ARIADNE

D16.3: Final Report on Data Mining

Authors:

W.X.Wilcke, VU University Amsterdam
H. Dimitropoulos, ATHENA RC

SEVENTH FRAMEWORK PROGRAMME

Ariadne is funded by the European Commission's 7th Framework Programme.

Version: 1.1 *(final*) **20th January 2017**

Authors: **W.X.Wilcke, VU University Amsterdam**

H. Dimitropoulos, ATHENA RC

Contributing Partners **V. de Boer, VU University Amsterdam**

F.A.H van Harmelen, VU University Amsterdam

M.T.M de Kleijn, VU University Amsterdam

R. van het Veer, VU University Amsterdam

M. Wansleeben, Leiden University

I. Foufoulas, ATHENA RC

A. Giannakopoulos, ATHENA RC

Quality review: **Holly Wright, ADS**

Table of Contents

List of Abbreviations **3**

Executive Summary **4**

1 Introduction and Objectives **6**

1.1. Summary of Previous Findings 7

1.2. Continuation 9

1.3. Structure of Report 10

2. Revisiting Semantic Web Mining **12**

2.1. Linked Data as a Graph 13

2.2. Challenges Involved 14

3. Case Studies **16**

3.1. ARIADNE Registry 18

3.2. OPTIMA 22

3.3. SIKB Dutch Archaeological Protocol 0102 38

4. Conclusion **51**

4.1. Recommendations 53

Appendix A SIKB Protocol 0102 Conversion to RDF **54**

A.1 Protocol Description 54

A.2 Protocol Conversion 55

A.2.1 Protocol Enrichment 62

A.3 Thesauri Conversion 64

A.4 Data Conversion and Publication 66

Appendix B Pipeline VUA/LU **67**

B.1 Data Preparation 68

B.2 Hypothesis Generator 70

B.3 Pattern Evaluator and Knowledge Consolidator 71

Appendix C Pipeline ATHENA RC **72**

C.1 Pipeline Components 72

Appendix D Expert Evaluation **77**

D.1 OPTIMA Case Evaluations 77

D.2 SIKB Protocol Case Evaluations 81

List of Abbreviations

The following abbreviations will be used in this report.

Abbreviation	*Full Term*
ACDM	ARIADNE Catalogue Data Model
ADS	Archaeological Data Service
API	Application Programming Interface
ARIADNE	Advanced Research Infrastructure for Archaeological Data set Networking in Europe
ARM	Association Rule Mining
BGV	Basic Geo Vocabulary
DM	Data Mining
GIS	Geographic Information System
KG	Knowledge Graph
LAD	Linked Archaeological Data
LD	Linked Data
LOD	Linked Open Data
ML	Machine Learning
NLP	Natural Language Processing
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SIKB	Dutch Foundation Infrastructure for Quality Assurance of Soil Management
SKOS	Simple Knowledge Organization System
SW	Semantic Web
SWM	Semantic Web Mining
UI	User Interface
URI	Universal Resource Indicator
W3C	World Wide Web Consortium
WGS	World Geodetic System
WP	Work Package

Executive Summary

ARIADNE, the Advanced Research Infrastructure for Archaeological Data set Networking in Europe, facilitates a central web portal that provides access to archaeological data from various sources. Parts of these data are being provided as Linked Data. Users can explore these data using traditional methods such as faceted browsing and keyword search, as well as the more-advanced capabilities that come with Linked Data such as semantic queries. A shared characteristic amongst these methods is that they allow users to explore *explicit* information present in the data. However, since the data is available in structured and explicit form, we can additionally use Machine Learning and Data Mining techniques to identify *implicit* patterns that exist within this explicit information. The work in this report aims to facilitate this.

This report (D16.3) is a direct continuation of D16.1, *First report on Data Mining.* D16.1 presented a detailed study into the theoretical background of both Linked Data, Data Mining, and on the novel area of research where both domains meet, known as Semantic Web Mining (SWM). It additionally discusses the application of SWM within the archaeological community from the perspective of both user needs and data characteristics. Finally, the report concludes by recommending several Data-Mining tasks that could be of use to the archaeological researcher. We have selected one of these tasks, namely Hypothesis Generation, as the primary focus of this report:

> Hypothesis Generation involves detecting interesting and potentially relevant patterns that can be presented to users as starting blocks for forming new research hypotheses. The researcher might already have a hypothesis, in which case found patterns may strengthen his or her belief in the hypothesis. Alternatively, the patterns may reveal something new to the researcher that he or she is interested in exploring further. The support and confidence of patterns will be generated algorithmically on the basis of predefined criteria and user feedback.

The Hypothesis Generation task has been implemented into an experimental pipeline named MinoS (Mining on Semantics). This pipeline has been integrated into a simple infrastructure consisting of a Data-Mining backend, running on a high-performance server, with access to the Linked Open Data (LOD) cloud. A subset of the LOD's data is retrieved via a SPARQL query which will be generated semi-automatically based on a certain task, and a set of provided constraints. Both the task and constraints will be provided by the user through a simple User Interface (UI).

In addition to Hypothesis Generation, we have also applied more traditional data mining methods during the experiments. These methods will involve 1) Semantic Content Mining to determine and compare project connotations, 2) relationship discovery among the authors of OpenAIRE and ARIADNE Reports, 3) the creation and analysis of ARIADNE's author networks, and 4) performing text mining on OpenAIRE publications to link them with ARIADNE metadata or other extracted objects from the ARIADNE reports.

To investigate the effectiveness of the pipeline with respect to the selected task, we have been running multiple experiments on three different data sets. These constitute 1) project-wide meta-data as available in the ARIADNE Registry, 2) project reports that have been semantically annotated using the OPTIMA pipeline developed for Semantic Technologies for Archaeological Resources

(STAR)[1], and 3) rich database extractions (SIKB Protocol 0102) from entire projects that have been converted by us into Linked Data. These three data sets have been specifically chosen for these experiments as each one differs significantly in its level of information granularity. This allowed us to test the effects of information granularity on the relevance of the results produced by the pipeline.

Each of the three data sets have been treated as separate and parallel case studies, to test the goal of producing results relevant to the archaeological community. Here, the first use case (ARIADNE Registry) differs from the other two in that it has been investigated by the ATHENA RC, whereas the other two have been investigated by the collaboration of the Vrije Universiteit Amsterdam and Leiden University (VUA/LU). Therefore, the first use case has used the pipeline developed by the UoA, whereas the remainder has used the pipeline developed by the VUA/LU.

Overall, the combined results of all three case studies can be seen as a mildly promising beginning in the use of Data Mining for Linked Archaeological Data. Moreover, the domain experts were surprised by the range of patterns discovered, despite most patterns describing either trivialities or tautologies. Nevertheless, there are still challenges to overcome for the field of SWM to mature, and to be of actual use to the archaeological community. Until that time arrives, it may be best to employ traditional Data Mining techniques that have been perfected over several decades, and which have proven to produce reliable and useful results.

[1] See http://hypermedia.research.southwales.ac.uk/kos/star/

[2] Bizer, C, T Heath, and T Berners-Lee. "Linked data-the story so far." *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 2009: 205-227.

1 Introduction and Objectives

ARIADNE, the Advanced Research Infrastructure for Archaeological Data set Networking in Europe, facilitates a central web portal that provides access to archaeological data from various sources. Some of these data are being provided as Linked Data (LD): an open format that is well established and a World Wide Web Consortium (W3C) standard. As such, it aims to propel the creation, management, and use of data towards a higher level of interoperability[2]. Users can explore these data using traditional methods, such as faceted browsing and keyword search, as well as more-advanced capabilities that come with Linked Data such as semantic queries. A shared characteristic amongst these methods is that they allow users to explore *explicit* information present in the data. However, since the data is available in structured and explicit form, we can use Machine Learning and Data Mining techniques to identify *implicit* patterns in this explicit information[3].

Data Mining (DM) provides tools and techniques to identify valid, novel, potentially useful, and ultimately understandable patterns in data[4]. These patterns represent implicit regularities within these data, and often point to some yet-unknown characteristic that is being shared amongst the different data points. Some of these characteristics may be of relevance to the users of the data, who may infer new theories from them. These theories can consequently be used to draw new conclusions, or to confirm or falsify existing ones. Ultimately, this may result in addition, removal, or correction of data points that either do or do not adhere to the discovered patterns.

Discovering patterns by conventional Data Mining requires data to be stored in tables, where individual entities are represented by the rows in such a table. In the Linked Data paradigm, information is not represented using tables, but rather in graphs, consisting of entities linked through RDF *triples*. Data Mining on graph data in general (Linked Data) is considered state-of-the-art research. This report (D16.3) will present the final work on our continuing effort as part of WP16.1 to explore the challenges and opportunities for Data Mining on archeological Linked Data. We build on lessons learned and the recommendations made within D16.1, *First report on Data Mining*[5]. and present three cases of Data Mining using different techniques and on different types of archeological Linked Data. We first summarize the previous findings and provide background on Semantic Web Mining.

---

[2] Bizer, C, T Heath, and T Berners-Lee. "Linked data-the story so far." *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 2009: 205-227.

[3] Hastie, T, R Tibshirani, J Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning.* New York: Springer, 2009.

[4] Fayyad, U M. "Data mining and knowledge discovery: Making sense out of data." *IEEE Intelligent Systems* 11, no. 5 (1996): 20-25.

[5] Wilcke, W X, V de Boer, M T M de Kleijn, F A H van Harmelen, and M Wansleeben. *D16.1: First Report on Data Mining.* Deliverable, ARIADNE, 2015.

1.1. Summary of Previous Findings

First report on Data Mining (D16.1) presented a detailed study into the theoretical background of both Linked Data, Data Mining, and the novel area of research where both domains meet, known as Semantic Web Mining.

Semantic Web Mining is a young area of research of which many aspects are still uncertain or unexplored, both from a technical and practical perspective. To this end, the study focussed on the more-prominent movements as seen in recent literature. In addition, the study analysed expected usage-patterns in order to determine users' needs and wishes with respect to Data Mining (1.1.1), as well as conducting a thorough exploration of the expected data's characteristics (1.1.2). The outcomes of these studies lead to the selection of two recommendations for potentially-interesting Data-Mining Tasks (1.1.3), namely *Hypothesis Generation* and *Data Quality Analysis*. We will provide a concise summarisation of these sections next.

1.1.1. Domain Understanding

The user-requirement study involved an analysis of the questionnaires and interviews that were conducted by work package 2.1[6] and 13.1[7], respectively. While providing valuable insight into the stakeholders of ARIADNE, both work packages only touched on the possibility of Data Mining. Therefore, several additional interview sessions with stakeholders were held, during which the possibility of Data Mining was more-actively explored. Regardless, of all the topics discussed, only a few were relevant with respect to Data Mining.

In its entirety, the requirement study indicated that the majority of difficulties experienced by stakeholders could be mitigated by the use of the explicit information in the Linked Data alone (for example, through directly querying this data). However, several of these issues could additionally be improved with the help of Data Mining. These issues involved knowing which data are available, how to locate relevant data, and how to distil the results. In addition, the quality of the data was mentioned prominently, thereby emphasising their (lack of) completeness and the (lack of) confidence in how trusted the data should be. Together, these were amongst the prime areas considered to which a Data-Mining solution could be applied.

1.1.2. Data Understanding

Exploring the data is an important early step within any Data-Mining process, during which the data's characteristics, quality, and abnormalities are inspected. Generally, a generous amount of data is provided from which conclusions can be drawn that influence choices made during the development of the eventual Data-Mining solution. Unfortunately, the minimal amount of Linked Data that has been made available through ARIADNE were insufficient, therefore, Linked Data from

---

[6] Selhofer, H, and Guntram Geser. *D2.1: First report on users' needs.* Deliverable, ARIADNE, 2014.

[7] Hollander, H, and M Hoogerwerf. *D13.1: Service Design.* Deliverable, ARIADNE, 2014.

several different archaeological repositories around the globe were inspected instead. These data were chosen for their almost-disjoint characteristics, thus hopefully providing good representations of the different facets that ARIADNE would produce. The analysis resulted in several observations. Most prominent, apart for the generally expected differences in ontologies and structure, was the Linked Archaeological Data examined were found to strongly depend on descriptive values. Moreover, all data sets were found to consist largely of relatively 'flat' data structures, where information relevant to a node in the graph is relatively few steps away in that graph. This is likely to have its origin in direct conversion from tabular data. These aspects of the data had to be considered during the development of the Data-Mining solution.

1.1.3. Recommendations Made

Based on the study of both domain and data, as well as on practical constraints with respect to time and resources, two Data-Mining solutions were chosen which were deemed the most feasible and suitable for implementation within the ARIADNE infrastructure. These constitute 1) the ability for users to generate potentially-relevant hypotheses, and 2) analysing the quality of data as well as helping to improve it. We will briefly touch on these two solutions next:

Hypothesis Generation

The official project proposal of ARIADNE mentions the ability to detect patterns in archaeological data or related data, and applications within the ARIADNE infrastructure. Data-Mining methods are capable of detecting such patterns. Interesting and potentially relevant subsets of these patterns can then be presented to users as starting points for forming new research hypotheses. The researcher might already have a hypothesis, or the patterns may reveal something the researcher is interested in exploring further. The interestingness of patterns will be generated algorithmically on the basis of predefined criteria and user feedback.

Data Quality Analysis

Two aspects that reflect poorly on the quality of data are the occurrence of gaps and errors in the knowledge contained therein. In case of the former, filling these voids involves predicting the most-likely resource, link, or literal. In the latter case, these errors typically contain anomalies within the data, which cannot be explained by any of the discovered patterns alone. Depending on the likelihood of them being erroneous, the detected errors could be suggested for removal or tagged as dubious. Alternatively, they could be replaced by a prediction on the correct value.

1.2. Continuation

This report discusses the technical feasibility and practical usability of Data-Mining on Linked Archaeological Data. As such, this report emphasises 1) data preparation, 2) implementation, 3) experimentation, and 4) evaluation:

Data Preparation
The work in this report will discuss the relevancy of mining Linked Archaeological Data for the archaeological community. This relevancy can only be measured meaningfully when it is reflected in the results produced by a Data-Mining experiment. For these results to posses such relevant characteristics, they must be mined from data that possesses these characteristics as well. Unfortunately, *First report on Data Mining* (D16.1) concluded with the observation that very little of this data exists at present. Therefore, part of this work will involve the creation of a dedicated Linked Archaeological Data cloud filled with fine-grained information.

Implementation
A considerable part of the work in this report will involve the development and implementation of two pipelines: ATHENA and VUA/LU. Both pipelines will facilitate Data-Mining on Linked Archaeological Data, with the sole difference being the approach. More precisely, the ATHENA pipeline specifically aimed to find correlations within and between different sets of coarse-grained information, whereas the VUA/LU pipeline aimed to detect domain-relevant regularities within fine-grained information. Both pipelines were used during the subsequent experimentation.

Experimentation
The experimental design covered three case studies with distinctly different levels of information granularity. As such, it was possible to ascertain the effect of these differences on the resulting output, as well as on the relevancy of this output towards the archaeological community. The three case studies encompass meta-data (coarse grained), annotated reports ('finer' grained), and rich database extracts (fine grained).

Evaluation
All output from the experimentations were evaluated algorithmically to ascertain a quantitative measure of relevancy. For this purpose, standard graph statistics were employed and ontological background knowledge was utilised. This evaluation step will result in a smaller output, of which its reduction rate will be dependent on the input constraints. The reduced output was subsequently evaluated qualitatively together with domain experts to ascertain the relevancy for the archaeological community.

1.1.1 Task Selection

The ARIADNE deliverable *First report on Data Mining* (D16.1) concluded with the recommendation of two Data Mining tasks: Hypothesis Generation and Data Quality Analysis. For the purpose of this deliverable efforts were focussed on Hypothesis Generation only. This decision was made for the following reasons:

1) Hypothesis Generation constitutes a novel method to detect domain-relevant regularities within data sets. To the best of our knowledge, this has never before been applied to archaeological data. In contrast, Data Quality Analysis has a long history with many applications in many different domains. If fact, most off-the-shelf statistical software packages already provide this form of automatic curation.
2) Hypothesis Generation allows exploitation of the relational structure within data sets, as well as their semantics. Hence, it is more likely that complex patterns that span over multiple related entities will be discovered. In contrast, Data Quality Analysis would largely focus on sets of values, hence gaining little benefit from Linked Data's relational structure and semantics. Differently put, Data Quality Analysis is likely to perform equally well on Linked Data as on tabular data, and thus has less academic value.
3) Hypothesis Generation is likely to yield archaeologically-interesting patterns that, if found relevant, may add to the knowledge base of the archaeological community. As such, new insights may be created that form the starting point of new research endeavours. In contrast, Data Quality analysis will, at best, result in the improvement of a data set's correctness. While valuable, this does not meet our goal of accentuating the relevancy of mining Linked Archaeological Data to the archaeological community
4) Hypothesis Generation produces results that are easily interpretable by individuals without a background in Data Mining or similar. The importance of this aspect became apparent following the user-requirement analysis of D16.1. More precisely, archaeological researchers have the need to clearly understand the generated hypotheses themselves, instead of needing to decipher a symbolic representation. Only if this criterion holds, will the researchers even consider them. In contrast, most methods capable of Data Quality Analysis apply a form of statistical analysis. As such, they are more likely to produce their results as a symbolic representation.

As postulated above, it is unlikely data sets with coarse-grained information will yield archaeologically-interesting results. Therefore, we additionally applied more traditional data mining methods during the corresponding experiments. More precisely, these methods involved 1) Semantic Content Mining to determine and compare project connotations, 2) relationship discovery among the authors of OpenAIRE and ARIADNE Reports, 3) the creation and analysis of ARIADNE's author networks, and 4) performing text mining on OpenAIRE publications to link them with ARIADNE metadata or other extracted objects from the ARIADNE reports.

1.3. Structure of Report

Following this section, a concise summary of the field of Semantic Web Mining and the challenges involved will be given in Section 2. Next, we will discuss three case studies with increasing granularity of knowledge representation in Section 3. In each of these three cases, we will start by going through the respective data set and task description, followed by a discussion of the design, results, and evaluation of their respective experiments. Section 4 will continue by looking at the aggregated results and their evaluation, followed by final remarks and several concluding recommendations. Finally, for those interested, Appendices A through C will discuss the details of the pipelines in more detail, whereas Appendix D will list the evaluation reports. A schematic visualisation of this structure is provided in Figure 1.

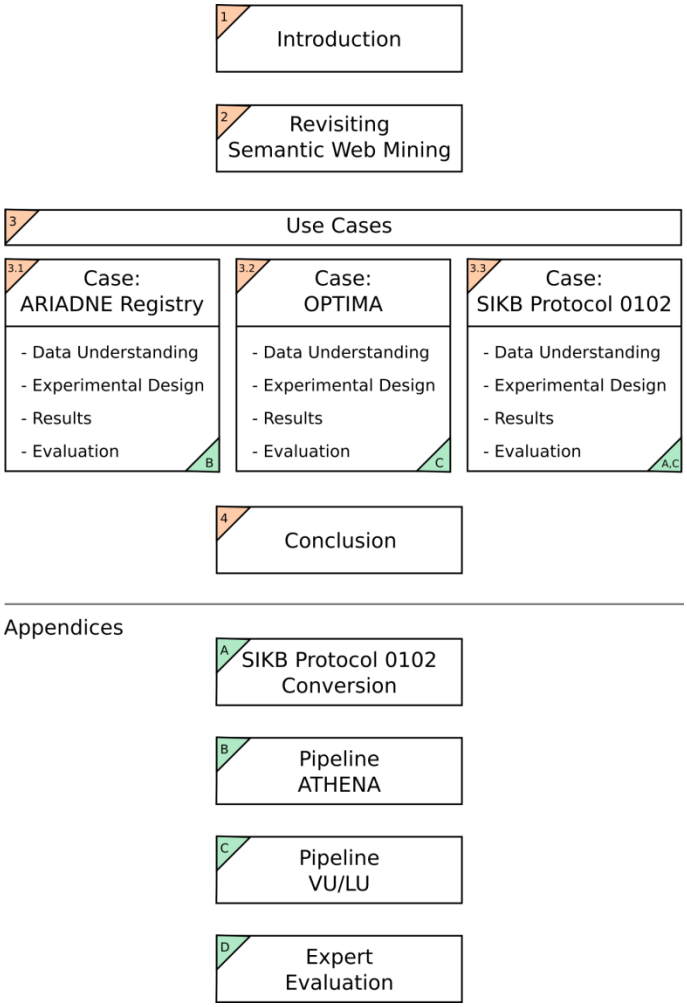

Figure 1: Structure of this report.

2. Revisiting Semantic Web Mining

Semantic-Web Mining (SWM) is an umbrella term which denotes the area of Data Mining that focuses on mining Linked Data found on the Semantic Web. It is a relatively new area of research of which many aspects are still uncertain or left unexplored, from both a technical and practical perspective. If fact, many tasks and learning methods are still under heavy development, with few of them having progressed outside the confines of academic research.

There are two prominent approaches to SWM, namely propositional learning and relational learning. Propositional, or "traditional", learning solely concerns propositional data formats such as tables (Figure 2-1 Left). Here, the graph structure of Linked Data is first converted to a tabular structure. As such, this approach makes several assumptions about the data that do not necessarily hold for Linked Data. For example, the assumption that each data point is independent of any arbitrary other data point in the data set. In contrast, relational learning *does* assume that such a dependency may exist.

Given the graph structure of Linked Data (Figure 2-1 Right), it would seem natural to focus on relational learning. A considerable disadvantage of this approach is that where propositional learning is a mature and established field of work, relational learning has hardly been applied outside of experimental research. Therefore, propositional learning is often still preferred when learning from Linked Data. This work's decision to employ such propositional methods is largely motivated by this aspect.

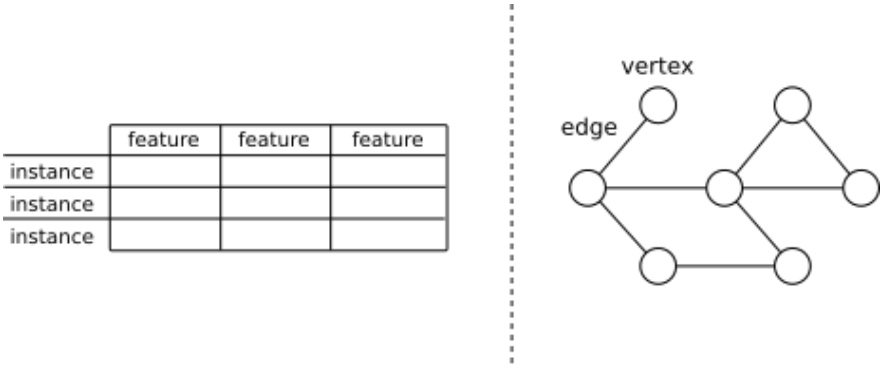

Figure 2-1: Two different data models. Left) Propositional (tabular data). Right) Relational (graph data).

Additional steps are needed when applying propositional learning to relational structures such as Linked Data, namely *context sampling* and *propositionalisation*. Context sampling involves the creation of individual units that represent instances (e.g. rows in a table) over which we want to learn. These individual units do not explicitly exist in Linked Data, and thus should be created by smartly sampling related concepts (i.e. directly and indirectly connected nodes). Once all instances have been sampled, they can be translated to a propositional format (e.g. a table) by a process known as propositionalisation. This will result in a data set that is suitable for propositional learning.

2.1. Linked Data as a Graph

LD data sets are represented using the Resource Description Format (RDF). In RDF, information is represented using triples (*subject*, *predicate*, *object*), which together form graphs. These graphs are sometimes referred to as Knowledge Graphs (KGs). In a KG, both *subjects* and *objects* are encoded as vertices, whereas the *predicates* are encoded as edges. Differently put, each triple (*subject*, *predicate*, *object*) in the data set is represented by an unique and directed edge between exactly two vertices. Finally, to map the semantics of the edges and vertices in the resulting graph, these elements are labelled with their identifier (for Linked Data this is a URI) or a simple label.

Depending on its purpose, a KG may differentiate between the instance level and the schema level of a given data set. The latter contains all information on classes and properties, whereas the former concerns all information on the instantiations of those classes and properties. This distinction is illustrated in Figure 2-2, which depicts a simple instance and schema graph on the left and right side, respectively.

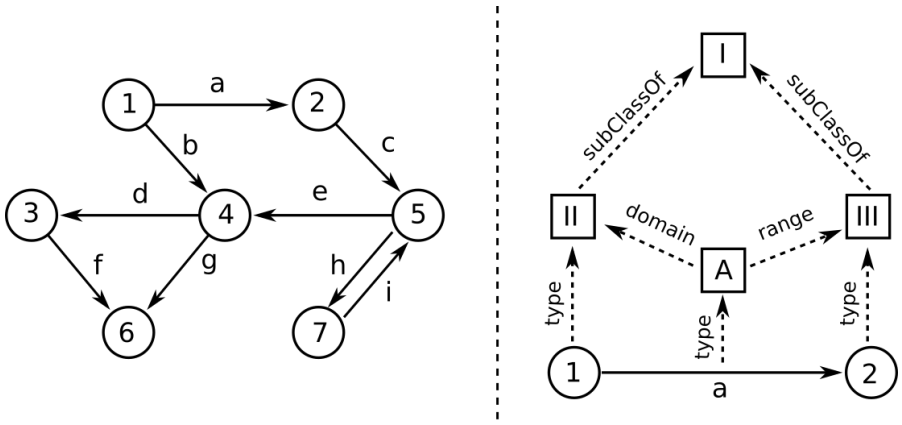

Figure 2-2: A graph representation of Linked Data. Left) An instance graph with numbers denoting resources and with letters denoting predicates. Right) A schema graph, projected over a triple in graph form, with Roman numbers denoting resource classes and with capital letters denoting predicate classes.

While intuitive and easily interpretable, the graph representation of Linked Data is a simplification of the actual data. Nevertheless, it offers a suitable approximation for data mining purposes in which one generalises over the data. Therefore, it is by far the most common representation used in the related literature. The remainder of this work will follow this view and thus refer to a Linked Data set as a Knowledge Graph.

2.2. Challenges Involved

Previous research done in D16.1 indicated a number of challenges that arise at the intersection of Data Mining and Linked Data. A concise description of these challenges is given next.

Different ontologies

Different KGs may employ very different ontologies to represent semantics (for example, in one KG, a person might be represented using the Friend-of-a-Friend ontology (FOAF)[8], while in another, persons are represented using a CIDOC CRM class[9]). Similarly, some KGs may employ different versions of the same ontologies. Either situation poses a significant challenge for any learning task that tries to exploit semantical relations. That is, patterns found to fit semantic constructs of ontology *A* cannot be used to derive new insights from a KG that employs ontology *B*.

Structural variation

KGs from different sources may vary considerably in their structure. This may be due to the ontology used or simply due to the creator's discretion. Alternatively, any structural peculiarity may be inherited from the original data set it was derived from. Irrespective of the cause, these variations pose a challenge when trying to generalize over KGs with different origins. More specifically, patterns found to describe certain substructures, e.g. a subgraph of a KG, are unlikely to perform well on KGs with a totally different structure.

Descriptive values

Archaeological data sets tend to have a frequent use of descriptive values (for example in free text). Often, these values provide crucial information and thus cannot simply be ignored without a significant loss of information. Including these values poses a considerable challenge as Data Mining is incapable of naturally processing free text. That is, any non-numerical value should be converted to an internal numerical representation prior to learning.

Concept ambiguity

In a KG, concepts are defined by their directly and indirectly related vertices. This becomes evident when viewing directly related vertices as the attributes of a concept, and when considering indirectly related vertices as *attributes of attributes* that might contribute to this concept. For instance, knowing the colour of a certain material is likely meaningful to the semantic representation of a find composed of said material. In contrast, knowing the number of atoms in that material's element is hardly relevant. This example already illustrates the challenge of choosing the subset of related vertices that most accurately define a certain concept.

[8] http://xmlns.com/foaf/spec/
[9] http://www.cidoc-crm.org/Entity/e21-person/version-6.2

Data model incompatibility

Propositional Data Mining is unable to process any data which does not conform to a propositional data model. Among those data models that do not conform is that of a KG. Therefore, to learn over a KG, they must first be translated to fit a propositional data model. An unfortunate side effect of this process is that complex relational structures are replaced by more simple propositional structures. Differently put, we lose potentially-relevant information by propositionalising a KG. To minimise this loss, a suitable strategy must be selected.

3. Case Studies

The informativeness of patterns found in Knowledge Graphs depends heavily on the contents of those graphs as well as on the structural features of these graphs. To increase the scope of the work in this task, three separate case studies in Semantic Web Data Mining on archaeological data were conducted. Each of the three studies used different types of Linked Data, produced within the ARIADNE project. These are:

1. Data from the ARIADNE Registry,
2. Grey literature that has been semantically-annotated using the OPTIMA text-mining pipeline,
3. Rich Linked Data database extractions that follow the SIKB Protocol 0102 specification.

One helpful way of organizing the three cases is to order them by `granularity' of knowledge. More fine-grained data sets have more specific information about archaeological finds and their contexts, whereas more coarse-grained data sets describe information at a higher level (metadata). Figure 3-1 shows the three data sets on this continuum.

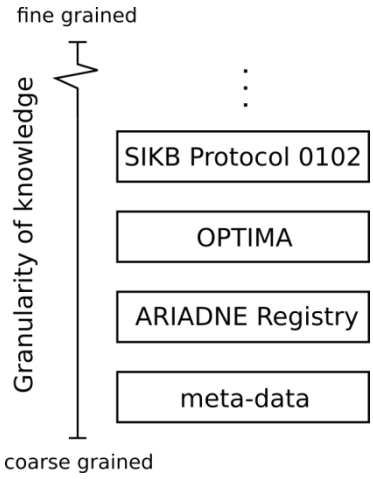

Figure 3-1: Relative differences between the granularity of knowledge seen with various forms of data.

The first case will involve the ARIADNE Registry (Section 3.1), which can be regarded as rich meta-data that has been extended with domain-specific elements. As such, these data hold information about documents, rather than about archaeological knowledge within those documents. Therefore, we expect it unlikely to yield results that are of interest to an archaeologist.

The second case will concern the combined output of the OPTIMA pipeline on a large corpus of British archaeological reports (Section 3.2). In contrast with the ARIADNE Registry, these data do contain archaeological knowledge in the form of semantic annotations. However, the context of these annotations is rather limited, spanning a paragraph at most. Therefore, the span of any potential pattern is also limited.

The third and final case will involve a cloud of SIKB Protocol 0102 instances (Section 3.3). These data concern many of a project's elements, ranging from reports and database extracts to media, finds, and the context of these finds. As such, these instances include meta-data that is similar to that of the ARIADNE Registry, as well as holding archaeological knowledge that is both deeper and better connected than that produced by the OPTIMA pipeline. Therefore, we expect these data likely to yield results that are of interest to an archaeological researcher.

Each of the three data sets will be treated as separate and parallel case studies. Here, the first use case (ARIADNE Registry) differs from the other two in that it will be investigated by the ATHENA RC, whereas the other two will be investigated by the Vrije Universiteit Amsterdam (VUA). Therefore, the first use case used the pipeline developed by ATHENA RC, whereas the remainder will use the pipeline developed by the VUA.

Each of these three cases and their potential use will now be discussed in more detail.

3.1. ARIADNE Registry

The ARIADNE registry described in this report uses the ARIADNE Catalogue Data Model (ACDM) and Catalogue system designed and created by ATHENA RC and CNR. The ARIADNE Catalogue is centred on the model of individual data sets, but as many of the partners hold data in the form of collections, the Catalogue has also been extended to handle collections. The Catalogue is also a metadata registry (for a discussion of metadata registries and standards, along with the technologies used to develop the metadata registry within the Catalogue, followed by a description of the ACDM and Catalogue system itself please refer to deliverable D3.1: Initial report on standards and on the project registry[10].

3.1.1. Data Description

This "Specification of the ARIADNE Catalogue Data Model (ACDM)" document describes in detail the data model underlying the catalogue developed by the ARIADNE project for describing the archaeological resources that are made available by the partners of the project for the purposes of discovery, access and integration. These resources include:

- data resources, such as data sets and collections;
- services; and
- language resources, such as metadata formats, vocabularies and mappings.

The model is addressed to cultural institutions, private or public, which wish to describe their assets in order to make them known to e-infrastructures.

One of the aims of T16.1 is to explore linkages between data and publications. An example given in the DoW suggests that ARIADNE provide metadata from the integrated data sets to OpenAIRE, which would then return back links to possibly relevant publications. Building a bridge between ARIADNE and OpenAIRE would be a positive step towards exploring the possibility of linking archaeological data with research publications.

For such a use-case, the Data Resources metadata of the ACDM are relevant, and more specifically the following classes:

Collection
This class is a specialisation of the class DataResource, and has as instances collections in the archaeological domain. In order to be as general as possible, archaeological collections are defined as an aggregation of resources, and named the items in the collection. Being aggregations, collections are akin to data sets, but with the following, important difference:

[10] Papatheodorou, C, et al. *D3.1: Initial report on standards and on the project registry.* Deliverable, ARIADNE, 2013.

the items in a data set are data records of the same structure (see definition of Data set below). In contrast, the items in a collection are individual objects which are different from records (e.g., images, texts, videos, etc.) or are themselves data resources such as collections, data sets, databases or GIS; for instance, a collection may include a textual document, a set of images, one or more data sets and other collections.

Database
This class is a specialisation of the class DataResource, and has as instances databases, defined as a set of homogeneously structured records managed through a Database Management System, such as MySQL.

Data set
This class is a specialisation of the classes DataResource. It has archaeological data sets as instances. An archaeological data set is defined as a set of homogeneously structured data records, consisting of fields carrying data values.

GIS
This class is a specialisation of the class DataResource, and has as instances data records, consisting of fields carrying data values, which are not managed through a Geographical Information Systems (GISs).

It was decided to focus on Databases and Data sets from the ARIADNE catalogue, as a good starting point for investigating relevant links to OpenAIRE publications.

3.1.2. Task Description

As stated in Section 1.2 that, in addition to Hypothesis Generation, it was also planned to apply more traditional data mining methods during the experiments. These methods involve 1) Semantic Content Mining to determine and compare project connotations, 2) relationship discovery among the authors and the citations of OpenAIRE and ARIADNE Reports, 3) the creation and analysis of ARIADNE's author networks, and 4) performing text mining on OpenAIRE publications to link them with ARIADNE metadata or other extracted objects from the ARIADNE reports. Apart from the first method (see Section 3.2 for Semantic Content Mining) the remaining methods are dealt with in this section.

3.1.3. Experimental Design

An algorithm for citation extraction and algorithm data set matching was implemented first. The implementation details can be found in Appendix C.

As a first experiment, arXiv open access publications were used and text mining techniques were applied to find references to ARIADNE collections or data sets. It was hoped that with arXiv, links to ARIADNE's dendrochronology data set records could be found, but unfortunately they were not.

For the second experiment, PubMed/ePMC publications were used, as ARIADNE contains some data sets relevant to the biomedical domain (mitochondrial DNA sequence, SNP data, dating techniques,

etc.). However, this set also did not produce any significant findings. Testing with other OpenAIRE repositories with PDF/full-text collections of their publications will continue (although note that arXiv and ePMC are two of the largest such collections).

Meanwhile, linkages between ARIADNE and OpenAIRE in the opposite direction are being explored. For example, starting with the ADS grey literature reports, citation extraction algorithm looking for hooks to OpenAIRE publications was run. Although no links to arXiv publications were found, when running the algorithm on ePMC publications, promising results were returned. The results are described in section 3.1.4, followed by their evaluation by an archaeologist in section 3.1.5.

Similar experiments on other repositories are now being run, including experiments using the ADS grey literature metadata, in order to:

(i) discover relationships among the authors of OpenAIRE and ARIADNE Reports,
(ii) create and analyse author networks,
(iii) do further text mining on OpenAIRE publications to link them with ARIADNE metadata or other extracted objects from the ARIADNE reports.

So far, (i) above has been pursued, finding thousands of single author links between ADS grey literature and OpenAIRE authors. The set of results is too large to manually evaluate and a large number of links are expected to be false matches, i.e. different people that happen to have the same surname and initials (some names are quite common). To make this more manageable, two-author links between ADS reports and OpenAIRE publications were looked for (see Appendix C for the algorithm), resulting in 131 ADS-OpenAIRE links of 54 distinct author-pair names. This set will continue to be evaluated until the end of the ARIADNE project, so the results are not presented in this deliverable. Any further findings before the end of the ARIADNE project will be included in the final reporting.

In parallel, clusters of similar records within the ARIADNE database are also being looked for. For this purpose, the keywords generated by text mining the abstracts and titles of ARIADNE's records were extracted and similarity algorithms in order to generate clusters of related records were run. A first set of results were produced (e.g. ~250K similar pairs with jaccard > 0.3), but keyword extraction is being modified to improve results, as there are many "automated" descriptions in the data that share common expressions with only one word differing, etc. The above algorithm will be used to find similarities between the 16M+ publications in OpenAIRE and ARIADNE records, using titles and abstracts.

3.1.4. Results

As described in the previous section, it was not possible to find references to ARIADNE collections or data sets when using the citation extraction and algorithm data set matching algorithm on two of the largest PDF/full-text corpora in OpenAIRE, the ArXiV and PubMed/ePMC repositories.

However, when exploring linkage between ARIADNE and OpenAIRE in the opposite direction, and specifically, when running the citation extraction algorithm on the ADS grey literature reports looking for hooks to OpenAIRE publications, the results were positive, as described below.

A large subset of the ePMC corpus was used and 216 direct citation links from ADR reports to PubMed publications were found, as well as 83 indirect links. The direct links were all high confidence links, the indirect links were generally medium confidence, and there were seven false positives which were all low confidence and were removed after manual curation. The false positives can be eliminated in future runs be selecting an appropriate confidence cut-off value.

In total, 299 valid citation links from ADS grey literature reports to PubMed were found, mostly references to publications in medical history journals/epidemiology (e.g. yellow fever in the 1790s), geochronology, paleopathology, dental anthropology, anatomy, DNA studies, and so on.

Indirect links are usually links from an ADS report to a review of the cited work in OpenAIRE. So, not a direct citation match, but very relevant. In other words, an ADS report is linked to an OpenAIRE publication which discusses, presents or reviews the publication cited by ADS, but is not the actual publication.

All links found were given for further assessment to an archaeologist to confirm their relevance to this project. The archaeologist's evaluation is presented in section 3.1.5 below.

Regarding the utilisation of the ADS grey literature metadata to discover relationships among ADS and OpenAIRE authors and to create and analyse author networks, etc., since at the time of writing the mining experiments are still being run and results are not complete or fully evaluated yet, any further findings before the end of the ARIADNE project will be included in the final reporting.

3.1.5. Evaluation

The 299 citation links from ADS grey literature reports to PubMed/ePCM publications in OpenAIRE mentioned in the previous section were evaluated by an archaeologist.

According to the summary evaluation report, "the vast majority of records are absolutely useful and relevant". The archaeologist also noted that topics were related to fields such as anthropology, biology, palaeontology, pottery, and also outside the UK, such as Oceania. With the exception of a few records, they were useable.

The only problem mentioned was a few multiple/duplicate entries. Upon further investigation, this appears to be caused by a small set of ADS reports that appear to have more than one (very similar in text) version.

In more detail, 292 of the links were judged as relevant, six links as possibly relevant, and eight links as irrelevant. This is in line with the internal evaluation (in section 3.1.4).

3.2. OPTIMA

Archaeological reports hold a wealth of information which software agents are unable to interpret. This can be partially solved by processing reports through a Natural Language Processing (NLP) pipeline. Such a pipeline, called OPTIMA, was developed by Andreas Vlachidis [11] in the context of the Semantic Technologies for Archaeological Resources (STAR) [12]project. For this purpose, it employs both the CIDOC CRM and CIDOC CRM-EH ontologies.

3.2.1. Data Description

Since its development, OPTIMA has processed numerous archaeological reports written in English. This has resulted in several thousand files full of triples which assign semantic annotations to their respective reports. These annotations may involve the following concepts:

- Physical objects and finds, as well as the materials of which they consist and the contexts in which they were discovered.
- Temporal information such as archaeological period appellations and time spans, as well as the events or objects to which they are assigned.
- Various types of events related to the production and deposition of finds, as well as to their place and time of occurrence.

A small excerpt, representative of all of OPTIMA's output, is shown in Excerpt 3-1. There, a certain find is declared as being of the type 'natural clay', thereby including a reference to a vocabulary. The occurrence of that type's name within the accompanying sentence is listed as a note belonging to this find. In addition, the find is declared as being moved to a context of type 'linear feature' by a certain deposition event.

Several observations can be made that influence the design of the Data-Mining pipeline. Firstly, the limits of OPTIMA's capabilities largely prevent inferences being made beyond the linguistic context of a term. As a result, there is hardly any cross-linking between the annotated concepts beyond their directly related attributes. Moreover, it is very likely for the same concept to be annotated more than once and with different attribute values if the corresponding term is mentioned more than once as well. Consequently, per processed document, the result is multiple small and possibly conflicting graphs.

A related observation is these small graphs translate to individuals with relatively few attributes. For instance, the find listed in Excerpt 3-1 has only has three potentially-relevant attributes. These concern the find's deposition event, as well as its type and context. All other attributes are either irrelevant, e.g. notes and type appellations, or have the same value throughout the file, e.g. sources. Consequently, a Data-Mining pipeline is only able to compare individuals using a handful of characteristics. As such, it may limit the predictive power of this pipeline.

[11] See http://www.andronikos.co.uk/
[12] See http://hypermedia.research.southwales.ac.uk/kos/star/

```
:contextFind_A
   a crmeh:EHE0009.ContextFind ;
   crm:P2F.has_type type_C ;
   crm:P3F.has_note "...alignment. 2.13 Linear feature [101] was observed cutting into the natural clay,
                     orientated NE - SW towards the south-east end of the trench. This
                     feature..."^^xsd:string ;
  dc:source source_X ,
            source_Y .

:contextFindDepositionEvent_A
  a crmeh:EHE1004.ContextFindDepositionEvent ;
  crm:P3F.has_note "13 Linear feature [101] was observed cutting into the natural clay"^^xsd:string ;
  crm:P25F.moved contextFind_A ;
  crm:P25F.moved_to context_A ;
  dc:source source_X ,
            source_Y .

:context_A
   a crmeh:EHE0007.Context ;
   crm:P2F.has_type type_D ;
   crm:P3F.has_note "... across the very south-east end of the trench on a NE - SW alignment. 2.13 Linear feature
                    [101] was observed cutting into the natural clay, orientated NE - SW towar..."^^xsd:string ;
  dc:source source_X ,
            source_Y .

:type_C
  a crm:E55.Type ;
  rdf:value "natural clay" ;
  crmeh:EXP10F.is_represented_by concept_C ;
  dc:source source_X ,
            source_Y .

:type_D
  a crm:E55.Type ;
  rdf:value "Linear Feature" ;
  crmeh:EXP10F.is_represented_by concept_D ;
  dc:source source_X ,
            source_Y .
```

Excerpt 3-1: An annotated context find (natural clay) and its linked resources.

A final observation is the total absence of continuous numerical attributes. Rather, all possible attributes hold nominal or free text values. Data-Mining only works with numerical values. For this reason, nominal or free text values are often represented by such numerical values. While this appears to nullify the initial observation, the now numerical values are simply a different representation of the nominal and free text values, and thus differ from 'real' numerical values. As a result, learning tasks that rely on 'real' numerical values, such as regression, cannot be used.

3.2.2. Task Description

Recall from Section 1.2 that Hypothesis Generation and Semantic Content Mining was chosen as the preferred task for the following experiments. Hypothesis Generation involves detecting interesting and potentially relevant patterns that can be presented to users as starting blocks for forming new research hypotheses. The researcher might already have a hypothesis, in which case found patterns may strengthen their belief in the hypothesis. Alternatively, the patterns may reveal something new to the researcher that they are interested in exploring further. The support and confidence patterns will be generated algorithmically on the basis of predefined criteria and user feedback.

Semantic Content Mining involves determining the optimal connotation of a data set by its semantics. In the case of a single semantically-annotated report, the result will likely be limited. This stems mainly from the low number of relevant attributes per graph. A different approach is to mine over more than one semantically-annotated report. For this purpose, the whole output of a processed report could function as a semantic bag of words. This would, for example, allow for more accurate clustering or classifying of reports. Consequently, it could minimize the time that researchers have to spend searching for relevant documents. In addition, it would allow for the categorisation of reports in different types of hierarchies, e.g., by period or by location. Yet another possibility is to rank reports based on how well their contents fits a certain criterion.

3.2.3. Experimental Design

The tasks discussed above were investigated using multiple experiments. Each experiment tested various sets of constraints on different subsets of the entire data set. To acquire these subsets, archaeological searches on topics of interest were translated into SPARQL queries. These queries were subsequently executed on a data set. Together with the remaining constraints, the subset of data was offered as input to the pipeline described in Appendix B. The results and evaluation of this specific case are provided in Section 3.2.4 and 3.2.5, respectively.

Hypothesis Generation

The following experiments involved the generation of hypotheses using ARM. Each of the experiment's runs used a subset of the data set described in Section 3.2.1. Textual descriptions of these subset's criteria are provided in Table 3-1, as well as additional constraints. Note that a detailed explanation of how these criteria come into play has been documented in Appendix B.

Table 3-1: Configurations of the Hypothesis Generation Experiment

ID	Data set Selection Criteria	Variation	Sampling Method	Generalization Factor
A	All facts related to artefacts (EHE0009_ContextFind)	1	Context Specification A°	0.1
		2	Context Specification A°	0.6
		3	Context Specification A°	0.9
B	All facts related to artefact deposit events	1	Context Specification B†	0.1

	(EHE1004_ContextFindDepositionEvent)	2	Context Specification B†	0.6
		3	Context Specification B†	0.9
C	All facts related to context events (EHE1001_ContextEvent)	1	Context Specification C‡	0.1
		2	Context Specification C‡	0.6
		3	Context Specification C‡	0.9
D	All facts related to artefact deposit events (EHE1002_ContextFindProductionEvent)	1	Context Specification D•	0.1
		2	Context Specification D•	0.6
		3	Context Specification D•	0.9

The next four tables denote the context definitions used by experiments A through D. These definitions describe one or more predicate paths that constitute an individual context. As such, they

Legend

ID	Identifier of experiment. This is tied to the selected subset.
Variation	Variation on experiment with a different set of constraints.
Data set Selection Criteria	Criteria used to generate a subset of the entire data cloud.
Sampling Method	Method to sample individuals within the subset.
Generalization Factor	How well rules generalize. Higher values imply less generalization.

contribute relevant information to an individual which is exploited in the pipeline. For instance, the context definition listed in *Table 3-2* will result in individuals of class *EHE0009_ContextFind* with each having as many attributes as there are rows (i.e. 3). The corresponding attribute values were retrieved by requesting the values at the end of the defined predicate paths. Note that these context definitions were created with the help of a domain expert.

*Table 3-2: Context definition for instances of type **EHE0009_ContextFind**, as used in experiment A.*

ID	Predicate Path
1	http://purl.org/dc/elements/1.1/source
2	http://www.cidoc-crm.org/cidoc-crm/P4F_has_time-span, http://www.cidoc-crm.org/cidoc-crm/P1F_is_identified_by, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value
3	http://www.cidoc-crm.org/cidoc-crm/P108F_has_produced, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value

*Table 3-3: Context definition for instances of type **EHE1004_ContextFindDepositionEvent**, as used in experiment B.*

ID	Predicate Path
1	http://purl.org/dc/elements/1.1/source
2	http://www.cidoc-crm.org/cidoc-crm/P25F_moved, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value
3	http://www.cidoc-crm.org/cidoc-crm/P25F_moved, http://www.cidoc-crm.org/cidoc-crm/P45F_consists_of, http://www.w3.org/1999/02/22-rdf-syntax-ns#value
4	http://www.cidoc-crm.org/cidoc-crm/P26F_moved_to, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value
5	http://www.cidoc-crm.org/cidoc-crm/P26F_moved_to, http://www.cidoc-crm.org/cidoc-crm/P45F_consists_of, http://www.w3.org/1999/02/22-rdf-syntax-ns#value

*Table 3-4: Context definition for instances of type **EHE1001_ContextEvent**, as used in experiment C.*

ID	Predicate Path
1	http://purl.org/dc/elements/1.1/source
2	http://www.cidoc-crm.org/cidoc-crm/P4F_has_time-span, http://www.cidoc-crm.org/cidoc-crm/P1F_is_identified_by, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value
3	http://www.cidoc-crm.org/cidoc-crm/P7F_witnessed, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value

*Table 3-5: Context definition for instances of type **EHE1002_ContextFindProductionEvent**, as used in experiment D.*

ID	Predicate Path
1	http://purl.org/dc/elements/1.1/source
2	http://www.cidoc-crm.org/cidoc-crm/P4F_has_time-span, http://www.cidoc-crm.org/cidoc-crm/P1F_is_identified_by, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value
3	http://www.cidoc-crm.org/cidoc-crm/P108F_has_produced, http://www.cidoc-crm.org/cidoc-crm/P2F_has_type, http://www.w3.org/1999/02/22-rdf-syntax-ns#value

Semantic Content Mining

The following three experiments – A, B, and C – involved Semantic Content Mining. Each of the experiment's runs will use the data set described in Section 3.2.1. Experiment A concerned the aggregated data of more than 50 annotated reports. To exemplify the difference between mining *within* and mining *between* reports, experiments B and C concerned one arbitrarily-selected report each. While the remaining reports were mined as well, they are omitted those from this document

for clarity. Note however, that the selected reports are a representive samples of all available reports. An overview of these experimental configurations is provided in Table 3-6.

All resources in the data set were of a type that has been defined with a string literal (Excerpt 3-1). This string literal equals the raw annotated term from the corresponding report. Hence, is it likely to have different types defined for various versions of the same term, for instance, due to singular or plural forms, due to qualifiers, or due to other linguistic symbols (e.g. a dash). Therefore, to prepare the annotated data for mining, we have first aligned all terms with the help of text mining tools. In addition, text cleaning tools were used to remove most erroneous terms.

Table 3-6: Configurations of experiments for Semantic Content Mining

ID	Data set Selection Criteria	# Contexts	# Finds	# Periods
A	Entire Data set consisting of more than 50 aggregated annotated reports.	26675	20984	32755
B	Annotated report on Lincolnshire project[†]	2906	1903	3401
C	Annotated report on Essex project[‡]	3005	2119	3788

[†] *ARCHAEOLOGICAL EVALUATION ON LAND AT MANOR FARM, SUDBROOK, LINCOLNSHIRE (SUMF06) Work Undertaken For HPC Homes Ltd. September 2006*

[‡] ARCHAEOLOGICAL EXCAVATION AND MONITORING, DAGNETS FARM, M. LANE, BRAINTREE, ESSEX. September 2005

Legend (Table 3-6**)**

ID	Identifier of experiment. This is tied to the selected subset.
# Contexts	Number of semantic annotations on contexts.
# Finds	Number of semantic annotations on finds.
# Periods	Number of semantic annotations on periods.

3.2.4. Results

In this section, the results of the two selected Data Mining tasks are presented: Hypothesis Generation and Semantic Content Mining. These results are evaluated in Section 3.2.5.

Hypothesis Generation

This section will present the output of the Hypothesis Generation experiments described above. In total, the combined raw results hold more than ten thousand potential hypotheses. To reduce this to a more-manageable amount, the raw results were first passed through an algorithmic filter. This filter (Table 3-7) applies 'common sense' heuristics to narrow the search. For instance, hypotheses with a minimal support apply only to a single resource and are therefore unsuitable for generalisation over other resources. In addition, hypotheses with a maximum confidence apply to all resources of a certain type, and can thus be regarded as trivial. Omitting these hypotheses will thus lead to a cleaner result with less noise.

Table 3-7: Table listing the conditions used to filter the raw output of the experiments.

FILTER	Condition
1	Hypothesis must not be too common, for else it describes common knowledge that does not contribute to the entity's semantics. Value depends on class frequency in dataset.
2	Hypothesis must not be too rare, for else it describes a peculiarity of a few distinct cases. Value depends on class frequency in dataset.
3	Hypothesis must not hold for only a single entity, for else it describes a unique characteristic of that entity. Value depends on class frequency in dataset.
4	Hypothesis must not be about any of the following irrelevant properties: - RDF label - OWL sameAs - SKOS prefLabel - SKOS note - SKOS scopeNote - SKOS inScheme - SKOS topConceptOf - DCTERMS issued - DCTERMS medium - PRISM versionIndentifier - CIDOC CRM preferred_identifier - CIDOC CRM identified_by - CIDOC CRM note - CIDOC CRM documents - GEOSPARQL asGML

Unfortunately, none of the hypotheses generated in any of the four experiments passed the criteria defined by the algorithmic filter. Specifically, the large majority of the hypotheses described a pattern of a single resource and its attributes, thus failing the third criterion. The remaining hypotheses described patterns that were too rare, thus failing the second criterion. An example of such a hypotheses is one that describes that a term in report *X* is found to be of type *Y*, and dates from period *Z*. Here, *X*, *Y*, and *Z* represent variables, specifically URIs. Therefore, this hypothesis, whilst correct, does not teach us anything which has not already been described in the data itself.

To allow us to analyse the generated hypotheses, we have rerun the algorithmic evaluation without upholding criteria two and three. This resulted in several thousand potential hypotheses for every one of the four experiments. For each of these, we have randomly sampled five hypotheses for expositional purposes. These hypotheses are listed in tables Table 3-8,

Table 3-9,

Table 3-10, and Table 3-11 for experiments A through D, respectively. All sample hypotheses have been manually transcribed to facilitate easy interpretation. Please note that the data set is lenient in its use of semantics. Hence, while some transcriptions might appear to extend the implication of a hypothesis, they are in fact an accurate description.

Table 3-8: Five representable results of experiment A (archaeological finds (EHE0009_ContextFind).

Ex.	Hypothesis
A1	[crmeh#EHE0009_ContextFind] IF ('P2F_has_type', 'torc') THEN ('DC:source', 'Suffolk') Support: < 0.001 Confidence: 1.000 For every find holds that, if they are of material type 'torc', then they originate within the reports on Suffolk.
A2	[crmeh#EHE0009_ContextFind] IF ('P2F_has_type', 'sherds') THEN ('P45F_consists_of', 'pottery') Support: < 0.001 Confidence: 1.000 For every find holds that, if they are of type 'sherds', then they consist of 'pottery'.
A3	[crmeh#EHE0009_ContextFind] IF ('P2F_has_type', 'ironstone') THEN ('P45F_consists_of', 'pottery') Support: < 0.001 Confidence: 1.000 For every find holds that, if they are of type 'ironstone', then they consist of 'pottery'.
A4	[crmeh#EHE0009_ContextFind] IF ('P2F_has_type', 'datable artefacts') THEN ('DC:source', 'Suffolk') Support: < 0.001 Confidence: 1.000 For every find holds that, if they are of type 'datable artefacts', then they originate within the reports on Suffolk.
A5	[crmeh#EHE0009_ContextFind] IF ('P2F_has_type', 'whitewashed bricks') THEN ('P45F_consists_of', 'building') Support: < 0.001 Confidence: 1.000 For every find holds that, if they are of type 'whitewashed bricks', then they consist of a 'building'.

Table 3-8 lists five representable results of experiment A (archaeological finds (EHE0009_ContextFind). Three of these results indicate a correlation between the material and category of a find. For instance, sample A2 states that finds of 'material' sherds consist of pottery. Note that, in this example, the property 'material' is used in its broadest sense. The remaining three hypotheses indicate a correlation between certain types of finds and the report they are from. For instance, hypothesis A4 states that all 'datable artefacts' have been documented in reports on Suffolk. It is easy to see, even without the help of domain experts, that this cannot be true. The reason that this hypothesis has been generated stems from the use of the joined term 'datable artefacts' in one report, whereas that joined term is not present in any of the other reports.

Table 3-9: Five representable results of experiment B on artefact deposit events (EHE1004_ContextFindDepositionEvent).

Ex.	Hypothesis
B1	[crmeh#EHE1004_ContextFindDepositionEvent] IF ('P25F_moved', 'bowl') THEN ('DC:source', 'Lincolnshire') Support: < 0.001 Confidence: 1.000 For every deposition event holds that, if it involved moving a 'bowl', then it originates from reports on Linconshire.
B2	[crmeh#EHE1004_ContextFindDepositionEvent] IF ('P25F_moved', 'iron') THEN ('P26F_moved_to', 'internal well) Support: < 0.001 Confidence: 1.000 For every deposition event holds that, if it involved moving 'iron', then it moved that find to an 'internal well'.
B3	[crmeh#EHE1004_ContextFindDepositionEvent] IF ('P25F_moved', 'pottery') THEN ('P26F_moved_to', 'rubbish pit') Support: < 0.001 Confidence: 1.000 For every deposition event holds that, if it involved moving 'pottery', then it moved that find to a 'rubbish pit'.
B4	[crmeh#EHE1004_ContextFindDepositionEvent] IF ('P25F_moved', 'clay') THEN ('P26F_moved_to', 'grey deposit') Support: < 0.001 Confidence: 1.000 For every deposition event holds that, if it involved moving 'clay, then it moved that find to a 'grey deposit'.
B5	[crmeh#EHE1004_ContextFindDepositionEvent] IF ('P26F_moved_to', 'wall') THEN ('DC:source', 'Headland') Support: < 0.001 Confidence: 1.000 For every deposition event holds that, if it involved moving a 'wall', then it originates from reports on Headland.

Table 3-9 lists five representable results of experiment B on artefact deposit events (EHE1004_ContextFindDepositionEvent). As in the sample of experiment A, two hypotheses indicate a correlation between a resource and the report in which it is documented. Here, these resources involve movement of finds, rather than their type of material. For instance, hypothesis B1 states that only reports on Lincolnshire mention bowls that have been deposited. All remaining hypotheses correlate the moving of a find with the context it was moved to by means of a deposition event. For instance, hypothesis B3 states that pottery was found moved to a 'rubbish pit'.

Table 3-10: Five representable results of experiment C on context events (EHE1001_ContextEvent).

Ex.	Hypothesis
C1	[crmeh#EHE1001_ContextEvent] IF ('DC:source', 'Thames') THEN ('P4F_has_time-span', 'iron age') Support: < 0.001 Confidence: 1.000 For every context event holds that, if it originates from reports on Thames, then it dates from the 'iron age'.
C2	[crmeh#EHE1001_ContextEvent] IF ('P7F_witnessed', 'structure X') THEN ('P4F_has_time-span', '17th century') Support: < 0.001 Confidence: 1.000 For every context event holds that, if it witnessed context 'structure X', then it dates from the '17th century'. Here, X represents a specific resource.
C3	[crmeh#EHE1001_ContextEvent] IF ('P7F_witnessed', 'enclosure X') THEN ('P4F_has_time-span', 'iron age') Support: < 0.001 Confidence: 1.000 For every context event holds that, if it witnessed context 'enclosure X', then it dates from the 'iron age'. Here, X represents a specific resource.
C4	[crmeh#EHE1001_ContextEvent] IF ('P7F_witnessed', 'town X') THEN ('P4F_has_time-span', 'medieval') Support: < 0.001 Confidence: 1.000 For every context event holds that, if it witnessed context 'town X', then it dates from the 'medieval'. Here, X represents a specific resource.
C5	[crmeh#EHE1001_ContextEvent] IF ('P7F_witnessed', 'street X') THEN ('P4F_has_time-span', 'medieval') Support: < 0.001 Confidence: 1.000 For every context event holds that, if it witnessed context 'street X', then it dates from the '17th century'. Here, X represents a specific resource.

Table 3-10 lists five representable results of experiment C on context events (EHE1001_ContextEvent). Once again, we can observe a hypothesis that correlates a resource to a report. In this sample, that resource involves contexts with time spans. For instance, hypothesis C1

states that all contexts documented in reports on Thames date from the Iron Age. The four other hypotheses correlate the finding of a context with a certain time span. For instance, hypothesis C2 states that a certain structure dates from the 17th century. Note that, in this sample, context lacks a specific context type. Hence, each context instance (variable *X*) is unique and thus prevents the hypothesis from generalising to other contexts.

Table 3-11: Five representable results of experiment D on artefact deposit events (EHE1002_ContextFindProductionEvent)

Ex.	Hypothesis
D1	[crmeh#EHE1002_ContextFindProductionEvent] IF ('P108F_has_produced', 'brooches') THEN ('P4F_has_time-span', 'Roman Period') Support: < 0.001 Confidence: 1.000 For every context production event holds that, if it has produced 'brooches', then it dates from the 'Roman Period'.
D2	[crmeh#EHE1002_ContextFindProductionEvent] IF ('P108F_has_produced', 'tempered sherds') THEN ('P4F_has_time-span', '14th to 15th century ') Support: < 0.001 Confidence: 1.000 For every context production event holds that, if it has produced 'tempered sherds', then it dates from the '14th to 15th century'.
D3	[crmeh#EHE1002_ContextFindProductionEvent] IF ('P108F_has_produced', 'brick') THEN ('P4F_has_time-span', 'modern') Support: < 0.001 Confidence: 1.000 For every context production event holds that, if it has produced 'bricks', then it dates from 'modern' times.
D4	[crmeh#EHE1002_ContextFindProductionEvent] IF ('P108F_has_produced', 'pottery') THEN ('P4F_has_time-span', 'late medieval') Support: < 0.001 Confidence: 1.000 For every context production event holds that, if it has produced 'pottery', then it dates from 'late medieval' times.
D5	[crmeh#EHE1002_ContextFindProductionEvent] IF ('P108F_has_produced', 'roman pottery') THEN ('P4F_has_time-span', 'Roman age') Support: < 0.001 Confidence: 1.000 For every context production event holds that, if it has produced 'roman pottery', then it dates from 'Roman age'.

Table 3-10 lists five representable results of experiment D on artefact deposition events (EHE1002_ContextFindProductionEvent). In contrast with the results of the other experiments, all generated hypotheses are of similar structure. That is, all hypotheses in the sample correlate the production of finds to a time span. For instance, hypothesis D2 states that tempered sherds have

been produced in the 14th to 15th century. Note however, that the term "14th to 15th century" is seen as different in the data set from synonyms such as "14th - 15th century" and "14th century to 15th century".

Semantic Content Mining

This section will present the output of the Semantic Content Mining experiments described above. For each experiment, the results were ordered on the normalised frequency of their semantic tags. For expositional purposes, these tags were separated into three categories: contexts, finds, and periods. Furthermore, the results for each of these three categories have been limited to their top-ten tags.

Experiment A mined the aggregated data from more than 50 reports. The top-ten semantic tags occurring in its results are shown in Figure 3-2 for contexts, finds, and periods. In the case of contexts, a gradual decrease in percentages can be observed from the most frequent to least frequent context. Moreover, we can observe the percentage range is rather limited, from three to eight percent. Both observations seem to indicate that none of the contexts occur significantly more than other contexts.

In the case of finds, we can observe a steep decrease in percentages over the four most-frequent tags. In these four steps, the percentage decreases from 22 to four percent, thus losing close to one-fifth in frequency. Together, these two observations may indicate that pottery, sherds, and brick are mentioned far more frequent than other types of finds.

Finally, in the case of periods, a sudden and strong decrease was observed after the third most-frequent tag. This is followed by a slowly decreasing right tail distribution. Both observations seem to indicate that the three most-frequent periods; medieval, Roman, and modern, are mentioned far more often than any of the other periods.

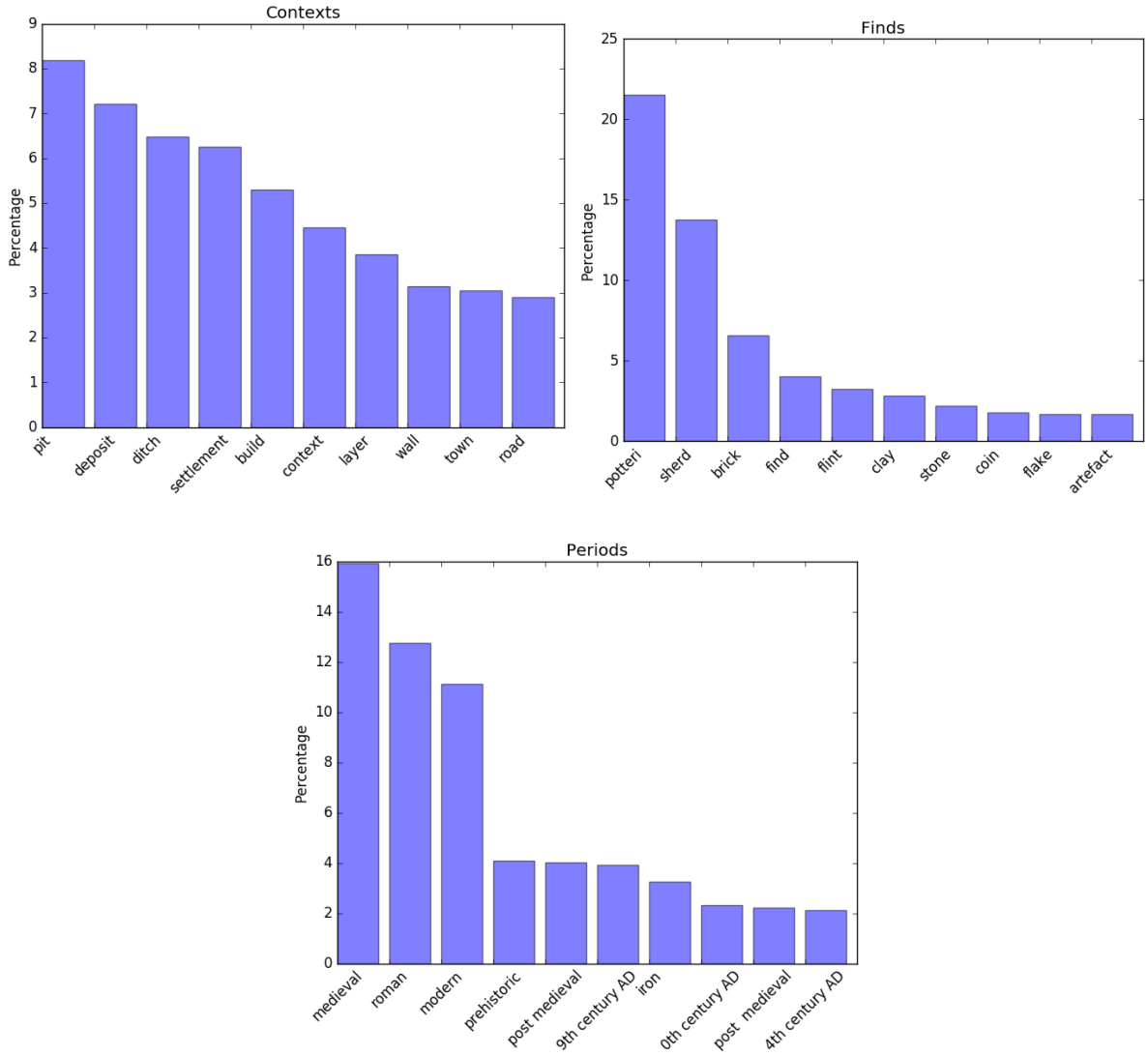

Figure 3-2: Top ten overall semantic content tags from experiment A on contexts (top left), finds (top right), and periods (bottom centre). Note that spelling errors in tags are inherited from the NLP output.

Looking at the results from experiments B and C on reports about Lincolnshire and Essex, respectively, these were arbitrarily selected from all available reports to serve as a representable sample of all project reports. As before, all results have been separated in three categories, and ordered by the normalised frequency of their semantic tags. Furthermore, the results have been limited to their top-ten tags.

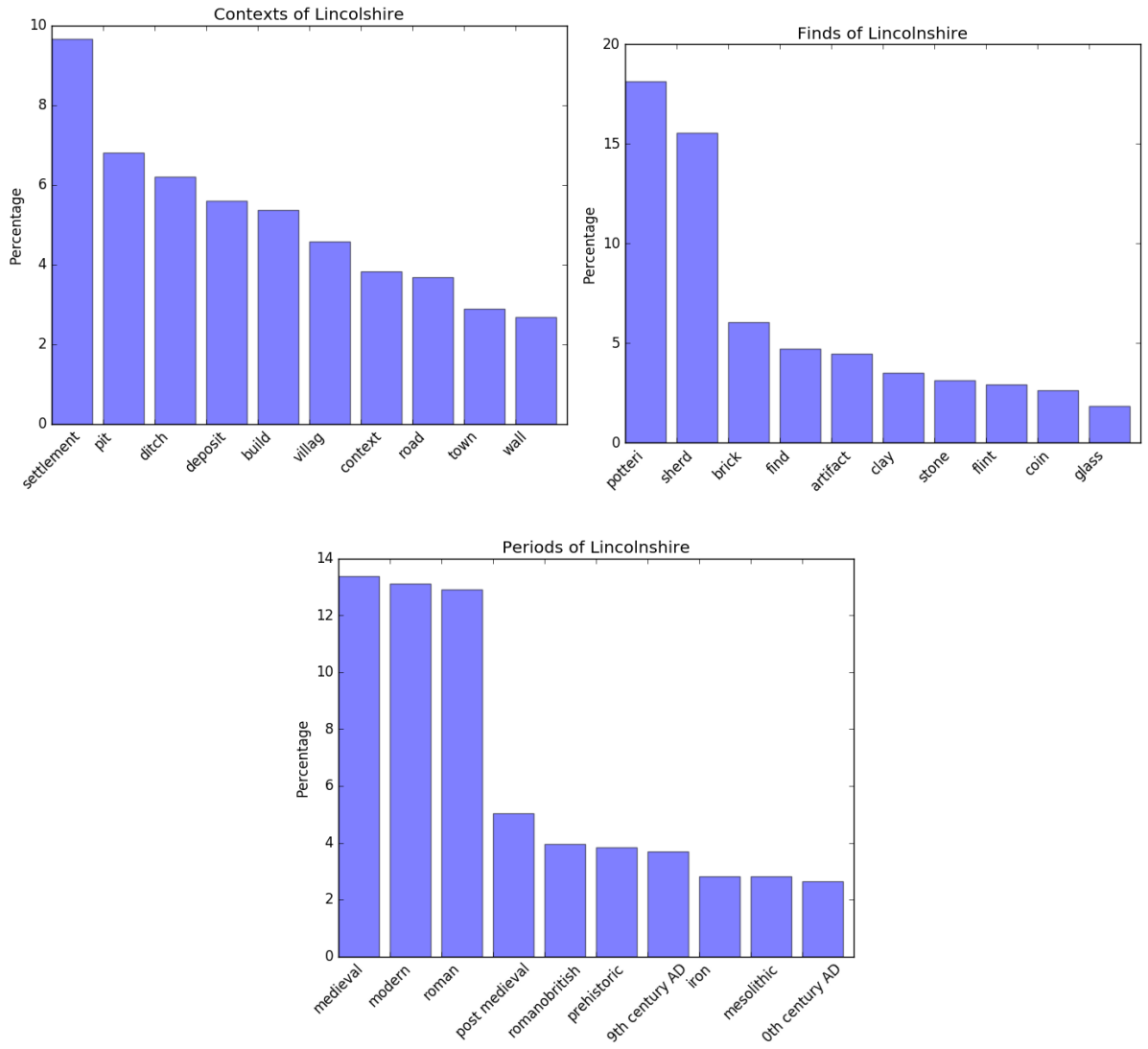

Figure 3-3: Top ten semantic content tags from experiment B on contexts (top left), finds (top right), and periods (bottom centre) from Lincolnshire project reports. Note that spelling errors in tags are inherited from the NLP output.

Experiment B mined the data from a report on Lincolnshire. The top-ten semantic tags occurring in its results are shown in Figure 3-3 for contexts, finds, and periods. In the case of contexts, a steady decrease can be observed that is largely similar to the results from experiment A. The sole exception involves 'settlements', which occur three percent more frequently than the second context 'pit'. In addition, the term 'village' appears to occur more frequently in this report than on average over all reports.

In the case of finds, both 'pottery' and 'sherds' occur considerably more frequently than any of the other finds. Nevertheless, the top four most-frequent terms follow the average find distribution over all reports. However, it can additionally be observed that the trend of the distribution is less smooth than the average distribution.

Finally, in the case of periods, three periods can be observed – medieval, modern, and Roman – roughly share the position of most-prominent term. These three pose a stark contrast to the

remaining periods, which form a slowly decreasing right tail. A similar tail can be observed in the average distribution.

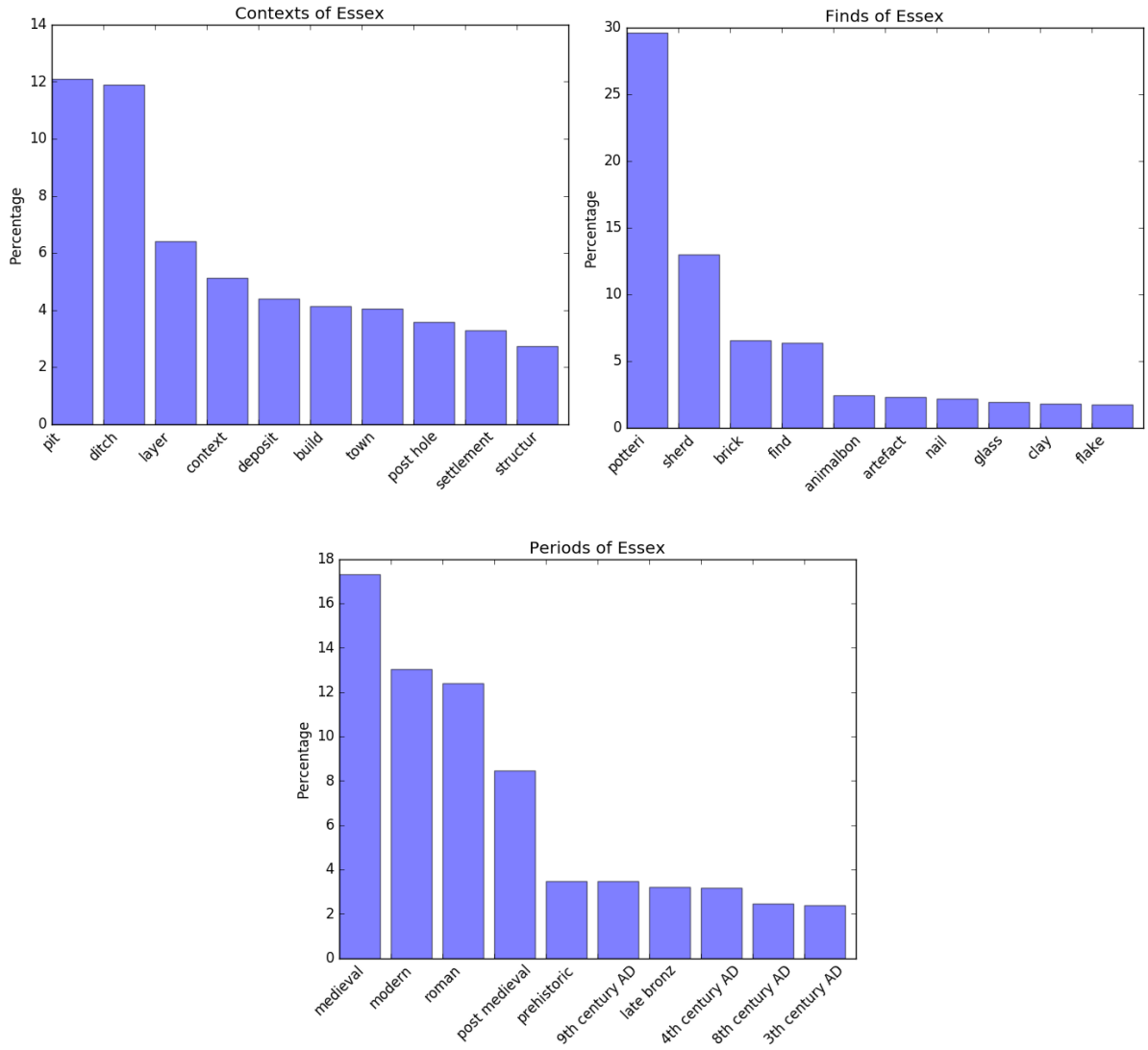

Figure 3-4: Top ten semantic content tags from experiment C on contexts (top left), finds (top right), and periods (bottom centre) from Essex project reports. Note that spelling errors in tags are inherited from the NLP output.

Experiment C mined the data from a report on Essex. The top-ten semantic tags occurring in its results are shown in Figure 3-4 for contexts, finds, and periods. In the case of contexts, it can be observed that both 'pit' and 'ditch' are featured most prominently in reports from Essex, whereas the remaining contexts occur considerably less frequent following a slowly decreasing right tail. In fact, it can be observed that the percentage scale has a maximum value that is several percentages higher than those in the results of both experiment A and B, hence amplifying the previous observation.

In the case of finds, we can observe that pottery occurs considerably more frequent than any of the other terms. In fact, it occurs almost twice as often as on average over all reports. It can additionally

be observed that the remaining finds drop relatively quickly into a frequency percentage of two to three percent. This is roughly similar to the right tail distribution over all reports

Finally, in the case of periods, it can be observed that 'medieval' is by far the most-prominently featured term in the reports, with its value exceeding the number one period in the distribution over all reports. In addition, the terms 'post medieval' occurs relatively frequent compared to the results of both experiment A and B.

3.2.5. Evaluation

This section will involve an evaluation on the results of the experiments. For this purpose, we will first discuss the task of Hypothesis Generation, followed by the task of Semantic Content Mining.

Hypothesis Generation

In total, the combined raw results contained more than ten thousand potential hypotheses. However, none of these were able to pass the criteria defined by the algorithmic filter. Specifically, the large majority of the hypotheses describe a pattern of a single resource and its attributes, whereas the remaining hypotheses describe patterns that, while larger, were still too rare. Analysis indicated that the structure, quality, and size of the data set is the likely reason for the unsatisfying results of this experiment.

As discussed in section 3.2.1, the OPTIMA data set used in the experiments is structurally rather flat. This is caused by OPTIMA's limitations in making inferences that go beyond the local linguistic context of a term. Therefore, there is hardly any cross-linking between the annotated concepts beyond their directly related attributes. Consequently, per processed document, there are multiple small and disconnected graphs, rather than one large and strongly interconnected graph.

Discovering patterns within multiple disconnected graphs proved to be difficult, as there is no explicit reference between two or more graphs that indicate the existence of a possible pattern. This problem is reflected in the resulting hypotheses, which barely spans beyond describing an entity's own attributes. As a result, these hypotheses are unlikely to describe patterns that are unknown or relevant to domain experts. Moreover, it additionally results in hypotheses with very low support values. This greatly decreases the effectiveness of any algorithmic evaluation which, in turn, leads to an explosive increase in the number of generated hypotheses. Often, as was the case in all four experiments, this will make manual evaluation infeasible.

This problem is further increased by the suboptimal quality of the data, as caused by the technical limitations of NLP. More specifically, it is very likely for the same concept to be annotated more than once and with different attribute values if the corresponding term is mentioned more than once as well. Moreover, variations on terms (e.g. singular or plural form, quantifiers, and other linguistic constructs) are seen as entirely different concepts. Each of these terms is subsequently assigned its own URI without reference (e.g. owl:sameAs) as its encountered variation.

A final remark concerns the dimensionality of the data set. In total, the data set was comprised of more than 50 annotated archaeological reports. After conversion, this resulted in nearly one million triples describing nearly 21,000 finds with 27,000 contexts. Despite these numbers, these entities

varied widely, partly due to term inconsistency. Therefore, the pipeline had trouble discovering patterns that can be generalised over the entire data, and instead began overfitting the data.

To ascertain the usefulness of the resulting hypotheses and of the method used to generate them, several domain experts were asked to evaluate the results listed in Table 3-8,

Table 3-9,

Table 3-10, and Table 3-11 on both their plausibility and their relevancy. The reports on this evaluation are listed in Appendix D Expert Evaluation. A shared opinion amongst the evaluation group was that the method produced mostly nonsense, and that none of the hypotheses were of actual use. This supports the outcome of our initial algorithmic evaluation, as well as confirming the theory about the usefulness of mining coarse-grained data.

Semantic Content Mining

The results of the experiments on Semantic Context Mining demonstrated a different method of visualising the semantics of one or more archaeological reports. Using statistical metrics, it was possible to compare the differences in connotation between and within the reports. More specifically, it was possible to analyse the term frequency distribution of finds, contexts, and periods, as well as compare them to those of other reports, or to the global average.

While content mining itself is not novel, exploiting semantic annotations for this purpose is, and give the added possibility to classify terms by their type hierarchy, as well as retrieve terms to which they have been linked. However, any imperfection in the NLP technique will be present in the resulting data. When left uncurated, as is the case in the OPTIMA data set, these imperfections may result in incorrect or inaccurate semantic associations. Therefore, drawing conclusions during subsequent analysis must be handled with care.

Content mining itself has limited use for archaeological research. It can, however, be used to facilitate more accurate clustering or classifying of reports. This could minimise the time that researchers have to spend searching for relevant documents. Furthermore, it would allow for the categorisation of reports in different types of hierarchies, e.g., by find, context, or period. Yet another possibility is ranking reports based on how well their contents fits a certain criterion.

3.3. SIKB Dutch Archaeological Protocol 0102

In an effort to standardise and smooth information exchange between data producers (excavating organisations) and receivers (physical and electronic depots), the Dutch Foundation Infrastructure

for Quality Assurance of Soil Management[13] (SIKB) has, in strong collaboration with the archaeological community, and has been developing the Archaeological Protocol 0102. This protocol, often referred to as *pakbon* (package slip), provides a formalised means to summarise many project elements – reports, databases, media, finds and their context, et cetera – into a single semi-structured file suitable for automated processing. As such, instances of this protocol offer a rich and varied source of archaeological knowledge.

SIKB Protocol 0102 specifies XML as the recommended data serialisation format. For instances of this protocol to be applicable to this research, first the protocol and its already-existing instances had to be translated, and its referenced thesauri transfored into Linked Data. For this reason, the protocol's relatively flat-structured schema has been converted to a more interconnected graph structure with the help of domain experts. In addition, referenced thesauri have been translated to their SKOS equivalent, and a parser has been written to facilitate automatic conversion of existing protocol instances. A set of forty of these instances have since been converted and are hosted via the ClioPatria triple store[14]. Note that a full account of this conversion process is documented in Appendix A.

3.3.1. Data Description

Each converted protocol instance, which will now be referred to as an *instance graph*, consists of the complete contents of the original file. As such, the following high-level groups can be distinguished:

- General information about the entire archaeological project, including people, companies and organisations involved, as well as the general location where the research took place.
- Final and intermediate reports made during the project, as well as different forms of media such as photos, drawings, and videos together with their meta-data and (cross) references to their respective files and subjects.
- Detailed information about the finds discovered during the project, as well as their (geospatial and stratigraphic) relations to each other and the archaeological context in which they were found.
- Accurate information about the locations and geometry where finds were discovered, archaeological contexts were observed, and media was created.

A small excerpt has been given in Excerpt 3-2. To emphasise the difference in the level of knowledge representation with Excerpt 3-1, the focus has been placed once again on an arbitrary find. This find can be seen to hold a relatively large number of attributes, most of which are resources themselves. Two of these are listed as well. Directly linked to the find is the group of finds with which it was discovered. The archaeological context in which this discovery has been made can be seen listed as the former location of said group, and is composed of yet another context.

[13] Stichting Infrastructuur Kwaliteitsborging Bodembeheer in Dutch.
[14] Currently hosted at pakbon-ld.spider.d2s.labs.vu.nl.

Several observations can be made that influence the design of the Data-Mining pipeline. Firstly, many elements have been defined using a custom ontology which utilised Dutch URIs following the protocol's naming scheme. The resources and properties to which these URIs belong have been defined as subclasses and subproperties of equivalent CIDOC CRM and CIDOC CRM-EH elements. This heritage may influence the number of steps (i.e. triples) required to state a fact. For example, when using the aforementioned ontologies, four triples are needed to assign a range of periods ("from *P1* to *P2"*) to a find. As a result, rather than looking one step ahead for direct attributes, a Data-Mining pipeline should look at several.

Another noteworthy characteristic of the *instance* graphs is their considerable portion of free text attributes. These attributes involve general comments, descriptions, various notes, as well as names, labels, and identifiers. A Data-Mining pipeline is, by default, unable to process these textual attributes. A possible solution is to incorporate a method to transform these to numerical attributes. However, as including the free-text attributes listed above is unlikely to increase the overall predictive performance, it might be more effective to simply ignore them.

```
:contextFind_A
   a crmeh:EHE0009_ContextFind ;
   rdfs:label "Vondst (TDS1431:V00960AML)"@nl ;
   :SIKB0102S_aantal numberOfPartsMeasurement_A ;
   :SIKB0102S_artefacttype SIKB_Code_Artefacttype_AWG ;
   :SIKB0102S_geconserveerd false ;
   :SIKB0102S_gedeselecteerd false ;
   :SIKB0102S_gewicht weightMeasurement_A ;
   :SIKB0102S_materiaalcategorie SIKB_Code_Materiaalcategorie_KER ;
   crm:P140i_was_attributed_by datingEvent_A ;
   crm:P1_is_identified_by "TDS1431:V00960AML"^^xsd:ID ;
   crm:P46i_forms_part_of bulkFind_X,
                          collection_Y ;
   crm:P48_has_preferred_identifier "contextFind_A"^^xsd:ID .

:collection_Y
   a crm:E78_Collection ;
   rdfs:label "Veldvondst (TDS1431:413)"@nl ;
   :SIKB0102S_verzamelwijze SIKB_Code_Verzamelwijze_SCHA ;
   crm:P1_is_identified_by "TDS1431:413"^^xsd:ID ;
   crm:P46_is_composed_of contextFind_A,
                          contextFind_B ;
   crm:P48_has_preferred_identifier "collection_Y"^^xsd:ID ;
   crm:P53_has_former_or_current_location context_Y .

:context_Y
   a crmeh:EHE0007_Context ;
   rdfs:label "Vondstcontext (type SPOOR, van Spoor (TDS1431:762))"@nl ;
   :SIKB0102S_contexttype :SIKB_Code_Contexttype_SPOOR ;
   crm:P48_has_preferred_identifier "context_Y"^^xsd:ID ;
   crm:P53_is_former_or_current_location_of collection_Y;
   crm:P89i_contains subContext_Y .
```

Excerpt 3-2: Small example of a context find and (part of) its context. For readability, both identifiers and URIs have been replaced by human-interpretable alternatives. Note that properties starting with "SIKB0102S" are subproperties of their CIDOC CRM equivalent.

3.3.2. Task Description

Recall from Section 1.2 that Hypothesis Generation was chosen as the preferred task for the following experiments. Hypothesis Generation involves detecting interesting and potentially relevant patterns that can be presented to users as starting points for forming new research hypotheses. The researcher might already have a hypothesis; in which case found patterns may strengthen their belief in the hypothesis. Alternatively, the patterns may reveal something new to the researcher that they are interested in exploring further. The support and confidence of patterns will be generated algorithmically on the basis of predefined criteria and user feedback.

3.3.3. Experimental Design

The task discussed above will be investigated using multiple experiments. Each experiment will test various sets of constraints on different subsets of the entire data cloud. To acquire these subsets, archaeological topics of interest will be translated into SPARQL queries. These queries will subsequently sent to the endpoint of the data cloud. Together with the remaining constraints, the subset of data will be offered as input to the pipeline described in Appendix B. Note that the results and evaluation of this specific case will be provided in Section 3.3.4 and 3.3.5, respectively.

The following experiments will involve the generation of hypotheses using ARM. Each of the experiment's runs will use a subset of the data cloud described in Section 3.3.1. Textual descriptions of these subset's criteria are provided in Table 3-12, as well as additional constraints. Note that a detailed explanation of how these criteria come into play has been documented in Appendix B.

Table 3-12: Configurations of the Hypothesis Generation Experiment

ID	Data set Selection Criteria	Variation	Sampling Method	Generalization Factor
A[15]	All facts related to archaeological contexts (EHE0007_Context).	1	L. Neighbourhood (depth = 2)	0.6
		2	L. Neighbourhood (depth = 2)	0.9
		3	L. Neighbourhood (depth = 3)	0.6
		4	L. Neighbourhood (depth = 3)	0.9
B	All facts related to archaeological cuts (EHE0007_Context with context type Cut)	1	Context Specification B†	0.1
		2	Context Specification B†	0.6
		3	Context Specification B†	0.9
C	All facts related to artefacts (EHE0009_ContextFind)	1	Context Specification C‡	0.1
		2	Context Specification C‡	0.6

[15] Experiment A is the only experiment in which Local-Neighbourhood sampling will be used because its data set selection criterion for archaeological contexts (EHE0007_Context) includes several subclasses, each with very different attribute sets. Hence, no single context definition would suffice.

		3	Context Specification C‡	0.9
D	All facts related to archaeological projects (EHE0001_EHProject)	1	Context Specification D•	0.1
		2	Context Specification D•	0.6
		3	Context Specification D•	0.9

Legend

ID	Identifier of experiment. This is tied to the selected subset.
Variation	Variation on experiment with a different set of constraints.
Data set Selection Criteria	Criteria used to generate a subset of the entire data cloud.
Sampling Method	Method to sample individuals within the subset.
Generalisation Factor	How well rules generalise. Higher values imply less generalisation.

The next three tables denote the context definitions used by experiment B, C, and D. These definitions describe one or more predicate paths that constitute an individual's context. As such, they contribute relevant information to an individual which we exploit in our pipeline. For instance, the context definition listed in *Table 3-13* will result in individuals of class *EHE0007_Context* with each having as many attributes as there are rows (i.e. 12). The corresponding attribute values will be retrieved by requesting the values at the end of the defined predicate paths. Note that these context definitions were created with the help of domain experts.

*Table 3-13: Context definition for instances of type **EHE0007_Context**, as used in experiment B.*

ID	Predicate Path
1	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_grondspoortype
2	http://www.cidoc-crm.org/cidoc-crm/P53i_is_former_or_current_location_of
3	http://www.cidoc-crm.org/cidoc-crm/P89_falls_within, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_contexttype
4	http://purl.org/crmeh#EHP3i, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_kleur
5	http://purl.org/crmeh#EHP3i, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_textuur
6	http://www.cidoc-crm.org/cidoc-crm/P53i_is_former_or_current_location_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_structuurtype
7	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_diepte, http://www.cidoc-crm.org/cidoc-crm/P40_observed_dimension, http://www.cidoc-crm.org/cidoc-crm/P90_has_value
8	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_diepte, http://www.cidoc-crm.org/cidoc-crm/P40_observed_dimension, http://www.cidoc-crm.org/cidoc-crm/P91_has_unit
9	http://www.cidoc-crm.org/cidoc-crm/P140i_was_attributed_by, http://www.cidoc-crm.org/cidoc-crm/P141_assigned, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_beginperiode
10	http://www.cidoc-crm.org/cidoc-crm/P140i_was_attributed_by, http://www.cidoc-crm.org/cidoc-crm/P141_assigned, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_eindperiode
11	http://www.cidoc-crm.org/cidoc-crm/P53i_is_former_or_current_location_of, http://www.cidoc-crm.org/cidoc-crm/P140i_was_attributed_by, http://www.cidoc-crm.org/cidoc-crm/P141_assigned, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_beginperiode
12	http://www.cidoc-crm.org/cidoc-crm/P53i_is_former_or_current_location_of, http://www.cidoc-crm.org/cidoc-crm/P140i_was_attributed_by, http://www.cidoc-crm.org/cidoc-crm/P141_assigned, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_eindperiode

*Table 3-14: Context definition for instances of type **EHE0009_ContextFind**, as used in experiment C.*

ID	Predicate Path
1	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_artefacttype
2	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_exposabel
3	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_geconserveerd
4	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_gedeselecteerd
5	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_materiaalcategorie
6	http://www.cidoc-crm.org/cidoc-crm/P3_has_note
7	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_gewicht, http://www.cidoc-crm.org/cidoc-crm/P90_has_value
8	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_gewicht, http://www.cidoc-crm.org/cidoc-crm/P91_has_unit
9	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_aantal, http://www.cidoc-crm.org/cidoc-crm/P90_has_value
10	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_aantal, http://www.cidoc-crm.org/cidoc-crm/P91_has_unit
11	http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_verzamelwijze
12	http://www.cidoc-crm.org/cidoc-crm/P140i_was_attributed_by, http://www.cidoc-crm.org/cidoc-crm/P141_assigned, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_beginperiode
13	http://www.cidoc-crm.org/cidoc-crm/P140i_was_attributed_by, http://www.cidoc-crm.org/cidoc-crm/P141_assigned, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_eindperiode
14	http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_bewaarTemperatuur
15	http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_bewaarVochtigheid
16	http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_breekbaar
17	http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_lichtgevoelig

*Table 3-15: Context definition for instances of type **EHE0001_EHProject**, as used in experiment D.*

ID	Predicate Path
1	http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_onderzoektype
2	http://www.cidoc-crm.org/cidoc-crm/P7_took_place_at, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_gemeentecode
3	http://www.cidoc-crm.org/cidoc-crm/P7_took_place_at, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_plaatscode
4	http://www.cidoc-crm.org/cidoc-crm/P7_took_place_at, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_provinciecode
5	http://www.cidoc-crm.org/cidoc-crm/P7_took_place_at, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_toponiem
6	http://www.cidoc-crm.org/cidoc-crm/P7_took_place_at, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_vindplaatstype
7	http://www.cidoc-crm.org/cidoc-crm/P7_took_place_at, http://www.cidoc-crm.org/cidoc-crm/P4_has_time-span, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_beginperiode
8	http://www.cidoc-crm.org/cidoc-crm/P7_took_place_at, http://www.cidoc-crm.org/cidoc-crm/P4_has_time-span, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_eindperiode
9	http://www.cidoc-crm.org/cidoc-crm/P9_consists_of, http://www.cidoc-crm.org/cidoc-crm/P108_has_produced, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_artefacttype
10	http://www.cidoc-crm.org/cidoc-crm/P9_consists_of, http://www.cidoc-crm.org/cidoc-crm/P108_has_produced, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_materiaalcategorie
11	http://www.cidoc-crm.org/cidoc-crm/P9_consists_of, http://www.cidoc-crm.org/cidoc-crm/P108_has_produced, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_verzamelwijze
12	http://www.cidoc-crm.org/cidoc-crm/P9_consists_of, http://www.cidoc-crm.org/cidoc-crm/P108_has_produced, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://www.cidoc-crm.org/cidoc-crm/P46_is_composed_of, http://www.cidoc-crm.org/cidoc-crm/P46i_forms_part_of, http://www.cidoc-crm.org/cidoc-crm/P53_has_former_or_current_location, http://pakbon-ld.spider.d2s.labs.vu.nl/ont/SIKB0102S_contexttype

3.3.4. Results

This section will present the output of the experiments described above. In total, the combined raw results hold more than hundred thousand potential hypotheses. To reduce this to a more-manageable number, the raw results were first passed through an algorithmic filter. This filter (Table 3-16) applies 'common sense' heuristics to narrow the results. For instance, hypotheses with minimal support apply only to a single resource and are therefore unsuitable for generalisation over other resources. In addition, hypotheses with a maximum confidence apply to all resources of a certain type, and can thus be regarded as trivial. Omitting these hypotheses will thus lead to a cleaner result with less noise.

Table 3-16: Table listing the conditions used to filter the raw output of the experiments.

FILTER	Condition
1	Hypothesis must not be too common, or else it describes common knowledge that does not contribute to the entity's semantics. Value depends on class frequency in dataset.
2	Hypothesis must not be too rare, or else it describes a peculiarity of a few distinct cases. Value depends on class frequency in dataset.
3	Hypothesis must not hold for only a single entity, or else it describes a unique characteristic of that entity. Value depends on class frequency in dataset.
4	Hypothesis must not be about any of the following irrelevant properties: - RDF label - OWL sameAs - SKOS prefLabel - SKOS note - SKOS scopeNote - SKOS inScheme - SKOS topConceptOf - DCTERMS issued - DCTERMS medium - PRISM versionIndentifier - CIDOC CRM preferred_identifier - CIDOC CRM identified_by - CIDOC CRM note - CIDOC CRM documents - GEOSPARQL asGML

A first pass through the algorithmic filter reduced the number of potential hypotheses to several thousands. For every hypothesis, both its (relative) support and (relative) confidence was calculated. These metrics were used to rank the hypotheses by quantitative relevancy for each of the four experiments: A, B, C, and D. We have semi-randomly sampled five hypotheses from the top hundred hypotheses for expositional purposes. These hypotheses are listed in tables Table 3-17, Table 3-18,Table 3-19, and Table 3-20 for experiments A through D, respectively. All sample hypotheses have been manually transcribed to facilitate easy interpretation.

Table 3-17: Five representable results of experiment A on archaeological contexts (EHE0007_Context).

Ex.	Hypothesis
A1	[crmeh#EHE0008_ContextStuff] IF (SIKB0102S_kleur, 'lichtgeel') THEN (SIKB0102S_grondspoortype, SIKB_Code_Grondspoortype_GRAF.DIER) Support: 0.001 Confidence: 1.000 For every context stuff holds that, if they have a bright yellow colour, then they are of specific type ANIMAL GRAVE.
A2	[crmeh#EHE0004_SiteSubDivision] IF (SIKB0102S_vindplaatstype, SIKB_Code_Complextype_NBAS) THEN (P4_has_time-span, "Tijdspanne (van periode NEOV tot periode NTL)") Support: 0.077 Confidence: 1.000 For every site holds that, if they are of specific type BASE CAMP, then they are dated from the NEOV to NTL period.
A3	[crmeh#EHE0004_SiteSubDivision] IF (SIKB0102S_vindplaatstype, SIKB_Code_Complextype_GVX) THEN (P4_has_time-span, "Tijdspanne (van periode ROM tot periode NT)") Support: 0.077 Confidence: 1.000

	For every site holds that, if they are of specific type GRAVEYARD, then they are dated from the ROM to NT period.
A4	[crmeh#EHE0004_SiteSubDivision] IF (SIKB0102S_vindplaatstype, SIKB_Code_Complextype_EIVB) THEN (P4_has_time-span, "Tijdspanne (van periode MESO tot periode NEO)") Support: 0.077 Confidence: 1.000 For every site holds that, if they are of specific type FLINT CARVING, then they are dated from the MESO to NEO period.
A5	[crmeh#EHE0004_SiteSubDivision] IF (SIKB0102S_vindplaatstype, SIKB_Code_Complextype_GVC) THEN (P4_has_time-span, "Tijdspanne (van periode BRONSM tot periode IJZL)") Support: 0.077 Confidence: 1.000 For every site holds that, if they are of specific type CREMATION, then they are dated from the MIDDLE BRONS to IRON period.

Table 3-1 lists five representable results of experiment A on all types of archaeological contexts (EHE0007_Context). Interestingly, none of the generated hypotheses apply to this class itself, but rather to a subclasses thereof. This is caused by the chosen sampling method (i.e. local neighbourhood). Irrespective of this, only one of the potential hypotheses applies to context stuff. Specifically, it positively correlates the colour of the stuff with the purpose of the context's contents. The low support of this hypothesis seems to indicate that this pattern has only been found to hold infrequently.

Additionally, all four remaining hypotheses are of similar form. That is, each one correlates with the category of the site where the context was found to the likely time span of that context. For instance, hypothesis A4 states that sites at which flint carvings occurred likely date to the Mesolithic to Neolithic period. Finally, note that each support has the same value. This may indicate that certain clusters in the data set exist where all contexts are of from the same location.

Table 3-18: Five representable results of experiment B on archaeological finds (EHE0009_ContextFinds).

Ex.	Hypothesis
B1	[crmeh#EHE0009_ContextFind] IF (SIKB0102S_artefacttype, SIKB_Code_Artefacttype_HOUTSKL') THEN (SIKB0102S_materiaalcategorie, SIKB_Code_Materiaalcategorie_OPHK') Support: 0.009 Confidence: 0.958 For every find holds that, if they are of specific type CHARCOAL, then they are of material type CHARCOAL.
B2	[crmeh#EHE0009_ContextFind] IF (SIKB0102S_materiaalcategorie, SIKB_Code_Materiaalcategorie_STU') THEN (P4_has_time-span, "Tijdspanne (van periode PALEOV tot periode NTL)"@nl) Support: < 0.001 Confidence: 1.000 For every find holds that, if they are of specific type TUFF, then they are dated from EARLY PALEO to LATE NT.
B3	[crmeh#EHE0009_ContextFind] IF (SIKB0102S_artefacttype, SIKB_Code_Artefacttype_RUWNIJM') THEN (P4_has_time-span, "Tijdspanne (van periode ROMV tot periode ROML)"@nl) Support: 0.002 Confidence: 1.000 For every find holds that, if they are of specific type RAW EARTHENWARE (Nimeguen), then they are dated from EARLY ROMAN to LATE ROMAN.

B4	[crmeh#EHE0009_ContextFind] IF (P4_has_time-span, "Tijdspanne (van periode NEOM tot periode NEOL)"@nl) THEN (SIKB0102S_materiaalcategorie, SIKB_Code_Materiaalcategorie_SDI') AND (SIKB0102S_artefacttype, SIKB_Code_Artefacttype_HAMERBL') Support: 0.001 Confidence: 1.000 For every find holds that, if they are dated from MIDDLE NEO to LATE NEO, then they are of specific type HAMMER AXE and of material type DOLERITE.
B5	[crmeh#EHE0009_ContextFind] IF (SIKB0102S_artefacttype, SIKB_Code_Artefacttype_DISSEL') THEN (P4_has_time-span, "Tijdspanne (van periode NEO tot periode BRONS)"@nl) Support: 0.001 Confidence: 1.000 For every find holds that, if they are of specific type ADZE, then they are dated from NEO to BRONS.

Table 3-18 lists five representable results of experiment B on archaeological finds (EHE0009_ContextFinds). Three hypotheses correlate the categorical or material type of a find to the time span it dates to. For instance, hypothesis B3 states that raw Nimegeun earthenware likely dates somewhere between early and late roman times. The two remaining hypotheses correlate the category of a find to its material. It is noteworthy to observe that, if we focus purely on their form, these hypotheses – B1 and B4 – are the inverse of each other. That is, B1 applies to finds of a certain category, whereas B4 applies to finds of a certain material.

Table 3-19: Five representable results of experiment C on archaeological cuts (EHE0007_Context).

Ex.	Hypothesis
C1	[crmeh#EHE0007_Context] IF (SIKB0102S_grondspoortype, SIKB_Code_Grondspoortype_HUTKOM') THEN (P4_has_time-span, "Tijdspanne (van periode ROM tot periode MEV)"@nl) AND (SIKB0102S_structuurtype, SIKB_Code_Structuurtype_PLATTEGR') Support: 0.001 Confidence: 1.000 For every context holds that, if they are of specific type DWELLING, then they are dated from ROMAN to MIDDLE DARK AGES, and have structure FLOOR PLAN.
C2	[crmeh#EHE0007_Context] IF (SIKB0102S_grondspoortype, SIKB_Code_Grondspoortype_GREPPEL.HUISGREP') THEN (P4_has_time-span, "Tijdspanne (van periode IJZV tot periode IJZV)"@nl) AND (SIKB0102S_structuurtype, SIKB_Code_Structuurtype_HUIS') Support: 0.001 Confidence: 1.000 For every context holds that, if they are of specific type HOUSE DITCH, then they are dated from EARLY IRON AGE, and have structure HOUSE.
C3	[crmeh#EHE0007_Context] IF (SIKB0102S_grondspoortype, SIKB_Code_Grondspoortype_WATERPUT') THEN (P4_has_time-span, "Tijdspanne (van periode BRONS tot periode IJZ)"@nl) AND (SIKB0102S_structuurtype, SIKB_Code_Structuurtype_WATERPUT') Support: 0.001 Confidence: 1.000 For every context holds that, if they are of specific type WELL, then they are dated from BRONS to IRON AGE, and have structure WELL.
C4	[crmeh#EHE0007_Context] IF (SIKB0102S_grondspoortype, SIKB_Code_Grondspoortype_GRAF.INHUMGRF) THEN (P4_has_time-span, "Tijdspanne (van periode PREH tot periode PREH)"@nl) AND (SIKB0102S_structuurtype, rdflib.term.URIRef('SIKB_Code_Structuurtype_GRAF) Support: 0.001

	Confidence: 1.000 For every context holds that, if they are of specific type BURIAL, then they are dated from PREH, and have structure GRAVE.
C5	[crmeh#EHE0007_Context] IF (SIKB0102S_diepte, "21 cm") AND (P4_has_time-span, "Tijdspanne (van periode PREH tot periode PREH)"@nl) THEN (SIKB0102S_grondspoortype, SIKB_Code_Grondspoortype_GREPPEL.STANDGRP') Support: < 0.001 Confidence: 1.000 For every context holds that, if they are 21 cm deep, then they are dated from PREH, and have specific type DITCH.

Table 3-19 lists five representable results of experiment C on archaeological cuts (EHE0007_Context). In contrast to the results from all other experiments, the generated hypotheses contain an additional element in either their antecedent (C5) or their consequent (C1 through C4). This is likely the result of having the same support and confidence as their fragments – if A → B & C holds, then its fragments A → B and A → C hold as well – but are favoured by the algorithmic filter due to their broader applicability.

Four of the generated hypotheses correlate the type of cut of a certain context to its type of structure and its likely time span. For instance, hypothesis C1 states that floor plans of dwellings likely date somewhere between Roman times and middle Dark ages. We can additionally observe that hypothesis C5 correlates a certain depth and likely time span of a context to the category of that context. It is noteworthy to remark the fixed value for depth of "21 cm" as antecedent. As the current method used to generate hypotheses is unable to compare literals, it is also unable to differentiate between closely related literals. Hence, a value of "21" is seen as different from "20" as it is from "100", "1000", and even from non-numerical literals such as "posthole".

Table 3-20: Five representable results of experiment D on archaeological Projects (EHE0001_EHProject).

Ex.	Hypothesis
D1	[crmeh#EHE0001_EHProject] IF (SIKB0102S_onderzoektype, SIKB_Code_Verwerving_AOP) THEN (SIKB0102S_vindplaatstype, SIKB_Code_Complextype_BEWV.X) Support: 0.014 Confidence: 1.000 For every projects holds that, if they are of specific type 'destructive excavation', then they have location type 'settlement with defences' .
D2	[crmeh#EHE0001_EHProject] IF (SIKB0102S_onderzoektype, SIKB_Code_Verwerving_AVE) THEN (SIKB0102S_vindplaatstype, SIKB_Code_Complextype_BEWV.X) Support: 0.014 Confidence: 1.000 For every project holds that, if they are of specific type 'destructive mapping', then they are of location type 'settlement with defences'.
D3	[crmeh#EHE0001_EHProject] IF (P7_took_place_at, location X) THEN (P7_took_place_at, location Y) Support: 0.029 Confidence: 0.500 For every project holds that, if they took place at location X, then they also too place at location Y. Here, X and Y are variables and Y lies within X.

D4	[crmeh#EHE0001_EHProject] IF (SIKB0102S_onderzoektype, type X) THEN (P9_consists_of, document Y) Support: 0.043 Confidence: 0.333 For every project holds that, if they are of specific type X, then is consists of (documentation) Y. Here, X and Y are variables.
D5	[crmeh#EHE0001_EHProject] IF (P9_consists_of, document X) THEN (P9_consists_of, container Y) Support: 0.014 Confidence: 1.000 For every project holds that, if they are documented in document X, then they consist of container Y. Here, X and Y are variables.

Table 3-20 lists five representable results of experiment D on archaeological projects (EHE0001_EHProject). Two of these hypotheses – D1 and D2 – correlate the type of project to the type of its location. For instance, hypothesis D1 states that destructive excavations occur at settlements with defences. It is unlikely that this hypothesis can be generalised over other, yet unseen, data sets. This is also indicated by the low support it has, and is probably caused by several protocol instances documenting different facets of the same project.

The three remaining hypotheses can be best viewed as templates of possible hypotheses. For this purpose, individual URIs with variables *X* and Y have been replaced. For instance, hypotheses D3 states that projects occurring at a certain location (i.e. area) has sites within that location. Note however, that this hypothesis has a confidence of 0.5, thus indicating it was found to be true in half the cases. Similarly, hypothesis D4 was found to be true in only one third of the cases.

3.3.5. Evaluation

In its entirety, the combined raw results contained more than a hundred thousand potential hypotheses. We were able to bring this down to several thousands by passing these results through an algorithmic filter. The remaining hypotheses were subsequently assigned support and confidence values. These values were used to rank the hypotheses by quantitative relevancy, after which we semi-randomly sampled five hypotheses from the top hundred for each of the four experiments.

Overall, it was observed that the generated hypotheses were able to correctly reflect patterns within the data set. However, the relevancy of these patterns, as well as their ability to generalise, has proven difficult to determine without manual evaluation. Sifting through candidate hypotheses by hand takes considerable time, and the metrics, the purpose of which is to guide this process more intelligently, were found to be inadequate. More specifically, both confidence and support are strongly influenced by the size, variety, and quirks of a data set. For instance, the uniqueness of entities in Linked Data can quickly lead to high confidence and low support.

From a data perspective, we observed difficulties with both the instance and ontology graph. In the case of the latter, it appears that the path length between related entities makes it difficult to discover over. This is largely a characteristic inherited from the CIDOC CRM ontology, which is known for its verbosity. Regardless, shorter paths (i.e. more local patterns) were computationally less costly, and were therefore preferred by the method employed. Looking at the instance graph, we could observe that common and omnipresent attributes clutter results. In this case, Linked Data's

characteristic of reusing properties amplified this issue. In addition, many of the interesting values for these properties consisted of textual or numerical literals. As the method used to generate hypotheses was unable to compare literals, it is also unable to differentiate between closely related literals. This became especially apparent in cases of geometries (WKT literal) and metrics.

To ascertain the usefulness of the resulting hypotheses and of the method that has been used to generate them, we have asked several domain experts to evaluate the results listed in Table 3-17, Table 3-18, Table 3-19, and Table 3-20 on both their plausibility and their relevancy. The reports on this evaluation are listed in Appendix D Expert Evaluation. A shared opinion amongst the evaluation group was that the method produced mostly plausible hypotheses, with a handful potentially being of actual use. The main criticism was that few hypotheses yielded novel insights into the data or were generalisable to other data sets. Instead most hypotheses were found to be either trivial, tautologies, or specific to a single project or area. Nevertheless, all experts were of the opinion that the method yielded promise, and that, with sufficient improvements, it may eventually result in a tool that is of actual use to the archaeological community.

4. Conclusion

This work set out to investigate the technical feasibility and practical usability of the recommendations made at the end of deliverable D16.1: *First Report on Data Mining*. For the purpose of this report, efforts were focussed primarily on one of these recommendations, specifically Hypothesis Generation, due to this task's novelty and potential usefulness to the archaeological community. As this task was not found suitable for all data sets, more traditional data mining methods were additionally applied as well. These involved 1) Semantic Content Mining to determine and compare project connotations, 2) relationship discovery among the authors of OpenAIRE and ARIADNE Reports, 3) the creation and analysis of ARIADNE's author networks, and 4) performing text mining on OpenAIRE publications to link them with ARIADNE metadata or other extracted objects from the ARIADNE reports.

To increase the scope of our work in this task, three separate case studies in Data Mining on archaeological data were conducted. These are 1) data from the ARIADNE Registry, 2) grey literature that has been semantically-annotated using the OPTIMA text-mining pipeline, and 3) rich Linked Data database extractions that follow the SIKB Protocol 0102 specification. Each of these three studies has a different granularity of knowledge. More fine-grained data sets have more specific information about archaeological findings and their contexts, whereas more coarse-grained data sets describe information at a higher level.

Overall, the combined results of all three case studies can be seen as a mildly promising for Data Mining for Linked Archaeological Data. From the results of the ARIADNE Registry case study, the most encouraging were the citation links found in ADS grey literature reports to PubMed/ePCM publications in OpenAIRE. Evaluation of these links by a domain expert indicated that a considerable number of them were indeed relevant. Note that these links covered various topics, including topics of the broader archaeological field such as anthropology, biology, paleontology, pottery, as well as outside the UK, e.g. Oceania.

From the results of the case studies on both semantically-annotated reports (OPTIMA) and rich Linked Data database extractions (SIKB Protocol 0102), the generated hypotheses were able to correctly reflect patterns within the data set. However, the relevancy of these patterns, as well as their ability to generalise, has proven to be very dependent on both the structure and quality of the data set. In the OPTIMA case study, the data set was of insufficient quality (uncurated NLP output) and structurally flat. Therefore, the generated hypotheses barely spanned beyond describing an entity's own attributes and, consequently, were found to be of little use to our evaluation group. In contrast, a shared opinion amongst the evaluation group was that the results of the SIKB Protocol case study were nearly all plausible, with only a handful potentially being of actual use. This, in combination with the results from the OPTIMA case study, supports the initial theory about the influence of a data set's granularity of knowledge on the usefulness of pattern mining.

There are numerous challenges still to overcome for Hypothesis Generation to mature and be of use to the archaeological community. For instance, sifting through candidate hypotheses by hand takes considerable time, and the metrics, the purpose of which is to guide this process more intelligently,

were found to be inadequate. Moreover, the main criticism from the evaluation group concerned the fact that few hypotheses yielded novel insights into the data or were generalisable to other data sets. Instead most hypotheses were found to be either trivial, tautologies, or specific to a single project. Nevertheless, all experts were responded positively to the range of patterns the method was able to generate.

At present, Data Mining for Linked Data still has a long way to go before it will transcend the academic stadium. Eventually, with sufficient improvements, continuing research may result in a tool that is of actual use to the archaeological community. Until that time arrives, it may be best to employ traditional Data Mining techniques that have matured over several decades, and which have proven to produce reliable and useful results.

4.1. Recommendations

Based upon the research done and the experience gained over the past years working on this deliverable, a series of recommendations can be made.

Granularity of knowledge

The usefulness of data mining to domain experts depends heavily on the data's granularity of knowledge. More fine-grained data have more-specific information, whereas more coarse-grained data describes information at a higher level (e.g. collection level metadata). Discovering domain-relevant patterns within coarse-grained data proves to be difficult, as these data simply do not contain such patterns, neither explicitly nor implicitly. Therefore, mining coarse-grained data will at best, result in high-level constructs, rather than yield new insight that can help further archaeological research. Hence, the creation and use of fine-grained data should be stimulated.

Choice of ontology

The informativeness of patterns depends heavily on the structural features of the data. These features are often inherited from the ontologies used to describe the data. More verbose ontologies, such as CIDOC CRM, require more steps (i.e. triples) to describe the same information. Discovering patterns over a larger number of steps is computational more difficult than if the number of steps are less. As a result, local patterns are preferred (less costly), thus potentially leading to more trivial and less complex patterns. Therefore, to discover more useful patterns, either an ontology with low verbosity should be used, or implement a penalty for local patterns so that more complex patterns will be favoured.

Data set quality

Discovering plausible patterns from data requires these data to be of a sufficient quality. Therefore, all data should be curated before a Data Mining method is applied. If this is not the case, as observed with the results of the uncurated OPTIMA text mining case study, then none of the patterns found will be archaeologically correct. This is a known weakness of many text mining efforts. Therefore, it is recommended to not use their output, except when it is curated.

Literal support

Most archaeological data sets consist of textual or numerical values that can not be represented by an entry from a thesaurus. In Linked Data, these values become literals holding a certain data type. As such, they differ from resources and are thus not well accounted for in the graph representation of a data set. Most Data Mining methods that can be applied to Linked Data exploit this graph representation, as does the one employed in this work, and thus often ignore literals. However, these literals typically contain very interesting values, such as geometries and metrics (e.g. depth and dimensions). We strongly recommend taking these values into consideration.

Appendix A SIKB Protocol 0102 Conversion to RDF

The SIKB Archaeological Protocol 0102, often referred to as *pakbon* (package slip), provides a formalised means to summarise many of a project's elements into a single semi-structured file suitable for sharing and automated processing. Due to the use of XML as the recommended data serialisation format, protocol, thesauri, and already existing *pakbon* files had to be translated to LD. This entire process has been documented in this appendix.

A.1 Protocol Description

A coarse-grained view of the SIKB Archaeological Protocol 0102 indicates the existence of several distinct groups of information. In short, these groups concern the following forms of data:

- General information about the entire archaeological project, including people, companies and organisations involved as well as the general location where the research took place.
- Final and intermediate reports made during the project, as well as different forms of media such as photos, drawings, and videos together with their meta-data and (cross) references to their respective files and subject.
- Detailed information about the finds discovered during the project, as well as their (geospatial and stratigraphic) relation to each other and the archaeological context in which they were found.
- Accurate information about the locations and geometry at which finds were discovered, archaeological contexts were observed, and media was created.

Despite their implicit relations, the structure holding these four groups is rather flat and disconnected (Figure A-1). In fact, all but five of the protocol's 43 elements are direct children of the root element (Figure A-2). To express relations that are more complex, the protocol employs identifiers and cross referencing. These identifiers are only guaranteed to be unique within a single instance of the protocol.

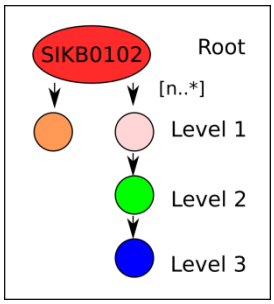

Figure A-1: Level hierarchy of the protocol's original data model. Note that the first level has been divided in meta-data (orange) and instance data (purple).

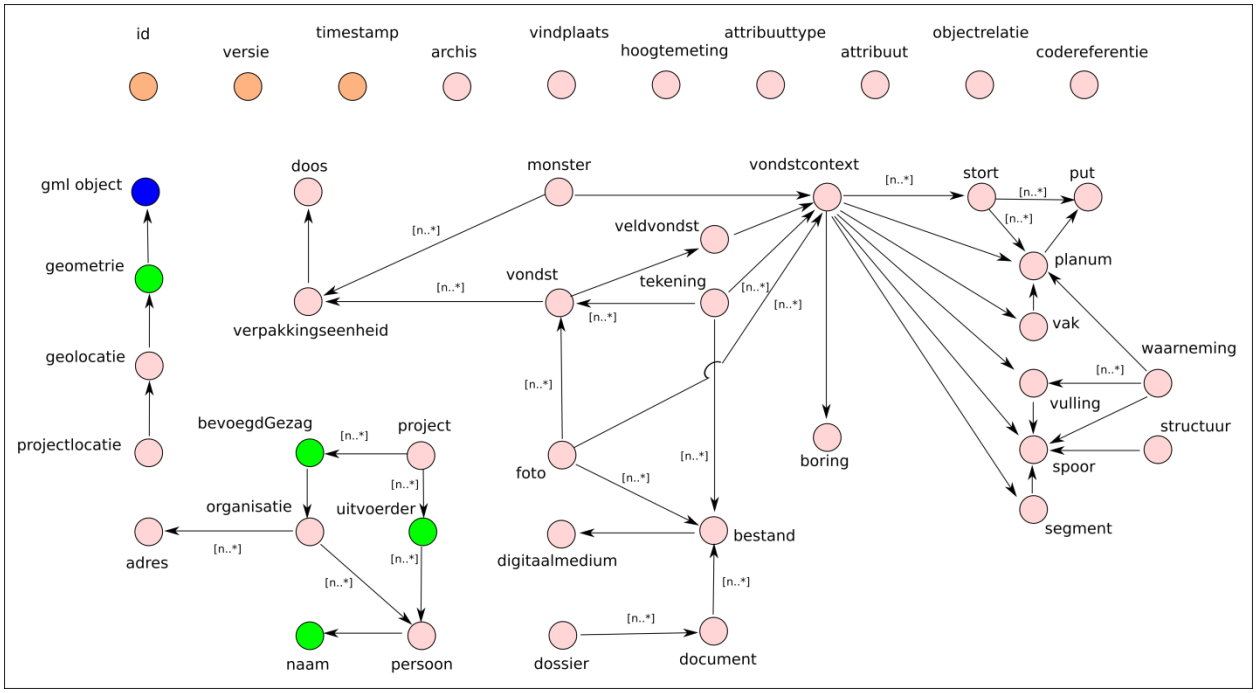

Figure A-2: Overview of the protocol's original data model using the original (Dutch) labels. Vertices have been colour-coded following the level hierarchy as depicted in Figure A-1. Note that connections between vertices of the same level are accomplished by cross referencing identifiers.

A.2 Protocol Conversion

Three high-level and sequential phases can be distinguished where the protocol has been converted. Firstly, the original data model had to be understood completely. A second phase involved a native translation from the tree-based model to a graph-based model. Finally, this graph-based model was restructured to make better use of its relational characteristics.
These steps will now be described in more detail.

Data Model Mapping

A complete understanding of the protocol's original model was achieved by several means. An initial step involved a study of the protocol's specification and the importance of the design choices. To this end, several archaeological researchers were invited to provide explanations and background information. Additionally, it was discussed with various commercial users of the protocol, as well as with one of the people who helped develop it. Finally, these efforts resulted in the understanding that has been partially described in Appendix A.1.

Native Translation

A native translation involves a direct conversion from the protocol's original tree-shaped model to a relational graph-shaped model. Hereto, all relations and concepts within the tree were converted to edges and vertices, respectively. That is, non-terminal leafs were linked to their children, and leafs referring to identifiers were linked to the items with those identifiers. In addition, the naming scheme of the tree's leafs was used to label both the edges and vertices of the graph model. Note that the resulting model would hold a structure similar to the schematic interpretation found in Figure A-2.

Following the naming scheme of the protocol resulted in Dutch labels without semantical references. Therefore, these semantics were added to the model in the form of a separate ontology by transforming these labels to URIs with a *pakbon-ld* base. This decision was made to prevent the alienation of the Dutch archaeological community, who is likely to have been accustomed to the Dutch naming scheme. To maximise the compatibility with other work within ARIADNE, the resources to which these URIs belong have been made subclasses of equivalent CIDOC CRM and CIDOC CRM-EH resources where applicable. Similarly, most properties with a Dutch URI have been made subproperties of the aforementioned ontologies.

To strengthen the translated ontology further a script was developed that automatically added *rdfs:domain* and *rdfs:range* constraints to each predicate based on their connections within the graph model. In addition, each resource and predicate was assigned a useful label that was generated based upon its URI as well as on connected vertices. Finally, descriptions were added to each element by scraping the protocol's specification.

Model Restructuring

While the initial native translation allows us to benefit from LD from a technological perspective, it does not exploit its relational characteristics' full potential. That is, the topology of the graph model is as flat and disconnected as that of the original tree model. In addition, it is evident that several relations and concepts existed purely for reason of convenience. Similarly, all relations are unidirectional with some of them appearing to expect a specific workflow. For this reason, we decided to restructure the model.

An initial step in the restructuring process was the removal of redundant concepts and relations. Instead, a single element remained to which all related elements referred. In addition, where applicable, inverse relations were added to facilitate easy browsing and querying. Finally, both new concepts and relations were added to accommodate the topology which the CIDOC CRM and CIDOC CRM-EH ontologies impose on the semantics of the data model. Examples of added concept types are *events*, *attributions*, and *productions*.

Another modification was the separation between the protocol meta-data and the instance data. Differently put, rather than each instance having a protocol version, each distinct version of the protocol holds zero or more instances made following that version of the protocol. As an example, consider an arbitrary version of the protocol defined as a document (E31) in Figure A-3. Save for its identifier, the protocol holds a version, a timestamp, and is specified to *document* (P70) any number of archaeological reports (EHE0001).

For those interested, figures Figure A-2 through Figure A-10 provide a complete overview on all parts of the new data model. Note that this model may alternatively be viewed in its entirety as a graph (png, 7mb) or as a file (.ttl).

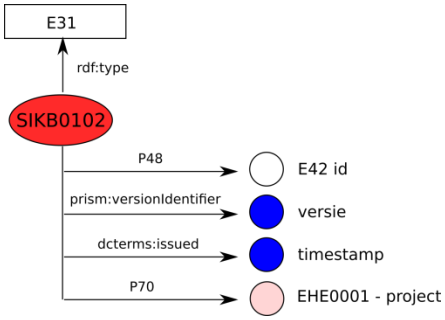

Figure A-3: Part of the protocol's converted data model concerning protocol meta data and project linking.

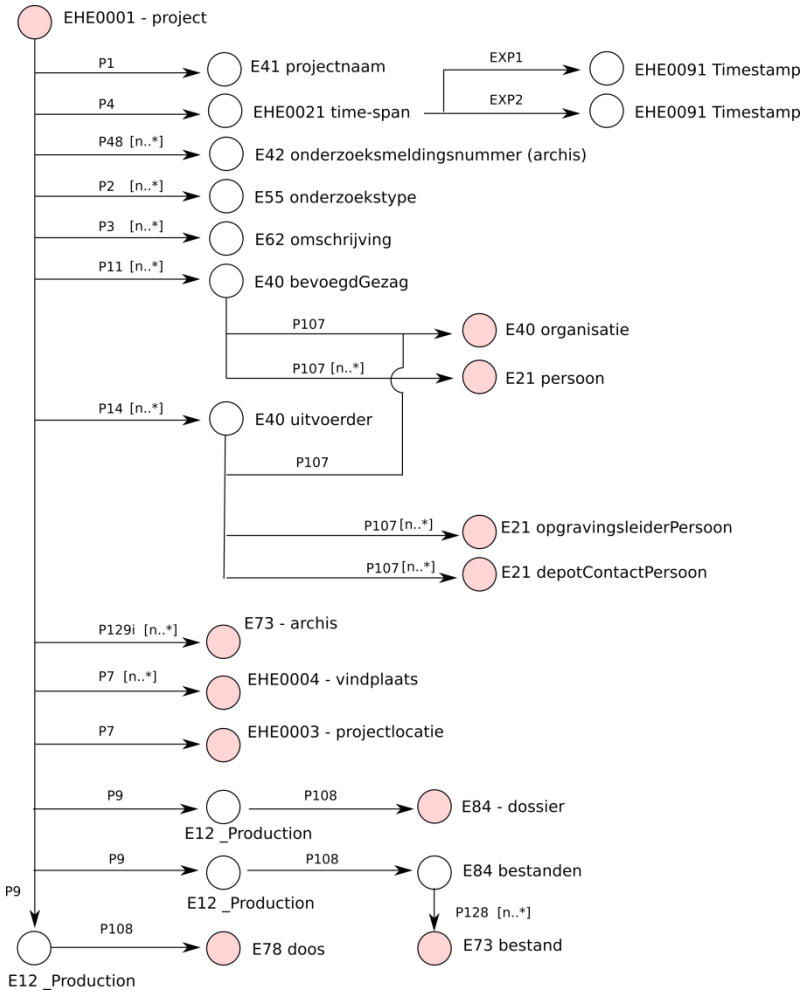

Figure A-4: Part of the protocol's converted data model concerning general information on a certain project.

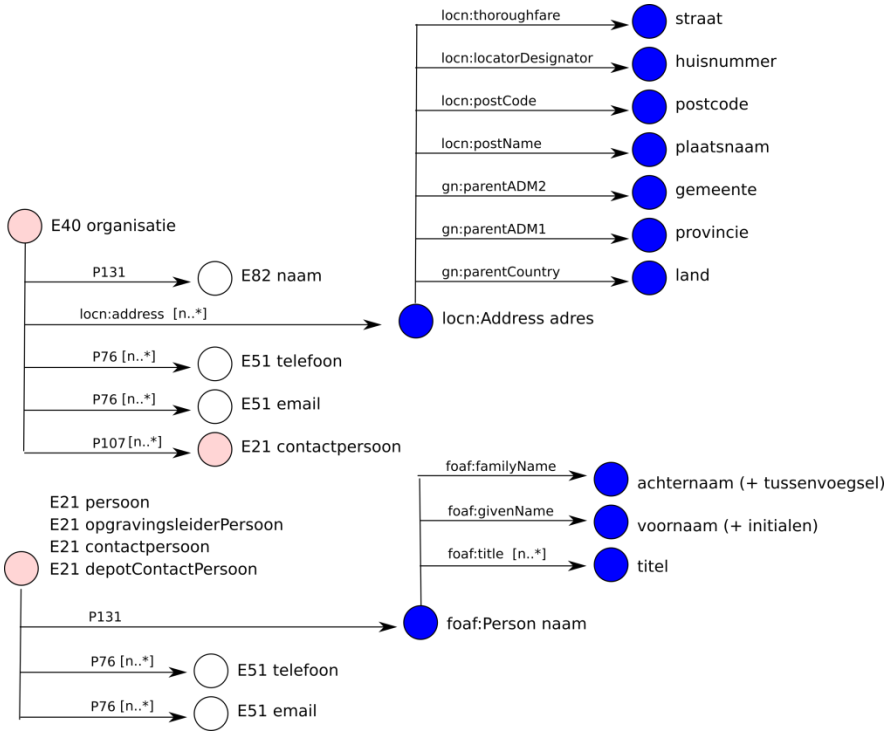

Figure A-5: Part of the protocol's converted data model concerning organizations and people.

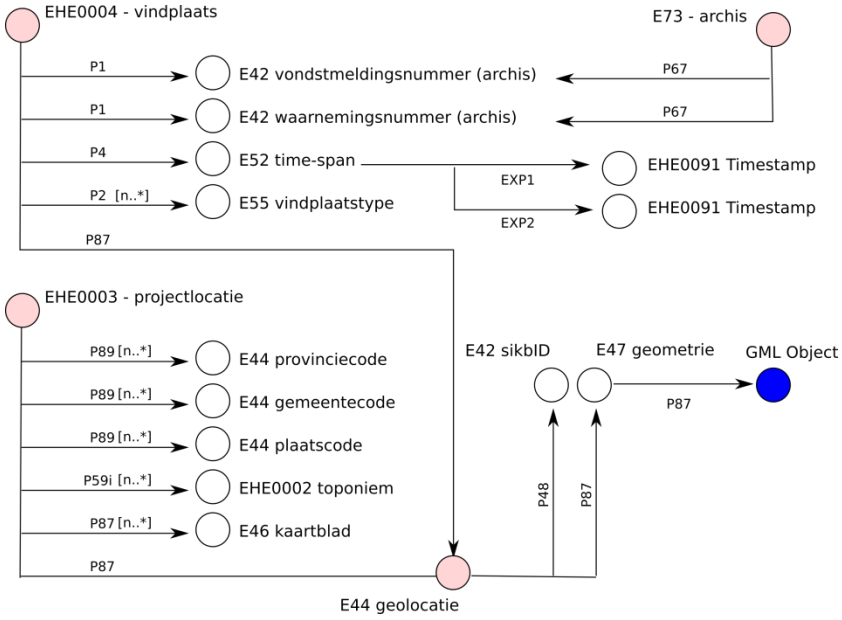

Figure A-6: Part of the protocol's converted data model concerning location information and geometries.

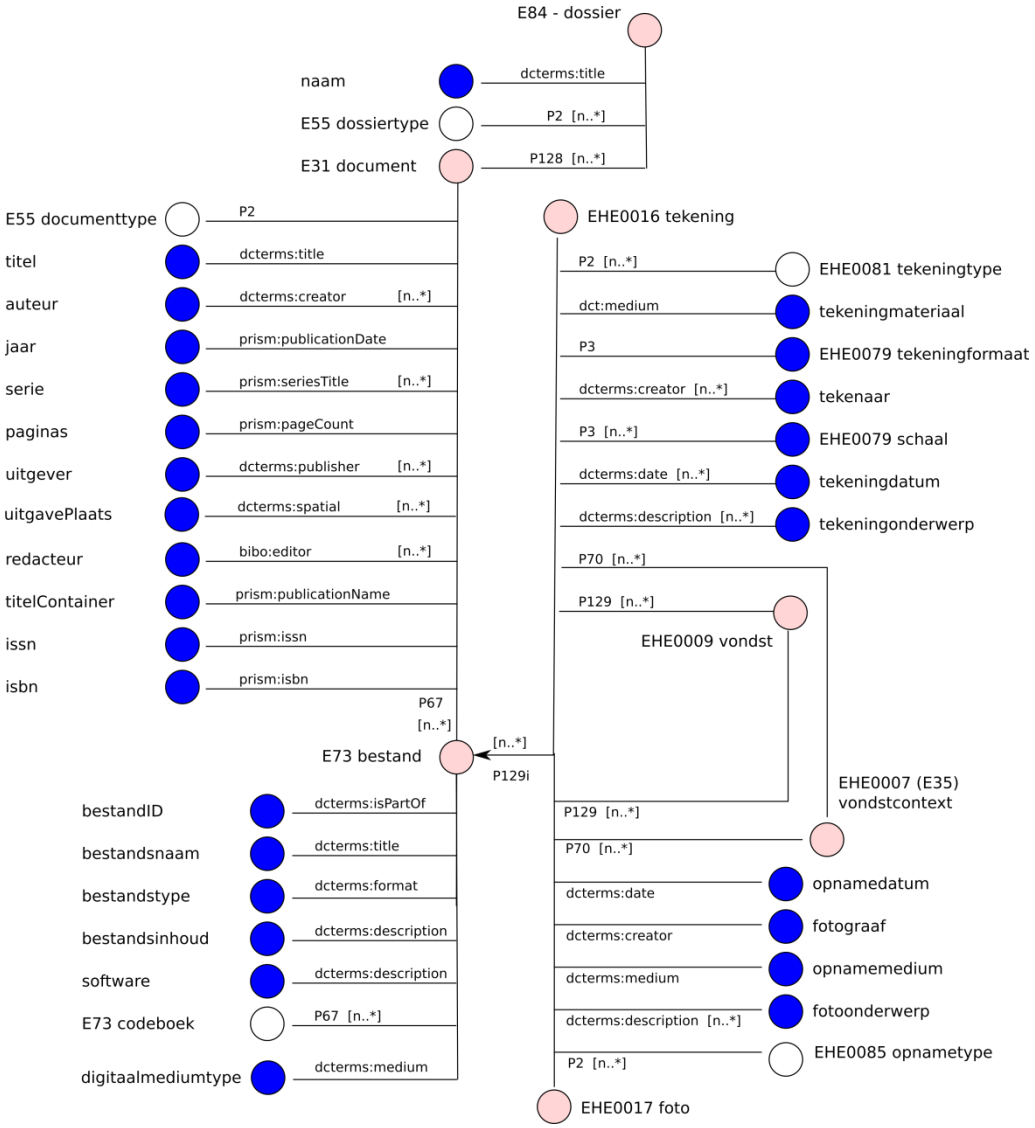

Figure A-7: Part of the protocol's converted data model concerning analogue and digital files and media.

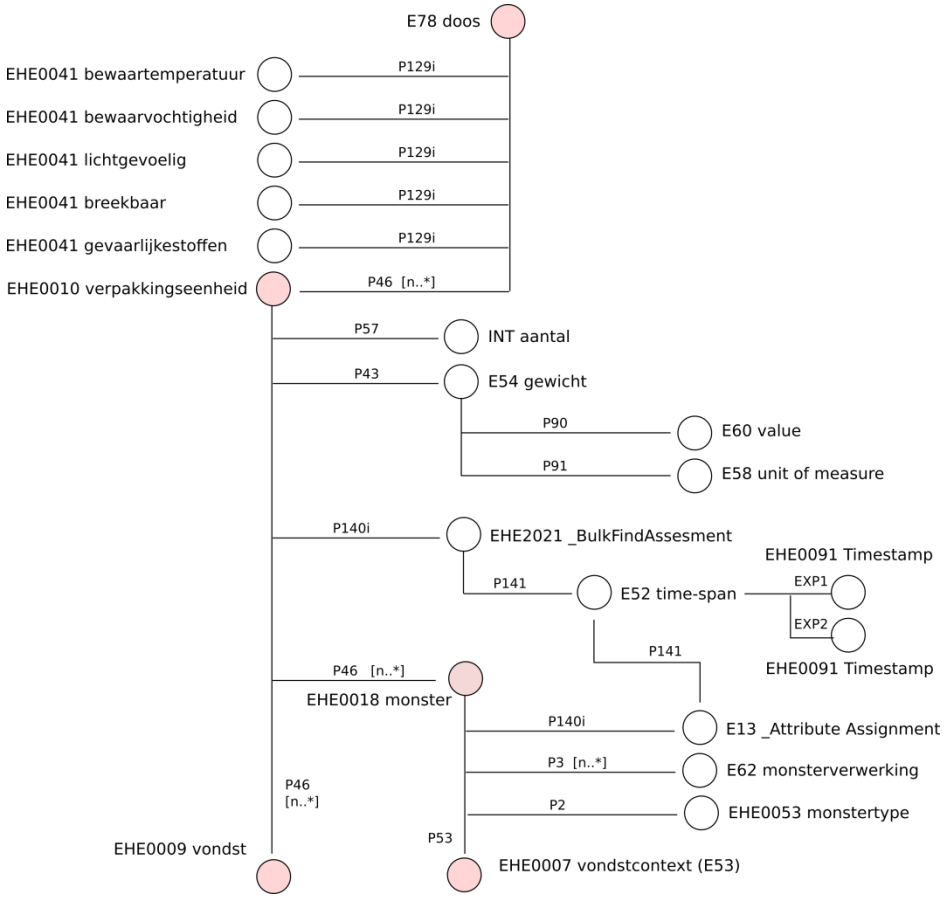

Figure A-8: Part of the protocol's converted data model concerning storage units and samples.

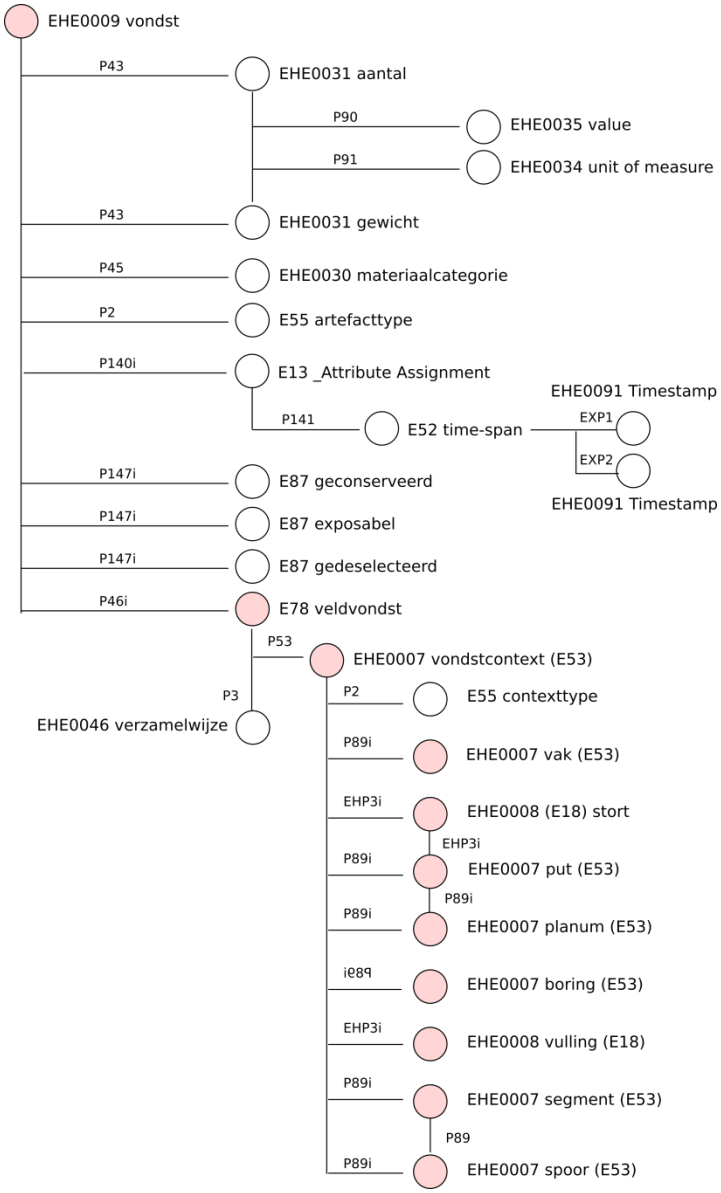

Figure A-9: Part of the protocol's converted data model concerning finds.

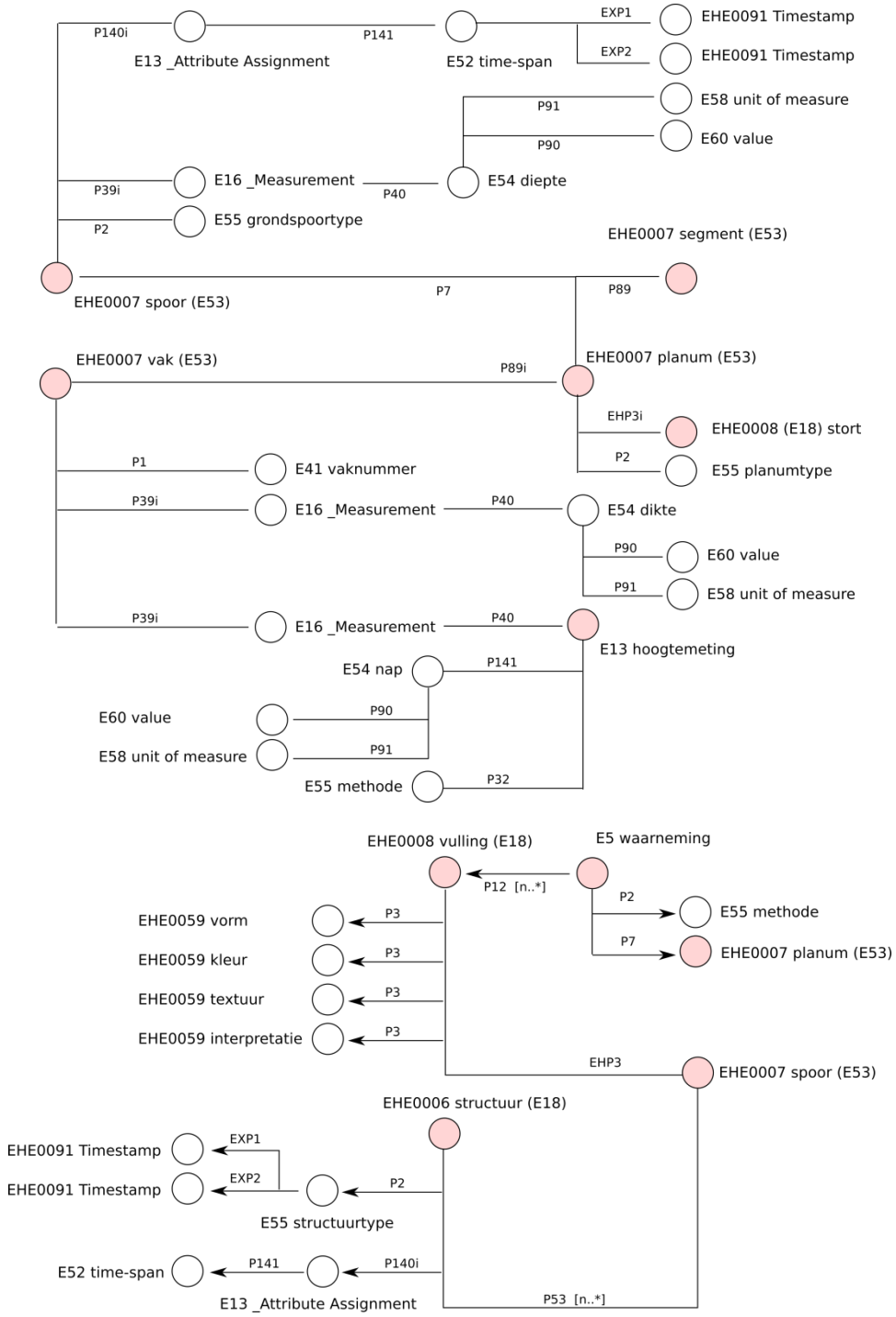

Figure A-10: Part of the protocol's converted data model concerning contexts.

A.2.1 Protocol Enrichment

Several steps were taken to further enrich the graph-based model of the protocol. Most prominent steps involved the reconciliation of two or more entities, and the inclusion of spatial geometries. Both approaches are briefly described below.

Entity Reconciliation

A considerable number of resources within and between protocol instances are a representation of the same entity. Most frequently, these concern locations, companies and institutes, and persons involved, as well as numerous vocabulary items such as period and find classification. All resources that are listed as vocabulary items are linked to a SKOS version of a suitable vocabulary, as is documented in Section A.2.1. For all remaining items, a simple form of entity reconciliation was applied.

Entity reconciliation is achieved by comparing the attributes of any two entities of the same type. Which attributes are compared and how differs by type. For instance, the name of an institute may differ slightly between entities as long as its address matches. This address is composed of multiple elements with varying levels of granularity. To allow for imprecisions, finer-grained levels are given less weight in the final match result. That is, house numbers may differ a bit as long as the county, city, and street fully match. Similarly, finer-grained elements that are missing from one of the two entities do not automatically result in a failed match.

Entities that are found to match are linked together using the *rdfs:sameAs* relation (Figure A-11). This relation strongly implies a bi-directional property, and should thus normally point both ways. However, a large number of duplicate entities were found to occur over a relatively low number of protocol instances. A considerable number of duplicates were (far) less informative than those remaining. Finding an informative entity from a less-informative one would therefore require continuously visiting a randomly-selected linked entity until one was found. To alleviate this inconvenience, linking was only bi-directional between equally-informative entities, or otherwise linking solely from the less-informative to the more-informative entity.

Figure A-11: Two resources are linked by the rdfs:sameAs relation if they 1) are of the same type, and 2) have (nearly) similar values for certain attributes.

Geospatial Extension

Several forms of geospatial referencing exist in the protocol's specification. Most prominent is the general location of the project, as well as the more specific locations of excavation sites. These are specified as geometries following the *GML* standard, and typically involve points or polygons. Other geospatial references concern location appellations and addresses that belong to the companies, institutes, and people involved. In contrast with the geometries, these references are in plain text and regularly incomplete. Therefore, where possible, these references were supplemented by point geometries (Figure A-12).

Geometries were acquired by resolving the appellations or addresses via the open API provided by OpenStreetMap. Once successful, this API returned a 2D-coordinate following the WGS84 standard. Depending on the completeness and accuracy of the input, this point geometry can range from an approximation to a perfect fit. However, to prevent the inclusion of truly erroneous geometries, references were omitted to anything broader than a place.

Found geometries are supplemented parallel to the appellation or address of an entity. This is achieved by pointing towards a geometry instance using the *locn:geometry* relation. The geometry

instance is seen as being independent of the entity referring to it. Hence, more than one entity may refer to the same geometry. In addition, such a geometry instance itself is specified as being both of the type *locn:Geometry* and *wgs84:Point*, and refers to the two coordinates using the corresponding relations from the *wgs84* ontology.

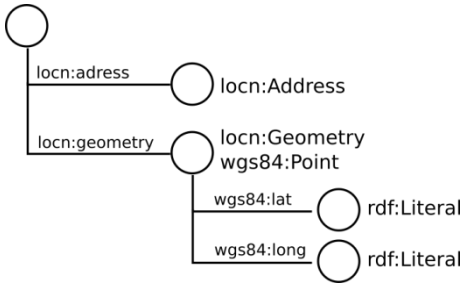

Figure A-12: Resources that link to an address are automatically provided with a geometry consisting of a 2D point coordinate in the WGS84 projection space.

A.3 Thesauri Conversion

A considerable number of entities link to concepts from the Dutch archaeological thesaurus ABR[16]. This thesaurus is maintained by the Dutch government, and is the common standard in Dutch archaeological research. The ABR was converted to allow the full use of this standard and its hierarchical properties. This is discussed briefly below.

The ABR consists of multiple thesauri covering various archaeological topics. These topics range from period classifications to artefact materials, as well as place appellations and map sections (Table A-1). In total, 24 of these topics exist, most of which hold several dozen concepts. Each of these concepts is linked to a short code, as well as to related concepts. These involve more narrowly defined concepts, if any, as well as alternative and deprecated versions of the current concept. Finally, each concept holds a short description and an optional note.

[16] Archaeologisch Basis Register (ABR). See abr.erfgoedthesaurus.nl

Table A-1: A table containing all code lists as an instance of the SKOS:Concept class (left), and the type of their respective entries (right).

SKOS:Concept	RDF:type of members
archistypeCodelijst	E55_type
artefacttypeCodelijst	E55_type
complextypeCodelijst	E55_type
contexttypeCodelijst	E55_type
documenttypeCodelijst	E55_type
gemeenteCodelijst	E44_Place_Appellation
grondspoortypeCodelijst	E55_type
hoogtemetingmethodeCodelijst	E55_type
kaartbladCodelijst	E46_Section_Definition
materiaalcategorieCodelijst	EHE0030_ContextFindMaterial
monstertypeCodelijst	EHE0053_ContextSampleType
objectrelatietypeCodelijst	E55_type
papierformaatCodelijst	EHE0079_RecordDrawingNote
periodeCodelijst	EHE0091_Timestamp
plaatsCodelijst	E44_Place_Appellation
planumtypeCodelijst	E55_type
provincieCodelijst	E44_Place_Appellation
structuurtypeCodelijst	E55_type
tekeningfototypeCodelijst	E55_type
uitvoerderCodelijst	E42_Identifier
verwervingCodelijs	E55_type
verzamelwijzeCodelijst	EHE0046_ContextNote
waardetypeCodelijst	E55_type
waarnemingmethodeCodelijst	E55_type

Each of the ABR's thesauri have been converted following the Simple Knowledge Organization System (SKOS) standard. As a result, the converted ABR consists of 24 entities of the type *SKOS:ConceptScheme*, with each one holding several dozen members of the type *SKOS:Concept*. Entities of either type link to their label and description using the *SKOS:prefLabel* and *SKOS:scopeNote* relations, respectively. Finally, hierarchical properties are implemented using *SKOS:broader* and *SKOS:narrower* relations.

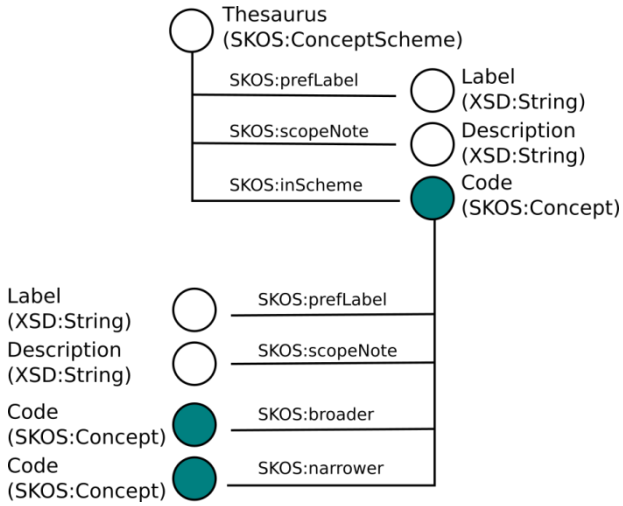

Figure A-13: The structure of a single code list and three of its entries (green vertices). For simplicity, only one entry is shown in detail. Note that non-SKOS predicates have been omitted.

A.4 Data Conversion and Publication

A conversion tool[17] has been developed which allows users to easily convert any number of protocol instances. For this purpose, the tool takes two passes through the provided protocol instance. During the first pass, all entities are converted and assigned a new randomly-generated and unique identifier. This identifier will additionally be used as the latter half of an entity's URI. Once completed, the second pass restructures the partially-converted protocol instance, as well as enrich the existing data with additional resources and relations.

Unique identifiers are created by generating a salted hash value based upon the attributes of an entity. Which attributes are used differs per type. For instance, the hash value generated for documents depends solely on its ISSB in combination with its assigned type. Other attributes, such as its title or authors, are not guaranteed to be unique for a document, and thus may result in clashing identifiers when used. Finally, note that identifiers of entities that belong to exactly one parent entity are preceded by their parent identifier. An example of such an entity is a dating event, as that specific event will never be performed again. A counter-example is the period assigned by said event, as other dating events might conclude with that same period.

During the conversion process, the tool will automatically perform entity reconciliation within and between provided protocol instances, as well as with the online data cloud holding previously-converted protocol instances. However, each converted instance will keep exactly one copy of all non-thesauri entities it describes to ensure that converted instances can serve as solitary and complete units. Note that these instances are given the same URI as their already-existing counterpart, hence preventing duplication in the data cloud.

All currently-converted protocol instances have been made available in an experimental data cloud[18]. Currently, the cloud is hosted via the ClioPatria[19] triple store, which is a free and open-source NO-SQL database build upon the PROLOG logic programming framework. In addition, all data has been made referenceable and queryable through both a web interface and a REST API. Finally, please note that the data within the cloud is purely experimental, and may hold bugs and errors that might not be fixed at present.

[17] Available at github.com/wxwilcke/pakbon-ld
[18] Live at pakbon-ld.spider.d2s.labs.vu.nl
[19] Available at cliopatria.swi-prolog.org

Appendix B Pipeline VUA/LU

A pipeline has been developed to facilitate easy use of the implemented rule mining algorithm by non-experts. The pipeline consists of two high-level components: a backend and a frontend. All processing and computational steps occur within the former, and includes a basic interface for the user. A schematic depiction of this structure is provided in Figure B-1, which includes a bidirectional connection between the backend and the Linked Open Data (LOD) cloud. As postulated in D16.1, it is assumed that at least part of ARIADNE's data will be included within the LOD cloud. Note that, alternatively, a local data set can be used as well.

To make use of the pipeline, users should first provide a set of constraints. These constraints are used to automatically generate a suitable SPARQL query that matches the users' current topic of interest. This query will subsequently be used to retrieve a subset of the LOD cloud, ensuring the data set will reflect the users' interests. Upon successful retrieval of the subset, its data will be made suitable for further processing by running it through a dedicated pre-processing module. The pre-processed data set will then be offered to the Hypothesis Generation module, which generates Semantic Association Rules and their algorithmic measures of relevance. These rules are then presented to the user for evaluation, who may separate narrow the search by adding various filters. Once satisfied, the user may choose to store any or all of the rules.
The remainder of this section will briefly discuss the inner workings of the separate modules.

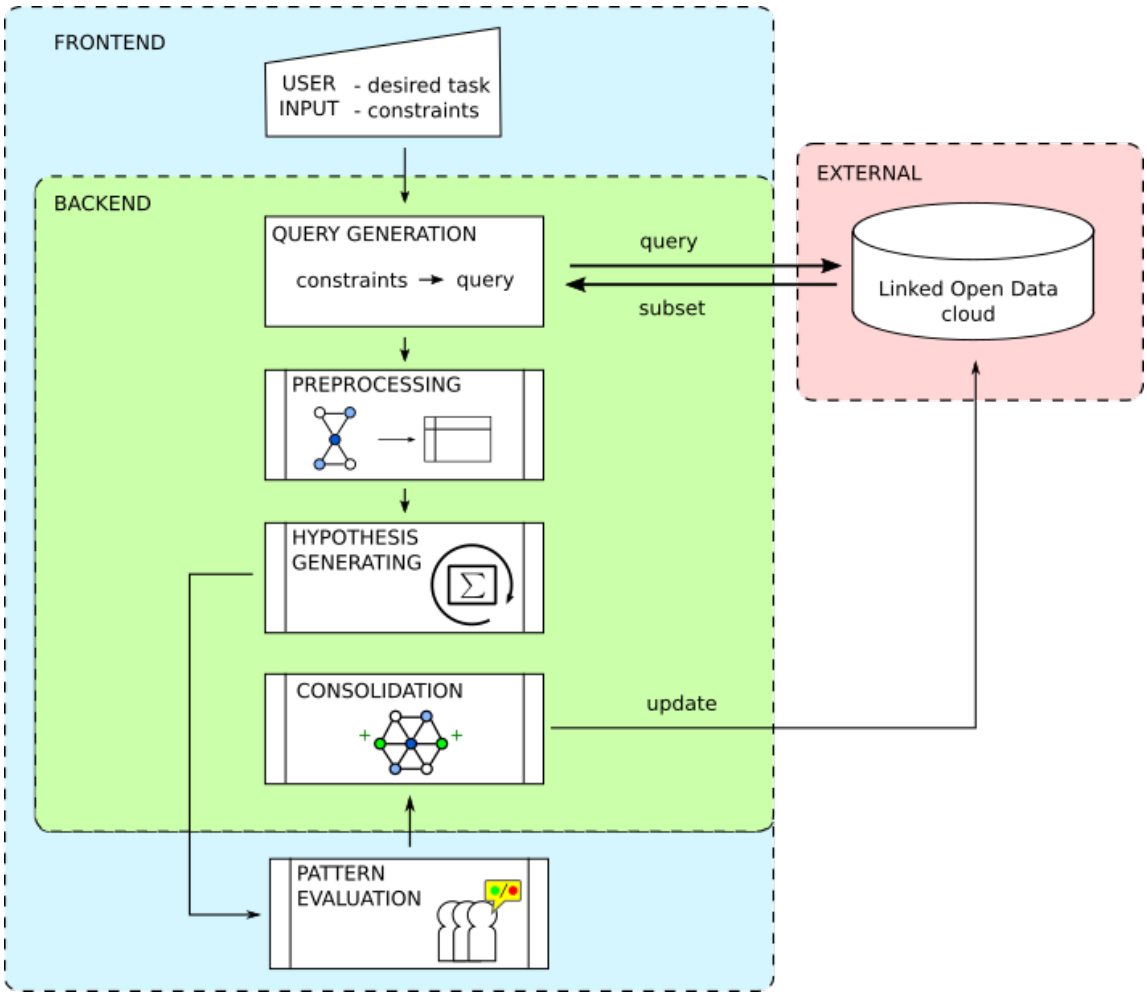

Figure B-1: Flowchart depicting the steps within the Rule Mining pipeline. The backend holds all processing and computational steps, whereas the frontend provides a basic interface to the user. All data is acquired from the LOD cloud, and the backend holds a bidirectional connection.

B.1 Data Preparation

Data preparation involves a two-step sampling technique to extract distinct individuals from a set of triples. Firstly, all entities in the set that match a certain input pattern are extracted. This pattern is intended to capture a user's topic of interest, and is constructed prior to sampling. The resulting entities form the basis for the individuals, and are the starting point of the second sampling step known as context sampling.

Context sampling involves finding a set of facts, related to a specific individual that optimally captures the semantic representation of that individual. Differently put, it concerns discarding any entity and path reachable from a certain individual that are deemed irrelevant to that individual. Hence, this step results in a set of individuals together with relevant information about these individuals.

As an example, consider the small KG as provided in Figure B-2: Four stages of context sampling from a small example Linked Data set.Four stages of context sampling from a small example Linked Data set. There, frame A depicts a single entity that has been selected as an individual following the

first sample step. At present, this individual holds no information apart from its URI. In frame B, the individual is assigned a context of four entities. They are directly connected to the individual and can thus be viewed as analogues to the attribute values in a row of a table. Additional entities have been added in frame C, perhaps because all directly-connected entities were blank nodes and that relevant information lies behind them. Finally, frame D depicts the optional addition of external information with the help of other data sets.

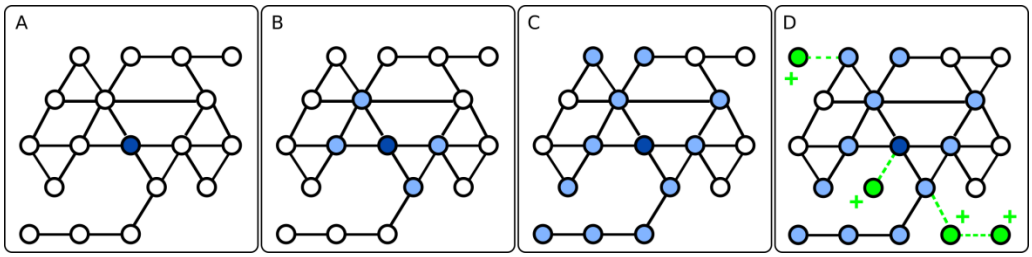

Figure B-2: Four stages of context sampling from a small example Linked Data set. The dark blue node represents the entity for which a context will be sampled. Light blue (local) and green (remote) nodes represent entities included in that context. In contrast, white nodes are excluded from that context.

Currently, the pipeline supports two context sampling strategies; by local neighbourhood and by definition. Both strategies are concisely discussed below.

Local Neighbourhood

Sampling by local neighbourhood makes the assumption that more directly connected entities are more relevant to a concept than those that are less-directly connected. Differently put, the more steps taken away from a certain concept, the more its semantic representation diffuses. To prevent this representation from diffusing too much, a local neighbourhood approach generates an individual's context by sampling depth first up to a certain depth. This creates a shell of a certain thickness around said individual.

A benefit of using a local neighbourhood for sampling is that it typically provided good approximations due to the intuitive correctness of the assumptions it makes. Additionally, its simplicity allows users to easily adjust the workings of the sampling method to fit different ontologies. That is, ontologies that are more verbose may require a deeper sample. Nevertheless, this method has the downside that it is a rather blunt instrument to sample with. For instance, it may return any number of entities. This number is even likely to vary between individuals within the same sample. In addition, it may include one or more irrelevant or duplicate attributes, as the method has no means of evaluating their usefulness.

Context Definition

Sampling by context definition assigns each individual with a context that is similar to or exactly equals this definition. This definition constitutes a set of predicate paths that have their origin at a specific individual. Depending on user-provided constraints, either single or parallel paths may be allowed. For example, when a storage container refers to several of the items it holds via separate *hasPart* relations. Additionally, it may or may not include individuals with a partially-satisfied context.

Using a context definition has the advantage of precise control on how individuals are generated. Hence, a domain expert may insert their knowledge into the process by only including predicate paths that they deem relevant. This additionally prevents the inclusion of irrelevant and duplicate entities. However, a self-defined context does have the disadvantage of requiring additional user input. To ensure correctness, this user should be somewhat knowledgeable with respect to the domain at hand.

B.2 Hypothesis Generator

Important criteria for the hypothesis generator are the interpretability and the believability of the generated hypotheses. To satisfy the former, users should be able to clearly understand the generated hypotheses, instead of needing to decipher a symbolic representation. To satisfy the believability, the returned hypotheses should additionally show the reasoning behind them, rather than merely a conclusion. Only if both criteria hold, will the users consider them. This aspect became apparent following the user-requirement analysis in D16.1.

A natural approach to tackle both the above criteria is to employ Association Rule Mining (ARM). In ARM, the emphasis lies in discovering rules that explain the patterns and regularities in the data[20]. Each of these rules is assigned a confidence score which is based on the frequency of the rule's antecedent, as well as on the number of true and false positives encountered in the data set. A domain expert can subsequently decide whether the provided confidence value is sufficiently high to deem the corresponding rule trustworthy.

As a trivial example, consider a subset of the data that concerns documented postholes with their precise location. By running through all data points, a pattern may emerge between the local geometry of the postholes and the structure they are thought to be the remains of. A generated hypothesis might then state that a series of postholes are likely the remains of a house if and only if they are organised in a certain shape.

Implementation

Traditional ARM operates on tabular data and hence is unsuitable for learning rules from LD. Therefore, this pipeline employs a recent adaptation of ARM for LD, known as SWARM[21]. The workings of SWARM are discussed briefly below.

A core notion in ARM is the *item set*. Items sets are sets of items that belong together following a certain pattern. SWARM extends this notion with *semantic item sets*, which are sets of entities that share a semantic pattern. An example of such a pattern may state that several entities are all made

---

[20] Fürnkranz, J, and T Kliegr. "A brief overview of rule learning." *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, 2015: 54-69.

[21] Barati, M, B Quan, and L Qing. "SWARM: An Approach for Mining Semantic Association Rules from Semantic Web Data." *PRICAI 2016: Trends in Artificial Intelligence: 14th Pacific Rim International Conference on Artificial Intelligence, Phuket, Thailand, August 22-26, 2016, Proceedings*, 2016: 30-43.

using a certain material. Once all such semantic patterns have been mapped, SWARM proceeds by generating *common behaviour sets*. These sets contain two or more joined semantic item sets of which the entities are roughly equal. The rationale behind this is that patterns of similar sets of entities are likely to be common amongst these entities. For instance, all entities made using a certain material were also used for the same purpose.

Multiple patterns shared amongst similar entities form the basis to generate rules that generalise these patterns over other similar entities. To this end, SWARM exploits the data's semantics on the level of RDF and RDFS. More specifically, class and property relations are being used to infer a more general class to which a joined pattern is attributed. For instance, patterns found to hold for a considerable number of finds of all subtypes, are generalised to hold for all finds independent of subtype. These generalised patterns are then rewritten as association rules.

B.3 Pattern Evaluator and Knowledge Consolidator

Upon completion of all prior processes, detected patterns are presented to the user for evaluation (Figure B-3). Hereto, the pipeline provides a simple interactive user interface. This interface presents a candidate solution together with its support and confidence values. The support indicates how common the antecedent is amongst entities of a certain class, whereas the confidence indicates what portion of those entities the consequents hold as well.

CANDIDATE 7 / 12

IF	has type (*RDF:type*)	Context Find (*CRM-EH:EHE0009*)
…AND	consists of (*CRM:P54*)	Material Ceramics (*VOC:MatCER*)
THEN	of subtype (*CRM:P2*)	Earthenware (*VOC:TypeETW*)

Confidence : 0.86
Support : 0.79

- Save Candidate 7? (**y**es / [**n**o] / **a**bort / **v**erbose / add **f**ilter)
> yes
- Candidate 7 has been stored

Figure B-3: Example interaction between the user and the interface.

The number of generated candidate rules is strongly dependent on the provided parameters and the size of the data set. Manually evaluating these rules becomes unfeasible when their number exceeds tens or hundreds of thousands. Therefore, the interface offers users the option to add one or more filters for the purpose of narrowing the search. Examples of such filters have minimal support or confidence, and the in- or exclusion of rules that aim at a specific class of entities. Once satisfied with the selection, the user may choose to store it in one or more of the provided serialisation formats. Depending on this format, users may reuse the selection at a later moment or share it with fellow researchers.

Appendix C Pipeline ATHENA RC

In this Appendix, the details and implementation issues regarding the various text and data mining solutions implemented by ATHENA are described and presented in section 3.1.

1) Algorithm for citation extraction and algorithm data set matching.
2) Author matching procedure.
3) Implementation details (madIS ETL+mining software system).

C.1 Pipeline Components

C.1.1 Citation extraction and data set matching

The main difficulty in extracting citations from a paper arises from the fact that citations to a given paper can be presented in many different formats and the publication's text is frequently unstructured. Citation extraction deals with the extraction of a citation along with its metadata (author names, journals, dates). Title and author matching is also a complicated task as titles have variable lengths, and matching everything within the text is highly complex. In addition, author names and other metadata are input in different ways and in different order.

An algorithm that deals with two major challenges was designed which isn't dependant on the publication's format. This means that the algorithm's precision remains the same regardless of the format or even the language of the document. This reduced the execution time and increased scalability using database techniques.

It was felt that given an efficient matching algorithm, it would be possible to match every possible trigram of a given publication's text with the trigrams in the publication titles. To achieve this, traditional database techniques (executing a JOIN operator between the text trigrams and the title trigrams) were used.

To reduce the computational load and increase the quality of the matching results, the publication's text was first filtered to extract sections with higher than average density of dates and URLs. These sections may contain references, so they were the only parts of the publication's text that were mined. One significant advantage of this approach is that it can also locate references that appear in the body or footnote(s) of a paper and not only in the references section.

We also preprocessed/normalised the metadata to apply the citation/data set matching faster, e.g. by reducing spaces between words, replacing punctuation marks with underscores, converting text in lower case, etc. The preprocessing phase made it possible to achieve higher recall rates, as it helped to overcome misspelling issues and other mistakes. Exactly the same preprocessing procedure was applied to the publication's full text.

After preprocessing, the titles of the data sets were used and one identifying trigram for each title was kept. Identifying trigrams are trigrams that appear in as few titles as possible. These trigrams

were matched to the extracted sections. If a trigram matched, then the full title and the other metadata (authors, dates etc.) were matched (we utilize a trigrams-based inverted index, but the details are omitted, as they are very low level).

For filtering out false matches, weighted metadata (author names, publication date, publisher, journal names) of the citation was used and a bag of words for each publication metadata was produced. This bag contained normalised author names, surnames, publishers, etc., along with a weight value (for example surnames weigh more than first names or publication years). The context of the title match was then searched for words that match. If a word matched, the confidence value was increased by its weight. The context used had a fixed length and consisted of 60 words before and after the title string. The title length was also used to calculate the confidence value. It was obvious that a lengthier title that matches is more likely to be a true match than a shorter one.

The calculation of the weights is an iterative process. We ran the algorithm on a set of publications, manually curated the results and changed the weights until the algorithms worked with high precision when processing publications from various repositories. This way, a threshold for the confidence value that leads to the most accurate results was decided, so when a citation match confidence value falls below this threshold then the corresponding link is filtered out from the results produced.

The drawback of this technique is that the context used for calculating confidence value has a stable length for speed purposes, so it may contain strings from previous or next citations. Nevertheless, manual curation of experimental results indicates that less than 1% of matches are false due to this.

C.1.2 Author Matching Procedure

The steps of the author matching procedure are as follows:

1. Create a list of pairs of authors for each publication in OpenAIRE.
2. Create a similar author list for each ARIADNE report. For documents with a single author,
   add the author to the list only if the *FullName* is longer than 10 characters.
3. Homogenise all the names in the above lists in the form [Surname, F], where F is the first
   letter of the *Firstname* author.
4. Return all links between an ARIADNE and an OpenAIRE documents where:
   a. either both authors are matched (if the ARIADNE document has at least 2 authors), or
   b. at least one author matches (if the ARIADNE document has only one author).

Regarding step two above: for each ADS grey literature report there is a single attribute (AUTHS) containing all of the documents' authors in one string, for example: "Sherlock, H, Pikes, P J and Newby-Vincent, J", so they are first separated at the commas and between the "and" strings.

C.1.3 Clustering and Similarity Algorithms

All the algorithms were implemented on top of madIS[22], a powerful extension of a relational DBMS with user-defined data processing functionality. MadIS is built on top of the SQLite API with several Python extensions. SQLite API allows the use of different DBMSes that may support it. So, madIS may work with SQLite or Oracle BerkeleyDB or any other DBMS which support this API. This API provides a powerful streaming interface, which is exploited in madIS via Python generators; a powerful language pattern that allows co-routines via a yield statement.

A Python program can be written as if it is in control of iteration (e.g., iterate over an external data source), yet yield values on demand, with control transfer to the DBMS engine for each produced value. In this way, madIS can process large amounts of heterogeneous, external data in a streaming way. Moreover, madIS uses Python UDFs in the same way as its native SQL functions. Both Python and the DBMS are executed in the same process, greatly reducing the communication cost between them. This is a major architectural element and has a positive impact on joint performance.

MadIS is highly scalable, easily handling 10s of Gigabytes of data on a single machine. This benefit transparently carries over to distributed systems (e.g., Hadoop[23], Exareme[24]) which can use madIS in each node. The main goal of madIS is to promote the handling of data-related tasks within an extended relational model. In doing so, it upgrades the database from having a support role (storing and retrieving data) to being a full data processing system on its own.

In madIS, queries are expressed in madQL: a SQL-based declarative language extended with additional syntax and user-defined functions (UDFs). One of the goals of madIS is to eliminate the effort of creating and using UDFs by making them first-class citizens in the query language itself. To allow easy UDF editing with a text editor, madIS loads the UDF source code from the file system. Whenever a UDF's source code changes, madIS automatically reloads the UDF definition. This allows rapid UDF development iterations, which are common in data exploration and experimentation.

The expressiveness and the performance of madIS along with its scalability features were compelling reasons for choosing it to implement our algorithm.

For example, the query shown in Figure C-1 is the final query (after calculating all the weights and thresholds) used to produce the citation links described earlier. The query extracts the references section, preprocesses the documents, matches them with the trigrams table, and calculates the confidence value:

[22] L. Stamatogiannakis, M. L. Triantafyllidi, Y. Foufoulas, M. Vayanou, and M. Kyriakidi. madIS - Extensible relational db based on sqlite. https://github.com/madgik/madis, (accessed January 10, 2017).
[23] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, pages 1–10. IEEE, 2010.
[24] Y. Chronis, Y. Foufoulas, V. Nikolopoulos, A. Papadopoulos, L. Stamatogiannakis, C. Svingos, and Y. Ioannidis. A relational approach to complex dataflows. MEDAL 2016.

```
select docID, citID from (
  select docID, citID,
  (length(title) + conf*10)/(length(context)*1.0) as nconfidence
  from (
    select docID, citID, regexpcountuniquematches(bag,context)
    + 2*regexprcountuniquematches(authorsurname,context)
    + 0.5*regexprcountuniquematches(authorname,context)
    + regexprcountuniquematches(publisher,context) as confidence,
    regexprmatches(title,lower(context)) as full_title_match,
    title, context
    from(
     select docid, lower(stripchars(middle,'_')) as doc_trigram,
     prev||' '||middle||' '||next as context
     from(
      select id as docid,
      textwindow2s(normalizetext(textreferences(text)), 30, 3, 30)
      from docs
     )
    ), trigrams
  where doc_trigram = trigram and full_title_match
  )
  where nconfidence > 0.28
)
```

Figure C-1: : Final madIS query producing citation links.

The use of the UDFs that appear in the query is shown in Table C-1. Note that due to the expressiveness of madIS, it is very easy to experiment with the algorithm either removing or adding new operators (i.e. by removing the *textreferences* operator we can simply make the algorithm run on the full publication text) or simply changing the weights while running the algorithm iteratively in order to decide the final weights and thresholds.

Table C-1: : UDFs used in citation mining algorithm

madIS UDF's	*Description*
normalizetext(text)	Normalises text (implements preprocessing steps.)
textwindow2s (text, prev, middle, next)	Returned schema : prev, middle, next Returns a rolling window over the text. In our algorithm, the window includes 30 words before the trigram (3 words) and 30 words after.
textreferences(text)	Implements the reference extraction phase of the algorithm. It returns the references section of the input text.
normregexprmatches(pattern,arg)	This function returns true if the pattern matches arg or false otherwise.

nregexpcountuniquematches (pattern, expression)	Returns the number of unique matches of pattern in expression.
stripchars(str[,stripchars])	Returns *str* removing leading and trailing whitespace characters or *stripchars* characters if given.

Appendix D Expert Evaluation

This section holds the raw evaluations reports are provided by the domain experts. Note that we have translated both hypotheses and remarks from Dutch to English.

D.1 OPTIMA Case Evaluations

Person 1

Hypothesis	Plausible [Y/N]	Valuable [Y/N]	Relevance [0-3]	Remarks
For every find holds that, if they are of material type 'torc', then they originate within the reports on Suffolk.	N	N	0	-
For every find holds that, if they are of type 'sherds', then they consist of 'pottery'.	Y	N	0	-
For every find holds that, if they are of type 'ironstone', then they consist of 'pottery'.	Y	N	0	-
For every find holds that, if they are of type 'datable artefacts', then they originate within the reports on Suffolk.	N	N	0	-
For every find holds that, if they are of type 'whitewashed bricks', then they consist of a 'building'.	Y	N	0	-
For every deposition event holds that, if it involved moving a 'bowl', then it originates from reports on Linconshire.	N	N	0	-
For every deposition event holds that, if it involved moving 'iron', then it moved that find to an 'internal well'.	N	N	0	-
For every deposition event holds that, if it involved moving 'pottery', then it moved that find to a 'rubbish pit'.	N	N	0	-
For every deposition event holds that, if it involved moving 'clay, then it moved that find to a 'grey deposit'.	N	N	0	-

For every deposition event holds that, if it involved moving a 'wall', then it originates from reports on Headland.	N	N	0	-
For every context event holds that, if it originates from reports on Thames, then it dates from the 'iron age'.	Y	N	2	-
For every context event holds that, if it witnessed context 'structure X', then it dates from the '17th century'. Here, X represents a specific resource.	N	N	0	-
For every context event holds that, if it witnessed context 'enclosure X', then it dates from the 'iron age'. Here, X represents a specific resource.	N	N	0	-
For every context event holds that, if it witnessed context 'town X', then it dates from the 'medieval'. Here, X represents a specific resource.	Y	N	0	-
For every context event holds that, if it witnessed context 'street X', then it dates from the '17th century'. Here, X represents a specific resource.	N	N	0	-
For every context production event holds that, if it has produced 'brooches', then it dates from the 'Roman Period'.	N	N	0	-
For every context production event holds that, if it has produced 'tempered sherds', then it dates from the '14th to 15th century'.	N	N	0	-
For every context production event holds that, if it has produced 'bricks', then it dates from 'modern' times.	Y	N	0	-
For every context production event holds that, if it has produced 'pottery', then it dates from 'late medieval' times.	N	N	0	-

For every context production event holds that, if it has produced 'roman pottery', then it dates from 'Roman age'.	Y	N	0	-

Person 2

Hypothesis	Plausible [Y/N]	Valuable [Y/N]	Relevance [0-3]	Remarks
For every find holds that, if they are of type 'torc', then they originate within the reports on Suffolk.	n	n	0	-
For every find holds that, if they are of type 'sherds', then they consist of 'pottery'.	y	n	1	-
For every find holds that, if they are of type 'ironstone', then they consist of 'pottery'.	n	n	0	-
For every find holds that, if they are of type 'datable artefacts', then they originate within the reports on Suffolk.	y	n	0	-
For every find holds that, if they are of type 'whitewashed bricks', then they consist of a 'building'.	y	n	0	-
For every deposition event holds that, if it involved moving a 'bowl', then it originates from reports on Linconshire.	y	n	0	-
For every deposition event holds that, if it involved moving 'iron', then it moved that find to an 'internal well'.	n	n	0	-
For every deposition event holds that, if it involved moving 'pottery', then it moved that find to a 'rubbish pit'.	n	n	0	-
For every deposition event holds that, if it involved moving 'clay, then it moved that find to a 'grey deposit'.	n	n	0	-
For every deposition event holds that, if it involved moving a 'wall', then it originates from reports on Headland.	y	n	0	-
For every context event holds that, if it originates	y	y	1	-

from reports on Thames, then it dates from the 'iron age'.				
For every context event holds that, if it witnessed context 'structure X', then it dates from the '17th century'. Here, X represents a specific resource.	y	n	0	-
For every context event holds that, if it witnessed context 'enclosure X', then it dates from the 'iron age'. Here, X represents a specific resource.	y	n	0	-
For every context event holds that, if it witnessed context 'town X', then it dates from the 'medieval'. Here, X represents a specific resource.	y	n	0	-
For every context event holds that, if it witnessed context 'street X', then it dates from the '17th century'. Here, X represents a specific resource.	y	n	0	-
For every context production event holds that, if it has produced 'brooches', then it dates from the 'Roman Period'.	y	y	1	-
For every context production event holds that, if it has produced 'tempered sherds', then it dates from the '14th to 15th century'.	n	n	0	All sherds are tempered.
For every context production event holds that, if it has produced 'bricks', then it dates from 'modern' times.	y	n	0	-
For every context production event holds that, if it has produced 'pottery', then it dates from 'late medieval' times.	y	n	0	-
For every context production event holds that, if it has produced 'roman pottery', then it dates from 'Roman age'.	n	n	0	Very trivial

D.2 SIKB Protocol Case Evaluations

Person 1

Hypothesis	Plausible [Y/N]	Valuable [Y/N]	Relevance [0-3]	Remarks
For every context stuff holds that, if they have a bright yellow colour, then they are of specific type ANIMAL GRAVE.	y	not	1	trivial and not very useful
For every site holds that, if they are of specific type BASE CAMP, then they are dated from the NEOV to NTL period.	y	not	1	This is true but a little rough and trivial. Not very useful
For every site holds that, if they are of specific type GRAVEYARD, then they are dated from the ROM to NT period.	y	not	1	Possibly interesting as the data set apparently only holds graveyards from those periods. However, the range (+-2000 years) is rather big.
For every site holds that, if they are of specific type FLINT CARVING, then they are dated from the MESO to NEO period.	y	y	2	Very trivial, but interesting that it can find such patterns.
For every site holds that, if they are of specific type CREMATION, then they are dated from the MIDDLE BRONZE to IRON period.	y	y	2	This is very interesting, although trivial. I would expect also cremation graves in the Roman period. Apparently, the data sample does not contain such contexts.
For every find holds that, if they are of specific type CHARCOAL, then they are of material type CHARCOAL.	y	not	1	Trivial, not very relevant.
For every find holds that, if they are of specific type TUFF, then they are dated from EARLY PALEO to LATE NT.	y	not	1	Period too broad. Not really knowledge.
For every find holds that, if they are of specific type RAW EARTHENWARE (Nimeguen), then they are dated from EARLY ROMAN to LATE ROMAN.	y	not	1	Trivial. Finds are dated according to their artefact type.
For every find holds that, if they are dated from MIDDLE NEO to LATE NEO, then they are of specific type HAMMER AXE and of material type DOLERITE.	y	y	2	This is interesting. This would mean that from NEOM-NEOL we only "found hammer axes". I would expect that we would find also other finds from this period(e.g. pottery or worked bones)
For every find holds that, if they are of specific type ADZE, then they are dated from NEO to	y	y	2	This is true. This is again archaeological knowledge, very nice that the algorithm got this.

BRONZE.				
For every context holds that, if they are of specific type DWELLING, then they are dated from ROMAN to MIDDLE DARK AGES, and have structure FLOOR PLAN.	y	y	2	This means that "dwellings" are only used from the Roman period to the Medieval period. This is archaeologically correct. This is typical specialized archaeological knowledge. Interesting that you found these relationships!
For every context holds that, if they are of specific type HOUSE DITCH, then they are dated from EARLY IRON AGE, and have structure HOUSE.	y	y	2	This is also true, "house ditches" are indeed known in the iron age for the Netherlands. Interesting that the algorithm get this info. I however think that such ditches would also be found in later periods.
For every context holds that, if they are of specific type WELL, then they are dated from BRONS to IRON AGE, and have structure WELL.	y	y	2	I would expect wells also to be present in later periods. Interesting knowledge, it could however mean that the sample is too small. But I am impressed.
For every context holds that, if they are of specific type INHUMATION, then they are dated from PREH, and have structure GRAVE.	y	y	2	Interesting, I would expect inhumation to also have happened in later periods. This could mean that the archaeologists have defined these differently. Rules like this could indicate differences over time in funerary practices.
For every context holds that, if they are 21 cm deep, then they are dated from PREH, and have specific type DITCH.	n	not	0	This one is a little too vague... I don't get this one completely, but is seems a bit far-fetched.
For every projects holds that, if they are of specific type 'destructive excavation', then they have location type 'settlement with defenses'.	y	a little	1	Interesting! This could indicate that settlements are archaeologically more findable/ visible than for example graves. However, the sample might be too small to say something like this.
For every project holds that, if they are of specific type 'destructive mapping', then they are of location type 'settlement with defenses'.	y	a little	1	same as previous
For every project holds that, if they took place at location X, then they also too place at location Y. Here, X and Y are variables and Y lies within X.	y	not	0	not very useful
For every project holds that, if they are of specific type X, then is consists of (documentation) Y. Here, X and Y are variables.	y	not	0	same
For every project holds	y	not	0	same

that, if they are documented in document X, then they consist of container Y. Here, X and Y are variables.				

Person 2

Hypothesis	Plausible [Y/N]	Valuable [Y/N]	Relevance [0-3]	Remarks
For every context stuff holds that, if they have a bright yellow colour, then they are of specific type ANIMAL GRAVE.	No	-	0	nonsense
For every site holds that, if they are of specific type BASE CAMP, then they are dated from the NEOV to NTL period.	Yes	Yes	2	within the available resources
For every site holds that, if they are of specific type GRAVEYARD, then they are dated from the ROM to NT period.	Yes	Yes	2	within the available resources
For every site holds that, if they are of specific type FLINT CARVING, then they are dated from the MESO to NEO period.	Yes	Yes	2	within the available resources
For every site holds that, if they are of specific type CREMATION, then they are dated from the MIDDLE BRONS to IRON period.	Yes	Yes	2	within the available resources
For every find holds that, if they are of specific type CHARCOAL, then they are of material type CHARCOAL.	Yes	No	1	
For every find holds that, if they are of specific type TUFF, then they are dated from EARLY PALEO to LATE NT.	Yes	No	1	trivial date, any ages…
For every find holds that, if they are of specific type RAW EARTHENWARE (Nimeguen), then they are dated from EARLY ROMAN to LATE ROMAN.	Yes	No	2	by default, thesaurus
For every find holds that, if they are dated from MIDDLE NEO to LATE NEO, then they are of specific type HAMMER AXE and of material type DOLERITE.	Yes	No	2	by default, thesaurus

For every find holds that, if they are of specific type ADZE, then they are dated from NEO to BRONS.	Yes	Yes	3	date by default = MESOL-NEOL
For every context holds that, if they are of specific type DWELLING, then they are dated from ROMAN to MIDDLE DARK AGES, and have structure FLOOR PLAN.	Yes	No	2	by default, thesaurus
For every context holds that, if they are of specific type HOUSE DITCH, then they are dated from EARLY IRON AGE, and have structure HOUSE.	Yes	No	2	within the available resources
For every context holds that, if they are of specific type WELL, then they are dated from BRONS to IRON AGE, and have structure WELL.	Yes	No	2	within the available resources
For every context holds that, if they are of specific type INHUMATION, then they are dated from PREH, and have structure GRAVE.	Yes	No	2	
For every context holds that, if they are 21 cm deep, then they are dated from PREH, and have specific type DITCH.	No	No	0	nonsense
For every projects holds that, if they are of specific type 'destructive excavation', then they have location type 'settlement with defenses' .	Yes	No	1	within the available resources, too general
For every project holds that, if they are of specific type 'destructive mapping', then they are of location type 'settlement with defenses'.	Yes	No	1	too general
For every project holds that, if they took place at location X, then they also too place at location Y. Here, X and Y are variables and Y lies within X.	Yes	-	0	trivial
For every project holds that, if they are of specific type X, then is	Yes	-	0	trivial

consists of (documentation) Y. Here, X and Y are variables.				
For every project holds that, if they are documented in document X, then they consist of container Y. Here, X and Y are variables.	Yes	-	0	trivial

Person 3

Hypothesis	Plausible [Y/N]	Valuable [Y/N]	Relevance [0-3]	Remarks
For every context stuff holds that, if they have a bright yellow colour, then they are of specific type ANIMAL GRAVE.	N	N	0	This would be very unlikely to be held as a general hypothesis
For every site holds that, if they are of specific type BASE CAMP, then they are dated from the NEOV to NTL period.	Y	N	3	The dating range is far too wide to hold any value
For every site holds that, if they are of specific type GRAVEYARD, then they are dated from the ROM to NT period.	N	N	2	What is a 'graveyard'? A cemetery or a burial site in general?
For every site holds that, if they are of specific type FLINT CARVING, then they are dated from the MESO to NEO period.	N	N	3	Paleolithic would be missing here
For every site holds that, if they are of specific type CREMATION, then they are dated from the MIDDLE BRONZE to IRON period.	N	N	2	Medieval and Roman cremations would be missing here
For every find holds that, if they are of specific type CHARCOAL, then they are of material type CHARCOAL.	Y	N	0	This is a tautology
For every find holds that, if they are of specific type TUFF, then they are dated from EARLY PALEO to LATE NT.	Y	N	3	The dating range is far too wide to hold any value
For every find holds that, if they are of specific type RAW EARTHENWARE (Nimeguen), then they are dated from EARLY ROMAN to LATE ROMAN.	Y	Y	3	This seems likely, although I am no specialist on this particular topic
For every find holds that, if they are dated from MIDDLE NEO to LATE	N	N	3	This is obviously not true

NEO, then they are of specific type HAMMER AXE and of material type DOLERITE.				
For every find holds that, if they are of specific type ADZE, then they are dated from NEO to BRONS.	Y	Y	3	This seems likely, although I am no specialist on this particular topic
For every context holds that, if they are of specific type DWELLING, then they are dated from ROMAN to MIDDLE DARK AGES, and have structure FLOOR PLAN.	N	N	0	This is obviously not true
For every context holds that, if they are of specific type HOUSE DITCH, then they are dated from EARLY IRON AGE, and have structure HOUSE.	N	Y	3	This does not account for house ditches after the Iron Age (of which there obviously are)
For every context holds that, if they are of specific type WELL, then they are dated from BRONS to IRON AGE, and have structure WELL.	N	N	2	This is a tautology in part
For every context holds that, if they are of specific type INHUMATION, then they are dated from PREH, and have structure GRAVE.	N	N	0	Burials are too generic to make statements of value on
For every context holds that, if they are 21 cm deep, then they are dated from PREH, and have specific type DITCH.	N	N	0	This is obviously not true
For every projects holds that, if they are of specific type 'destructive excavation', then they have location type 'settlement with defenses'.	N	N	0	This is obviously not true
For every project holds that, if they are of specific type 'destructive mapping', then they are of location type 'settlement with defenses'.	N	N	0	This is obviously not true
For every project holds that, if they took place at location X, then they also too place at location Y. Here, X and Y are variables and Y lies	N	N	0	This holds no value

within X.				
For every project holds that, if they are of specific type X, then is consists of (documentation) Y. Here, X and Y are variables.	N	N	0	This isn't stating anything, as far as I can see
For every project holds that, if they are documented in document X, then they consist of container Y. Here, X and Y are variables.	N	N	0	This isn't stating anything, as far as I can see