journal of computer applications in archaeology

Automatic Normalization of Temporal Expressions

RESEARCH ARTICLE

CERI BINDING
DOUGLAS TUDHOPE
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Dates, periods and timespans are described in archaeological datasets using a number of different textual patterns for which myriad variations exist, rendering direct automated comparison difficult. The issue can occur even within records from the same dataset and is further compounded when attempting to integrate multilingual data – particularly where dates may be expressed in words rather than numbers. The same problem can be found in temporal metadata, whether manually entered or generated via Natural Language Processing (NLP) techniques from reports and grey literature. Resolving and normalizing dates and periods to internationally agreed standard formats enables efficient data integration, interchange, search, comparison and visualization. This paper reports on the design and implementation of a tool to normalize temporal expressions to a numerical time axis and reflects on key issues.

Textual patterns for seven categories of temporal expression have been normalized: Ordinal named or numbered centuries; Year spans; Single year (with tolerance); Decades; Century spans; Single year with prefix; Named periods. The following languages are currently supported: Dutch, English, French, German, Italian, Norwegian, Spanish, Swedish, Welsh. Methods are described together with an (open source) normalization tool developed in Python and four applications of the method are discussed, together with limitations and future work. Results are presented from diverse data sets and languages. The input is a temporal text string and a language code (ISO639-1). The output is a tab delimited text file with start/end years (in ISO 8601 format), relative to Common Era (CE). The normalized outputs are provided as additional attributes along with the original text expression for consuming software to employ in end-user applications.

CORRESPONDING AUTHOR:
Douglas Tudhope
Hypermedia Research Group, University of South Wales, GB
douglas.tudhope@southwales.ac.uk

KEYWORDS:
temporal expressions; dating; time periods; semantic integration; software; multilingual

TO CITE THIS ARTICLE:
Binding, C and Tudhope, D. 2023. Automatic Normalization of Temporal Expressions. *Journal of Computer Applications in Archaeology*, 6(1): 24–39. DOI: https://doi.org/10.5334/jcaa.105

1 INTRODUCTION

Archaeology and related heritage disciplines have a particular preoccupation with time since dating and placing into temporal sequences are key outcomes of many investigations. However, the systematic digital processing of dates and temporal expressions poses significant practical challenges due to the variety of expressions encountered (in different languages) and also due to the wide variety of dating methods employed with differing characteristics (for an overview of methods, see for example, Renfrew & Bahn 2020). Sometimes only approximate dates are appropriate – a wide variety of terms indicating different kinds of approximation can be found. There is a distinction between absolute and relative dating methods. Absolute dating methods (independently measurable) include radiometric measurement – e.g. C14 radiocarbon dating, dendrochronology, thermoluminescence, numismatics. Relative dating methods (by reference to other known dates) include seriation, typology, stratigraphic sequencing, pollen analysis. Methods from both categories are often used in conjunction; absolute dating may determine the age of certain objects or stratigraphic units and relative dating can then tie associated artefacts and layers into the sequence provided by absolute dates. Some dating methods carry inherent probabilities or may be only applicable for certain periods, such as radiocarbon absolute dating.

Various calendar systems (with fixed reference points) can be used for absolute dating. The output of archaeological dating is often specified as a range of dates rather than a single absolute date. In some cases, it may be appropriate to refer to periods and at other times numeric date ranges, of which there are a plethora of expressions in any language. The issue is compounded by the question of precision; sometimes a very precise date can be given (perhaps by historical methods), while in other cases the period might stretch over millennia. Categorisation itself is rendered difficult by the overlapping and palimpsest character of many archaeological remains (Bailey 2007).

This paper addresses a recurring issue encountered when attempting to integrate archaeological and cultural heritage datasets, where a number of different textual patterns and conventions have been used to represent dates, periods and timespans in data fields. The myriad variations make direct automated comparison of the dates all but impossible. The issue can occur even within records from the same dataset and is further compounded when attempting to integrate multilingual data, presenting a significant practical hindrance to semantic interoperability. The same problem can be found in temporal metadata, whether manually entered or produced via Natural Language Processing (NLP) techniques.

Initiatives concerning FAIR (findability, accessibility, interoperability, and reuse) and Open Data emphasise the need to make data and reports arising from archaeological and heritage investigations freely available to support reuse and comparison of published data (e.g. Evans 2015; Wilkinson *et al.* 2016). In order to support cross search, comparison and meta research, there is a need to resolve and normalize diverse textual expressions of time to more structured, consistent and language neutral formats and, if possible, a consistent numerical time axis.

Section 2 discusses the background context and Section 3 outlines the aims of the paper. Section 4 summarises various issues encountered in normalizing temporal expressions. The methods adopted in this paper are described in Section 5 and Section 6 outlines their implementation in a software tool. Section 7 reports on different test results and applications of the normalization tool. Section 8 discusses reflections on the work, together with limitations and future work, while Section 9 draws conclusions.

2 BACKGROUND

The systematic expression of temporal information has been a recurring theme in digital archaeology and cultural heritage over the years. A detailed description of the data model for a project to digitise the existing card index of archaeological radiocarbon dates can be found as early as 1982 (Moffett & Webb 1982). Castleford (1992) discusses the difficulties posed by temporal data for digital representation in archaeology and reviews the literature on time and Geographic Information Systems (GIS). A follow up review is provided by Daly and Lock (1991) who highlight the effect of different theoretical notions of time for archaeology. They outline three potential possibilities for exploiting temporal data: object orientated, animation and 3D techniques. All have been pursued, particularly if we consider ontological modelling of temporal relationships (e.g. Binding 2010) to be a form of object orientation. For example, Johnson's (2002, 2003) influential *Time* Map project included exemplars of the combination of spatial data with chronological animation to illustrate change over time and provide interactive educational overviews of archaeological and historical museum datasets.

A formalism for defining dates, both precise and imprecise expressions, together with a syntax and associated temporal algebra was proposed by Signore *et al.* (1997) in the context of historical geography. Grandi (2006) discusses an XML based prototype implementation to encode indeterminate temporal expressions for an Italian language historical dictionary, following a probabilistic approach with indeterminate dates represented by a probability distribution over a time interval with different shapes for expressions of *during, early, late,* and *around.* Green (2011) discusses a GIS extension to encompass the inherent uncertainties of archaeological temporal data, with various methods

to relate the inherent probability of an uncertain date (e.g. derived from radiocarbon dating) fitting a given time period, including standard, normal, radiocarbon and *terminus post quem* probability.

Named periods are complex in that they combine a spatial and temporal component and sometimes a cultural practice (or style). The timespans associated with periods vary with spatial extent. They are often provided by national authorities – examples include English/Scottish/Welsh periods (HeritageData 2023) while Dutch language periods are given in the Archeologisch Basis Register (ABR) thesaurus (Erfgoedthesaurus 2023) implemented in the national data interchange standard (Boasson & Visser 2017). The discussion on periods in Section 8 includes future work plans to take advantage of work in Linked Open Data that has created an international platform (PeriodO 2023) for sharing period definitions along with the underlying date range (Shaw *et al.* 2016).

These issues are brought into relief by work on semantic integration for purposes of cross search and potentially synthetic research, typically involving temporal information and sometimes archaeological reports as well as data. Examples or discussions with English language data include Kansa (2014); Richards & Hardman (2008); Tudhope *et al.* (2011). Dutch language examples include Brandsen *et al.* (2019), Brandsen & Lippok (2021), Vlachidis & Tudhope (2022), while the ARIADNE archaeological infrastructure project integrated data in multiple European languages (Aloia *et al.* 2017; Binding *et al.* 2019). Making the case for the application of knowledge organization approaches, Lee (2017) highlights how the application of Bayesian modelling, together with the collection of the existing (uncertain) radiocarbon dates for Neolithic 'causewayed enclosures', significantly refined the construction dates and facilitated a historical narrative.

3 AIMS

Different sources of temporal expressions may be used for computer-based analysis. Data fields may be controlled, semi-controlled or free text fields requiring data cleaning. Archives and repositories will probably include temporal metadata. The outcomes of NLP on reports and grey literature may include a variety of temporal expressions. In some cases, temporal data is expressed as numeric date ranges, sometimes in terms of decades or centuries or named periods. For international data integration, the temporal expressions will occur in different languages.

In previous work, the authors have explored semantic integration of data and grey literature reports with different structures, languages and terminology via a common conceptual framework. This previous work (e.g. Binding 2010; Binding *et al.* 2015; Tudhope *et al.* 2011) had to encompass a wide variety of expressions of temporal data, expressing dates, timespans and periods (e.g. "MLC2-C3", "AD341-6", "Early fifth century", "Georgian" etc.). These values were manually normalized to a consistent format to enable the relative comparison of timespans, in a time consuming, case by case approach. This paper reports on subsequent work that investigated the feasibility of an automated process and the resulting techniques for the normalization of date ranges and periods to a numerical date range.

This paper has a specific focus on the systematic digital representation of temporal expressions; it is not concerned with archaeological models of time or dating methods except in their consequence for consistent and appropriate digital representation. The intention is not to make arguments on relative merits of investigation strategies or dating methods but rather to explore possibilities for consistent, language neutral expressions to improve semantic integration possibilities. The paper describes a method and its implementation (as an open source application tool) to provide normalized numeric start/end timespan attributes for temporal expressions (in different languages) in addition to the original metadata.

4 DESIGN ISSUES

Various issues must be addressed in order to perform effective normalization of dates, periods and timespans. A consistent epoch or anchor point must be established in order to relate one date with another. A variety of textual temporal patterns and variants (including prefixes and suffixes) should be accommodated for different languages. The underlying digital representation must provide an appropriate level of accuracy for archaeological temporal granularity. The first two issues are discussed in this section while the digital representation is discussed as part of the methodology in Section 5.

4.1 ANCHORING DATES AND TIMES

An epoch may be either a period of time or a fixed point in time about which other dates are relatively measured. Analogous to geographical data where the information space is addressed through the use of coordinates relative to a fixed datum or zero point, the epoch of a calendar era is a temporal anchor point or origin, relative to which date/time values are expressed. Roe (2023) provides a software utility with classes to define an era and functions for combining and converting between eras. Many different epochs have been used to anchor the dates mentioned in historical sources, such as Ab Urbe Condita, Anno Diocletiani, Anno Hijra, Anno Mundi, etc. Calendar citations commonly encountered in archaeological literature include:

- BC/AD – Before Christ/Anno Domini – used by the Gregorian calendar, the predominant calendar currently applied to dating in Western cultures.
- BCE/CE – Before Common Era/Common Era – exactly aligned to BC/AD.

As dates are expressed relative to an epoch, so times (of the day) are similarly anchored relative to a fixed zero point – Coordinated Universal Time (UTC) aligns with the former Greenwich Mean Time (GMT). The world is divided into a series of *time zones* having a time offset to be applied relative to this fixed point to give the local time for locations within each zone. For the purposes of this paper, the level of granularity appropriate for archaeological dating information is taken as a year; finer intervals of time (days, hours, etc.) are considered out of scope for the present paper. See Sugimoto (2021) for an initiative developing Linked Open Data at the granularity level of a single day.

4.2 IDENTIFYING TEMPORAL PATTERNS

Analysis of the temporal expressions encountered in previous work reveals that the textual patterns mostly fell into the seven general categories outlined below. In addition, there are many possible prefixes and suffixes that may trivially or fundamentally alter what is meant in terms of a date or date range e.g. "Circa/Ca/Cca/C", "Early", "Mid", "Late", "BC" etc. – see Section 8 for further discussion. Examples of the output (labelled timespan with numeric start year and end year) produced by the normalization tool for each category are provided in Section 6 for the languages implemented to date.

4.2.1 Ordinal named or numbered centuries

Text string such as "*19th century*", "*Nineteenth century*". Assigning start/end dates is fairly straightforward, although sometimes there is a prefix: early/mid/late (see section 5).

4.2.2 Year spans

Numeric years in a general *{from year}{separator}{to year}* format e.g. "*1715–1825*". Sometimes the 'to year' component may be shortened e.g. "*1485–92*" so the normalization process needs to identify this case in order to calculate the correct dates.

4.2.3 Single year with tolerance

A single numeric year, sometimes appended with a specific positive/negative tolerance e.g. "*1666*", "*1485 + 5–10*", "*1540 ± 9*". The normalization process must isolate the quoted year and apply any specified tolerances to give the correct start/end years.

4.2.4 Decades

Text string indicating a period of 10 years e.g. "the *1880s*". There is some ambiguity (in English at least) whether "*the 1900s*" refers to the decade or the century and sometimes it is impossible to know without further context. Note "*the 1900s*" as a decade means 1900–1909 CE but the century started in the year 1901 and so 1900 belongs in the previous century (see the discussion on century boundaries in Section 5).

4.2.5 Century spans

Text string such as "*Late 15th–Mid 17th century*" – a span of ordinal named/numbered centuries. The start point of the first part and the end point of the second part give the overall timespan.

4.2.6 Month/Season prefix and single year

Text string such as "*Summer 1615*", "*January 1725*". The granularity of most dates/periods encountered was at the level of years rather than fractions of years, hence the level of granularity is taken as a year. Currently month/season prefixes are matched but not reflected in the resulting timespan.

4.2.7 Named periods

Normalization of references to named periods (E.g. "*Georgian*", "*Iron Age*", "Romeins tot middeleeuws", "Romersk til middelalderen", "Canoloesol i Edwardaidd") involves lookup tables of period labels with associated start/end dates derived from national authorities in different languages. Compared to the other textual patterns this is a different (but frequently encountered) use case and requires knowledge of the spatial context and date information. More information is provided in the discussion in Section 8.

5 METHODOLOGY

The approach compares each textual expression of time encountered against a series of regular expression patterns (elaborated over successive projects) to identify a pattern match and then determines a date range (start/end years) for the timespan based on the pattern matched. If no match is identified, the text value is compared against a lookup list of known named time period labels with associated date ranges. Table 1 shows regular expression examples. Note there is some variation in the regular expression syntax implemented by different programming libraries – the example syntax shown here is compatible with the *Python re library*, but would require slight revision for use with e.g. the *.NET Framework Regex class*, or *JavaScript RegExp object* (the latter having no support for named capture groups). As regular expressions can rapidly become complex a modular approach was taken for greater reusability and consistency – concatenation of modular units then allowed more complex expressions to be constructed.

PYTHON REGULAR EXPRESSION EXAMPLE	INPUT PHRASE	OUTPUT VALUES
^(?P<yearMin>\d+)(?:\s?\-\s?)(?P<yearMax>\d+)$	*"1450–1460"*	yearMin = 1450, yearMax = 1460
^(?P<decade>\d+0)\'?s$	*"1850s"*	decade = 1850
^(?:C(?:irca|\.)\s*)?(?P<year>\d+)$	*"C. 1485"*	year = 1485

Table 1 Examples of regular expression matching on typical date patterns.

For some cases the start/end dates are present and are extracted directly from the textual string. In most cases, some additional processing is required after the initial pattern match. For example, the pattern for ordinal centuries (e.g. *Early 2nd Century AD*) is matched via regular expressions, but deriving the actual start/end dates requires further processing. The patterns also observe the possible presence of prefix and/or suffix modifiers and modify the output dates in order to take account of the meaning of phrases, such as *BC, BCE, BP, Early, Mid, Late* etc. Offsets are applied when tolerances occur, (e.g. *1540 ± 9* = start year 1531, end year 1549).

5.1 CENTURY CONVENTIONS

Century boundaries have been the subject of recurring debate over many years, bibliographic evidence demonstrating similar arguments being rekindled on the cusp of each of the previous three centuries (e.g. Freitag 1995). The issue concerns whether centuries start at year 0 or year 1. The Gregorian calendar does not include a year 0 (the year directly preceding 1 AD is 1 BC) so it would seem logical that AD/CE centuries start at year 1 and end at year 100, although this view is not universally shared. The Getty Art & Architecture Thesaurus (AAT 2022) concept for *twentieth century* acknowledges this ambiguity within the scope note that accommodates both viewpoints: "*Century in the proleptic Gregorian calendar including the years 1900 to 1999 (or 1901 to 2000)*".

For the purposes of this work, centuries are assumed to start at (and include) year 1 and end at (and include) year 100. E.g. "*13th–14th century AD*" starts at (and includes) the year 1201 and finishes at (and includes) the year 1400. Prefix modifiers for centuries are given the meaning described in Table 2. The boundaries of *Early*, *Mid* and *Late* overlap, reflecting a level of approximation when using such terms.

AD years are represented as positive numbers, BC years are represented as negative numbers. Astronomical years and the ISO international standard for representation of dates and times (ISO 8601) consider the year 0 as meaning 1 BCE, so –1 represents 2 BCE, –2 represents 3 BCE etc. This means BCE/CE duration calculations would be correct, but the value present in the year position for any date prior to 1 CE is represented by the BCE year plus 1 (e.g. 400 BCE is represented in ISO 8601 as "–0399").

PREFIX	START	END
Early	1	40
Mid	30	70
Late	60	100
First Half	1	50
Second Half	51	100
First Quarter	1	25
Second Quarter	26	50
Third Quarter	51	75
Fourth Quarter	76	100

Table 2 Numerical meaning assigned to century modifiers.

5.2 TIME REPRESENTATION IN PROGRAMMING ENVIRONMENT DATA TYPES

In the choice of underlying data types for implementation, it may seem initially attractive to utilize programming environment data types to represent dating information. However, while such data types may incorporate millisecond accuracy for use cases where that is useful, they have unworkable limitations in the range of years accommodated for archaeological time, as outlined in this section.

Computer operating systems measure elapsed time relative to their own *epoch*. For example, UNIX Time is the number of *seconds* elapsed since Thursday 1st January 1970 00:00:00 UTC, the UNIX epoch (ICU 2023). Database field data types also have significant limitations when it comes to storing archaeological or geological date information and therefore implementations often eschew the inbuilt data types in favour of numeric years or ISO strings. The Microsoft SQL Server database has a *datetime* data type accommodating dates ranging from January 1st 1753 to December 31st 9999 (Microsoft SQL 2022). The significance of this seemingly arbitrary start year is that 1753 was the year after Britain stopped using the Julian calendar in favour of the Gregorian calendar, 'losing' 12 days in the process but avoiding potential complications in the calculation of accurate durations by not accommodating any earlier dates. The SQL Server *datetime2* data type accommodates a date range between January 1st 0001 to December 31st 9999, using the *proleptic* Gregorian calendar to extend backwards past 1753 to the year 0001. ORACLE date values follow the Julian calendar and range from January 1, 4712 BCE to December 31, 9999 CE (Oracle Database 2023). Postgres database date fields (Postgres 2023) can accommodate years ranging from 4713 BC to 5874897 AD (4713 BC marks the start of the Julian calendar). The Microsoft .NET framework *System.DateTime* class accommodates a year range of 0001 CE to 9999 CE (Microsoft.NET 2023). The System.DateTime class gives an accuracy of 100 nanoseconds but does not cater for negative (BCE) dates – consequently it is not an appropriate candidate for representing many historical dates.

The Java language counts the number of milliseconds from midnight on the epoch of January 1st 1970 UTC. The Java date class is a *signed long* value (64 bits = -2^{63} to $+2^{63}-1$), accommodating an overall year range of approximately 290,000,000 years forwards and backwards (Nielsen 2000). The comparatively newer java.time.Year class supports a larger range of dates, from a minimum year of -999,999,999 to a maximum year of +999,999,999 (Oracle Java 2023). The ECMAScript/JavaScript specification for the Date object (ECMA 2011) also uses the 01/01/1970 epoch, accommodating an overall year range of approximately 285,616 years forwards and backwards (slightly smaller for ECMAScript Date objects). The JSON specification does not prescribe a serialized format for the representation of dates (Newtonsoft 2023).

In summary, programming environment data types are not always appropriate for the representation of the range of archaeological or geological time. The approach adopted in this paper is to employ a standardised literal syntax textual representation, as described below.

5.3 TEXTUAL REPRESENTATION AND INTERCHANGE OF DATE/TIME INFORMATION

A common internationally agreed standard format can be utilized for the purposes of efficient data interchange, integration, search and visualization. The ISO 8601 standard (ISO 2019a) sets out internationally agreed ways to represent dates and time, and covers how to serialize dates, times and intervals in text consistently and unambiguously, using the proleptic Gregorian Calendar.

An example serialization format for a time interval including a starting date/time and ending date/time is "1998-12-01T12:03/2004-04-02T14:12". Depending on the use case, the granularity may be reduced to just years e.g. "1998/2004".

The standard covers numeric dates only and does not cover the language specific word-based temporal category types, outlined in Section 4.2, involving references to centuries, named periods, etc. Therefore, in the method described here, temporal values employing words are converted to year spans (the original strings are retained for any consuming application purposes) and the timespan is output as ISO 8601 conformant year span values. The advantage is that temporal expressions are converted to a standard format that can be directly and consistently parsed, manipulated and compared. Thus, for example, "5th Century" is converted to "0401/0500". Similarly, named time periods are converted to year values via lookup. Named periods are not as clear cut as (say) century or decade formats because their meaning (as a span of years) is not implicit in the string itself and this issue is further discussed in Section 8. The original period labels are retained, with the timespan output provided as additional attributes.

These attributes are expressed conformant to XML Schema datatypes as a subset of ISO 8601 that employs the 'seven-property model': i.e. Year, Month, Day, Hour, Minute, Second, Time-zone designator. Years may be negative and more than 4-digit years are accommodated. Years less than 4 digits must be zero-padded.

xsd:datetime: [-][Y*]YYYY-MM-DDThh:mm:ss[.s[s*]][TZD]
e.g. "2018-05-26T12:02:08.125Z"

Elements of the seven-property model may be omitted to reflect decreased precision, in strict granularity order – from the least to the most significant. This is illustrated by the xsd:gYear datatype which omits five properties in granularity order. The time zone designator remains optional though the granularity expressed or suggested by inclusion of time zones is not appropriate for our use cases (where the level of granularity is taken as a year):

xsd:gYear: [-][Y*]YYYY[TZD]
e.g. "–10500", "–0034", "0420", "1066"

The XML Schema datatypes do not accommodate the expression of time intervals (only durations). However, these datatypes were derived from ISO 8601 – which does allow for time intervals to be expressed using a forward slash character, so intervals as year spans may be serialized as a consistently formatted string e.g. "–0034/0420".

6 IMPLEMENTATION

The *Yearspans* software application tool (Binding 2023) has been developed to perform bulk matching of data against temporal patterns in different languages, following the methods described in Section 5. Textual expressions are normalized to a timespan with numeric year start/end boundaries. The application matches patterns corresponding to the seven general categories of temporal expressions described in Section 4.2 (see Table 3).

The application combines regular expression pattern matches and lookups. Small modular regular expressions are used to build more complex expressions. Years are expressed relative to Common Era (CE). Centuries are considered to start at year 1 and end at year 100. For matches on named periods (e.g. *Georgian*, *Victorian*

1.	Ordinal named or numbered centuries
2.	Year spans (from year – to year)
3.	Single year (possibly with a tolerance)
4.	Decades
5.	Century/millennium spans
6.	Month (or season) and single year
7.	Named periods

Table 3 Categories of temporal expression.

etc.) the start/end years are derived from standard authority lists (e.g. the FISH periods list (2023) and see the discussion in Section 8).

The application is developed in Python (and is an evolution of an earlier C# .NET version). The input for a single match is a temporal text string and a language code (ISO639-1:2002). The output is a class instance having properties of *label*, *minYear* and *maxYear*. The class can also produce ISO 8601 conformant formatted strings representing the years and the year span, e.g.

input = value: "410–1065", language: "en"
output = label: "410–1065", minYear: 410,
maxYear: 1065, isoSpan: "0410/1065"

A bulk processing Python script encompassing the same functionality can be used to process a text file containing a list of temporal values. The input is a file name and a language code; the output is a delimited text file having the columns as the properties shown above (i.e. *label, minYear, maxYear, isoSpan*).

The Yearspans application software is freely available via a source code repository (Binding 2023). Language-specific regular expression patterns used for matching are included for Dutch, English, French, German, Italian, Norwegian, Spanish, Swedish and Welsh. While there is some variation in the range of textual patterns covered, the GitHub repository contains a suite of unit tests using the Python *unittest* framework (280 tests in all) covering the categories outlined in Section 4.2 in each supported language.

Examples of the categories of temporal expressions and corresponding results obtained from the application are shown in Table 4.

7 APPLICATION

Test results and working applications with diverse forms of input data are described below.

7.1 DATASETS FROM ADS ARCHIVE

In this test, six English language data files were downloaded from project files deposited in the Archaeology Data Service archives and freely available (ADS Archives 2023). The project sources for the test datasets were:

CATEGORY	INPUT	LANGUAGE	MIN YEAR	MAX YEAR
1	Early 2nd Century BC	en	–0199	–0159
1	Inizio undicesimo secolo d.C.	it	1001	1040
1	Begin 11e eeuw voor Christus	nl	–1099	–1059
1	Tidlig på 1100-tallet e.Kr.	no	1001	1040
1	Principios del siglo XI a.C.	es	–1099	–1059
1	Frühes elfte Jahrhundert n. Chr	de	1001	1040
2	1839–75	en	1839	1875
2	140–144 d.C.	it	0140	0144
2	Tidigt 1100-tal f.Kr.	sv	–1099	–1059
2	1250 – 57 e.Kr.	sv	1250	1257
3	*1540 ± 9*	en	1531	1549
4	1950-tallet	no	1950	1959
4	intorno al decennio 1910	it	1910	1919
5	5th – 6th century AD	en	401	600
5	finales del 1° a principios del 2° milenio d.C.	es	0600	1400
5	inizio del undicesimo alla fine del dodicesimo secolo d.C.	it	1001	1200
5	laat 1e tot begin 2e millennium na Christus	nl	0600	1400
5	III e lo II secolo a.C.	it	–0299	–0100
5	début 11ème à fin 12ème siècle après JC	fr	1001	1200
6	July 1855	en	1855	1855
6	Estate 1855	it	1855	1855
7	Victorian	en	1837	1901
7	georgienne à victorienne	fr	1714	1901
7	Canoloesol i Edwardaidd	cy	1066	1910
7	Völkerwanderungszeit	de	0375	0586

Table 4 Normalized output for different temporal categories and languages.

- VAG Dendrochronology Database – Vernacular Architecture Group, 2000 (updated 2022). A database (two datasets) of tree-ring dates based on samples taken from the structural elements of over 4500 traditional timber structured buildings (doi:10.5284/1091408)
- Cloakham Lawns, Axminster, Devon. Archaeological excavation undertaken by Cotswold Archaeology (doi:10.5284/1042741)
- The Staffordshire Hoard: an Anglo-Saxon Treasure – Barbican Research Associates, 2017 (updated 2019) (doi:10.5284/1041576)
- Excavation at Goldsmith Street, Exeter 1971 (Exeter archive site 37) Exeter City Council, Cotswold Archaeology, 2015. Dates extracted from downloadable context dating PDF (doi:10.5284/1035174)

For each test dataset the particular column containing the date information was isolated and extracted as a simple UTF-8 text file. In the case of the Goldsmith Street data, the temporal expressions were present as a listing within a PDF file and were manually extracted. Initial tests, with refinement of the regular expression patterns, were performed on the files from the VAG Dendrochronological Database, which accordingly is not an 'unseen' dataset. The results of the tests are summarised in Table 5, together with notes on the reasons for non-matches. The discussion in Section 8 outlines some approaches for further refinement.

7.2 ITALIAN LANGUAGE ARCHAEOLOGICAL EXPRESSIONS

The Hypermedia Research Group collaborated in a pilot NLP development for Italian archaeological reports, as part of a European Open Science Cloud for Research (EOSC) Pilot Project (Felicetti *et al.* 2018). This work employed the GATE Natural Language Processing framework (Cunningham et al. 2013) to perform Named Entity Recognition (NER) on Italian archaeological reports. One aspect involved the extraction of temporal expressions from the documents, drawing on a sample of Italian expressions (with English translations). In testing the NER pipeline against Italian archaeological reports a total of 142 (unique) Italian expressions of dates, timespans or periods were extracted. These were subsequently reused as input to the Yearspans application for testing and improving the performance of the Italian expression patterns. Table 6 illustrates the variation in the textual expressions encountered, including the presence of prefixes, suffixes, separators, punctuation, and other different ways of expressing date spans. Note that the Actual Output from Yearspans mirrors the Anticipated Output, other than the two phrases shown in italics, arguably lacking century designators. The issue of incorrect or idiosyncratic source textual expression is discussed further in Section 8.

7.3 NORMALIZATION FOR ARIADNE MULTILINGUAL DEMONSTRATOR

The FP7 ARIADNE archaeological research infrastructure project (Aloia *et al.* 2017) produced a number of research

DATASET (AND FILE)	MATCHED	NO MATCH	TOTAL	NOTES
VAG Dendrochronology Database VA_Dend_2021_ADS-IND_FULL_TO_52.csv Col: Date	4,471 98.61%	63 1.39%	4,534 (non-blank values only)	Non-matches included unexpected suffixes e.g. *"1600+(soon)", "1238–1264?", "1660 Â ± 9", "1690+(close to edge)"*
VAG Dendrochronology Database VA_Dend_2021_ADS-IND_FULL_TO_52.csv Col: Date2	720 64.98%	388 35.02%	1,108 (non-blank values only)	Many of the non-matches are multi value fields (semicolon delimited) e.g. *"1528–1559; 1586/7; 1596; 1604/5; 1340+"*. Pre-processing could separate these values.
Cloakham Lawns, Axminster RAMM_66_2016_spot_dates.csv Col: F_Spot_date	7 19.44%	29 80.56%	36	Non-matches were on unfamiliar patterns e.g. *"RB"*, *"MLIA"*, *"LC17-C18"*. Incorrect match on *"LIA-C1"*
Staffordshire Hoard Appendix 1a SHRR24_App1a_SHER.csv col: Dates	43 91.49%	4 8.51%	47 (non-blank values only)	Non-match on *"c 10000 BC- c AD1799"* as second 'circa' not included in the matching pattern
Staffordshire Hoard Appendix 1b SHRR24_App1b_WWHER.csv	11 100%	0	11	Small dataset, all known patterns e.g. *"c.410–1065"*, *"c.4000 BC-AD 1539"*
Goldsmith St Exeter Context dating and finds listing GS1_finds_listing_71.pdf	120 86.33%	19 13.67%	139	Non-matches were on additional punctuation or text e.g. "?12th century", "after 1300"
Totals	**5,372 91.44%**	**503 8.56%**	**5,875 100%**	

Table 5 Summary of results from testing on ADS archive datasets.

TEXTUAL PHRASE	APPROXIMATE TRANSLATION	ANTICIPATED OUTPUT		ACTUAL OUTPUT	
		MINYEAR	MAXYEAR	MINYEAR	MAXYEAR
Dal 1925	*"Since 1925"*	1925	1925	1925	1925
Nel 1897	*"In 1897"*	1897	1897	1897	1897
211 a.C.	*"211 BC"*	–211	–211	–211	–211
in 1191	*"in 1191"*	1191	1191	1191	1191
575–520 a.C.	*"575–520 BC"*	–575	–520	–575	–520
II a.C.	*"2nd BC"*	–200	–101	–	–
II-III secolo d.C.	*"2nd–3rd century AD"*	101	300	101	300
intorno a VI sec. d.C.	*"around 6th century AD"*	501	600	501	600
tra IV e III secolo a.C.	*"between 4th and 3rd century BC"*	–400	–201	–400	–201
tra IV e V sec. d.C.	*"between 4th and 5th century AD"*	301	500	301	500
tra lo 141 ed lo 91 a.C.	*"between 141 and 91 BC"*	–141	–91	–141	–91
tra lo 1790 ed lo 1797	*"between 1790 and 1797"*	1790	1797	1790	1797
tra lo I e lo VI sec.	*"between the 1st and the 6th century"*	1	600	1	600
tra lo III e lo IV sec. d.C.	*"between the 3rd and 4th centuries AD"*	201	400	201	400
tra lo III ed lo II	*"between the 3rd and the 2nd"*	101	300	200	201
tra VII e VI a.C.	*"between 7th and 6th BC"*	–700	–501	–700	–501

Table 6 Example of results obtained for extracted Italian temporal expressions originating from EOSC pilot data.

demonstrators exploring the semantic integration of detailed archaeological data, which included a case study by the Hypermedia Research Group exploring the integration of archaeological reports and datasets in Dutch, English and Swedish (selection of reports only) languages via a broad theme of wooden objects and their dating via dendrochronological techniques. Relevant selections from datasets were mapped to a common semantic framework and spine vocabulary, while natural NLP information extraction techniques were applied to grey literature reports. An integrated RDF dataset was produced along with an interactive research demonstrator query builder application, which shielded the user from the complexity of the semantic framework. The Demonstrator is available for use (ARIADNE Demonstrator 2017). The case study, together with reflections on the methods and use cases, is discussed by Binding, Tudhope & Vlachidis (2019) but the timespan normalization aspect was not addressed in any detail in that paper.

Date spans in a variety of text formats were present in both dataset fields and the NLP output from the reports. The (earlier C# incarnation) timespan tool was used to process the temporal textual values against the regular expressions (as described in Section 5) and generate start/end years. A selection of the results from the NLP output in the three languages is given in Table 7, which shows the input string and resulting timespans. Each unique timespan had its own URI and several records might be connected with a particular timespan in the semantic framework.

Figures 1–4 show a selection of example screen dumps from the Demonstrator illustrating use of the timespan normalization. The query builder is on the left with results

TIMESPAN-LABEL	MINYEAR	MAXYEAR
jaar 1302	1302	1302
16e eeuw	1501	1600
14® eeuw	1301	1400
derde eeuw	201	300
tweede eeuw	101	200
55 en 69 na Chr	55	69
124 ± 6 en 125 ± 6 na Chr	118	130
125 ± 6 na Chr	119	131
nineteenth century	1801	1900
AD 1424–89	1424	1489
AD 1487	1487	1487
8th century AD	701	800
AD 458–704	458	704
eighth century	701	800
AD 1395/6	1395	1396
1635–1972 AD	1635	1972
9 AD	9	9
26 BC	–26	–26
AD 1353 to AD 1579	1353	1579
AD 1579/80	1579	1580
3e eeuw	201	300
1730-talet	1730	1739
vintern 1677–1678	1677	1678
1600-talet	1600	1699
30 år	30	30
1800-tal	1800	1899

Table 7 Selection of output from timespan normalization in ARIADNE Demonstrator.

appearing on the right. In the results, we see different categories of temporal textual patterns in different languages. We see that the normalized timespans make possible a common numeric time axis (slider interface). Figure 1 shows a query on records (including NLP output from reports) referring to *pine* within a specific time period with results from two Swedish reports concerning samples dated 1813/14. Figure 2 shows a query, this time over data fields, showing dendrochronological dating results with a variety of English language expressions for the 17th century. Figure 3 shows results from Dutch reports including a specific year with tolerance, while Figure 4 shows that tolerance in ISO 8601 representation (as discussed in section 5.3), with output from the normalization process expressed as CIDOC CRM property values.

7.4 ARIADNEPLUS DATA AGGREGATION

Temporal normalization was also employed in the successor H2020 ARIADNEplus project (2022), which enhanced and expanded the digital infrastructure for archaeological data and reports (Niccolucci & Richards 2019). Considerable effort in the project was devoted to the data aggregation process, which transforms source data and metadata to the common framework

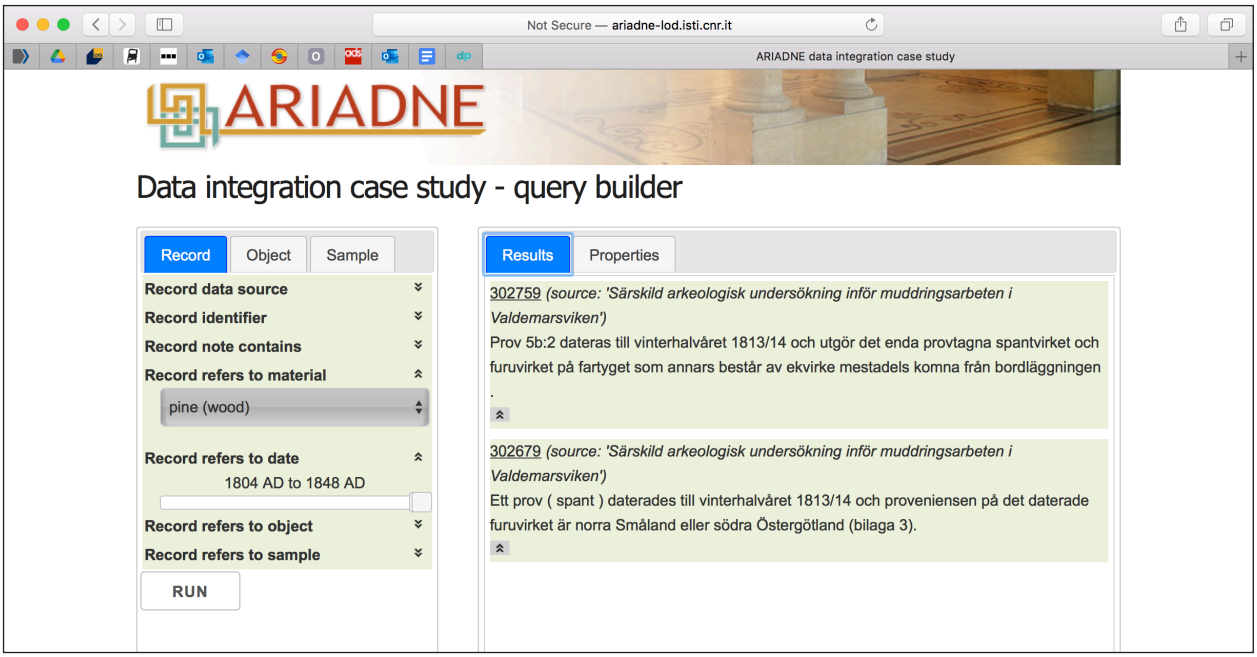

Figure 1 ARIADNE Demonstrator example with results from Swedish language reports.

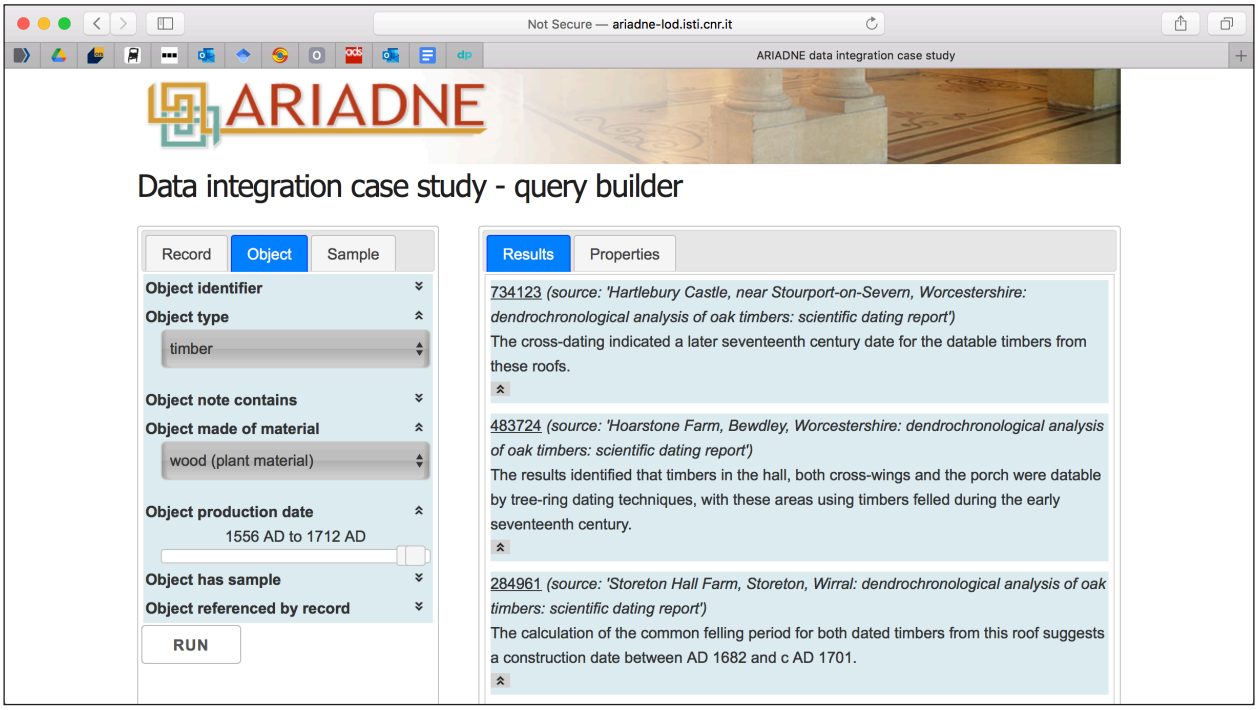

Figure 2 ARIADNE Demonstrator example with English data results.

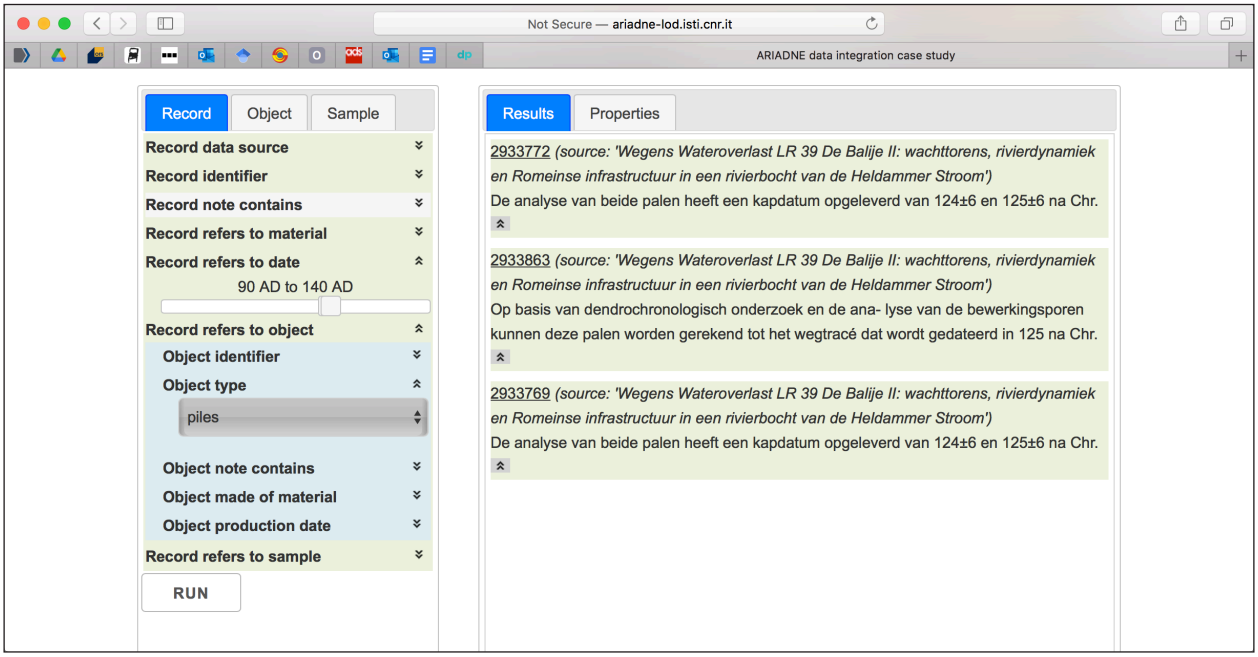

Figure 3 ARIADNE Demonstrator example with results from Dutch language reports.

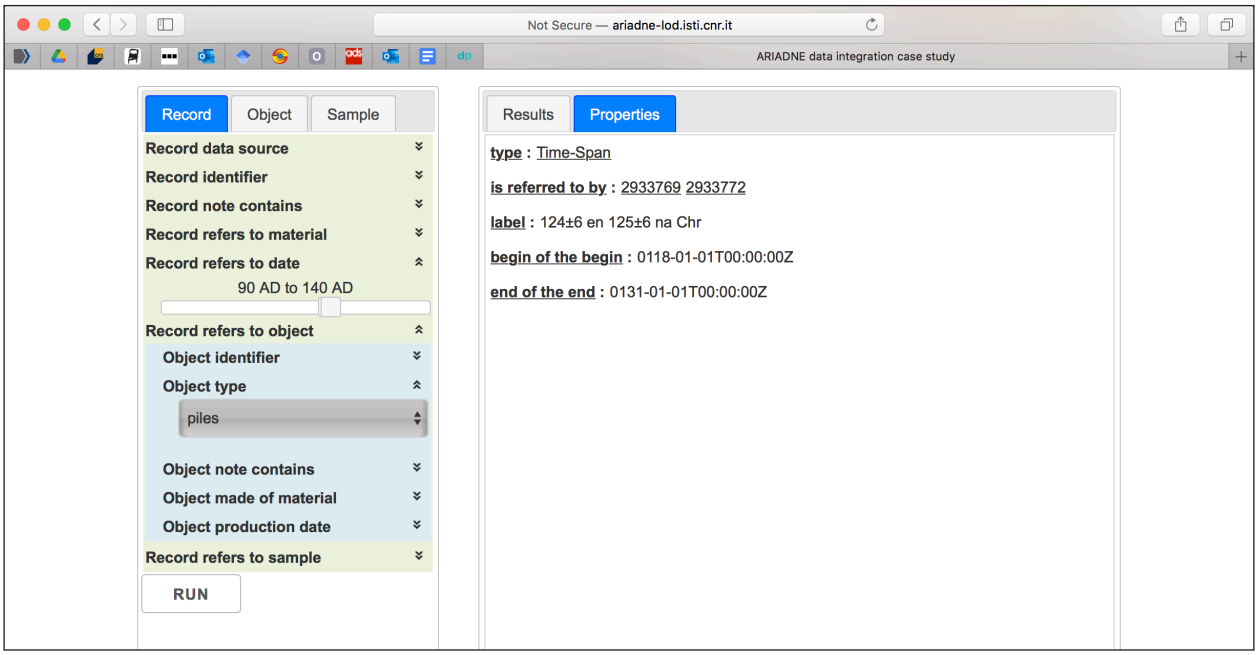

Figure 4 ARIADNE Demonstrator example with output from the normalization process on year tolerance expression.

and terminology for semantic integration. Temporal expressions are represented where possible by source labels and numeric year-spans. ARIADNEplus data partners provided PeriodO named periods in order to add *From* and *Until* dates in absolute years BCE and CE during the metadata enrichment stage of the workflow to yield a common time axis. However, some project datasets hold timespans in a wide variety of other textual formats. Accordingly, a modified version of the Yearspan normalization utility was made available in the aggregation workflow (and GitHub repository) in order to supplement the original text values with additional attributes defining the start and end years of the span.

In a challenging case for the aggregation process, source XML data records were imported from a data provider (Archaeology Data Service) and named centuries and year spans instances were supplemented with derived start/end year elements. The process was applied to XML files, containing a mixture of named periods and year spans in over 100,000 temporal records. Figure 5 shows example output, where a record with a "dcterms:temporal" element *(850–1068)*, has been supplemented with elements for "minYear" and "maxYear" as xsd:gYear values.

Figure 6 shows another example, where a "dcterms:temporal" element *(19th century)*, has been supplemented with elements for the start and end of the century (following the conventions discussed in Section 5).

```
<record type="record">
<dc:title>ST. MARTIN'S CHURCH</dc:title>
<dc:creator>Exeter City Council</dc:creator>
<dc:subjectPeriod>
<dc:subject>CHURCH</dc:subject>
<dcterms:temporal>850 - 1068</dcterms:temporal>
<minYear xsi:type="http://www.w3.org/2001/XMLSchema#gYear">0850</minYear>
<maxYear xsi:type="http://www.w3.org/2001/XMLSchema#gYear">1068</maxYear>
</dc:subjectPeriod>
...
</record>
```

Figure 5 ARIADNEplus example with numeric year span.

```
<record type="record">
<dc:title>Milestone, village centre, jct of Church Street &amp; Bedford
Street</dc:title>
<dc:creator>Historic Milestone Society</dc:creator>
<dc:subjectPeriod>
<dc:subject>MILESTONE</dc:subject>
<dcterms:temporal>19th century</dcterms:temporal>
<minYear xsi:type="http://www.w3.org/2001/XMLSchema#gYear">1801</minYear>
<maxYear xsi:type="http://www.w3.org/2001/XMLSchema#gYear">1900</maxYear>
</dc:subjectPeriod>
...
</record>
```

Figure 6 ARIADNEplus example with century textual expression.

8 DISCUSSION, LIMITATIONS AND FUTURE WORK

While the temporal patterns implemented cover the vast majority of patterns encountered in our work to date, we hope to extend the range and also language coverage in future projects and collaborations (the tools and patterns are available as open source and contributions are welcome). One limitation is that while the patterns include a date with tolerances, probabilistic temporal formats are not covered and specific formats for scientific dating methods are not currently covered. For example, both calibrated and uncalibrated results should be reported for radiocarbon dating; care is needed to distinguish calibrated from uncalibrated results in order to yield comparable dates. Similarly, different dendrochronological dating techniques and interpretations of missing tree-rings may make it appropriate to report raw dates for the tree-ring series in combination with the estimated felling dates. Specialised database initiatives are emerging for such purposes (e.g. Edvardsson *et al.* 2022).

Data integration problems can arise where the epoch is unspecified (or unknown). For example, assuming a default epoch of AD/CE should be based on background knowledge about the origins and meaning of the original data. Some textual patterns involving numeric values apply to multiple languages, depending on assumptions about the epoch in use (see section 4.1). The languages implemented so far are European, with an assumption of the same underlying epoch ("5th Century" and "V secolo" both result in dates relative to BCE/CE). In the interests of future expansion of scope, the epoch should probably be a configurable option. Some form of visualization of the output would also be a useful future extension of this work, perhaps including geographical context.

The treatment of qualifiers indicating uncertainty around dates is a complex issue requiring further work. The approach sometimes used in semantic integration is to ignore all qualifiers – however that ignores potentially useful information. Prefixes and suffixes often express fuzziness or uncertainty (e.g. *around | approx. | circa | after | +– | ? | early | mid | late | first half etc.*). For data integration purposes, it is desirable to quantify what is meant, although reaching consensus may be a long-term community effort. For example, "circa" may refer to a specific tolerance – if so should that be a proportional tolerance for dates exhibiting lesser granularity ("circa 5000 BC" vs. "circa 1753 AD")? Section 5.1 outlines where we have taken a tentative step with century subdivisions. Similarly, dual named periods have been encountered in existing data e.g. "Neolithic/Bronze Age". It is unclear whether this refers to a larger overarching period or whether it means a smaller transitional period between the two periods. In some legacy datasets, without recourse to the original context, it may be impossible to know the original intent of some temporal expressions. The extension to ISO 8601 (ISO 2019b)

allows the expression of uncertain or approximate dates and initiatives are exploring practical tools and modelling based on the standard (Shaw 2023).

A pre-processing stage for data cleaning has not been included but might improve the prospects for matching against the predefined textual patterns – e.g. normalizing whitespace, internationalized characters and capitalization. The extent to which it is possible to cater for localised conventions, idiosyncrasies and common errors in textual expressions of time merits further exploration and discussions. This crosses the boundary of NLP, beyond the scope of the work reported here on the normalization of (identified) temporal representations, as opposed to the identification and extraction of that information from textual reports. Our original use case was to support the comparison of date spans as commonly expressed in data records but it was subsequently realised that the normalization could also be applied to the outcomes of NLP information extraction (or NER), as discussed in section 7. However, it must be appreciated that the information automatically extracted from documents (even with the most advanced NLP systems) is probably less reliable than that derived from metadata or designated data fields, where some interpretative judgment has already been applied. Consideration should be given as to whether an indication of the provenance should be included as an additional attribute. We are currently investigating the application of NER to extract subject metadata from text. The integration of the temporal normalization techniques reported here with prior information extraction via NLP techniques forms part of this future work.

Section 3 outlines the broad aim of the temporal normalization techniques as supporting the comparison of diverse kinds of temporal expressions, via a common numerical time axis, for purposes of semantic integration. This integration involves the repurposing of legacy data and textual accounts often intended as self-contained, descriptive records. With the desire to integrate or combine datasets and increasingly textual reports for purposes of meta research comes the need to impose a systematic treatment and standardisation of data and metadata. However, it is important to bear in mind the original context (and provenance) of the dating information when aggregating dates deriving from different investigation strategies or dating methods. Similarly, periods may express a broader or more uncertain timespan than numeric date ranges, relying on intellectual judgment in the assignment of date ranges to periods. Care should be taken with purely automatic methods of subsequent analysis to ensure that the underlying data and investigation methods are comparable. Contextual metadata (paradata) should accompany integrated data collections, including the enrichment and integration strategies followed and version information for the conversion software (as well as original datasets) and the authority sources for periods and other vocabularies.

The first six of the temporal categories outlined in Section 4.2 afford general patterns and rules, broadly similar across the languages covered to date. However, named periods are a different case, requiring a lookup associating period labels with date ranges. We have drawn on standard period authorities for the languages covered. These tend to draw on vocabularies designed for information retrieval rather than NLP and may require adjustment and enrichment to reflect the range of terminology found in free text reports. Periods vary with geographical extent; the same period label string (e.g. 'Iron Age') often denotes a different time span in different countries and even regions. Additionally, the time spans associated with a period label may be revised by authorities, as scholarship develops over time. Thus, care should be taken when applying the date ranges associated with different period authorities. In future work, we intend to make the selection of period authorities a configurable parameter, specifying a particular period authority from PeriodO, a gazetteer of period definitions in different languages with authorities (and publication dates when available). Here each period is modelled as a spatio-temporal concept with a unique identifier, representing an intersection of subject, place and time, and parsing of period labels in the user interface yielding time spans as ISO 8601 years (Rabinowitz *et al*. 2016). Our approach extends the normalizing of period definitions to a wide range of categories of temporal expressions and languages, as found within data fields and NLP output generally. The use cases for such normalization include cleaning during interactive data entry and enriching existing legacy data for semantic integration and cross-search of multilingual datasets.

9 CONCLUSIONS

This paper shows how commonly recurring time-related textual patterns in different languages can be identified, matched, and normalized automatically via a software tool. Textual patterns for seven categories of temporal expression have been normalized: Ordinal named or numbered centuries; year spans; single year (with tolerance); decades; century spans; single year with prefix; named periods. Results from different datasets and languages have been presented that would have been a time-consuming exercise to achieve manually. The normalization of the seven categories of temporal expression is a novel contribution to the authors' knowledge. The normalized outputs are provided as additional attributes along with the original text expression. The tool is available as open source and consuming applications may select all or particular categories of temporal expressions appropriate for their purposes.

Time is a key dimension that must be addressed in the integrated data (and report) archives emerging from the application of FAIR and Open Data principles. Search via period names is important although carrying a specific semantics due to the inter-connection of temporal, geographical and cultural aspects. The aggregation of multilingual data presents an additional level of complexity given the wide variety of temporal expressions. A common numerical timeline is a pre-requisite for effective semantic integration of archaeological data. The example scenarios presented in Section 7.3 illustrate the potential of such a timeline for interrogation of an integrated archive of diverse datasets and archaeological reports in different languages.

The flexibility afforded by free text data entry accumulates a technical debt that must eventually be repaid by normalizing the data at a later stage. Automatically resolving and normalizing expressions of timespans, dates and periods to a common numerical timeline facilitates the semantic integration that makes possible efficient data interchange, cross search, comparison and synthetic research.

ACKNOWLEDGEMENTS

Parts of this work were supported by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRASTRUCTURES-2012-1-313193 (the ARIADNE project) and the H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914 (the ARIADNEplus project). Thanks are due to project partners generally and to the reviewers for helpful comments. The views and opinions expressed in this article are the sole responsibility of the authors.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Ceri Binding orcid.org/0000-0002-6376-9613
Hypermedia Research Group, University of South Wales, GB
Douglas Tudhope orcid.org/0000-0002-5222-0430
Hypermedia Research Group, University of South Wales, GB

REFERENCES

AAT. 2022. *Getty Art & Architecture Thesaurus*. Getty Research institute. Available at https://www.getty.edu/research/tools/vocabularies/obtain/download.html [Last accessed 25 Feb 2023].

ADS Archives. 2023. *Archive of Archaeology Data Service*. Available at https://archaeologydataservice.ac.uk/archive/archives.xhtml [Last accessed 25 Feb 2023].

ARIADNE Demonstrator. 2017. *ARIADNE Data Integration Demonstrator*. Available at http://ariadne-lod.isti.cnr.it [Last accessed 25 Feb 2023].

ARIADNEplus. 2022. *ARIADNEplus H2020 project*. Available at https://ariadne-infrastructure.eu [Last accessed 25 Feb 2023].

Aloia, N, et al. 2017. Enabling European Archaeological Research: The ARIADNE E-Infrastructure. *Internet Archaeology,* 43. DOI: https://doi.org/10.11141/ia.43.11

Bailey, G. 2007. Time perspectives, palimpsests and the archaeology of time. *Journal of Anthropological Archaeology,* 26(2): 198–223. DOI: https://doi.org/10.1016/j.jaa.2006.08.002

Binding, C. 2010. Implementing archaeological time periods using CIDOC CRM and SKOS. In: Aroyo, L, et al. (eds.), *Proc. 7th Extended Semantic Web Conference, Heraklion, Part I, Lecture Notes in Computer Science 6088*. Springer-Verlag, pp. 273–287. Available from https://link.springer.com/book/10.1007/978-3-642-13486-9 [Last accessed 25 Feb 2023]. DOI: https://doi.org/10.1007/978-3-642-13486-9_19

Binding, C. 2023. *Yearspans sourcecode repository*. Available at https://github.com/cbinding/yearspans [Last accessed 25 Feb 2023].

Binding, C, Charno, M, Jeffrey, S, May, K and **Tudhope, D.** 2015. Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data. *International Journal on Semantic Web and Information Systems*, 11(1): 1–29. DOI: https://doi.org/10.4018/IJSWIS.2015010101

Binding, C, Tudhope, D and **Vlachidis, A**. 2019. A study of semantic integration across archaeological data and reports in different languages. *Journal of Information Science*, 45(3): 364–386. DOI: https://doi.org/10.1177/0165551518789874

Boasson, W and **Visser, R.** 2017. SIKB0102: Synchronizing Excavation Data for Preservation and Re-Use. *Studies in Digital Heritage,* 1(2): 206–224. DOI: https://doi.org/10.14434/sdh.v1i2.23262

Brandsen, A, Lambers, K, Verberne, S and **Wansleeben, M.** 2019. User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. *Journal of Computer Applications in Archaeology,* 2(1): 21–30. DOI: https://doi.org/10.5334/jcaa.33

Brandsen, A and **Lippok, F.** 2021. A burning question – Using an intelligent grey literature search engine to change our views on early medieval burial practices in the Netherlands. *Journal of Archaeological Science,* 133: 105456. DOI: https://doi.org/10.1016/j.jas.2021.105456

Castleford, J. 1992. Archaeology, GIS, and the Time Dimension: an Overview, In: Lock, G and Moffett, J (eds.), *Proc. Computer Applications and Quantitative Methods in Archaeology*. Tempus Reparatum, Oxford, 1991, pp 95–106. Available from https://proceedings.caaconference.org/files/1991/13_Castleford_CAA_1991.pdf [Last accessed 25 Feb 2023].

Cunningham, H, Tablan, V, Roberts, A and **Bontcheva, K.** 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2). DOI: https://doi.org/10.1371/journal.pcbi.1002854

Daly, P and **Lock, G.** 1999. Timing is Everything: Commentary on Managing Temporal Variables in Geographic Information Systems. In: Barceló, J, Briz, I and Vila, A (eds.), *New Techniques for Old Times. Proc. 26th Conference on Computer Applications and Quantitative Methods in Archaeology*. Barcelona, 1998. Archaeopress, Oxford, pp 287–296. Available from https://proceedings.caaconference.org/files/1998/45_Daly_Lock_CAA_1998.pdf [Last accessed 25 Feb 2023].

ECMA. 2011. *ECMAScript® Language Specification*. Available at http://ecma-international.org/ecma-262/5.1/#sec-15.9.1.1 [Last accessed 25 Feb 2023].

Edvardsson, J, Hansson, A, Sjölander, M, von Boer, J, Buckland, P, Linderson, H, Gunnarson, B, Linderholm, H, Drobyshev, I and **Hammarlund, D.** 2022. Old Wood in a New Light: An Online Dendrochronological Database. *International Journal of Wood Culture* (published online ahead of print 2022). DOI: https://doi.org/10.1163/27723194-bja10009

Erfgoedthesaurus. 2023. *Archeologisch Basis Register*. Available at https://thesaurus.cultureelerfgoed.nl [Last accessed 25 Feb 2023].

Evans, T. 2015. A reassessment of archaeological grey literature: semantics and paradoxes. *Internet Archaeology*, 40. DOI: https://doi.org/10.11141/ia.40.6

Felicetti, A, Williams, D, Galluccio, I, Tudhope, D and **Niccolucci, F.** 2018. NLP Tools for Knowledge Extraction from Italian Archaeological Free Text. In: Addison, A and Thwaites, H (eds.), *Proc. 3rd Digital Heritage International Congress*. Institute of Electrical and Electronics Engineers, Digital Heritage 2018, San Francisco. DOI: https://doi.org/10.1109/DigitalHeritage.2018.8810001

FISH. 2023. FISH Periods List. Available from http://www.heritage-standards.org.uk/chronology [Last accessed 25 Feb 2023].

Freitag, R. 1995. The Battle of the Centuries – A List of References. Available from https://www.loc.gov/rr/scitech/battle.html [Last accessed 25 Feb 2023].

Grandi, F. 2006. XML Representation and Management of Temporal Information for Web-Based Cultural Heritage Applications. *Data Science Journal*, 1: 68–83. DOI: https://doi.org/10.2481/dsj.1.68

Green, C. 2011. It's about Time: Temporality and Intra-Site GIS. In: Jerem, E, Redő, F and Szeverényi, V (eds.), *On the Road to Reconstructing the Past. Proc. 36th International Conference Computer Applications and Quantitative Methods in Archaeology (CAA)*. Budapest, 2008. Archeaeolingua, pp 206–211. Available at https://proceedings.caaconference.org/files/2008/CD28_Green_CAA2008.pdf [Last accessed 25 Feb 2023].

HeritageData. 2023. *HeritageData Vocabularies*. Available at https://www.heritagedata.org/blog/vocabularies-provided [Last accessed 25 Feb 2023].

ICU. 2023. *ICU Documentation. Universal Time Scale*. Available at https://unicode-org.github.io/icu/userguide/datetime/universaltimescale.html [Last accessed 25 Feb 2023].

ISO. 2019a. Date and time — Representations for information interchange — Part 1: Basic rules. *Standard ISO 8601–1: 2019*. Available at https://www.iso.org/standard/70907.html [Last accessed 25 Feb 2023].

ISO. 2019b. Date and time — Representations for information interchange — Part 2: Extensions. *Standard ISO 8601–2: 2019*. Available at https://www.iso.org/standard/70908.html [Last accessed 25 Feb 2023].

Johnson, I. 2002. Contextualising Archaeological Information Through Interactive Maps. *Internet Archaeology*, 12. DOI: https://doi.org/10.11141/ia.12.9

Johnson, I. 2003. Integrating Databases With Maps: The Delivery Of Cultural Data Through TimeMap. In: Bearman, D and Trant, J (eds.), *Proc. Museums and the Web*. Charlotte. Available at https://www.museumsandtheweb.com/mw2003/papers/johnson/johnson.html [Last accessed 25 Feb 2023].

Kansa, E. 2014. Open Context and Linked Data. *ISAW Papers* 7(10), In: Elliott, T, Heath, S and Muccigrosso, J (eds.), *Current Practice in Linked Open Data for the Ancient World*, Available at http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/kansa [Last accessed 25 Feb 2023].

Lee, E. 2017. "Knowledge was their treasure": applying KO approaches to archaeological research. *Knowledge Organization*, 44(8): 644–655. DOI:https://doi.org/10.5771/0943-7444-2017-8-644

Microsoft NET. 2023. *Datetime Struct (.NET)*. Available at https://learn.microsoft.com/en-us/dotnet/api/system.datetime [Last accessed 25 Feb 2023].

Microsoft SQL. 2022. *Datetime (SQL Server)*. Available at https://docs.microsoft.com/en-us/sql/t-sql/data-types/datetime-transact-sql [Last accessed 25 Feb 2023].

Moffett, J and **Webb, R.** 1982. Database Management Systems and Radiocarbon Dating. In: Laflin, S (ed.), *Proc. Computer Applications in Archaeology Conference*. Birmingham, pp. 76–78. Available at https://proceedings.caaconference.org/files/1982/08_Moffett_Webb_CAA_1982.pdf [Last accessed 25 Feb 2023].

Niccolucci, F and **Richards, J.** (eds) 2019. *The ARIADNE Impact*. Hungary: Archaeolingua Foundation. DOI: https://doi.org/10.5281/zenodo.4319058

Nielsen, R. 2000. Calculating Java dates. *Infoworld*. Available at https://www.infoworld.com/article/2076270/calculating-java-dates.html [Last accessed 25 Feb 2023].

Newtonsoft. 2023. *Serializing Dates in JSON*. Available at https://www.newtonsoft.com/json/help/html/DatesInJSON.htm [Last accessed 25 Feb 2023].

Oracle Database. 2023. *Oracle Database Concepts, Date datatype*. Available at https://docs.oracle.com/cd/B28359_01/server.111/b28318/datatype.htm#CNCPT413 [Last accessed 25 Feb 2023].

Oracle Java. 2023. *java.time.Year*. Available at https://docs.oracle.com/javase/8/docs/api/java/time/Year.html [Last accessed 25 Feb 2023].

Postgres. 2023. *Postgres Date/time Types*. Available at https://www.postgresql.org/docs/current/datatype-datetime.html [Last accessed 25 Feb 2023].

PeriodO. 2023. *A gazetteer of period definitions for linking and visualizing data*. Available at http://perio.do/ [Last accessed 25 Feb 2023].

Renfrew, C and **Bahn, P.** 2020. *Archaeology: Theories, methods and practice*. London: Thames & Hudson.

Richards, J and **Hardman, C.** 2008. Stepping Back from the Trench Edge: an archaeological perspective on the development of standards for recording and publication. In: Greengrass, M and Hughes, L (eds.), *The Virtual Representation of the Past*. Farnham: Ashgate, pp. 101–112. Available at https://eprints.whiterose.ac.uk/7795/ [Last accessed 25 Feb 2023]. DOI: https://doi.org/10.4324/9781315551753-8

Rabinowitz, A, Shaw, R, Buchanan, S and **Golden, P.** 2016. Making sense of the ways we make sense of the past: the PeriodO project. *Bulletin of the Institute of Classical Studies,* 59(2). DOI: https://doi.org/10.1111/j.2041-5370.2016.12037.x

Roe, J. 2023. *Era software utility*. https://era.joeroe.io/ [Last accessed 25 Feb 2023].

Shaw, R. 2023. *EDTF in RDF/OWL*. Available at https://github.com/periodo/edtf-ontology [Last accessed 25 Feb 2023].

Shaw, R, Rabinowitz, A, Golden, P and **Kansa, E.** 2016. A sharing-oriented design strategy for Networked Knowledge Organization Systems. *International Journal on Digital Libraries*, 17(1): 49–61. DOI: https://doi.org/10.1007/s00799-015-0164-0

Signore, O, Bartoli, R, Fresta, G and **Marchetti, A.** 1997. Issues in historical geography. In: Bearman, D and Trant, J (eds.), *Proc. 4th International Conference on Hypermedia and Interactivity in Museums*. Paris, pp 32–37. Available at http://www.archimuse.com/publishing/ichim97/bartoli.pdf [Last accessed 25 Feb 2023].

Sugimoto, G. 2021. Building Linked Open Date Entities for Historical Research. In: Garoufallou, E and Ovalle-Perandones, M (eds.), *Metadata and Semantic Research – 14th International Conference, MTSR 2020,* Revised Selected Papers (pp. 323–335). Communications in Computer and Information Science, vol. 1355. Springer. DOI: https://doi.org/10.1007/978-3-030-71903-6_30

Tudhope, D, May, K, Binding, C and **Vlachidis, A.** 2011. Connecting archaeological data and grey literature via semantic cross search. *Internet Archaeology,* 30. DOI: https://doi.org/10.11141/ia.30.5

Vlachidis, A and **Tudhope, D.** 2022. A Method for Archaeological and Dendrochronological Concept Annotation using Domain Knowledge in Information Extraction. *International Journal of Metadata, Semantics and Ontologies,* 15(3): 192–203. DOI: https://doi.org/10.1504/IJMSO.2021.123042

Wilkinson, M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1504/IJMSO.2021.123042

TO CITE THIS ARTICLE:
Binding, C and Tudhope, D. 2023. Automatic Normalization of Temporal Expressions. *Journal of Computer Applications in Archaeology,* 6(1): 24–39. DOI: https://doi.org/10.5334/jcaa.105

Submitted: 24 November 2022 **Accepted:** 03 March 2023 **Published:** 27 March 2023