Extracting Art Style Periods from the Web

Viktor de Boer, Maarten van Someren, and Bob J. Wielinga

Human-Computer Studies Laboratory, Informatics Institute, Universiteit van Amsterdam, email: {vdeboer,maarten,wielinga}@science.uva.nl

Abstract. A subtask of Ontology Enrichment is the extraction of time periods for temporal concepts. In this paper we describe a method for this period extraction. We show the working of the method in the cultural heritage domain, extracting periods for art styles. We also discuss a number of possible ways to transform the uncertain knowledge that this method provides into a semantic representation in an ontology.

1 Introduction

The Semantic Web needs automatic methods that extract ontology constructs. Ontology Learning attempts to extract classes and relations in an ontology, whereas Ontology Population is the task of extracting instances of both relations and classes. If the extraction of the constructs is done for an already existing ontology, we use the term Ontology Enrichment. Ontology Learning, Population and Enrichment all use Information Extraction (IE) methods to learn the target knowledge from a corpus.

These IE methods usually output the information with a degree of (un)certainty. In most ontologies, knowledge is represented in discrete form and that any statement can be determined as either true or false. In this paper we examine a specific example of this discrepancy and show a way of providing an interface between the uncertain output of an IE technique and the discrete representations in an ontology.

As our domain we chose the Cultural Heritage domain, where a number of relatively large taxonomies are in actual use by experts. This research takes place in a context where the aim is to integrate these taxonomies and enrich them in such a way that expert users are able to browse art repositories in a semantically rich system.

The core taxonomy we use as an ontology is the Art and Architecture Thesaurus[1] (AAT), an taxonomy containing over 133.000 concepts from the Cultural Heritage domain. A part of the AAT describes different art styles such as 'Baroque', 'Impressionism' etcetera. In previous experiments[2], we extracted relations between these art styles and periods and artists. In this paper, we will describe a method to extract periods for the individual art styles. This information is currently not present in the AAT. This Ontology Enrichment allows for temporal reasoning about art styles and related concepts.

We have split up our method into an Information Extraction phase and a Semantic Representation phase. In Section 3, we describe the first phase, where

we extract art style periods from the World Wide Web. In Section 4, we describe the different ways of converting this uncertain information into ontological constructs. In the next section, we define both tasks more extensively.

2 Problem Definition

We first give a general problem definition for the extraction of periods for an arbitrary domain and will then focus on the Cultural Heritage domain.

We have a partly populated ontology with a concept C that has a relation to a time interval (a period). This period is represented in some way in the ontology. We have a number of instances of C: $i_1, i_2, ...$ with one or more description labels $l_{1,1}, l_{1,2}, ..., l_{2,1}, ...$ For each of these instances, the task is to extract from the web the correct period related to it. We make two assumptions.

- In the ontology, there is only one period related to each instance.
- 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. In the case of art styles, these events might be creation dates of paintings, for which usually a single year is noted in the text.

We split the problem up into two phases:

1. In this first Information Extraction phase, we extract the distribution of individual moments (e.g. years) that characterize an instance i.
2. In the Semantic Representation phase, The distribution is then used to formalize the period in the form in which it is to be represented in the ontology.

In the rest of this paper, we use a RDF version of the AAT as our partly populated ontology. The concept C is `aat:Styles and Periods` and the instances are the individual art styles (`aat:Baroque`, `aat:Rococo`, etc). For our purposes, we expanded the AAT with a relation `has_period` with `aat:Styles and Periods` as the domain and a time interval as the range. Instances of this relation are used to denote the historical periods of the different art styles.

3 The Information Extraction Phase

For the extraction of the periods, we do not use domain- or structure-specific Information Extraction techniques such as the use of patterns or Natural Language processing techniques. The effectiveness of these methods are 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 (eg. years) from documents on the Web about the temporal concept. We assume 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 period by comparing the two. To further clarify this, we present an outline of the method in the next section.

3.1 The Information Extraction method

In Figure 1, we present the method used to extract the distribution of years. In Section 3.3, we describe the steps for the specific art style example discussed .

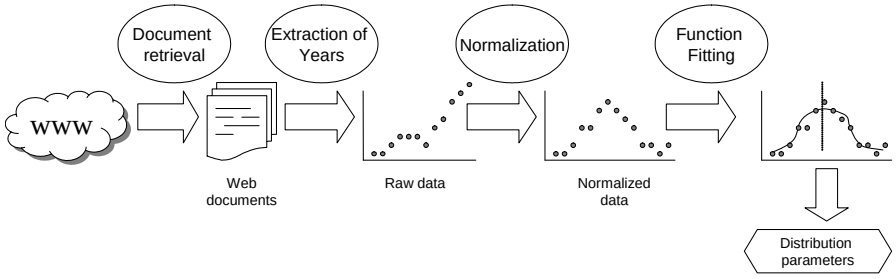

Fig. 1. Outline of the Information Extraction phase

In the first step, we use the full description of the art style instance to retrieve a working corpus of documents from the Web. For the retrieval step the search engine Google[1] is used. We construct a search string by taking all labels of the instance and connecting them with Google's binary search operator "OR". In the case of the instance `aat:Dada` this results in the search string *'"Dada" OR "Dadaism" OR "Dadaist"'*. This string is used to query Google. From the web pages retrieved by Google, we use the first 1000 pages.

In the next step of the algorithm, we extract all strings denoting years from the documents. For this, we use a single simple regular expression (four consecutive integers, followed by a punctuation mark or a whitespace). Note that more elaborate methods could be used to extract occurrences of strings denoting moments in time. However, we assume that the use of a simple method and a large number of documents leads to the same results and is less domain-, language- and structure-specific. For our purposes, we only gather year strings denoting years between 1250 and 2100 A.D. For each of these extracted years, we denote the frequency of the occurrences in all documents of the working corpus. This makes up our raw data.

Of course, a lot of year strings extracted from the working corpus will not be related to the concept but rather are used to denote things such as copyright dates. However, we assume that these dates occur in the working corpus and the other documents on the Web with the same frequency. To eliminate these years, we normalize the raw data. We divide the frequency of each year by the Google hit count for that year. In Figure 2, we show the effect normalization has on the data. This normalized data is used to estimate a distribution in the next step.

[1] http://www.google.com/

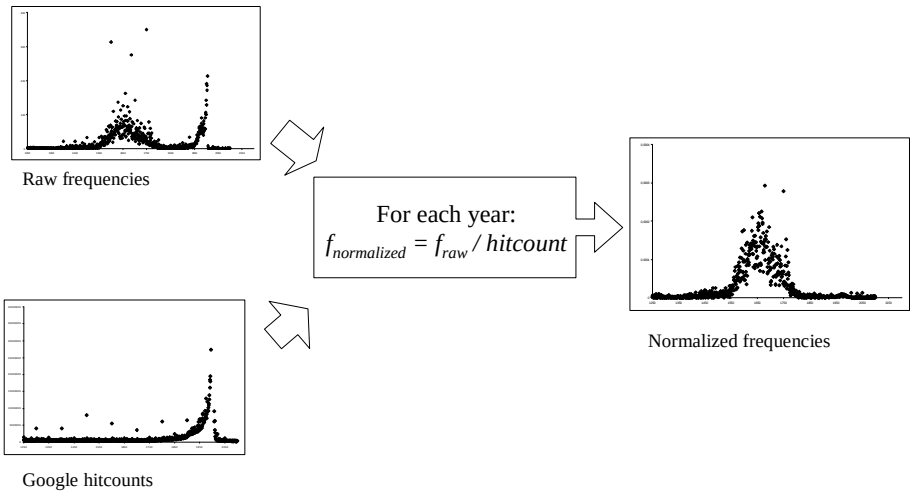

Fig. 2. The normalization of the raw frequency data

In the function fitting step, we fit a function to the data using a numerical fitting procedure. We minimize the sum of square root errors between the data points and the function value. We tested our method with four different functions. We describe these tests in the next section.

3.2 Choosing the Fitting Function

We considered four different function as our distribution function in the fitting step. All four functions originate from a intuitive notion of how moments are related to the periods. Examples of these functions are shown in Figure 3

- A block function, with three parameters (μ_b, σ_b, f_b).

$$g_b(x) = \begin{cases} f_b \text{ if } \mu_b - \sigma_b \leq x \leq \mu_b + \sigma_b \\ 0 \quad \text{otherwise} \end{cases}$$

- A normal distribution, with three parameters (μ_n, σ_n, f_n).

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

- A triangular 'fuzzy' function, with four parameters $(\mu_{f1}, \sigma_{l,f1}, \sigma_{r,f1}, f_{f1})$, where $\sigma_{l,f1}$ and $\sigma_{r,f1}$ denote the left and right base of the triangle (see Figure 3(c)).
- A trapezoid 'fuzzy' function, with five parameters $(\mu_{f2}, \sigma_{l,f2}, \sigma_{r,f2}, \sigma_{c,f2}, f_{f2})$, where $\sigma_{c,f2}$ is the length of the 'plateau' of the trapezoid and $\sigma_{l,f1}$ and $\sigma_{r,f1}$ denote the length of the left and right base sides (see Figure 3(d)).

To find out which of these functions produced the best result, we fitted the functions to nine art styles[2]. The fitting procedure was done using the numerical gradient descent method from MS Excel, minimizing the sum of the Squared Errors between the function value and the normalized frequency data. In Figure 3, we show the best fitting functions applied to the data of a single instance, the art style 'Baroque'.

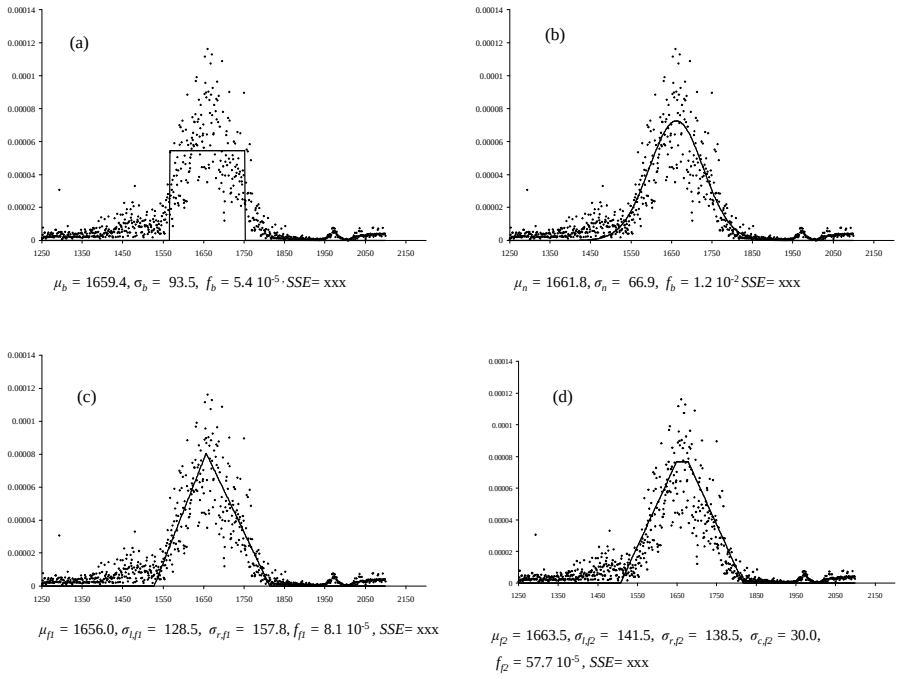

Fig. 3. The best fits for the Block (a), Normal distribution (b), Triangular Fuzzy (c) and Trapezoid Fuzzy (d) functions on the frequency data for the instance 'Baroque' (which approximately lasted during the 17th Century). For each function, the optimal parameter values are shown.

In Table 1, we list the averages of the Sum of Squared Error for the nine art styles that were considered. As can be seen, the Block function performs significantly worse than the other three functions. Of these three the Normal distribution has the best average errors. Because of this and considering the fact that the Normal distribution has only three parameters, as opposed to the five parameters of the Trapezoid Fuzzy function, we decided to use the Normal distribution as the function we use in the fitting step used in the Information

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

Extraction phase. The output of the Information Extraction phase consists of the values of two of the three 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 amount of years found and not the distribution across the time line.

Table 1. Averages of the Sum of Squared Errors for the four tested functions.

Block Function	4.18
Normal Distribution	2.70
Triangular Fuzzy	2.99
Trapezoid Fuzzy	2.77

3.3 Experiments

Setup To test how well our IE method is able to extract the distributions of the periods, we applied our method to extract periods on a set of manually selected art styles from the AAT. We selected the art styles on a number of different criteria. The first criterium was that the target period of the art style occurred within the limits of our algorithm (between 1250 and 2100). A more important selection criterium was that one or more labels of the art style should be unambiguous. For example, we discarded the art style `aat:Symbolist`, since its labels 'Symbolism' and 'Symbolist' do not produce an unambiguous corpus that is about the art style referred when presented to Google. This problem could be solved with simple query expansion techniques using either concepts higher up in the taxonomic structure of the ontology or by consulting other sources such as WordNet.

Selection along these criteria left us with 17 art styles[3]. For each of these art styles, we retrieved a corpus, extracted the year occurrences and fitted the normal distribution to the data. The AAT styles are found in the first column of Table 2. 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 and Results For the evaluation of the distribution we constructed a gold standard. For this, 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. 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

[3] These include the nine art styles from Section 3.2

[4] www.wikipedia.org, www.artcyclopedia.com, www.artlex.com, www.artchive.com, www.encyclopedia.com and www.britannica.com

century...' (www.artchive.org). To 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: First of all, 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 3, we list this gold standard mean and the error. We also list the error relative to the total length of the period (according to our gold standard). Intuitively, 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 with our method and the mean from our gold standard is 17% of the total period length.

Table 2. The 17 Art Styles, with the values for μ and σ that produced the best fit and evaluation of μ to the gold standard means

AAT 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
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:	0.17

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 art style: respectively $\mu - (\tau \cdot \sigma)$ and $\mu + (\tau \cdot \sigma)$. To find the optimal value for τ, 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 was relatively constant with an average of 0.81. In Table 3, we show the start and end dates we obtained with $\tau = 0.81$. These 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 how big this error is relative to the total length of the art style according to our gold standard. These values are also shown in Table 3. Averaged over all art styles the relative error is 0.23, indicating that on average, the start and end dates differ by 23% of the total length from the gold standard.

Table 3. Start and end dates for the 17 Art Styles ($\tau = 0.81$) compared to gold standard

AAT 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
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

To test how concept- and domain-independent the method and the value for τ is, we tested the method on extracting periods on a different domain (wars) and on a different concept in the same domain: artists. We tested the method with four different wars (World War I, World War II, the Hundred Year War and the Eighty Year War). As the labels we used to construct the corpus for the wars we used the aliases listed on their Wikipedia web pages. Also, a gold standard was constructed using these pages. The average relative error of μ was 0.16, comparable to the art style example. The optimal value for τ turned out to exactly be 0.81 also. With this value the average relative error for the start and end dates was 0.47.

For the artists, we selected eight well-known artists[5] and their labels from the Unified List of Artist Names (ULAN)[3] from different time periods. We constructed the gold standard (years of birth and death of the artist) from the Wikipedia pages. We found an average relative error for μ of 0.20. Here the optimal value of τ is different from the art styles and wars examples: 1.12. This however produced a lower average relative error of 0.27.

Discussion In the experiments for the art styles, the errors found are relatively low and especially if we consider the artificial nature of the gold standards (a discretization of very vague notions) we can say that the results are encouraging.

The fact that the optimal value for τ and the relative average errors vary very little between the art style and wars experiments show that the method can be used to extract periods for comparable concepts in different domains.

The differing optimal value for τ in the 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. The higher value of τ that we found indicates that for the method to find the entire lifespan, the start and end years must be found further away from the value of μ.

4 The Semantic Representation Phase

We have acquired a distribution for the periods associated with the art styles from the Information Extraction phase. We now would like to represent this knowledge in an ontology. This can be done in a number of different ways, but this choice is ultimately a modeling decision to be made by the ontology engineer. Different modeling decisions can be made according to the purpose of the ontology or the available and desired reasoning capabilities. In this section, we describe four different possible ways of representing periods in an ontology. For each of these possibilities, we determine the transformation that is to be used to rewrite the distributions found in the first phase to the ontological constructs. As an example we show the results of each of the four modeling choices for the art style 'Baroque' in Figure 4.

Discrete Start and End Dates An intuitive possible representation of a period is a start and end date. In this case, the periods associated with an art style have two relations defined: (eg. `hasStartYear` and `hasEndYear`). These two dates have to be determined from the Normal Distribution found in the IE phase. The most straightforward method has been introduced in the previous section, through the use of the factor τ. If we assume that the optimal value for τ is 0.81, we can determine the start and end dates for each of the art styles (these can be found in Table 3).

[5] Hieronymus Bosch, Vincent van Gogh, Francisco Goya, Gustav Klimt, Edouard Manet, Edvard Munch, Pablo Picasso and Leonardo da Vinci

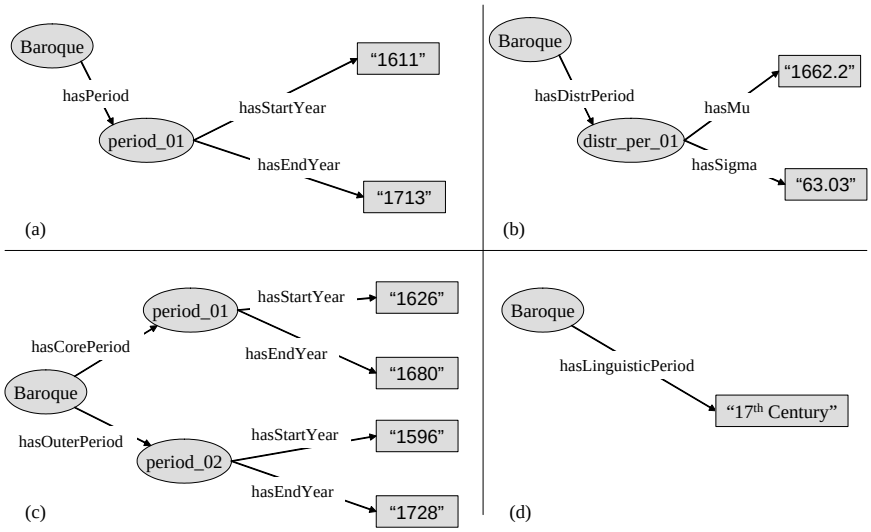

Fig. 4. The Baroque example for the four possible modeling choices discussed: discrete start and end dates (a), the distribution (b), multi-concept definition (c) and linguistic variables (c)

Distribution The representation that is most true to the data from the Information Extraction phase would be using the parameters of the normal distribution 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 (these can be found in Table 2).

Multi-concept definition An ontology engineer can also choose a hybrid solution, that on the one hand uses discrete values, but also takes into account the vagueness of the domain. An example could be to define two periods for each art style, a 'core' period and an 'outer' period, which both have discrete start and end dates. In this case, an extra step is needed as these four values need to be determined from the values of μ and σ. To do this, two factors are needed τ_c and τ_o, where 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 our experiments plus and minus a percentage. In Figure 4, we show the result for 'Baroque' with $\tau_c = 0.57$ and $\tau_o = 1.05$ (plus and minus 30%).

Linguistic Variables The fourth possibility we consider is the use of linguistic values for the periods, such as 'first half of the 18th Century' or '1960s'. This

would be closest to the way experts talk about periods in general and more specifically of art style periods. An ontology then allows for defining semantics of these individual values in such a way that there can be reasoned with it. We do not extensively list all possible rewriting rules from the distribution parameter values to such linguistic values. Instead we give three examples of such possible rules below:

- if $\mu \approx 1650$ and $\sigma \approx 50$ then use 'the 17th Century'
- if $\mu \approx 1725$ and $\sigma \approx 25$ then use 'first/second half of the 18th Century'
- if $\mu > 1900$ and $\sigma \approx 5$ then use 'around μ'

Note that 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 in [4] using the different representations.

5 Related Work

A subtask of Information Extraction and Question Answering (QA) is the identification and extraction of temporal data from textual corpora. Examples are [5] and [6]. The methods used look for patterns in the corpora and extract the periods accordingly. However, in our research we found that in our domain, these types of patterns (such as '...A lasted from X to Y...') often do not occur. Our method is not dependent on these language- and domain-specific phrases as it considers individual moments linked to the concept.

A lot of research has been done on representing time and time intervals. Allen[4] 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. A very popular way of representing time intervals uses of fuzzy intervals (such as in [7]). However, since the broader context of this research is the Semantic Web, we chose ontology constructs to represent the periods. In the context of the Semantic Web, work on constructing ontologies of time is also being undertaken [8] which we might use in the future.

6 Conclusions and Further Research

We have presented a method for the extraction of periods for temporal concepts and shown that this worked reasonably well. In this paper, we have shown that the method works for a number of instances of one concept. The method itself is domain-independent and we indeed have shown that for two other concepts, preliminary results are also promising.

We have also shown a way to convert the found distributions of the periods to ontological constructs. The choice for which constructs to use is a modeling choice to be made by an ontology engineer. Currently, the knowledge extracted

by our method is used in the MultimediaN e-culture browser[9] to facilitate richer browsing of Cultural Heritage repositories.

There are some obvious issues to be solved for further research to test the limitations of the method. We would like to know how well the Information Extraction Phase works with other concepts in other domains. The method can also be expanded with different patterns to test for moments, so that not only years will be extracted, but also the more vague notions such as 'end of the 1960s'. The range of years that can be extracted (currently 1250-2100) can also be expanded. To further improve the IE phase, we would also like to further investigate the fitting procedure. The dominance of positive errors in Table 2 indicates that the method consequently places a period too early and more insight in the reasons for this could lead to further improvements for the method.

Also, in the future, we plan to propose a framework for the integration of different modules that extract knowledge for a single ontology. For instance, in [2], we describe a system that can extract the instances of artist-art style relations. We assume that this knowledge can aid the extraction of periods for the same art styles and vice versa. For example, if we know that an artist linked to an art style produced a painting in a certain year, this might suggest that the art style was still active in that year.

Acknowledgements

This research was supported by the MultimediaN project (www.multimedian.nl) funded through the BSIK programme of the Dutch Government.

References

1. The Getty Foundation: Aat: Art and architecture thesaurus. http://www.getty.edu/research/tools/vocabulary/aat/ (2000)
2. de Boer, V., van Someren, M., Wielinga, B.J.: Extracting instances of relations from web documents using redundancy. submitted (2001)
3. The Getty Foundation: Ulan: Union list of artist names. http://www.getty.edu/research/tools/vocabulary/ulan/ (2000)
4. Allen, J.: Maintaining knowledge about temporal intervals. Communications of the ACM **26**(11) (1983) 832–843
5. Llido, D., Llavori, R.B., Cabo, M.J.A.: Extracting temporal references to assign document event-time periods. In: Proceedings of the 12th International Conference on Database and Expert Systems Applications. (2001) 62–71
6. Ahn, D., Schockaert, S., de Rijke, M., De Cock, M., Kerre, E.E.: Extracting, representing, and grounding events for temporal question answering. In: 6th Meeting of Computational Linguistics in the Netherlands (CLIN 2005). (2005)
7. Ohlbach, H.: Relations between fuzzy time intervals. In: Proceedings 11th. International Symposium on Temporal Representation and Reasoning. (2004) 44–51
8. Hobbs, J.R., Pan, F.: An ontology of time for the semantic web. CM Transactions on Asian Language Information Processing **3**(1) (2004)
9. Multimedian E-Culture Browser : (online demo). http://e-culture.multimedian.nl/demo/search (2006)