98	wide range
108	south african	98	thaba sione
108	national parks	98	stick figures
106	so-called	98	sea-level
106	more common	98	keep river
105	stone arrangements	98	art images
105	radiocarbon determinations	98	aboriginal land
105	national university	97	over years
105	contact period	97	more one
105	backed blades	97	far more
104	cross-cultural	97	early man
103	pre-historic	96	victoria river
103	more recently	96	see figure
102	twentieth century	96	over past
102	papua new	96	kimberley rock
102	even more	96	art assemblages
102	art studies	95	period time
102	ams radiocarbon	95	archaeological deposits
102	aboriginal culture	94	more years
101	states consciousness	94	distinction between
101	more detailed	94	direct dating
101	dating rock	94	between rock
100	similar those	93	see fig
100	sexual selection	93	parks wildlife
100	past years	93	language groups

93	great basin
92	stone age
92	south-west
92	art produced
92	art panel
91	visitor books
91	spencer gillen
90	well known
90	very little
90	sites such
90	hundred years
89	very small
89	strait islander
89	link between
89	langford ginibi
89	cultural material
88	time space
88	time period
88	social relations
88	site use
88	radiocarbon ages
88	archaeological site
87	south coast
87	land rock
87	kakadu national
87	human figure
87	australian prehistory
86	same way
86	red yellow
86	production rock
86	one more
86	fish remains
86	fine-grained
86	calvert ranges
86	aboriginal studies
85	very few
85	rock face
85	palaeolithic art
85	organic carbon
85	local aboriginal
85	figure shows
85	difference between
85	aboriginal rock
84	weathering rind
84	very similar
84	standing stones
84	palaeolithic cave
84	calcium oxalate
84	aboriginal sites
84	aboriginal heritage
83	sites located
83	kabadul kula
83	double woman
82	sites within
82	signi cance
82	motif types
82	contact rock
81	number june
81	case study
80	see table

80	recent times	76	archaeological data
80	painted rock	76	amorphous silica
79	south australian	76	aboriginal community
79	rock markings	75	silica skins
79	red white	75	parietal art
79	recent years	75	hand prints
79	parts world	75	geometric motifs
79	e-mail	75	different types
79	art motifs	75	david chant
78	south-western cape	75	art style
78	socio-cultural	74	use-wear
78	sites recorded	74	site types
78	radiocarbon age	74	shown figure
78	presented here	74	petroglyph sites
78	men women	74	personal communication
78	discussed above	74	open sites
78	cation-ratio	74	mask cave
78	art traditions	74	dynamic style
78	aboriginal occupation	74	artefact discard
77	quite different	74	ams dating
77	identi cation	73	two groups
77	black white	73	spatial distribution
77	art dating	73	shell midden
77	age determinations	73	olary domain
76	world heritage	73	naturalistic style
76	within last	73	human body
76	taken place	73	dated years
76	lower palaeolithic	73	correlation between
76	each site	73	art sequence
76	dates obtained	73	archaeological excavations

72	western australian
72	sargah-sar
72	river valley
72	region western
72	land plateau
72	european contact
72	duffer formation
72	aesthetic appreciation
71	west coast
71	wardaman country
71	use term
71	relatively recent
71	quartz grains
71	paintings engravings
71	one another
71	new zealand
71	art panels
71	art found
70	noted above
70	north-western
70	bradshaw figures
70	age years
69	united states
69	terminal pleistocene
69	similarities between
69	relations between
69	group identity
69	cal cal
69	bark-painting
69	artefacts recovered
68	spirit world
68	radiocarbon date
68	past present
68	mcdonald veth
68	artefact scatters
67	western new
67	two different
67	rock-shelter
67	number sites
67	during time
67	clothes peg
67	bradshaw paintings
67	bird tracks
67	association between
66	tigershark rockshelter
66	pressure flaked
66	kimberley points
66	ice age
66	homo sapiens
66	historical archaeology
66	groups people
66	engraving sites
66	discussed here
66	concentric circles
66	between people
66	between groups
66	art objects
66	archaeological investigations
65	such sites
65	social economic

65	ritual activities
65	pre-contact
65	pleistocene rock
65	more intensive
65	groote eylandt
65	different times
65	deposition rates
65	art rock
65	art made
65	around world
64	wildlife service
64	south-eastern
64	social interaction
64	sea levels
64	points out
64	geometric designs
64	discussed below
64	connection between
64	close proximity
64	charcoal samples
64	calibrated age
64	being used
64	base camps
64	art gallery
64	artefact assemblages
63	well-known
63	upper paleolithic
63	social context
63	semi-arid
63	present day
63	people living
63	figurative motifs
63	environmental conditions
63	cupule sites
63	cultural significance
63	bark paintings
63	art australia
62	southern african
62	social groups
62	small tool
62	sample size
62	pre-existing
62	pleistocene early
62	occupation site
62	nawarla gabarnmang
62	much larger
62	long time
62	less years
62	large-scale
62	interaction between
62	institute aboriginal
62	greenstone belt
62	calc-silicate
62	australian institute
61	years before
61	very large
61	total number
61	site located
61	samples collected
61	rock-art sites

61	one site	60	two three
61	non-human	60	three sites
61	much same	60	pleistocene age
61	last few	60	north-east
61	gwion paintings	60	non-iconic
61	environmental change	60	leroi-gourhan
61	devil lair	60	distinguish between
61	department archaeology	60	carnarvon gorge
61	contact between	60	carbon dioxide
61	caves rockshelters	60	canberra act
61	bronze age	59	willandra lakes
61	berndt berndt	59	university canberra
61	art form	59	pecos river
61	analysis rock	59	nsw australia
60	within site	59	mount alexina
60	watchman watchman	59	motifs produced

APPENDIX IV: COMPARING RELATED TERMS

Query: `java pitt.search.semanticvectors.CompareTermsBatch -searchvectorfile drxntermvectors.bin -queryvectorfile drxntermvectors.bin`

Search type: COMPARE (batch comparison)

Results:

ID #01: ROCK MARKING VS. ROCK ART VS. ROCK IMAGE

Cosine Score	Term 1	Term 2
0.857125	ROCK MARKING	ROCK ART
0.928918	ROCK MARKING	ROCK IMAGE
0.857125	ROCK ART	ROCK MARKING
0.847732	ROCK ART	ROCK IMAGE
0.928918	ROCK IMAGE	ROCK MARKING
0.847732	ROCK IMAGE	ROCK ART

ID #02: PICTOGRAM VS. PICTOGRAPH VS. ROCK PAINTING

Cosine Score	Term 1	Term 2
0.442097	PICTOGRAM	PICTOGRAPH
0.467603	PICTOGRAM	ROCK PAINTING
0.442097	PICTOGRAPH	PICTOGRAM
0.677837	PICTOGRAPH	ROCK PAINTING
0.467603	ROCK PAINTING	PICTOGRAM
0.677837	ROCK PAINTING	PICTOGRAPH

ID #03: PETROGLYPH VS. ROCK CARVING VS. ROCK ENGRAVING

Cosine Score	Term 1	Term 2
0.694527	PETROGLYPH	ROCK CARVING
0.800107	PETROGLYPH	ROCK ENGRAVING
0.694527	ROCK CARVING	PETROGLYPH
0.879005	ROCK CARVING	ROCK ENGRAVING
0.800107	ROCK ENGRAVING	PETROGLYPH
0.879005	ROCK ENGRAVING	ROCK CARVING

APPENDIX V: BATCH VECTOR COMPARISONS

Query: `java pitt.search.semanticvectors.CompareTermsBatch -searchvectorfile drxntermvectors.bin -queryvectorfile drxntermvectors.bin`

Search type: COMPARE (batch comparison)

Results:

ID #01: ROCK ART

Cosine Score	Term 1	Term 2
0.911308	ROCK ART	ROCK PAINTING
0.937888	ROCK ART	CAVE ART
0.908715	ROCK ART	ROCK SHELTER
0.961564	ROCK ART	PALEOLITHIC ART
0.893692	ROCK ART	ABORIGINAL ART
0.863062	ROCK ART	ART IMAGE
0.857125	ROCK ART	ROCK MARKING
0.854077	ROCK ART	PAINTED ROCK
0.979605	ROCK ART	PARIETAL ART

ID #02: ROCK PAINTING

Cosine Score	Term 1	Term 2
0.911308	ROCK PAINTING	ROCK ART
0.893399	ROCK PAINTING	CAVE ART
0.957439	ROCK PAINTING	ROCK SHELTER
0.888487	ROCK PAINTING	PALEOLITHIC ART
0.845403	ROCK PAINTING	ABORIGINAL ART
0.843777	ROCK PAINTING	ART IMAGE
0.943727	ROCK PAINTING	ROCK MARKING
0.952342	ROCK PAINTING	PAINTED ROCK
0.90312	ROCK PAINTING	PARIETAL ART

ID #03: CAVE ART

Cosine Score	Term 1	Term 2
0.937888	CAVE ART	ROCK ART
0.893399	CAVE ART	ROCK PAINTING
0.900682	CAVE ART	ROCK SHELTER
0.932674	CAVE ART	PALEOLITHIC ART
0.937717	CAVE ART	ABORIGINAL ART
0.948233	CAVE ART	ART IMAGE
0.8313	CAVE ART	ROCK MARKING
0.819991	CAVE ART	PAINTED ROCK
0.936952	CAVE ART	PARIETAL ART

ID #04: ROCK SHELTER

Cosine Score	Term 1	Term 2
0.908715	ROCK SHELTER	ROCK ART
0.957439	ROCK SHELTER	ROCK PAINTING
0.900682	ROCK SHELTER	CAVE ART
0.879054	ROCK SHELTER	PALEOLITHIC ART
0.816901	ROCK SHELTER	ABORIGINAL ART
0.825792	ROCK SHELTER	ART IMAGE
0.901688	ROCK SHELTER	ROCK MARKING
0.930858	ROCK SHELTER	PAINTED ROCK
0.91077	ROCK SHELTER	PARIETAL ART

ID #05: PALEOLITHIC ART

Cosine Score	Term 1	Term 2
0.961564	PALEOLITHIC ART	ROCK ART
0.888487	PALEOLITHIC ART	ROCK PAINTING
0.932674	PALEOLITHIC ART	CAVE ART
0.879054	PALEOLITHIC ART	ROCK SHELTER
0.902452	PALEOLITHIC ART	ABORIGINAL ART
0.878525	PALEOLITHIC ART	ART IMAGE
0.830876	PALEOLITHIC ART	ROCK MARKING
0.822866	PALEOLITHIC ART	PAINTED ROCK
0.955964	PALEOLITHIC ART	PARIETAL ART

ID #06: ABORIGINAL ART

Cosine Score	Term 1	Term 2
0.893692	ABORIGINAL ART	ROCK ART
0.845403	ABORIGINAL ART	ROCK PAINTING
0.937717	ABORIGINAL ART	CAVE ART
0.816901	ABORIGINAL ART	ROCK SHELTER
0.902452	ABORIGINAL ART	PALEOLITHIC ART
0.924195	ABORIGINAL ART	ART IMAGE
0.811518	ABORIGINAL ART	ROCK MARKING
0.769286	ABORIGINAL ART	PAINTED ROCK
0.888375	ABORIGINAL ART	PARIETAL ART

ID #07: ART IMAGE

Cosine Score	Term 1	Term 2
0.863062	ART IMAGE	ROCK ART
0.843777	ART IMAGE	ROCK PAINTING
0.948233	ART IMAGE	CAVE ART
0.825792	ART IMAGE	ROCK SHELTER
0.878525	ART IMAGE	PALEOLITHIC ART
0.924195	ART IMAGE	ABORIGINAL ART
0.779429	ART IMAGE	ROCK MARKING
0.743174	ART IMAGE	PAINTED ROCK
0.868171	ART IMAGE	PARIETAL ART

ID #08: ROCK MARKING

Cosine Score	Term 1	Term 2
0.857125	ROCK MARKING	ROCK ART
0.943727	ROCK MARKING	ROCK PAINTING
0.8313	ROCK MARKING	CAVE ART
0.901688	ROCK MARKING	ROCK SHELTER
0.830876	ROCK MARKING	PALEOLITHIC ART
0.811518	ROCK MARKING	ABORIGINAL ART
0.779429	ROCK MARKING	ART IMAGE
0.934176	ROCK MARKING	PAINTED ROCK
0.851105	ROCK MARKING	PARIETAL ART

ID #09: ROCK MARKING

Cosine Score	Term 1	Term 2
0.854077	PAINTED ROCK	ROCK ART
0.952342	PAINTED ROCK	ROCK PAINTING
0.819991	PAINTED ROCK	CAVE ART
0.930858	PAINTED ROCK	ROCK SHELTER
0.822866	PAINTED ROCK	PALEOLITHIC ART
0.769286	PAINTED ROCK	ABORIGINAL ART
0.743174	PAINTED ROCK	ART IMAGE
0.934176	PAINTED ROCK	ROCK MARKING
0.848891	PAINTED ROCK	PARIETAL ART

ID #10: PARIETAL ART

Cosine Score	Term 1	Term 2
0.979605	PARIETAL ART	ROCK ART
0.90312	PARIETAL ART	ROCK PAINTING
0.936952	PARIETAL ART	CAVE ART
0.91077	PARIETAL ART	ROCK SHELTER
0.955964	PARIETAL ART	PALEOLITHIC ART
0.888375	PARIETAL ART	ABORIGINAL ART
0.868171	PARIETAL ART	ART IMAGE
0.851105	PARIETAL ART	ROCK MARKING
0.848891	PARIETAL ART	PAINTED ROCK

APPENDIX V: BIGRAM VECTORS

ID #01

Search terms: aboriginal, art

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin aboriginal art`

Search type: SUM

Results:

Cosine Score	Term
0.902793	PAINTINGS
0.900456	VARNISH
0.886207	KANGAROOS
0.885108	ENGRAVINGS
0.882568	SHELTERS
0.875491	ART
0.875491	ABORIGINAL
0.87052	PAINTING
0.870376	IMAGERY
0.865578	MARKINGS
0.860821	INDIGENOUS
0.847627	RESEARCH
0.836783	HUMAN
0.836287	PIGMENT
0.836166	SPECIES
0.835889	IMAGES
0.834523	SITES
0.834088	CONSERVATION
0.83018	RITUAL
0.828461	LIFE

ID #02

Search terms: art, image

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin art image`

Search type: SUM

Results:

Cosine Score	Term
0.919869	PAINTING
0.916766	PAINTINGS
0.901692	VARNISH
0.888631	SHELTER
0.881626	ENGRAVINGS
0.878493	SURFACE
0.875171	IMAGE
0.875171	ART
0.874014	MARKINGS
0.871202	IMAGERY
0.862641	KANGAROOS
0.862306	PROGRAM
0.856425	SHELTERS
0.85582	PANEL
0.852148	TRADITION
0.846146	SEQUENCE
0.845437	IMAGES
0.842373	AND
0.841941	OUTCROP
0.840035	ASSEMBLAGE

ID #03

Search terms: cave, art

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin cave art`

Search type: SUM

Results:

Cosine Score	Term
0.909109	VARNISH
0.905467	PAINTINGS
0.904595	SHELTERS
0.90139	SHELTER
0.886774	PAINTING
0.880793	CAVE
0.880793	ART
0.877554	ENGRAVINGS
0.867516	KANGAROOS
0.865722	MARKINGS
0.859847	IMAGERY
0.849455	PIGMENT
0.845643	CRUST
0.840243	SURFACE
0.832606	KANGAROO
0.82714	POPULATION
0.827009	LANGUAGE
0.826722	PREHISTORIC
0.825827	CONSERVATION
0.823787	COUNTRY

ID #04

Search terms: painted, rock

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin painted rock`

Search type: SUM

Results:

Cosine Score	Term
0.918334	ROCK
0.918334	PAINTED
0.895769	CAVE
0.894458	PIGMENT
0.883984	PARIETAL
0.877532	SAN
0.875211	ENGRAVED
0.870112	PRIMITIVE
0.862571	PALEOLITHIC
0.85664	PREHISTORIC
0.851581	PRODUCING
0.847976	ANCIENT
0.844365	FREMONT
0.843111	EUROPEAN
0.834551	MEGALITHIC
0.822954	MOTIF
0.821575	SECULAR
0.820923	TREE
0.818879	CEREMONY
0.818802	IMAGERY

ID #05

Search terms: paleolithic, art

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin paleolithic art`

Search type: SUM

Results:

Cosine Score	Term
0.871194	VARNISH
0.857543	PALEOLITHIC
0.857543	ART
0.853288	SHELTERS
0.844526	KANGAROOS
0.835763	MARKINGS
0.832059	PAINTINGS
0.822228	PIGMENT
0.813968	PAINTING
0.805646	ENGRAVINGS
0.79948	IMAGERY
0.798293	SHELTER
0.796259	PALAEOLITHIC
0.794136	ROCK
0.793784	PREHISTORIC
0.788822	PRODUCING
0.785443	CAVE
0.782054	EUROPEAN
0.775165	AESTHETICS

ID #06

Search terms: parietal, art

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin parietal art`

Search type: SUM

Results:

Cosine Score:	Term
0.864724	SHELTERS
0.84789	VARNISH
0.843866	ART
0.843866	PARIETAL
0.838803	ROCK
0.838037	KANGAROOS
0.828983	PIGMENT
0.823982	PAINTINGS
0.811079	SHELTER
0.806818	CONSERVATION
0.806655	CAVE
0.804452	MARKINGS
0.802989	ENGRAVINGS
0.799934	PREHISTORIC
0.795785	PAINTING
0.795695	PALEOLITHIC
0.793412	PRODUCING
0.790709	SAN
0.790161	AESTHETICS
0.787317	KANGAROO

ID #07

Search terms: rock, art

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin rock art`

Search type: SUM

Results:

Cosine Score	Term
0.858851	ROCK
0.858851	ART
0.853726	SHELTERS
0.852288	VARNISH
0.829566	KANGAROOS
0.828551	PIGMENT
0.815796	PAINTINGS
0.798599	MARKINGS
0.794459	PARIETAL
0.794131	PREHISTORIC
0.79332	CAVE
0.790557	PAINTING
0.79053	ENGRAVINGS
0.790314	PALEOLITHIC
0.787309	SHELTER
0.786765	PRIMITIVE
0.785152	PRODUCING
0.779743	PRODUCING
0.774956	IMAGERY
0.773324	PORTABLE

ID #08

Search terms: rock, marking

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin rock marking`

Search type: SUM

Results:

Cosine Score	Term
0.899688	PIGMENT
0.881673	ROCK
0.881673	MARKING
0.873794	CAVE
0.857119	PRIMITIVE
0.855013	PREHISTORIC
0.85486	PRODUCING
0.850001	AGENCY
0.845823	PARIETAL
0.844704	EUROPEAN
0.843683	MAKING
0.841158	ANCIENT
0.841156	SAN
0.840004	POWER
0.839287	WORKS
0.838143	ENGRAVED
0.836046	SPECIES
0.834411	PALEOLITHIC
0.834098	PAINTED
0.830628	VISUAL

ID #09

Search terms: rock, shelter

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin rock shelter`

Search type: SUM

Results:

Cosine Score:	Term
0.931491	CAVE
0.905762	ROCK
0.905762	SHELTER
0.886796	SAN
0.886321	PIGMENT
0.881997	PARIETAL
0.862707	LANDSCAPE
0.862287	PRIMITIVE
0.859595	SITE
0.856021	DEAD
0.855045	DREAMING
0.854474	PREHISTORIC
0.853876	FRENCH
0.853239	ANCIENT
0.853058	BRADSHAW
0.852515	PALEOLITHIC
0.849466	KURU
0.846225	CRUST
0.844711	ENVIRONMENT
0.84306	REGION'S

ID #09

Search terms: rock, painting

Query: `java pitt.search.semanticvectors.Search -queryvectorfile drxntermvectors.bin -searchvectorfile drxntermvectors.bin rock painting`

Search type: SUM

Results:

Cosine Score	Term
0.927759	PIGMENT
0.91341	CAVE
0.904967	ROCK
0.904967	PAINTING
0.894379	PREHISTORIC
0.888719	SAN
0.87769	PRIMITIVE
0.877636	ANCIENT
0.875513	PRODUCING
0.87025	EUROPEAN
0.863836	PARIETAL
0.863445	PALEOLITHIC
0.86276	CEREMONY
0.861126	MOTIF
0.860948	VISUAL
0.859277	ENGRAVED
0.857901	BODY
0.857358	PAINTINGS
0.856934	STYLE
0.853099	IMAGERY

REFERENCES

Aboriginal Hertiage Act 2006 (Cth) http://www.austlii.edu.au/au/legis/vic/consol_act/aha2006164/. Accessed: 17 October 2013

"About Armadillo." http://www.hrionline.ac.uk/armadillo/armadillo.html. Accessed: 27 September, 2013.

Aitchison, Kenneth. "Supply, Demand and a Failure of Understanding: Addressing the Culture Clash between Archaeologists' Expectations for Training and Employment in 'Academia' Versus 'Practice'." *World Archaeology* 36, no. 2 (2004): 203-19.

"Archaeology Data Service/Digital Antiquity: Guides to Good Practice." Archaeology Data Service & Digital Antiquity, http://guides.archaeologydataservice.ac.uk/. Accessed: 31 July, 2013. Last Updated: Unknown.

"Archaeotools: Data Mining, Facetted Classification and E-Archaeology." Archaeology Data Service, http://archaeologydataservice.ac.uk/research/archaeotools. Accessed: 4 October, 2013. Last Updated: Unknown.

Arroyo-Bishop, D. "The Archéodata Project." In *Computer Applications and Quantitative Methods in Archaeology 1989*, edited by S. P. Q. Rahtz and J. D. Richards. 548, 69-86. Oxford: BAR International Series 1989.

Austin, David. "Archaeology, Funding and the Responsibilities of the University." In *The Responsibilities of Archaeologists: Archaeology and Ethics*, edited by Mark Pluciennik. *BAR International Series 981*, 31-38. Oxford: Archaeopress, 2001.

Baines, A., and K. Brophy. "Archaeology without -Isms." *Archaeological Dialogues* 13 (2006): 69-91.

———. "What's Another Word for Thesaurus?: Data Standards and Classifying the Past." In *Digital Archaeology: Bridging Method and Theory*, edited by T. L. Evans and P. Daly. 236-50. London: Routledge, 2006.

Ballsun-Stanton, Brian. "Asking About Data: Exploring Different Realities of Data Via the Social Data Flow Network Methodology." University of New South Wales, 2012.

Barrett, John C. "Fields of Discourse: Reconstituting a Social Archaeology." *Critique of Anthropology* 7, no. 5 (1988): 5-16.

Bednarik, Robert G. "IFRAO." http://home.vicnet.net.au/~auranet/ifrao/web/. Accessed: 2 October, 2013. Last Updated: Unknown.

———. "IFRAO 1988-2000: The First Dozen Years." http://home.vicnet.net.au/~auranet/ifrao/web/hist.html. Accessed: 2 October, 2013. Last Updated: December 2000.

———. "IFRAO Rock Art Glossary." http://home.vicnet.net.au/~auranet/glossar/web/glossary.html. Accessed: 8 September, 2013.

Bernier, Daniel. "Wordcram." http://wordcram.org/. Accessed: 24 September, 2013. Last Updated: 24 August, 2013.

Biber, Douglas, Susan Conrad, and Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

Borgman, Christine L., Jillian C. Wallis, and Noel Enyedy. "Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital

Libraries." [In English]. *International Journal on Digital Libraries* 7, no. 1-2 (2007/10/01 2007): 17-30.

Bradley, Richard. *The Significance of Monuments: On the Shaping of Human Experience in Neolithic and Bronze Age Europe*. London: Routledge, 1998.

Brank, Janez, Marko Grobelnik, and Dunja Mladenić. "Automatic Evaluation of Ontologies." In *Natural Language Processing and Text Mining*, edited by Anne Kao and Stephen R. Poteet. 193-220. London: Springer-Verlag, 2007.

Brock, Mark N. "Computerised Text Analysis: Roots and Research." *Computer Assisted Language Learning* 8, no. 2-3 (1995/06/01 1995): 227-58.

Chenhall, R. G. "The Archaeological Data Bank: A Progress Report." *Computers and the Humanities* 5, no. 3 (1971): 159-69.

Chippindale, Christopher, and Paul S. C. Taçon. "What's in a Word, What's in a Hyphen? A Modest Proposal That We Abandon the Words 'Petroglyph' and 'Pictograph', and Hyphenate 'Rock-Painting', 'Rock-Engraving', 'Rock-Art' among the Words We Use." *Rock Art Research* 23, no. 2 (2006): 254-57.

"The CIDOC Conceptual Reference Model: Applications." http://www.cidoc-crm.org/uses_applications.html. Accessed: September 24, 2013.

"Cosine Similarity." http://en.wikipedia.org/wiki/Cosine_similarity. Accessed: 2 October, 2013. Last Updated: Unknown.

Dagan, Ido. "Contextual Word Similarity." In *Handbook of Natural Language Processing*, edited by Robert Dale, Herman Moisl and Harold Somers. 477-506. New York: Marcel Dekker, Inc., 2000.

Davies, Mark. "The Corpus of Contemporary American English." http://corpus.byu.edu/coca/. Accessed: 1 October 2013, 2013. Last Updated: Unknown.

———. "Google Books." http://googlebooks.byu.edu/. Accessed: 1 October, 2013. Last Updated: Unknown.

Debachere, Marie-Claire. "Problems in Obtaining Grey Literature." *IFLA Journal* 21, no. 2 (1996): 94-98.

Dictionaries, Oxford. "About the Oxford English Corpus." http://www.oxforddictionaries.com/words/about-the-oxford-english-corpus. Accessed: 1 October, 2013. Last Updated: Unknown.

Doerr, Martin. "The CIDOC Conceptual Reference Model: An Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine* 24, no. 3 (2003): 75-92.

———. "The CIDOC Conceptual Reference Model: Who We Are." http://www.cidoc-crm.org/who_we_are.html. Accessed: 13 September, 2013. Last Updated: Unknown.

Doyle, David E. *Lake Hohokam Prehistory in Southern Arizona*. Scottsdale: Gila Press, 1981.

dstromfe. "Acrobat OCR: Make Your Scanned Documents Searchable." http://blogs.adobe.com/acrobat/acrobat_ocr_make_your_scanned/. Accessed: 14 October, 2013. Last Updated: 4 November 2009.

Dumais, Susan T., George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. "Using Latent Semantic Analysis to Improve Access to Textual Information." Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Washinton, DC, 1988.

"English Heritage Thesauri." http://thesaurus.english-heritage.org.uk/frequentuser.htm. Accessed: 13 October, 2013. Last Updated: Unknown.

Fernie, Kate. "Launch of Fp7 Infrastructure Project - Ariadne." http://www.ariadne-infrastructure.eu/News/Press-release. Accessed: 11 October, 2013. Last Updated: 17 April, 2013.

Foundation, The Apache Software. "Apache Lucene." http://lucene.apache.org/. Accessed: 14 October, 2013. Last Updated: 5 October 2013.

Google. "About Google Ngram Viewer." https://books.google.com/ngrams/info. Accessed: 1 October, 2013. Last Updated: Unknown Date, 2013.

Department of Planning and Community Development. "Guidelines for Investigating Historical Archaeological Artefacts and Sites." State of Victoria, 2012. 1-53.

Harping, Patricia. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, Cultural Works*. New Hampshire: Odyssey Press, Inc., 2010.

Harris, Z. S. *Mathematical Structures of Language*. New York: Wiley, 1968.

Hearst, Marti A. "Untangling Text Data Mining." Paper presented at the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, 1999.

"Heritage Victoria." Department of Planning and Community Development, http://www.dpcd.vic.gov.au/heritage. Accessed: 31 July 2013, 2013. Last Updated: 16 July 2013.

Holmen, Jon, Christian-Emil Ore, and Øyvind Eide. "Documenting Two Histories at Once: Digging into Archaeology." Paper presented at the Beyond the

artefact: Digital Interpretation of the Past, Proceedings of CAA 2004, Budapest, Hungary, 13-17 April 2004.

"Iso 21127:2006." http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=34424. Accessed: September 15, 2013.

Jeffrey, Stuart, Julian Richards, Fabio Ciravegna, Stewart Waller, Sam Chapman, and Ziqi Zhang. "The Archaeotools Project: Faceted Classification and Natural Language Processing in an Archaeological Context." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367, no. 1897 (2009): 2507-19.

Kansa, Eric C., and Ahrash Bissell. "Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains." *Data Science Journal* 9 (2010): 42-53.

Kintigh, Keith W. "The Challenge of Archaeological Data Integration." In *Technology and Methodology for Archaeological Practice: Practical Applications for the Past Reconstruction*. 1-8: British Archaeological Reports, 2009.

———. "The Promise and Challenge of Archaeological Data Integration." *American Antiquity* 71, no. 3 (2006): 567-78.

Kintigh, Keith W., Francis P. McManamon, and Katherine Spielmann. "Enhancing Data Comparability and Enabling Synthesis with tDAR (the Digital Archaeological Record)." Paper presented at the Towards a Data Standard for Paleolithic Archaeology, 78th Annual Meeting of the Society for American Archaeology, Honolulu, Hawaii, 3-7 April 2013.

Kubicek, Herbert, Ralf Cimander, and HansJochen Scholl. "Layers of Interoperability." Chap. 7 In *Organizational Interoperability in E-Government*. 85-96: Springer Berlin Heidelberg, 2011.

Landauer, Thomas K., and Susan T. Dumais. "Latent Semantic Analysis." http://www.scholarpedia.org/article/Latent_semantic_analysis. Accessed: 8 October, 2013. Last Updated: 21 October 2011.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "Introduction to Latent Semantic Analysis." *Discourse Processes* 25 (1998): 259-84.

Lebart, Ludovic, and Martin Rajman. "Computing Similarity." In *Handbook of Natural Language Processing*, edited by Robert Dale, Herman Moisl and Harold Somers. 477-505. New York: Marcel Dekker, Inc., 2000.

"Legacy Data." http://www.businessdictionary.com/definition/legacy-data.html. Accessed: 31 August, 2013.

Limp, W. Fredrick. "Web 2.0 and Beyond, or on the Web Nobody Knows You're an Archaeologist." In *Archaeology 2.0: New Tools for Communication and Collaboration*, edited by Eric C. Kansa. 265-79: Cotsen Institute of Archaeology, 2011.

McGuinness, Deborah L., and Frank van Harmelen. "Owl Web Ontology Language Overview." W3C, http://www.w3.org/TR/owl-features/. Accessed: 9 September, 2013. Last Updated: 12 September 2009.

McManamon, Francis P., and Keith W. Kintigh. "Digital Antiquity: Transforming Archaeological Data into Knowledge." *The SAA Archaeolgical Record* 10, no. 2 (2010): 37-40.

McWhorter, John. "David Brooks' Favorite New Theory of Language Is Wrong." http://www.newrepublic.com/article/113274/david-brooks-language-our-words-dont-reveal-our-worldview. Accessed: 4 October, 2013. Last Updated: 23 May 2013.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, Joseph P. Pickett, *et al.* "Quantitative

Analysis of Culture Using Millions of Digitized Books." *Science* 331, no. 6014 (2011): 176-82.

Nelson, Bryn. "Empty Archives." *Nature* 46 (2009): 160-63.

Niven, Kieron. "OASIS: Online Access to the Index of Archaeological Investigations." http://oasis.ac.uk/england/. Accessed: 31 July, 2013. Last Updated: 24 May 2012.

Noy, Natalya F., and Deborah L. McGuinness. "Ontology Development 101: A Guide to Creating Your First Ontology." Stanford University: Stanford, 2001.

Patrik, Linda E. "Is There an Archaeological Record?". *Advances in Archaeological Method and Theory* 8 (1985): 27-62.

"Protege." http://www.upriss.org.uk/awt/s12.html. Accessed: 4 October, 2013. Last Updated: Unknown.

Richards, Julian D. "From Anarchy to Good Practice: The Evolution of Standards in Archaeological Computing." *Archeologia e Calcolatori* 20 (2009): 27-35.

Richards, Julian D., and Catherine Hardman. "OASIS: Dealing with the Digital Revolution." Paper presented at the Digital Heritage of Archaeology, Proceedings of CAA 2002, Heraklion, Crete, 2-6 April 2002.

———. "Stepping Back from the Trench Edge: An Archaeological Perspective on the Development of Standards for Recording and Publication." In *The Virtual Representation of the Past*, edited by M. Greengrass and L. Hughes. 101-12. Farnham, Surrey: Ashgate Publishing Company, 2008.

Richards, Julian D., Stuart Jeffrey, Stewart Waller, Fabio Ciravegna, Sam Chapman, and Ziqi Zhang. "The Archaeology Data Service and the Archaeotools Project: Faceted Classification and Natural Language

Processing." In *Archaeology 2.0*, edited by Eric C. Kansa. 31-56: Cotsen Institute of Archaeology, 2011.

Sahlgren, Magnus. "An Introduction to Random Indexing." In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, 1-5, 2005.

———. "The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Spaces." Stockholm University, 2006.

Sahlgren, Magnus, Anders Holst, and Pentti Kanerva. "Permutations as a Means to Encode Order in Word Space." Paper presented at the The 30th Annual Meeting of the Cognitive Science Society, Washington D.C., USA, 23-26 July 2008.

Shwartz, B. K. "Further Comments on Christopher Chippindale and Paul S. C. Taçon's 'What's in a Word, What's in a Hyphen?': On 'Rock Art' History and Terminology." *Rock Art Research* 24, no. 1 (2007): 124-25.

Siegler, M. G. "Twitter Quietly Launched a New Search Backend Weeks Ago." http://techcrunch.com/2010/10/06/new-twitter-search/. Accessed: 14 October, 2013. Last Updated: 6 October 2010.

"Society for American Archaeology Principles of Archaeological Ethics." *American Antiquity* 61, no. 3 (1996): 451-52.

Spasić, Irena, Sophia Ananiadou, John McNaught, and Anand Kumar. "Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text." *Briefings in Bioinformatics* 6, no. 3 (2005): 239-51.

"tDAR: The Digital Archaeological Record." Digital Antiquity, http://www.tdar.org/. Accessed: 31 July, 2013. Last Updated: Unknown.

Turney, Peter D., and Patrick Pantel. "From Frequency to Meaning:Vector Space Models of Semantics." *Journal of Articial Intelligence Research* 37 (2010): 141-88.

Vasko, Stephanie E. "Examing Trends Via Google Ngram: Taking a Closer Look." https://sites.psu.edu/stephanievasko/2013/09/29/examining-trends-via-google-ngram-taking-a-closer-look/. Accessed: 14 October, 2013. Last Updated: 29 September 2009.

Voß, Jakob. "Revealing Digital Documents: Concealed Structures in Data." Cornell University, 2011.

"What Is Vector Space Model." http://www.igi-global.com/dictionary/vector-space-model/31436. Accessed: 14 September, 2013. Last Updated: Unknown.

Whitely, David S., ed. *Handbook of Rock Art Research*. Walnut Creek: AltaMira Press, 2001.

Whitley, David S. *Introduction to Rock Art Research*. Walnut Creek: Left Coast Press, 2011.

Widdows, Dominic. "Latent Semantic Analysis." https://code.google.com/p/semanticvectors/wiki/LatentSemanticAnalysis. Accessed: 14 October, 2013. Last Updated: 16 November 2012.

———. "Semantic Vectors." https://code.google.com/p/semanticvectors/. Accessed: 14 October, 2013. Last Updated: Feburary, 2013.

———. "Semantic Vectors: Permutation Indexing and Search." https://code.google.com/p/semanticvectors/wiki/PermutationSearch. Accessed: 4 October, 2013. Last Updated: 30 April, 2013.

———. "Semantic Vectors: Positional Indexes." https://code.google.com/p/semanticvectors/wiki/PositionalIndexes. Accessed: 4 October, 2013. Last Updated: 4 February, 2013.

Widdows, Dominic, and Trevor Cohen. "The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics." Paper presented at the Fourth IEEE International Conference on Semantic Computing, Pittsburgh, Pennsylvania, 22-24 September 2010.

Wielinga, B. J., A. Th. Schreiber, J. Wielemaker, and J. A. C. Sandberg. "From Thesaurus to Ontology." Paper presented at the Formal Ontology in Information Systems, Trento, Italy, 6-8 June 1998.

Wilcock, John. "On the Importance of High-Level Communication Formats in World Archaeology." In *Archaeology and the Information Age: A Global Perspective*, edited by Sebastian Rahtz and Paul Reilly. 69-75. London: Routledge, 2004.