3

EXTRACTING HISTORICAL TIME PERIODS FROM THE WEB

In this chapter we present an automatic method for the extraction of time periods related to ontological concepts from the web. The method consists of two parts: an Information Extraction phase and a Semantic Representation phase. In the Information Extraction phase, temporal information about events that are associated with the target instance are extracted from web documents. The resulting distribution is normalized and a model is fitted to it. This distribution is then converted into a semantic representation in the second phase. We present the method and describe experiments where time periods for four different types of concepts are extracted and converted to a time representation vocabulary, based on the TIMEX2 annotation standard.

The first part of this chapter is based on a paper coauthored with Maarten van Someren and Bob Wielinga, "Extracting Art Style Periods from the Web" [de Boer et al., 2006c], which was presented at the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, Budva, Montenegro, June 12, 2006. A version of this chapter, coauthored with Maarten van Someren and Bob Wielinga, has been accepted for publication in the Journal of the American Society for Information Science and Technology (JASIST).

3.1 INTRODUCTION

Temporal knowledge is of great value for many types of information systems [Alonso et al., 2007]. Often the information need of a user is subject to temporal constraints (e.g. "retrieve all newspaper headlines concerning financial crises between the First and Second World War"). Systems that employ some form of temporal reasoning are capable of retrieving the right results for such queries. Especially in the cultural heritage domain, a number of concepts can be identified that have associated time periods such as art styles and artists. Temporal knowledge can here for instance be used in a system to verify relations (if Picasso is born in 1881 it is unlikely that he was associated with the 17th Century 'Baroque' art style). Other concepts that can also be enriched with temporal information include wars and general historical periods. Enriching the existing ontologies with temporal information will allow more elaborate temporal reasoning, retrieval and browsing in systems using these ontologies.

Such explicit temporal knowledge is not always present in the background information used by the system. Therefore methods that automatically enrich this background information with temporal information are needed. Most current temporal information extraction systems aim to identify specific temporal expressions in the natural language text and relate these to events [Mani, 2003]. These systems focus mainly on the English language news domain, where temporal information is very often and explicitly included in the documents. They are not designed for using temporal information to enrich ontologies.

In this chapter, we present an automatic approach for extracting temporal information for arbitrary concepts. More specifically, we extract the time period associated with a historical concept. We do not aim to extract this period from specific documents or from a certain corpus, but rather from the Web as a whole. In our approach we aggregate the temporal information from different documents to produce a description of the time period associated with a historical concept. Our temporal Information Extraction method that again exploits redundancy of information about a target time period for a concept. In a second phase, this temporal information is translated into semantic ontology constructs denoting the time period.

The aim of this investigation is to develop a language-independent method for the extraction of time periods, rather than a language-specific method. For this reason we avoid more elaborate patterns and use short, language-independent regular expressions to extract a large number of years from a larger working corpus, loosely describing the target instance. The extracted years are combined to determine the resulting time period of the target instance. This coarse method of extraction ensures that the method is language, structure- and domain-independent. Through the combination of redundant temporal information, we can compensate for the initial drop in precision resulting from the coarse methods.

A problem that exists for automatic ontology enrichment in general is how to translate extracted information into the predefined vocabulary of the target ontology. A lot of Information Extraction methods output the retrieved information with a degree of (un)certainty. At the same time, in most basic ontologies, thesauri or structured vocabularies, knowledge is represented in discrete form so that any statement can be determined as being either true or false. There exists a 'Semantic Gap' between the statistical and often inconsistent information in corpora and the certain and consistent knowledge of the target ontologies. In this chapter we show how the statistical model resulting from our temporal Information Extraction phase can be transformed to discrete time period representations that still capture the fuzzy nature of the extracted time period.

In Section 3.2, we describe the temporal ontology enrichment task and give an overview of the whole method. In Sections 3.3 and 3.4, we describe and evaluate the two different parts of the method. In Section 3.5 we discuss related work. The conclusions are presented in Section 3.6.

3.2 TASK AND METHOD

We first describe our specific temporal ontology enrichment task in more detail and then provide the general approach and our main assumptions for this task.

We have a partly populated ontology with a concept C that has a specific relation to a time interval (a time period). This period is represented in some way in the ontology. We have a number of subconcepts or instances of C: $i_1, i_2, ...$, each having one or more description labels. For each of these instances, the task is to extract from the web the correct period related to it and to add it to the ontology in the representation used by the target ontology. We make three main assumptions.

- In the ontology, there is only one period related to each instance. With the method described below, we can only extract one time period per instance. In some cases, when one related period falls within a second, larger related period, we can use a method parameter to identify either the former or the latter. In Section 3.3.5.4, we give an example of this.
- We can extract a relatively unambiguous working corpus from the Web using the instance's labels in the ontology. This working corpus must contain a significant amount of documents about the target instance and its associated events.
- For each period it is possible to extract moments within this interval (e.g. years, days or minutes) from the Web. We assume that C is characterized by a number of events that are associated with these moments (these events need not to be specified in the ontology). The granularity of the moments that need to be extracted depends on the length and nature of the time period. In the case of art styles, the events are creation dates of paintings, exhibitions or similar events, for which we can extract a single year from documents. For shorter periods, months, days or even hours or minutes might be used. We here focus only on larger periods and therefore extract only years to determine the time periods.

We identify two main phases in this temporal ontology enrichment task. The first is an Information Extraction phase in which temporal information is extracted from the web. In the second phase, this information is translated to the representation for time periods that is used in the target ontology. We therefore also split up our approach into two separate phases:

1. In the *Information Extraction phase*, we extract the distribution of individual moments through the use of simple regular expressions. From the results, we identify the years that are particularly correlated with the target instance i by normalizing the extracted frequencies through the use of a general distribution of these moments. On these normalized data points, we then fit a simple statistical model. This fitted model describes the time period for the target instance and is the output of the Information Extraction phase. In Section 3.3, we describe the Information Extraction phase in more detail and give a quantitative evaluation of this phase.

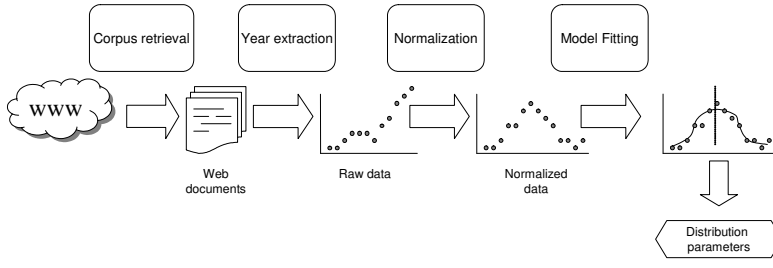

Figure 9: Outline of the Information Extraction phase

2. In the *semantic representation phase*, this model is then used to formalize the period in the form in which it is to be represented in the target ontology. The exact method of producing these representations from the extracted information is dependent on the chosen temporal representation. In Section 3.4, we describe a number of possible representations and elaborate on the methods to rewrite the extracted model to these representations. For a representation of time periods based on the TIMEX2 standard[Ferro et al., 2005], we provide a small set of transformation rules. We then evaluate the ontology enrichment process as a whole by generating representations in the TIMEX2 format and having them assessed by non-expert evaluators.

3.3 THE INFORMATION EXTRACTION PHASE

For the extraction of the periods, we do not employ domain- or structure-specific Information Extraction techniques such as the use of patterns or Natural Language Processing techniques. The effectiveness of these methods is often dependent on the specific domain and the structure of the documents in the corpus used. Instead, we use a very simple domain- and structure-independent method that extracts occurrences of moments (in our case years) from documents on the Web that are retrieved by querying the Google search engine with the labels of the concept that is to be enriched. We exploit the assumption that the distribution of moments in these documents is different from the distribution of moments in the whole of the Web and that we can find the target time period by comparing the two distributions. To further clarify this, we present a detailed description of the Information Extraction phase in the next section.

3.3.1 *Information Extraction method overview*

In Figure 9, we present the method used to extract the distribution of years.

To construct a working corpus from Web documents, we use the labels (using synonyms when available) of the instance that is to be enriched (the target

instance). For the retrieval step we use the search engine Google[1]. We construct a disjunctive search query by taking all labels of the instance and connecting them with Google's binary search operator "OR". This query is passed to Google and we download the first N documents as our working corpus. N is a parameter of the method, in the experiments described below, we used N=1000.

3.3.2 *Year extraction*

In the next step of the algorithm, we extract occurrences of moments from the documents using a single simple regular expression. As we have discussed earlier, we here only consider the case where the level of granularity is years, as we will be extracting periods spanning multiple years. We extract four consecutive integers, followed by a punctuation mark or a whitespace. In the experiments described below, we only extracted the years between 1000 and 2000 AD since we did not experiment with target instances with an expected time period before 1000 AD. For each of the extracted years, we denote the frequency of the occurrences in all documents of the working corpus. These frequencies make up the 'raw data' for the target instance.

This simple, coarse extraction rule will extract years that are not part of the target period such as years in copyright notices or publication years of books about a concept. In other words it will retrieve a lot of false positives. At the same time, the simple regular expression will miss more elaborate temporal declarations that more fine-grained methods might be able to extract. Of course, different additional patterns that extract temporal descriptions can be used to extract even more instances. Examples are patterns like "the NNth Century" or "the NNNNs". However, these more elaborate patterns will make the method more language-dependent and will require additional processing, making the method more complex. We assume however, that the use of a simple method and the use of a normalization procedure, combined with fitting an appropriate statistical model on the data will result in acceptable performance. We do not make any assumptions about the language or the structure of the documents, the nature of the target concepts or the domain. The exploitation of the redundancy of information on the Web through the use of this simple method on a large number of web documents makes the whole approach independent on the language and structure of the corpus.

3.3.3 *Normalization*

In the normalization step, we attempt to lower the frequencies of years that have a high frequency but do not belong to the target time period. Inspection of the raw data shows that besides years from the target period, very recent years (2005, 2006, etc.) and years that are 'round numbers' (1500, 1600, etc) generally have a very high frequency. These high frequencies can be considered as noise. The top left graph of Figure 10 shows an example frequency distribution of years for

1 http://www.google.com/

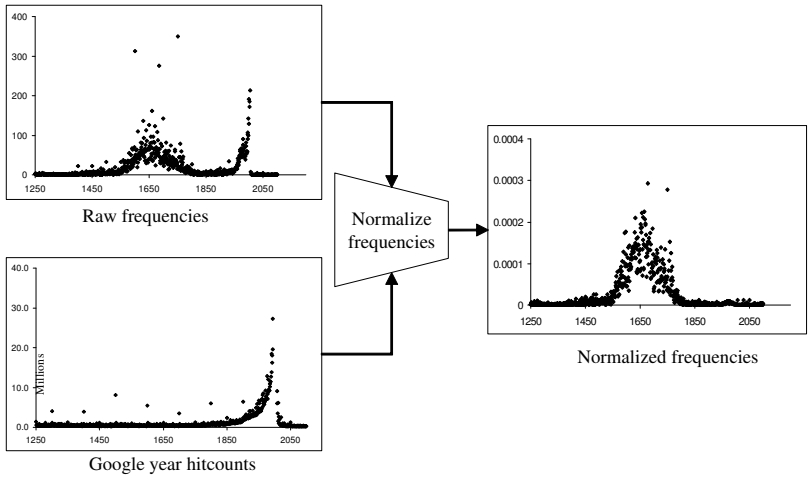

Figure 10: The normalization process of the raw frequency data for the example 'Baroque'

the concept 'Baroque': besides the higher frequencies for years around 1650, the noise can also be seen in the second peak.

We assume that this noise is evenly distributed among random web pages. We can therefore remove this noise by normalize the data by the distribution of years on the whole Web. For this we use the Google hit count distribution for the years. This distribution is shown in the bottom-left graph of Figure 10. As can be seen in this figure, the Google hit count distribution of years does indeed have the expected high peaks at recent years and round numbers.

We normalize the data by dividing the frequency of each year by the Google hit count for that year. This effectively filters out the noise peaks. The rightmost graph in Figure 10 shows the resulting distribution after the normalization step for the 'Baroque' example. The noise peak is removed and the correct high frequencies around the year 1650 are retained.

3.3.4 *Model fitting*

In the final step, we fit a model to the data using a numerical fitting procedure. Since we know that periods are sets of connected years, fitting an appropriate model to the data will eliminate any 'outliers' and provide us with a short description of the period (the optimal model parameters). To fit the model to the normalized data, we minimize the sum of square root errors between the data points and the model function value. To determine an appropriate model, we experimented with four different candidate models. All four models originate from an intuitive notion of how moments could be related to the periods.

Examples of these models for the target instance 'Baroque' are shown in Figure 11. The candidate models are:

- A 'block' model. The block model corresponds with the intuitive notion that time periods are discrete intervals with a start and an end year (e.g. a moment is either inside or outside the time period). The block model has three parameters: μ_b and σ_b, which correspond to center of the period and half of the period's length respectively and f_b, which is used to fit the height of the block function to the specific frequencies (see Figure 11(a)). The block model is defined by:

$$g_b(x) = \begin{cases} f_b & \text{if } \mu_b - \sigma_b \leqslant x \leqslant \mu_b + \sigma_b \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

- A normal distribution. This corresponds to the notion that time periods can be more or less 'fuzzy' and that the target instance has a temporal 'peak' and a vague start and end year (see Figure 11(b)). This model also has three parameters (μ_n and σ_n for the mean and standard deviation of the normal distribution and f_n, to again fit the height of the distribution to the frequencies) and is given by:

$$g_n(x) = f_n \cdot \text{norm}(\mu_n, \sigma_n) \tag{3.2}$$

- A 'triangular' model. This model also allows for more 'fuzzy' time periods. But unlike the previous model, the start and end of a period can have different 'slopes'. This can be used to also model time periods that have a gradual beginning but a more definite end or vice versa. This model has four parameters: μ_{f1} for the 'peak' of the period, $\sigma_{l,f1}, \sigma_{r,f1}$ determine the base of the left and right slopes respectively and f_{f1} is used to vary the height of the peak (see Figure 11(c)).

- A 'trapezoid' model, which is essentially the triangular model with the addition of a central period. With this most elaborate model, we are able to model time periods that have starts and ends with different fuzziness independent of the total length of the period. This trapezoid model has five parameters: μ_{f2} for the center of the central period, $\sigma_{c,f2}$, which determines the length of the central period (in the same way as the sigma parameter in the block model, $\sigma_{l,f2}$ and $\sigma_{r,f2}$ which are used to determining the base of the left and right slope respectively and finally the f_{f2} parameter to determine the height of the central plateau (see Figure 11(d)).

To find out which of these models produced the best result, we selected ten art styles[2], extracted the frequency data and fitted the models to this data. The fitting procedure was done using MS Excel's LP-solver, minimizing the sum of the Squared Errors between the model function value and the normalized

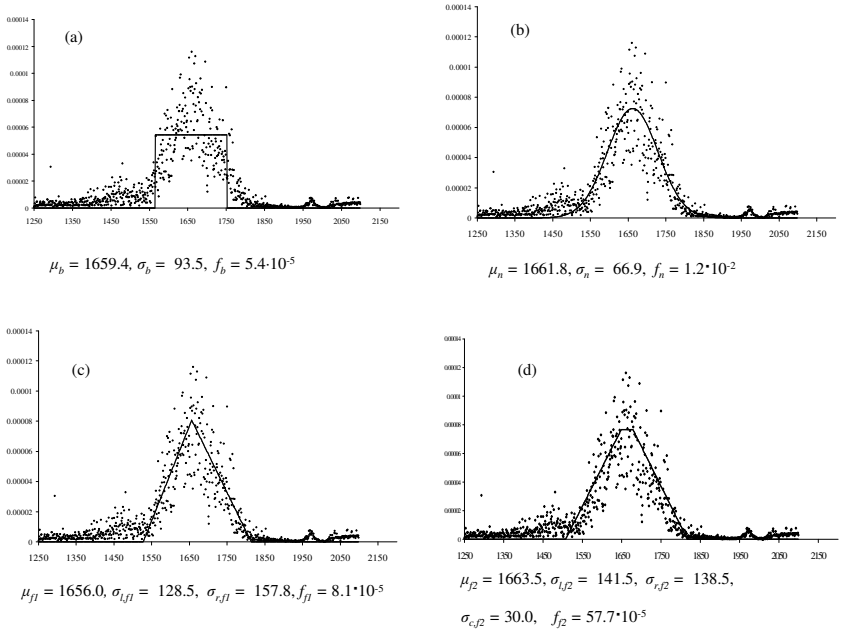

Figure 11: The best fits for the block model(a), the normal distribution (b), triangular (c) and trapezoid (d) models on the frequency data for the instance 'Baroque' (which approximately lasted during the 17th Century). For each model, the optimal parameter values are shown.

Table 14: Averages of the Sum of Squared Errors for the four tested models.

Block model	4.18
Normal distribution	2.70
Triangular model	2.99
Trapezoid model	2.77

frequency data. In Figure 11, we show the best fitting models for the data of a single instance, the art style 'Baroque'.

In Table 14, we list the averages of the Sum of Squared Error for the ten art styles. The relatively high error for the Block model is in contrast with the initial intuition that most time periods are discrete and have a definite start and end date. When we examine the web pages from the working corpus, we found that many time periods are indeed very fuzzily defined. The exact periods of many of the cultural heritage concepts or historical periods are often under discussion and do not have discrete start and end dates. This holds even for wars. For example, proposed starting years for World War II include 1931, 1937 and 1939[3].

Of the three fuzzy models, the normal distribution has the best average errors although the differences are very small. The models' ability to accurately model the data is about equal. However, considering that the normal distribution is with only three parameters the simplest of the three, we decided to use the normal distribution as the model used in the fitting step of the Information Extraction phase. The output of the Information Extraction phase therefore consists of the values of two of the model parameters resulting from the best fit: μ and σ. The value for the third parameter, the factor f is discarded, since it only indicates the number of years found and not the distribution across the time line.

3.3.5 *Experiments*

To evaluate the performance of the Information Extraction phase, we performed experiments where we extracted time periods for four different concepts with historical time periods: *art styles*, *historical periods*, *wars* and *artists*. In each of the following sections, we describe the extraction phase, the evaluation and the results for these concepts.

3.3.5.1 *Art styles*

We chose to use art styles defined in the Art and Architecture Thesaurus (AAT) [The Getty Foundation, 2000a], a cultural heritage vocabulary used in the MultimediaN E-Culture tool [Schreiber et al., 2008]. We manually selected 20 instances of the ART STYLE concepts from the AAT to be enriched with a time period. The 20 selected art styles include the ten art styles from Section 3.3.4. For each of

2 Art Deco, Art Nouveau, Baroque, Cubist, Dada, Expressionism, Impressionism, Neo-Impressionism, Neue Sachlichkeit and Surrealism

3 found on one web page: http://en.wikipedia.org/wiki/World_War_II, retrieved May 2006

these art styles, we retrieved a working corpus, extracted the year occurrences and fitted the normal distribution to the data.

Of the 20 art styles, we discarded two after the fitting phase since the resulting model fit was very obviously unfeasible. This bad fitting model had different causes for the two art styles. The art style `aat:Symbolist` did not produce a corpus with pages about the art style Symbolism. Instead, the extracted corpus consisted mainly of documents describing other concepts than the art style, for example cultural symbolism, symbolism in dreams and so on. Here, the concept labels ('Symbolism' and 'Symbolist') are clearly not specific enough to extract a good corpus. The ambiguous corpus leads to an 'unfocused' set of extracted years, which led to an obvious low quality fit of the model to the data. The second problematic art style, 'Neue Sachlichkeit', produced a bad fit because of the very low amount of web pages with years about that art style on the Web. This again led to an unfocused distribution of extracted years and a very bad model fit. In both cases, the quality of the corpus could be improved by query expansion techniques using either concepts higher up in the taxonomic structure of the ontology or by consulting other sources such as WordNet [Fellbaum, 1998]. In our experiment, we manually removed these two art styles, but an automated approach to this filtering would be relatively easy to implement, since the resulting model fit is very obviously unfeasible in the sense that it either has a very high error or it has extreme values for the model parameters.

The results for the remaining 18 AAT styles can be found in Table 15. The distributions found for each art style are listed in terms of μ and σ in the second and third column of that table.

Evaluation For the evaluation of the found distributions we constructed a gold standard. We consulted the art style pages of six different encyclopedic web sites[4]. From these pages, we manually extracted the period of the art style as determined in those documents. However, for most art styles, the web sites do not provide clear start and end dates. Usually only vague indications are given. As an example: The period of Baroque was noted as 'The style started around 1600,...' (www.wikipedia.org) and '... originated in Rome at the beginning of the 17th century...' (www.artchive.org). To quantitatively evaluate our method we need clear start and end dates. We therefore used a fixed set of rules for rewriting these vague notions to clear start and end dates. For instance the phrase 'started at the beginning of the 17th century' was interpreted as 'has start date 1600' etc. If a page did not list any duration for an art style, we did not consider that page for the art style. For each art style, we took the average of the start and end dates collected in this way as our 'gold standard'.

We then compared our distribution to this gold standard in two ways. We firstly compared the found mean μ to the gold standard mean. Secondly we compared the period length, determined by the found σ, to the start and end dates of the period. For readability, these two evaluations are shown in separate tables.

First, we compared for each art style the found mean, μ, to the mean of the start and end date of the gold standard. In Table 15, we list this gold standard

4 www.wikipedia.org, www.artcyclopedia.com, www.artlex.com, www.artchive.com, www.encyclopedia.com and www.britannica.com, all retrieved May 2006

Table 15: The 18 Art Styles, with the values for μ and σ that produced the best fit and evaluation of μ to the gold standard (GS) means

art style	μ	σ	GS: mean	error	rel. error
Abstract Exp.	1951.0	14.5	1955.0	-4.00	0.22
Art Deco	1928.1	9.3	1929.6	-1.44	0.08
Art Nouveau	1893.0	22.9	1894.1	-1.15	0.05
Baroque	1662.2	63.0	1663.1	-0.90	0.01
Bauhaus	1923.4	12.1	1926.0	-2.60	0.19
Counter-Reformation	1569.6	58.1	1576.4	-6.74	0.07
Cubist	1910.7	5.0	1914.3	-3.58	0.26
Dada	1917.5	3.5	1919.7	-2.13	0.29
Expressionist	1912.1	31.3	1921.8	-9.64	0.32
Fauve	1906.2	2.3	1905.8	0.44	0.06
Impressionist	1875.4	26.2	1877.2	-1.86	0.07
Mannerist	1552.6	53.2	1559.6	-6.96	0.10
Neo-Impressionist	1886.0	4.7	1885.5	0.46	0.07
Post-Impressionism	1881.2	31.5	1888.2	-6.95	0.30
Pre-Raphaelites	1857.9	31.9	1867.2	-9.28	0.25
Reformation	1541.4	33.9	1547.6	-6.22	0.08
Rococo	1738.7	41.0	1750.1	-11.31	0.14
Surrealist	1928.3	18.2	1937.0	-8.72	0.34
				average:	0.16

mean and the error. We also list the error relative to the total length of the period (according to our gold standard). This error measure corresponds to the intuition that it is less severe to make an error of one year for longer periods than it is for shorter ones. Averaged over all art styles, the error between the mean found in our Information Extraction phase and the mean from our gold standard is 16% of the total period length.

To evaluate the values of σ (the spread of the years associated with the art styles), we also needed to construct a strict start and end date from this value and compare it to the start and end dates of the gold standard. For this purpose, we introduce a factor τ that we use to provide a single start and end date for each extracted period of the art style: respectively $\mu - (\tau \cdot \sigma)$ and $\mu + (\tau \cdot \sigma)$.

This τ is a parameter of the method. To find the optimal value for τ in this experiment, we again used a numeric optimization procedure to minimize the error between start and end dates constructed using this factor and those from the gold standard. This optimal value of τ found for each art style ranged from 0.5 to 1.1 with an average of 0.81. In Table 16, we show the start and end dates we obtained with $\tau = 0.81$. These start and end years were compared to the start and end dates from the gold standard. For each art style, we calculated the average of the absolute errors for both dates. As with the means, we also calculated this error relative to the total length of the art style according to our gold standard. These values are also shown in Table 16. Averaged over all art styles the relative

Table 16: Start and end dates for the 18 Art Styles ($\tau = 0.81$) compared to the gold standard (GS)

art style	start	end	GS: start	GS: end	error	rel. error
Abstract Exp.	1939.3	1962.7	1946.0	1964.0	4.00	0.22
Art Deco	1920.6	1935.7	1920.2	1939.0	1.90	0.10
Art Nouveau	1874.4	1911.6	1882.0	1906.3	6.45	0.27
Baroque	1611.1	1713.3	1593.0	1717.0	10.90	0.09
Bauhaus	1913.6	1933.2	1919.0	1933.0	2.80	0.20
Counter-Reformation	1522.5	1616.8	1528.8	1624.0	6.74	0.07
Cubist	1906.6	1914.7	1907.3	1921.2	3.58	0.26
Dada	1914.7	1920.4	1916.0	1923.3	2.13	0.29
Expressionist	1886.8	1937.5	1906.5	1937.0	10.11	0.33
Fauve	1904.4	1908.0	1902.0	1909.5	1.92	0.26
Impressionist	1854.1	1896.6	1864.2	1890.3	8.21	0.32
Mannerist	1509.5	1595.7	1526.2	1593.0	9.70	0.15
Neo-Impressionist	1882.1	1889.8	1882.0	1889.0	0.46	0.07
Post-Impressionism	1855.7	1906.7	1876.7	1899.7	14.00	0.61
Pre-Raphaelites	1832.0	1883.7	1848.3	1886.0	9.28	0.25
Reformation	1513.9	1568.8	1508.4	1586.8	11.73	0.15
Rococo	1705.5	1772.0	1709.3	1790.8	11.31	0.14
Surrealist	1913.5	1943.1	1924.0	1950.0	8.72	0.34
					average:	0.23

error is 0.23, indicating that on average, the start and end dates differ by 23% of the total length from the gold standard.

3.3.5.2 *Historical Periods*

For our second experiment, we chose 20 historical periods, which we selected from the Wikipedia page about historical periods[5]. This page also provided us with a gold standard to which our results can be evaluated. After the model fitting step, we removed five of the periods since they showed a very bad fitting model with extreme values for μ and σ. Again, this was caused by either ambiguous concept labels ('Age of Discovery', 'Early Modern Period' and 'Modern Era') or not enough relevant documents that could be retrieved ('Qing Dynasty' and 'Mughal Empire'). Of the 15 remaining historical periods, we again show first the results of comparing the found means to the gold standard means in Table 17. We optimized the value for τ over all historical periods. For this experiment, we found that a $\tau = 0.89$ produced the overall best match to the gold standard. This value is relatively close to the optimal τ found for art styles in the previous section. In Table 18, we show the results of evaluating the found σ with respect to the gold standard using this value for τ.

5 http://en.wikipedia.org/wiki/List_of_time_periods, retrieved May 2006

Table 17: The 15 historical periods, with the values for μ and σ that produced the best fit and evaluation of μ to the gold standard means

hist. period	μ	σ	GS: mean	error	rel. error
Edwardian period	1906.7	15.3	1905.5	1.22	0.14
Elizabethan period	1585.7	37.4	1580.5	5.17	0.11
Georgian Era	1806.7	67.2	1772.0	34.71	0.30
Industrial Revolution	1809.8	63.5	1799.5	10.27	0.05
Interwar period	1930.9	13.3	1928.5	2.44	0.12
Jacobean Era	1628.3	25.4	1614.0	14.27	0.65
Late Middle Ages	1377.6	131.5	1400.0	-22.43	0.11
Machine Age	1929.2	31.4	1922.5	6.72	0.15
Meiji era	1879.5	27.4	1890.0	-10.51	0.24
Napoleonic Era	1809.1	8.4	1807.0	2.06	0.13
The Age of Enlightenment	1769.7	50.7	1749.5	20.22	0.20
The Protestant Reformation	1535.6	25.6	1549.5	-13.92	0.14
The Renaissance	1529.2	78.4	1449.5	79.72	0.27
Tokugawa shogunate	1763.8	192.4	1735.5	28.33	0.11
Victorian era	1863.8	32.6	1869.0	-5.16	0.08
				average:	0.19

Table 18: Start and end dates for the 15 historical periods ($\tau = 0.89$) compared to gold standard

hist. period	start	end	GS: start	GS: end	error	rel. error
Edwardian period	1892.1	1921.3	1901	1910	12.84	1.00
Elizabethan period	1550.0	1621.3	1558	1603	12.95	0.23
Georgian Era	1742.7	1870.7	1714	1830	51.57	0.30
Industrial Revolution	1749.3	1870.3	1700	1899	70.70	0.22
Interwar period	1918.3	1943.6	1918	1939	3.08	0.12
Jacobean Era	1604.1	1652.5	1603	1625	15.75	0.65
Late Middle Ages	1252.4	1502.8	1300	1500	41.38	0.11
Machine Age	1899.3	1959.1	1900	1945	7.55	0.15
Meiji era	1853.4	1905.6	1868	1912	16.79	0.24
Napoleonic Era	1801.0	1817.1	1799	1815	3.39	0.13
The Age of Enlightenment	1721.5	1818.0	1700	1799	32.84	0.20
The Protestant Reformation	1511.2	1560.0	1500	1599	33.57	0.27
The Renaissance	1454.5	1603.9	1300	1599	160.80	0.27
Tokugawa shogunate	1580.6	1947.1	1603	1868	40.56	0.14
Victorian era	1832.8	1894.9	1837	1901	6.04	0.08
					average:	0.23

3.3.5.3 *Wars*

For the third experiment we selected 22 wars from the Wikipedia page listing a number of wars[6]. Again, this also provided us with a gold standard. To test the range of our extraction method, we also included a number of wars with a 'Before Christ' period. For these wars, we used a different regular expression to extract the years: three consecutive integers, followed by a whitespace, followed by "BC". This introduces a language-specific element into the extraction procedure since different languages have different ways of denoting these 'B.C.' dates. For these wars, we extracted all B.C. dates between 1500 B.C. and 1 AD. The raw years were normalized by their Google hitcounts as normal. We discarded a total of seven wars after the model fitting step for the reasons described in the previous sections. This leaves us with 15 wars, three of which are 'B.C. wars'. We show the results of the evaluation of the means in Table 19. The discrepancy between our found mean and the gold standard mean is on average 19% of the length of the war time period.

We optimized the value for τ for the A.D. and B.C. wars separately. The global optimal value for A.D. wars was established at 0.72 whereas that of the B.C. wars was found to be 0.78. Both values are slightly lower than the values found in the previous experiments. In Table 20, we show the results of evaluating the found σ with respect to the gold standard using these values for τ for all 15 wars. The average error for the found start and end moments is 26% of the length of the war.

3.3.5.4 *Artists*

The goal of the last experiments is to extract time periods for artists. Here we can identify the problem that the first assumption from Section 3.2, that there is only one period related to the instance is not satisfied. The time period for a historical artist could refer to his or her life period but it might also refer to the period he or she was active as an artist. The years that will be extracted will probably be associated with events that occur during the active period of the artist. In general, when extracting the time period of a historical person, we will extract years that occur in the period of the life during which most of that person's important events will occur. In general, these events will occur in the middle of the person's life and will probably not involve the first or last 20 years or so of a persons life. We therefore can expect a higher optimal value for τ than the previously found values.

We manually chose 15 well-known artists from the Union List of Artist Names (ULAN) [The Getty Foundation, 2000c], of which we had to discard three after the model fitting step. For evaluation purposes, we assumed that the extracted period corresponded to the life span of the artist. We constructed a gold standard based on the artists' pages from Wikipedia. In Table 21 we show the results of evaluating the means.

As was expected, the optimal value of τ is indeed significantly larger than in the previous experiments: 1.42. This indicates that, compared to the other three

6 http://en.wikipedia.org/wiki/List_of_wars, retrieved May 2006

Table 19: The 15 wars, with the values for μ and σ that produced the best fit and evaluation of μ to the gold standard means

war	μ	σ	GS: mean	error	rel. error
A.D. wars					
Eighty Years' War	1593.0	67.4	1608	-15.02	0.19
Hundred Years' War	1379.0	61.3	1395	-16.01	0.14
Hussite Wars	1424.3	11.4	1428	-3.74	0.23
Italian Wars	1524.1	43.2	1527	-2.43	0.04
Napoleonic Wars	1810.7	5.1	1809	1.75	0.15
Seminole Wars	1842.8	24.6	1838	5.25	0.13
Thirty Years' War	1631.9	17.5	1633	-1.15	0.04
Uruguayan Civil War	1850.6	31.7	1845	5.65	0.47
Vietnam War	1967.0	4.7	1967	0.01	0.00
War of the Roses	1467.1	20.4	1470	-2.95	0.10
World War I	1916.5	2.4	1916	0.45	0.11
World War II	1942.4	2.6	1942	0.40	0.07
B.C. wars					
Persian Wars	491.9 BC	9.6	489 BC	-2.92	0.15
Punic Wars	228.8 BC	37.7	205 BC	-23.77	0.20
Syrian Wars	262.8 BC	68.6	237 BC	-25.84	0.35
				average:	0.19

Table 20: Start and end dates for the 12 A.D. wars ($\tau = 0.72$) and 3 B.C. wars ($\tau = 0.78$) compared to gold standard

war	start	end	GS: start	GS: end	error	rel. error
A.D. wars						
Eighty Years' War	1544.7	1641.3	1568	1648	26.67	0.19
Hundred Years' War	1335.1	1422.9	1337	1453	16.97	0.14
Hussite Wars	1416.1	1432.4	1420	1436	5.67	0.23
Italian Wars	1493.1	1555.0	1494	1559	2.88	0.04
Napoleonic Wars	1807.1	1814.4	1803	1815	4.38	0.19
Seminole Wars	1825.1	1860.4	1817	1858	9.32	0.13
Thirty Years' War	1619.3	1644.4	1618	1648	3.12	0.08
Uruguayan Civil War	1827.9	1873.4	1839	1851	22.26	1.39
Vietnam War	1963.7	1970.3	1959	1975	7.01	0.29
War of the Roses	1452.5	1481.6	1455	1485	4.22	0.10
World War I	1914.7	1918.2	1914	1918	0.80	0.11
World War II	1940.5	1944.3	1939	1945	1.92	0.19
B.C. wars						
Persian Wars	499.4 BC	484.4 BC	499 BC	479 BC	3.13	0.15
Punic Wars	258.2 BC	199.3 BC	264 BC	146 BC	32.41	0.25
Syrian Wars	316.4 BC	209.2 BC	274 BC	200 BC	47.05	0.35
					average:	0.26

Table 21: The 12 artists, with the values for μ and σ that produced the best fit and evaluation of μ to the gold standard means

artist	μ	σ	GS: mean	error	rel. error
El Greco	1587.3	26.0	1577.5	9.83	0.13
Frans Hals	1631.2	32.8	1624.5	6.71	0.08
Johannes Vermeer	1663.6	7.7	1653.5	10.07	0.23
Leonardo da Vinci	1496.1	21.1	1485.5	10.59	0.16
Henri Matisse	1907.1	10.4	1911.5	-4.41	0.05
J.E. Millais	1861.3	26.0	1862.5	-1.19	0.02
Claude Monet	1876.5	15.2	1883.0	-6.52	0.08
Otto Dix	1921.9	8.4	1930.0	-8.11	0.10
Pablo Picasso	1909.3	20.0	1927.0	-17.68	0.19
Paul Klee	1922.7	21.1	1909.5	13.17	0.22
Rembrandt van Rijn	1641.4	18.4	1638.0	3.45	0.06
Georges Seurat	1885.8	4.1	1875.0	10.80	0.34
				average:	0.14

concepts, for artists, the found years are located more closely to the mean. This results in a narrow optimal normal distribution with a relatively low value for σ. This is compensated by higher value of τ, when comparing it to the 'lifespan' gold standard. In Table 22, we show the resulting start and end years using this value of τ and compare them to the gold standard for each of the 12 artists.

When the gold standard does not concern the lifespan of the artist but rather his 'active period' we can assume that the extracted dates that correspond to events concerning the artist are indeed distributed more like the previous concepts. To test this, we also fitted the τ parameter to an 'active period' gold standard. For this gold standard we assumed that the first and last 20 years of the artist's life are not a part of his active period (e.g. the active period for Frans Hals is 1603-1646). With this gold standard, the best value for τ was found to be 0.74, which is indeed comparable to the best τ values of the art styles, historical periods and wars.

3.3.5.5 *Discussion*

In the experiments for the four concepts, the errors found are relatively low and especially if we consider the artificial nature of the gold standards we can say that the performance is acceptable. Determining the 'middle' of a target time period proved to be more successful than extracting the start and end-times as the relative errors of the means were lower than the relative errors of the edges of the period for all four concepts. This can partly be explained by the quality of the gold standard. The start- and end-years of a lot of the concepts are more vague than we can express in the gold standard. As we have already argued in Section 3.3.4, even concepts that appear to have clearly defined periods will on closer inspection be much more vague. A second reason is that for determining the start and end years, we include a τ parameter that is the same for all instances of a concept. For certain instances, this value of τ might be too small (when the events

Table 22: Start and end dates for the 12 artists compared to gold standard ($\tau = 1.42$)

artist	start	end	GS: start	GS: end	error	rel. error
El Greco	1550.3	1624.4	1541	1614	9.83	0.13
Frans Hals	1584.4	1678.0	1583	1666	6.71	0.08
Johannes Vermeer	1652.6	1674.5	1632	1675	10.55	0.25
Leonardo da Vinci	1466.0	1526.2	1452	1519	10.59	0.16
Henri Matisse	1892.3	1921.9	1869	1954	27.68	0.33
J.E. Millais	1824.2	1898.4	1829	1896	3.59	0.05
Claude Monet	1854.9	1898.1	1840	1926	21.39	0.25
Otto Dix	1909.9	1933.9	1891	1969	26.98	0.35
Pablo Picasso	1880.8	1937.8	1881	1973	17.68	0.19
Paul Klee	1892.5	1952.8	1879	1940	13.17	0.22
Rembrandt van Rijn	1615.3	1667.6	1607	1669	4.83	0.08
Georges Seurat	1880.0	1891.6	1859	1891	10.80	0.34
					average:	0.20

with which the extracted years are associated occur mainly in the center of the period) or too large (when the years are more spread out over the period.

The fact that the optimal value for τ and the relative average errors vary very little between the art style, historical periods, wars and active periods of artists show that the performance of the method does not vary very much across these tasks. For these types of concepts, we have determined that a τ of around 0.8 produces good results.

The different optimal value for τ (1.42) in the 'lifespan of artists' example can be explained by considering the type of years that appear in the extracted corpora for the artists. Most probably the years found will co-occur with the active period of an artist. Using the same value for τ as in the other two examples will not indicate the entire lifespan of the artists but rather his or her active period. In general, we can expect this to be true for people in general: most life-defining events will not occur in the first few or last few years of one's life. For art styles, historical periods, wars and indeed the active periods of artists the events are much more evenly distributed. Further research can determine whether this value of 1.42 for τ is also found for other person-related time periods such as, for example historical leaders.

3.4 THE SEMANTIC REPRESENTATION PHASE

We have acquired a distribution for the periods associated with the concepts from the Information Extraction phase. In the Semantic Representation phase, these results are to be translated to ontology constructs. There are numerous ways of representing an associated period for a concept. This choice is ultimately a modeling decision to be made by the ontology engineer based on the application of the ontology and the type of temporal reasoning required.

In this section, we describe four different ways of representing periods in an ontology. This is not intended to be an exhaustive listing of all possible representations, but it serves as an illustration of different possibilities. In two of the four modeling options, a completely discrete representation is used. In this case, the probabilistic distribution resulting from the Information Extraction phase needs to be discretized. A third option is to directly use the results from the Information Extraction phase, here the probabilistic nature of the extracted information is retained. In the fourth option, a structured vocabulary is used. In this case we use a discrete (ontological) representation and at the same time are able to capture the fuzzy nature of the actual period, as given by the best fitting normal distribution. In the next sections we determine the transformation that is to be used to rewrite the distributions found in the Information Extraction phase to the ontological constructs for each of these representations. We discuss the fourth option in more detail, providing a specific structured vocabulary and an experimental evaluation. In Figure 12, we give an overview of the different modeling options as discussed in this section for the example art style 'Baroque'.

These four representations do not contradict each other and could easily co-exist within a single ontology. It is a task for the designer of the reasoning engine to enable valid temporal reasoning, such as suggested by Allen [1983] using the different representations.

3.4.1 *Discrete Start and End Dates*

Arguably the most straightforward and most intuitive possible representation of a period is the use of discrete start and end dates. In this case, the period associated with an art style is defined by two relations: hasStartYear and hasEndYear. This representation has the advantage that the semantics of the temporal relations are clear and that reasoning with discrete time intervals is well understood [Allen, 1983]. This also implies its main disadvantage in that this discrete representation might not represent the underlying data correctly. As we have seen in the previous sections, historical time periods are hardly ever completely discrete. To say that the art style 'Baroque', for example, was present from the 1611 (or any other date), but not one year before is a simplification that corresponds very poorly to the more fuzzy use of the term in domain literature.

To transform a distribution resulting from the Information Extraction phase to this simple representation, the start and end year need to be determined. A very straightforward method has been introduced in the previous section, through the use of the factor τ. The start and end dates are determined at $\mu - (\sigma \cdot \tau)$ and $\mu + (\sigma \cdot \tau)$ respectively. For the value of τ, we can use 0.8 for art styles, wars, historical periods or similar concepts. For the lifespan of artists and other historical figures, a higher value (1.8) can be used.

3.4.2 *Multiple discrete intervals*

The second representation we discuss here is also a discrete one, but here the fuzzy nature of the concept is taken into account by using multiple discrete

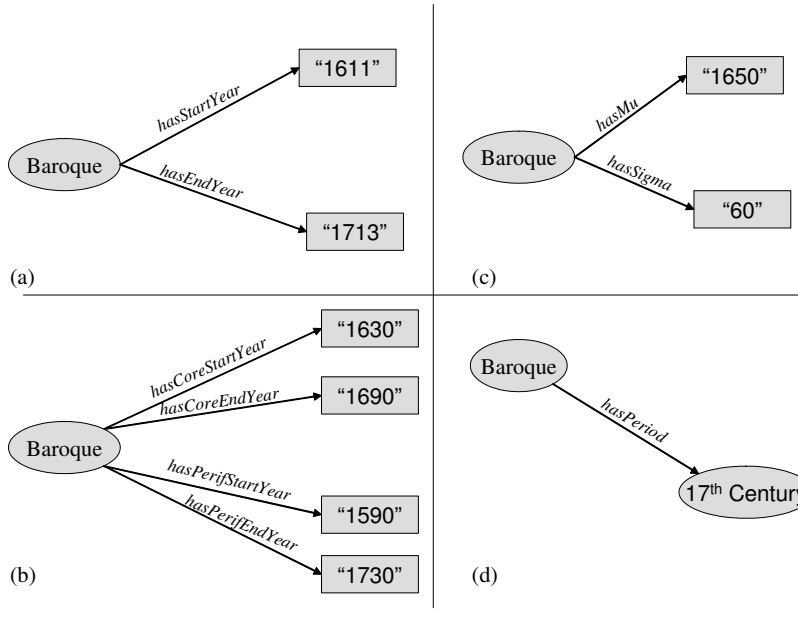

Figure 12: The Baroque example for four possible modeling choices discussed: discrete start and end dates (a), multiple discrete intervals (b), the model parameters(c) and the use of a structured vocabulary(c)

intervals. For this example we use two such intervals: a 'core' period and an 'outer' period, which both have discrete start and end dates. Here, reasoning is also still relatively easy, the choice of using either the core period or the outer period for a reasoning depends on the type of query. For instance, for visualization of the concept on a time-line, the core period might be used. On the other hand, when the application using the extracted knowledge wishes to retrieve whether a painting can belong to an art style given its creation date, the more relaxed notion of the time period in the form of the 'outer' period can be used. Although the semantics of the two different types of relations are less clear, the benefits of having a discrete representation are retained while some of the fuzziness of the period is also represented.

To convert the results from the Information Extraction phase into this representation, not one but two parameters are needed. These two factors, τ_c and τ_o, are used to determine the core and outer period. The core start and end dates are defined as $\mu + (\tau_c \cdot \sigma)$ and $\mu - (\tau_c \cdot \sigma)$ and the outer start and end dates are defined as $\mu + (\tau_o \cdot \sigma)$ and $\mu - (\tau_o \cdot \sigma)$. The optimal values of τ_c and τ_o could be determined using a gold standard that uses the notion of core and outer periods or they could be based on the value for τ we found in the previous section plus and minus a percentage. In Figure 12(b), we show the result for 'Baroque' with $\tau_c = 0.57$ and $\tau_o = 1.05$ (plus and minus 30%).

3.4.3 *Distribution*

The representation that is closest to the results from the Information Extraction phase is to use the parameters of the normal distribution model itself. In this case, every period linked to an art style has a mean and a standard deviation, which can be directly copied from the optimal distribution discovered in the Information Extraction phase. For the instances of the four concepts (art style, artist, historical period and war), these can be found in the tables of the previous section showing the best fitting models. Contrary to the other three representations, no information is lost with this representation. The drawback of this method, however, is that its semantics are very unclear and that temporal reasoning will require probabilistic methods. For human users of the ontology, the meaning of the values of a time period will also be unclear.

3.4.4 *Structured Vocabulary*

The fourth possibility we consider is the use of a structured vocabulary for the periods, which is to be included in the final ontology. This structured vocabulary will have instances that represent fuzzy periods such as 'first half of the 18th Century' or '1960s'. This representation is closest to the manner in which experts talk about periods of art styles, wars and historical periods. If done properly, the fuzziness of the domain is retained, while at the same time, a discrete representation is used in the form of instances of the structured vocabulary. Using a structured vocabulary would also allow for extensive re-use and sharing of the knowledge. Reasoning with the structured vocabulary elements is however not

straightforward and will require a clear semantics of the structured vocabulary as well as a clearly defined reasoning method. For the MultimediaN E-culture ontology, we chose the TIMEX2 standard [Ferro et al., 2005] to serve as a structured vocabulary. We then used the elements from the vocabulary to represent the periods associated with the instances from the four concepts. In Section 3.4.4.1, we give an overview of the structured vocabulary. In the subsequent sections we give our conversion rules and describe an experiment evaluating the structured vocabulary and the translation of the Information Extraction results to this representation.

3.4.4.1 *The TIMEX2 standard*

One of the main design principles used in the construction of the MultimediaN E-culture tool is to re-use as much existing structured vocabularies, taxonomies and ontologies that are in use in the cultural heritage domain as possible. For the representation of temporal information, we did not find an off-the-shelf structured vocabulary. Formal temporal ontologies such as described by Hobbs and Pan [2004] exist, but they do not offer a set of instances that correspond to intuitive notions of time periods, which is what we are looking for. Rather, they specify how different moments in time relate to each other. Whereas this could be used as a basis for a reasoning method for the extracted time periods, it does not offer the fuzzy concepts.

The need for a structured vocabulary to represent dates does not only come from these automatically extracted time periods. In the cultural heritage repositories that are imported into the MultimediaN E-culture browser's knowledge base, a large amount of cultural heritage objects are initially annotated with unnormalized dates. As an example, within the imported art objects from the ARIA database of the Rijksmuseum Amsterdam, we find 740 VRA:DATE annotations. Of these 740, less than half (310) are either a three or four-digit number, indicating a simple creation year. 104 date fields are left empty. 171 date fields are of the form "c. NNNN" and another 37 are of the form "c. NNNN-MMMM" describing approximate creation dates. The remaining 109 date fields are filled with fuzzy textual descriptions such as "mid-16th century", "2nd 1/2 18th century" etc. To make matters worse, in different cultural heritage repositories, different annotation standards or dates are used. A structured vocabulary for the representation of both dates and time periods will facilitate the normalization of these dates and thereby improve the usability of temporal annotations both within repositories and across different repositories.

In our search for a structured vocabulary that satisfies this need we turned to the field of Information Extraction. In general, the task of Information Extraction is to extract entities from a text corpus. A subtask is Temporal Information Extraction, where the specific goal is to extract phrases denoting statements about time. The detected temporal phrases are either automatically annotated or annotated by hand to serve as a gold standard. For this task, the TIDES annotation standard [Ferro et al., 2005] has been developed. TIDES was used as the temporal annotation scheme in the MUC-7 Information Extraction task [Chinchor, 1998] and the TERN Time Expression Recognition and Normalization

Evaluation extraction task [Ferro, 2004]. It is also a key component in the more recent TimeML representation language [Pustejovsky et al., 2003]. The annotation tag used in the TIDES annotation guidelines is the TIMEX2 tag. The TIMEX2 XML tags act as a normalization of the tagged natural language phrases and as such can be used as a structured vocabulary. We here give a short overview of the TIMEX2 annotation standard insofar it is used in the MultimediaN E-culture ontology.

For each temporal expression, one TIMEX2 tag is used with one of the following attributes, all of which are optional: VAL, MOD, ANCHOR_VAL, ANCHOR_DIR, SET, NON_SPECIFIC, COMMENT.

- The main attribute, VAL, is a normalization of an expression of a calendrical time or a duration. For its possible value it uses an extended ISO 8601 scheme, the international standard for denoting date and time. Extensions include the broader capture of time by means of omitting too specific information. Four digit values are used to denote a whole year, three digit values are used for decades (e.g. VAL="196" means 'the 1960's') and two digit values are used for centuries (e.g. VAL="18" means 'the 19th Century'). To avoid ambiguity, the year 196 AD is coded as VAL="0196". TIMEX2 additions to the ISO scheme that we also use are the suffix "H1" or "H2" to denote the first or second half of a year, decade or century and the prefix "BC" to denote B.C. dates. Also, arbitrary periods can be described with the "P" prefix, followed by a number and a unit suffix ("W" for weeks, "Y" for years etc.): VAL="P10Y" can be translated to 'for one decade'.

- The second XML attribute, MOD, is used as a normalization for temporal modifiers such as 'circa', or 'at the end of ...'. They act as a modifier to the period given by the VAL value. The possible values for MOD are:

  1. START, MID, END, APPROX. These values are used to indicate the start, middle or end of the period defined by VAL or that the VAL date is an approximate one.
  2. BEFORE, AFTER, ON_OR_BEFORE, ON_OR_AFTER. In our conversion rules, we do not use these values;
  3. LESS_THAN, MORE_THAN, EQUAL_OR_LESS, EQUAL_OR_MORE for "duration" expressions. These values are also not used in our method;

- The attribute ANCHOR_VAL is used in combination with a VAL period, denoting either the starting or end point of a period, depending on the value of ANCHOR_DIR. The possible values of ANCHOR_VAL are ISO 8601 dates, again with the extension that too specific information may be omitted.

- ANCHOR_DIR determines which direction the VAL period has with respect to ANCHOR_VAL. It has the possible values WITHIN, STARTING, ENDING, AS_OF, BEFORE and AFTER. We only use the latter to denote a period. (e.g. VAL="P5Y" ANCHOR_VAL= "1958" ANCHOR_DIR="AFTER" means 'the five-year period starting from 1958').

- The attribute SET is a boolean attribute (with values YES or NO) which identifies expressions denoting sets of times such as 'Every year'. Since we assume to deal with a single period for a concept, we assume the value NO for all annotations and we will leave out this attribute.

- The same holds for the NON_SPECIFIC boolean attribute, which is used to determine whether or not a temporal expression represents a whole class of temporal values rather than a specific time. It is also used to tag indefinite expressions such as 'The election took place *on a Tuesday.*' Since we assume to deal with a definite periods for a concept, we assume the value NO for all annotations.

- The final attribute, COMMENT, simply holds any comments. We do not use it in our structured vocabulary or the experiments described below.

TIDES is an annotation standard and as such every natural language temporal expression description has a single correct TIMEX2 tag. In our case, we lose the original natural language context and aim to map extracted numerical periods to TIMEX2 tags. In this case one period can in theory be described by more than one TIMEX2 tag. For example, the tags ⟨TIMEX2 VAL= "196"⟩ and ⟨TIMEX2 VAL= "P10Y" ANCHOR_VAL= "1960" ANCHOR_DIR= "AFTER"⟩ can both be used to represent the extracted period (1960-1970). In natural language, periods are usually either described by a start- and end year ("from 1916 to 1920") or by one or more terms describing a whole period (e.g. "the 17th century"). In our conversion, we prefer to represent a period using these latter period concepts. In some cases an extracted period has to be described by more than one of these expressions, for instance periods spanning multiple centuries. In this case, we use a conjunction of TIMEX2 annotations to represent a single period. For example, the period from 1675-1799 will be stored as '⟨TIMEX2 VAL= "16" MOD="END"⟩ AND ⟨TIMEX2 VAL= "17"⟩', corresponding to the natural language statement "End of the 17th Century and the (entire) 18th Century".

Most of the periods extracted can be described by a TIMEX2 tag using only VAL and MOD attributes. For example, the tag ⟨TIMEX2 VAL="16"⟩ corresponds to '17th century' and the tag ⟨TIMEX2 VAL= "198" MOD= "START"⟩ corresponds to 'start of the 1980's'. In the MultimediaN E-culture ontology, we implemented a TIMEX2 period description in RDF(S) as an XML literal. This means that the XML tag as such is the object of the hasPeriod property for an specific instance.

3.4.4.2 *TIMEX2 conversion rules*

To acquire TIMEX2 values from the distribution resulting from the Information Extraction phase, we use a set of conversion rules. These rules take as input the distribution parameter μ and the standard deviation modified by the value of τ used: $\sigma \cdot \tau$. The output of the rules is a TIMEX2 XML literal. For rewriting the results from the Information Extraction phase presented in Section 3.3.5, we used 13 rules for periods with a starting date 'Anno Domini' and another 11 rules for periods 'Before Christ'. We designed the rules such that for a single input, only one rule is applicable, so every input results in a single output TIMEX2 string.

Rule 1:

```
IF (Mu = 0  AND  (5 ≤ SigmaTau < 25)  AND  (Mu MOD 100 ≤ 30)) THEN
      RETURN "<TIMEX2 VAL=\"" +Substring(Mu, 0, 2) + "\" MOD="START">"
```

Rule 2:

```
IF (-1000 = Mu < -100  AND  (SigmaTau < 5)  AND  (4 = Mu MOD 10 < 6)) THEN
      RETURN "<TIMEX2 VAL=\"0" +Substring(Mu, 0, 3) + "BC\">"
```

Figure 13: Two example TIMEX2 rewriting rules. Rule 1 is used to form TIMEX2 tags for 'Start of the NNth century'. Rule 2 is used to form TIMEX2 tags for decades B.C. The VAL value starts with a "0" since we are dealing with three-digit years (between 999 and 99 BC) that are converted to decade notation (VAL="0NN BC").

We constructed the rules by hand to correspond to an intuitive notion of the time period that can be determined by μ and $\sigma \cdot \tau$. As an example, we list two of the rewriting rules in pseudocode in Figure 13. The complete rule set can be found in Appendix 1. The expressions in the rules contain numeric ranges of which the chosen values are based on heuristics and are certainly debatable. Using different values or using additional rules will produce different evaluation results.

3.4.4.3 *Evaluation*

Evaluation of the rewriting rules and the resulting TIMEX2 literals is not trivial. At the moment, the TIMEX2 representation is one of the representations for time that are used in the MultimediaN E-culture browser and efforts are being made to facilitate time reasoning and indexing using the TIMEX2 standard.

Here, we choose to evaluate the rules by applying them to the results for the four concepts from the Information Extraction phase presented in Section 3.3.5 and have people manually judge the results with respect to a gold standard. To make this evaluation task easier, we generated a natural language description for each of the resulting TIMEX2 XML tags. Generating these natural language descriptions from the TIMEX2 tags is a relatively trivial task as the tags themselves were initially designed to annotate natural language time descriptions. We simply chose for each rule the most typical English description of that period. For example, a tag resulting from rule 1 from Figure 13 would get the natural language description 'Start of the NNth Century', where NN is the value of the VAL attribute from the TIMEX2 tag plus one (the 17th Century runs from 1600 to 1699).

For each of the 60 instances from Section 3.3.5, we asked four non-expert evaluators to rate the natural language description with respect to the gold standard that was also used for the numerical evaluation in that section. For

each instance, the evaluator was instructed to select one of three categories: A natural language description was to be evaluated *Completely Incorrect* if it did not correspond to the period defined by the gold standard start and end year. If the natural language period description correctly corresponded to the period defined by the gold standard start and end year, the category *Completely Correct*. If the evaluator deemed the period description to be neither completely correct nor incorrect with respect to the gold standard period (for instance, the described time period is too broad or too narrow), the evaluator was asked to evaluate the instance with the category *Partially Correct*.

In Table 23 and Table 24, we show all the resulting TIMEX2 tags, the natural language descriptions and the original gold standard for the 60 instances from Section 3.3.5. The TIMEX2 tags were not presented to the evaluators.

Averaged over the four evaluators, 11.25% of the 60 instances were classified as being 'completely incorrect', 38.75% as 'completely correct' and the remaining 50.0% as 'partially correct'. To further show the performance, we gave for each evaluator a score of 1 to the 'completely correct' instances, 0.5 to the 'partially correct' instances, and a score of 0 to 'completely incorrect' instances. In the last column of tables 23 and 24, we also show the average *evaluation scores* for each of the instances. The average evaluation score for all instances is 0.64. Interestingly enough, the average evaluation scores do not vary very much for the four different concepts: the historical periods scored 0.60 on average, wars 0.69, art styles 0.62 and artists 0.66. The evaluation shows that the resulting natural language statements are of acceptable quality: 88.75% of the natural language descriptions of the instances were deemed to either be partially or completely correct with respect to the gold standard.

There is no instance that was evaluated as 'completely incorrect' by all four evaluators. One instance, 'The Renaissance', received a score of 0.125 (one 'partially correct', three 'completely incorrect' evaluations) and four instances scored 0.25: 'Industrial Revolution', 'Punic Wars', 'Modern Era' and 'Surrealism'. On the other side of the spectrum, six instances scored a 1, being classified as 'completely correct' by all evaluators.

Intuitively, we would expect that instances with high absolute errors will receive better evaluations and vice versa (garbage-in garbage-out). To verify this, we look at the absolute errors of the period edges as shown in Tables 16, 18, 20 and 22. Indeed, we find that instances with high absolute errors such as 'Industrial Revolution' or 'The Renaissance' have very poor evaluations. On the other hand, instances that have a low absolute error, have a high evaluation score (e.g. 'Neo-Impressionism'). The value of the Pearson correlation coefficient for the absolute error and the evaluation score is -0.52, indicating a substantial negative correlation between the two. The fact that the absolute error found in the evaluation of the Information Extraction results has a negative correlation with the evaluation score tells us that the quality of the extraction largely determines the quality of the resulting TIMEX2 tags and resulting natural language statements.

A number of instances do not adhere to this correspondence. In particular instances that received very specific period descriptions (such as 'Uruguayan Civil War'–'Approximately from 1825 to 1876') can have a relatively low absolute error, but still receive a low evaluation. Where more 'fuzzy' descriptions were

Table 23: The values of μ and $\sigma \cdot \tau$, the resulting TIMEX2 tags and natural language representation, the gold standard start- and end-years and the evaluation scores for the extracted time periods for Historical Periods and Wars

instance Label	μ	$\sigma \cdot \tau$	TIMEX2	nat. lang.	GS: start	GS: end	eval.
Historical Periods							
Edwardian period	1906.7	15.3	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1901	1910	1
Elizabethan period	1585.7	37.4	⟨ TIMEX2 VAL="15H2" ⟩	Second half of the 16th Century	1558	1603	0.75
Georgian Era	1806.7	67.2	⟨ TIMEX2 VAL="17H2" ⟩ AND ⟨ TIMEX2 VAL="18H1" ⟩	Second half of the 18th Century - First half of the 19th Century	1714	1830	0.625
Industrial Revolution	1809.8	63.5	⟨ TIMEX2 VAL="17H2" ⟩ AND ⟨ TIMEX2 VAL="18H1" ⟩	Second half of the 18th Century - First half of the 19th Century	1700	1899	0.25
Interwar period	1930.9	13.3	⟨ TIMEX2 VAL="19" MOD="MID" ⟩	Mid 20th Century	1918	1939	0.5
Jacobean Era	1628.3	25.4	⟨ TIMEX2 VAL="16H1" ⟩	First half of the 17th Century	1603	1625	0.875
Late Middle Ages	1377.6	131.5	⟨ TIMEX2 VAL="12H2" ⟩ AND ⟨ TIMEX2 VAL="13" ⟩ AND ⟨ TIMEX2 VAL="14" ⟩	Second half of the 13th Century - 14th Century - 15th Century	1300	1500	0.625
Machine Age	1929.2	31.4	⟨ TIMEX2 VAL="19H1" ⟩	First half of the 20th Century	1900	1945	1
Meiji Era	1879.5	27.4	⟨ TIMEX2 VAL="18H2" ⟩	Second half of the 19th Century	1868	1912	0.625
Modern Era	1933.3	67.6	⟨ TIMEX2 VAL="18" MOD="END" ⟩ AND ⟨ TIMEX2 VAL="19" ⟩	End of the 19th Century - 20th Century	1700	1999	0.25
Napoleonic Era	1809.1	8.4	⟨ TIMEX2 VAL="18" MOD="START" ⟩	Start of the 19th Century	1799	1815	0.875
Age of Enlightenment	1769.7	50.7	⟨ TIMEX2 VAL="17" ⟩ AND ⟨ TIMEX2 VAL="18" MOD="START" ⟩	18th Century - Start of the 19th Century	1700	1799	0.375
The Protestant Ref.	1535.6	25.6	⟨ TIMEX2 VAL="15H1" ⟩	First half of the 16th Century	1500	1599	0.375
The Renaissance	1529.2	78.4	⟨ TIMEX2 VAL="14" MOD="END" ⟩ AND ⟨ TIMEX2 VAL="15" ⟩	End of the 15th Century - 16th Century	1300	1599	0.125
Tokugawa shogunate	1763.8	192.4	⟨ TIMEX2 VAL="15" MOD="END" ⟩ AND ⟨ TIMEX2 VAL="16" ⟩ AND ⟨ TIMEX2 VAL="17" ⟩ AND ⟨ TIMEX2 VAL="18" ⟩ AND ⟨ TIMEX2 VAL="19H1" ⟩	End of the 16th Century - 17th Century - 18th Century - 19th Century - First half of the 20th Century	1603	1868	0.75
Victorian Era	1863.8	32.6	⟨ TIMEX2 VAL="18H2" ⟩	Second half of the 19th Century	1837	1901	0.625
Wars							
Eighty Years' War	1593	67.4	⟨ TIMEX2 VAL="15H2" ⟩ AND ⟨ TIMEX2 VAL="16H1" ⟩	Second half of the 16th Century - First half of the 17th Century	1568	1648	0.75
Hundred Years' War	1379	61.3	⟨ TIMEX2 VAL="13" ⟩ AND ⟨ TIMEX2 VAL="14" MOD="START" ⟩	14th Century - Start of the 15th Century	1337	1453	0.75
Hussite Wars	1424.3	11.4	⟨ TIMEX2 VAL="14" MOD="START" ⟩	Start of the 15th Century	1420	1436	0.625
Italian Wars	1524.1	43.2	⟨ TIMEX2 VAL="14" MOD="END" ⟩ AND ⟨ TIMEX2 VAL="15H1" ⟩	End of the 15th Century - First half of the 16th Century	1494	1559	0.75
Napoleonic Wars	1810.7	5.1	⟨ TIMEX2 VAL="18" MOD="START" ⟩	Start of the 19th Century	1803	1815	0.875
Seminole Wars	1842.8	24.6	⟨ TIMEX2 VAL="18" MOD="MID" ⟩	Mid 19th Century	1817	1858	0.75
Uruguayan Civil War	1850.6	31.7	⟨ TIMEX2 VAL="P64Y" ANCHOR_DIR="AFTER" ANCHOR_VAL="1818" MOD="APPROX" ⟩	Approximately from 1818 to 1882	1839	1851	0.5
Vietnam War	1967	4.7	⟨ TIMEX2 VAL="P9Y" ANCHOR_DIR="AFTER" ANCHOR_VAL="1962" MOD="APPROX" ⟩	Approximately from 1962 to 1971	1959	1975	0.5
War of the Roses	1467.1	20.4	⟨ TIMEX2 VAL="14" MOD="MID" ⟩	Mid 15th Century	1455	1485	1
World War I	1916.5	2.4	⟨ TIMEX2 VAL="191" MOD="MID" ⟩	Mid 1910's	1914	1918	0.75
World War II	1942.4	2.6	⟨ TIMEX2 VAL="194" MOD="START" ⟩	Start of the 1940's	1939	1945	0.875
Persian Wars	-491.9	9.6	⟨ TIMEX2 VAL="BC04" MOD="START" ⟩	Start of the 5th Century BC	499 BC	479 BC	0.625
Punic Wars	-228.8	37.7	⟨ TIMEX2 VAL="BC02H1" ⟩	Second half of the 3th Century BC	264 BC	146 BC	0.25
Syrian Wars	-262.8	68.6	⟨ TIMEX2 VAL="BC02" ⟩	3th Century BC	274 BC	200 BC	0.625

Table 24: The values of μ and $\sigma \cdot \tau$, the resulting TIMEX2 tags and natural language representation, the gold standard start- and end-years and the evaluation scores for the extracted time periods for Art Styles and Artist

instance Label	μ	$\sigma \cdot \tau$	TIMEX2	nat. lang.	GS: start	GS: end	eval.
Art Styles							
Baroque	1662.2	63	⟨ TIMEX2 VAL="16" ⟩	17th Century	1593	1717	0.625
Counter-Reformation	1569.6	58.1	⟨ TIMEX2 VAL="15"⟩ AND ⟨ TIMEX2 VAL="16" MOD="START" ⟩	16th Century - Start of the 17th Century	1529	1624	0.625
Mannerist	1552.6	53.2	⟨ TIMEX2 VAL="15" ⟩	16th Century	1526	1593	0.5
Reformation	1541.4	33.9	⟨ TIMEX2 VAL="P68Y" ANCHOR_DIR="AFTER" ANCHOR_VAL="1507" MOD="APPROX" ⟩	Approximately from 1507 to 1575	1508	1587	0.625
Rococo	1738.7	41	⟨ TIMEX2 VAL="17" ⟩	18th Century	1709	1791	0.375
Art Deco	1928.1	9.3	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1920	1939	0.375
Art Nouveau	1893	22.9	⟨ TIMEX2 VAL="18" MOD="END" ⟩	End of the 19th Century	1882	1906	0.75
Cubism	1910.7	5	⟨ TIMEX2 VAL="P10Y" ANCHOR_DIR="AFTER" ANCHOR_VAL="1905" MOD="APPROX" ⟩	Approximately from 1905 to 1915	1907	1921	0.625
Dada	1917.5	3.5	⟨ TIMEX2 VAL="191" MOD="END" ⟩	End of the 1910's	1916	1923	0.625
Expressionism	1912.1	31.3	⟨ TIMEX2 VAL="19H1" ⟩	First half of the 20th Century	1907	1937	0.75
Impressionism	1875.4	26.2	⟨ TIMEX2 VAL="18H2" ⟩	Second half of the 19th Century	1864	1890	0.75
Neo-Impressionism	1886	4.7	⟨ TIMEX2 VAL="188" ⟩	1880's	1882	1889	1
Surrealist	1928.3	18.2	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1924	1950	0.25
Abstract Expressionism	1951	14.5	⟨ TIMEX2 VAL="19" MOD="MID" ⟩	Mid 20th Century	1946	1964	1
Bauhaus	1923.4	12.1	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1919	1933	0.375
Post-Impressionism	1881.2	31.5	⟨ TIMEX2 VAL="18H2" ⟩	Second half of the 19th Century	1877	1900	0.625
Pre-Raphaelites	1857.9	31.9	⟨ TIMEX2 VAL="P63Y" ANCHOR_DIR="AFTER" ANCHOR_VAL="1826" MOD="APPROX" ⟩	Approximately from 1826 to 1889	1848	1886	0.75
Fauve	1906.2	2.3	⟨ TIMEX2 VAL="190" MOD="MID" ⟩	Mid 1900's	1902	1910	0.5
Artists							
El Greco	1587.3	26	⟨ TIMEX2 VAL="15H2" ⟩	Second half of the 16th Century	1541	1614	0.75
Frans Hals	1631.2	32.8	⟨ TIMEX2 VAL="16H1" ⟩	First half of the 17th Century	1585	1666	0.375
Johannes Vermeer	1663.6	7.7	⟨ TIMEX2 VAL="16" MOD="MID" ⟩	Mid 17th Century	1632	1675	1
Leonardo da Vinci	1496.1	21.1	⟨ TIMEX2 VAL="14" MOD="END" ⟩	End of the 15th Century	1452	1519	1
Henri Matisse	1907.1	10.4	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1869	1954	0.375
John Everett Millais	1861.3	26	⟨ TIMEX2 VAL="18H2" ⟩	Second half of the 19th Century	1829	1896	0.625
Claude Monet	1876.5	15.2	⟨ TIMEX2 VAL="18" MOD="END" ⟩	End of the 19th Century	1840	1926	0.5
Otto Dix	1921.9	8.4	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1891	1969	0.625
Pablo Picasso	1909.3	20	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1881	1973	0.5
Paul Klee	1922.7	21.1	⟨ TIMEX2 VAL="19" MOD="START" ⟩	Start of the 20th Century	1879	1940	0.875
Rembrandt van Rijn	1641.4	18.4	⟨ TIMEX2 VAL="16" MOD="MID" ⟩	Mid 17th Century	1607	1669	0.5
Georges Seurat	1885.8	4.1	⟨ TIMEX2 VAL="188" ⟩	1880's	1859	1891	0.75

used, the evaluation can still be good even though larger errors have been found. An example of this is 'Johannes Vermeer', with an evaluation score of 1 and an absolute error of 14.53 years.

3.5 RELATED WORK

Temporal Information Extraction is a subtask of Information Extraction where the goal is to identify and annotate temporal natural language expressions in corpus documents (usually news corpora). As such it is a form of Named Entity Recognition where the entities that are to be tagged are textual time expressions. The large number of systems that perform this temporal tagging make extensive use of hand-crafted rules (c.f [Llido et al., 2001], [Negri and Marseglia, 2004], [Mani and Wilson, 2000]). Approaches that use machine learning to automatically tag time expressions include the system developed by Boguraev and Ando [2005] which uses a set of cascaded finite-state grammars together with a classifier trained on a large annotated corpus. Another recent example of a system that uses machine learning to automatically tag time expressions is described by Ahn et al. [2007]. In this chapter, the authors learn a number of classifiers on an annotated corpus to identify and normalize temporal expressions. They report results comparable to state-of-the-art systems that employ hand-crafted rules.

The method presented in this chapter differs from these systems in that its goal is not to extract all temporal expressions from a document, but to infer the temporal information -more specific the associated time period- by aggregating temporal information from Web documents. It is an example of ontology enrichment in the sense that for every term one period needs to be extracted and that we are not interested in the individual temporal expressions. The extracted information itself can be used as background information for use in information retrieval and question answering systems. The method presented here is limited in the type of temporal information it can extract, but it is less language- and structure-dependent than methods that use rules or derive models from an annotated corpus. Furthermore, since we aggregate over different sources (web pages), we get a better overall representation of the shared view of a time period. The time periods associated with cultural heritage and historical concepts are often fuzzy and open to debate. By combining information from different sources we are able to extract a period that corresponds to the average of multiple points of view. A similar approach was taken by Schockaert et al. [2008], where the authors also aim to extract fuzzy time periods for terms from the web. They compare a number of heuristic strategies that use the relation between a term and it's sub-events (for examples wars and battles). The system relates the document structure to the temporal structure. The authors report good precision with respect to a ground truth. A significant difference is that in our case, the relation between the term and the sub-events is implicit: the sub-events do not need to be recognized in the text for the method to work.

Research has been done on representing time and time intervals. Allen [1983] introduced the notion of discrete time intervals and presented a set of fixed relations between these intervals. Since Allen, different representations of these

intervals have been studied. Fuzzy intervals are also being used to represent and reason with more vague time intervals (such as described in work by Ohlbach [2004] or Schockaert and De Cock [2008]). However, since the broader context of this research is the Semantic Web, we chose discrete constructs in the form of a structured vocabulary to represent the periods. In the context of the Semantic Web, work on constructing ontologies of time is also being undertaken (most prominently by Hobbs and Pan [2004]). However, this research also focuses on the formal representation of temporally related events. It does not provide a normalization or structured vocabulary for representation of actual time periods.

3.6 CONCLUSIONS

In this chapter, we described a method for the extraction of time periods for the enrichment of an ontology containing cultural heritage and historical concepts. The method consists of two separate phases that we have separately evaluated.

We have presented the Information Extraction phase, which exploits the redundancy of information on the World Wide Web to extract temporal information about a concept. This information is then aggregated and normalized and we fit a statistical model onto the extracted data. This model describes the distribution of the years associated with a concept. The resulting model serves as input for the semantic representation phase. We performed a quantitative evaluation of the Information Extraction phase by extracting distributions for 60 instances of four different historical concepts. The method requires no adjustments for the four concepts. However, if for a new concept additional background knowledge about the form of the frequency distribution is available, a different and appropriate model can be used in the fitting step.

The method itself is domain- and language independent (although for BC dates, a limited language dependency was introduced). The evaluation showed that the method performs at an acceptable level that is approximately the same for all four concepts.

Of course, the method itself could be improved upon: different models than the four models tested in Section 3.3.4 could be used and a method that could fit different models for different types of data could achieve a higher performance. Also, the extraction process itself could be improved: for more ambiguous concepts such as 'Symbolism' or 'The Modern Era', further query disambiguation will lead to a working corpus of a higher quality and therefore to a better distribution.

The method could be combined with a number of the systems mentioned in the previous section. More specifically, in the year extraction step of the Information Extraction phase, more sophisticated temporal extraction methods could be employed. The extracted timestamps could then again be combined, to construct an aggregated time period.

Another expansion of the method could involve extracting periods of a different (finer) granularity. In Section 3.2, we have stated our general assumption that the target concept must have extractable individual moments related to that concept. We here have only experimented with the extraction of time periods spanning multiple years, using a simple regular expression extract individual

years. Although for historical concepts, a lot of concepts will have periods spanning multiple years, in other domains this might not be the case. For this, not only years, but more fine-grained temporal expressions should be extracted and combined into a distribution. An example would be to identify the time period associated with the Paris student riots in 1968, requiring the extraction of more fine-grained dates (day-month-year). This will require a new set of extraction rules in the form of regular expressions. At the same time, the method is highly reliant on a large amount of temporal phrases on the Web. If there are only a few of such textual descriptions found in the working corpus, the performance of this method will drop considerably.

The second phase of the method involves constructing a Semantic Representation from the found distributions. An important issue here is that the choice of the representation is to be made by an ontology engineer rather than by the person or program that fills the knowledge base. In this chapter, we describe four possible representations: three of these are essentially numerical. The fourth representation is through the use of a structured vocabulary, based on the TIMEX2 annotation standard. With the use of this standard, the fuzziness of the extracted periods can be preserved in abstracted form while the resulting TIMEX2 values are still 'discrete'.

We have provided a number of simple and intuitive rewriting rules that can be used to rewrite normal distributions of time periods into the TIMEX2 representations. For different types of distributions, similar rewriting rules can be constructed with different parameters in the premises of the rules. To evaluate the rules and the TIMEX2 standard, we applied them to the distributions found in the Information Extraction phase. We then had four different evaluators evaluate the results by comparing the TIMEX2-related natural language descriptions to the gold standard that was also used in the Information Extraction phase evaluation. The results show that the rules preserve the quality of the distributions. Of course, this evaluation does not give a good indication of how well the TIMEX2 representation is for use in an ontology. This is much harder to evaluate. The TIMEX2 representation is currently in use in the MultimediaN E-culture ontologies for describing temporal information. For the art styles used here, TIMEX2 representations for the extracted time periods are stored in the RDF database after being inspected by hand. We are in the process of implementing temporal reasoning and indexing using the TIMEX2 representation. Further evaluation of the TIMEX2 standard and the extracted time periods for the cultural heritage concepts will be done within the context of its use in the MultimediaN E-culture browser.

APPENDIX I: TIMEX2 REWRITING RULES

In this appendix, we list the rewriting rules that we used for the conversion of time period distribution parameters to TIMEX2 instances, as implemented in Visual Basic 6.5. The function 'fit2Timex' has two input values: μ and the product of σ and τ. The output consists of the TIMEX2 XML string plus the natural language form of the TIMEX2 instance.

```
Function fit2Timex(Mu, SigmaTau) As Variant
 Dim Result(2) As String

 'This holds the TIMEX2 'XML' string
 Dim Timex2Str As String

 'This holds the Natural Language string
 Dim NLStr As String

 Dim Val1 As String

 If (Mu >= 0) Then
 'AD dates

  If (SigmaTau <= 1) Then
  'In this case, use approximately Mu
   Val1 = Left(CStr(Mu), 4)
   Timex2Str = "VAL=""" & Val1 & """ MOD=""APPROX"""

  ElseIf (SigmaTau <= 3.5 And SigmaTau > 1) Then
  'Start/Mid/End of Decade
   Val1 = Left(CStr(Mu), 3)
   If (Mu Mod 10 < 3) Then
    Timex2Str = "VAL=""" & Val1 & """ MOD=""START"""
   ElseIf (Mu Mod 10 > 7) Then
    Timex2Str = "VAL=""" & Val1 & """ MOD=""END"""
   Else
    Timex2Str = "VAL=""" & Val1 & """ MOD=""MID"""
   End If

  ElseIf (SigmaTau > 3 And SigmaTau <= 5 And (Mu Mod 10 >= 4) And
                                          (Mu Mod 10 <= 6)) Then
  'In this case, use NNNs (eg. 1920's)
   Val1 = Left(CStr(Mu), 3)
   Timex2Str = "VAL=""" & Val1 & """"

  ElseIf (SigmaTau <= 25 And SigmaTau > 5) Then
  'Start/Mid/End of Century
   Val1 = Left(CStr(Mu), 2)
   If (Mu Mod 100 < 30) Then
    Timex2Str = "VAL=""" & Val1 & """ MOD=""START"""
   ElseIf (Mu Mod 100 > 70) Then
    Timex2Str = "VAL=""" & Val1 & """ MOD=""END"""
   Else
    Timex2Str = "VAL=""" & Val1 & """ MOD=""MID"""
   End If

  ElseIf ((SigmaTau <= 40 And SigmaTau > 20) And (Mu Mod 100 >= 60
                                       Or Mu Mod 100 <= 40)) Then
  'First/Second Half of Century
   Val1 = Left(CStr(Mu), 2)
   If (Mu Mod 100 < 40) Then
    Timex2Str = "VAL=""" & Val1 & "H1"""
   ElseIf (Mu Mod 100 > 60) Then
    Timex2Str = "VAL=""" & Val1 & "H2"""
   End If
```

```
 ElseIf ((SigmaTau > 40 And SigmaTau < 70) And (Mu Mod 100 >= 35
                                        And Mu Mod 100 < 65)) Then
 'Xth Century
  Val1 = Left(CStr(Mu), 2)
  Timex2Str = "VAL=""" & Val1 & """"

 ElseIf (Not (Left(CStr(Mu - SigmaTau), 2) = Left(CStr(Mu +
 SigmaTau), 2))) Then
 'across multiple centuries (both)

  Start0 = Mu - SigmaTau
  End0 = Mu + SigmaTau
  NextCen = (Start0 \ 100 + 1) * 100

  TS1 = fit2Timex(0, (NextCen + Start0)/2, (NextCen - Start0)/2)
  TS2 = fit2Timex(0, (NextCen + End0)/2, (End0 - NextCen)/2)
  Timex2Str = TS1(0) & " AND " & TS2(0)

 Else
 'Default: Duration
  Start1 = Left(CStr(Mu - SigmaTau), 4)
  End1 = Left(CStr(Mu + SigmaTau), 4)
  Dur1 = End1 - Start1
  Timex2Str = "VAL=""P" & Dur1 & "Y"" ANCHOR_DIR=""AFTER""
                    ANCHOR_VAL=""" & Start1 & """ MOD=""APPROX"""
  NLStr = "Approximately from " & Start1 & " to " & End1
 End If

ElseIf Mu < 0 Then
'BC dates
 'First take care of 3-digit dates
 If (Mu > -1000 And Mu <= -100) Then
  MuStr = "0" & CStr(Abs(Mu)) & """"
 Else
  MuStr = CStr(Abs(Mu)) & """"
 End If

 If (SigmaTau <= 3) Then
 'In this case, use approximately Mu
  Val1 = Left(MuStr, 4)
  Timex2Str = "VAL=""BC" & Val1 & """ MOD=""APPROX"""

 ElseIf (SigmaTau <= 5 And ((Mu Mod 10 > 4) And
                                        (Mu Mod 10 < 6))) Then
 'In this case, use NNNs (eg. 1920's)
  Val1 = Left(MuStr, 3)
  Timex2Str = "VAL=""BC" & Val1 & """"

 ElseIf (SigmaTau <= 20 And SigmaTau > 5) Then
 'Start/Mid/End of Century
  Val1 = Left(MuStr, 2)
  If (Abs(Mu) Mod 100 < 30) Then
   Timex2Str = "VAL=""BC" & Val1 & """ MOD=""END"""
  ElseIf (Abs(Mu) Mod 100 > 70) Then
   Timex2Str = "VAL=""BC" & Val1 & """ MOD=""START"""
  Else
```

```
    Timex2Str = "VAL=""BC" & Val1 & """ MOD=""MID"""
   End If

  ElseIf ((SigmaTau <= 40 And SigmaTau > 20) And
                        (Mu Mod 100 > 60 Or Mu Mod 100 < 40)) Then
   'Start/Mid/End of Century
    Val1 = Left(MuStr, 2)
    If (Abs(Mu) Mod 100 < 40) Then
     Timex2Str = "VAL=""BC" & Val1 & "H1"
    ElseIf (Abs(Mu) Mod 100 > 60) Then
     Timex2Str = "VAL=""BC" & Val1 & "H2"
    End If

  ElseIf ((SigmaTau > 40 And SigmaTau < 70) And (Abs(Mu)
                       Mod 100 > 35 And Abs(Mu) Mod 100 < 65)) Then
   'Xth Century
    Val1 = Left(MuStr, 2)
    Timex2Str = "VAL=""BC" & Val1 & """"

  Else
   'No rule found, default
    Timex2Str = ""
    NLStr = ""
  End If
 End If

 fit2Timex = Result
End Function
```