XML REPRESENTATION AND MANAGEMENT OF TEMPORAL INFORMATION FOR WEB-BASED CULTURAL HERITAGE APPLICATIONS

Fabio Grandi

C.S.I.TE.-C.N.R. and Dipartimento di Elettronica, Informatica e Sistemistica
Alma Mater Studiorum – Università di Bologna, Viale Risorgimento 2, I-40136 Bologna, Italy
Email: fgrandi@deis.unibo.it

ABSTRACT

In this paper we survey the recent activities and achievements of our research group in the deployment of XML-related technologies in Cultural Heritage applications concerning the encoding of temporal semantics in Web documents. In particular we will review "The Valid Web", which is an XML/XSL infrastructure we defined and implemented for the definition and management of historical information within multimedia documents available on the Web, and its further extension to the effective encoding of advanced temporal features like indeterminacy, multiple granularities and calendars, enabling an efficient processing in a user-friendly Web-based environment. Potential uses of the developed infrastructures include a broad range of applications in the cultural heritage domain, where the historical perspective is relevant, with potentially positive impacts on E-Education and E-Science.

Keywords: XML, Temporal data management, Cultural Heritage, Digital libraries, Semantic Web

1 INTRODUCTION

The eXtensible Markup Language (W3C Consortium, 1997) is becoming the new emerging standard for data management and exchange over the Internet (Abiteboul, Buneman & Suciu, 1999). In particular, a great deal of interest concerns its adoption for the representation and integration of structured and unstructured data. Moreover, an outstanding (and very appealing for Cultural Heritage applications) XML feature is the capability of easily encoding semantic information in Web documents as *metadata*, which can automatically be used by advanced computer tools, like "intelligent" search engines, in the direction of the so-called Semantic Web (Semantic Web Agreement Group, 2001).

In this context, the CSITE-CNR database research group (CSITE-CNR DataBase and KnowledgeBase Group, 2001) has been interested in recent years in the introduction of *temporal* aspects to the Web, by adapting and extending concepts and techniques derived from the experience gained in more than ten years of temporal database research (Soo, 1991; Kline, 1993; Tsotras & Kumar, 1988; Tansel, Clifford, Gadia, Jajodia, Segev & Snodgrass, 1993; Etzion, Jajodia & Sripada, 1998). In particular, the pioneering work of our group for temporally extending the World Wide Web dates back to 1997. In fact, we presented our basic ideas during the Dagstuhl Seminar on temporal databases in June 1997, as witnessed by (Etzion, Jajodia & Sripada, 1997) and "Summaries of Current Work" in (Etzion et al., 1998). In such research, we first explored the applicability to the Web of the basic notions of *transaction time* and *valid time*. According to (Jensen, Clifford, Elmasri, Gadia, Hayes, Jajodia et al., 1998), transaction time is the time that some fact is *current in a database*, from when it is stored in the system to when it is deleted, whereas valid time corresponds to the time that some fact is *valid in the real world*. With respect to the Web (Grandi & Scalas, 1998), transaction time concerns the availability and versioning of Web resources, whereas valid time concerns the temporal validity of the information carried by the contents of a Web resource. For example, by moving back along transaction time we could access previous versions of the same Web page. However, for Cultural Heritage applications, the dimension of interest is mainly valid time, as it allows the explicit encoding of historical information within Web documents and the imposition of a "temporal view" for selective navigation and browsing of the Web contents.

Our early work was mainly developed in the context of the national research project INTERDATA ("Methodologies and technologies for data and process management on Internet and Intranet networks" (Project INTERDATA, 1997), project co-funded by the Italian Ministry of the University and Scientific Research). In particular, in the first part of the project we studied the problems related to the introduction of transaction time. We investigated several techniques for implementing a Web site with versioned resources, aimed at reducing data duplication, and extensions to the HTTP protocol for negotiating transaction-time versions (Cristofori, Grandi, Mandreoli & Scalas, 1998). We also developed a prototype Web site (Cristofori, Grandi, Mandreoli & Scalas, 1999) with temporal navigation facilities along the transaction-time axis. As to valid time, we proposed in (Grandi & Scalas, 1998) the first timestamping scheme for the explicit encoding of historical information in Web pages, based on non-standard HTML and custom browser extensions. At the same time, we also investigated "killer applications" for the best exploitation of valid-time temporal semantics on the Web. As a result, we focused on Cultural Heritage applications with our participation in another national research project, called "Cultural Heritage" (Project Cultural Heritage 2001), funded by the Italian National Research Council.

Subsequently, in order to put into practice the valid-time dimension in the Web, we combined our basic ideas with the emergence of the XML-related technologies and developed an XML/XSL infrastructure, named "The Valid Web", for the management of temporal documents and data (Grandi & Mandreoli, 1999; Grandi & Mandreoli, 2000c). The proposed techniques enable the explicit encoding of distinguished temporal/historical information within XML (or even legacy HTML) documents, whose contents can then be selectively accessed according to their temporal validity with any XML-compliant Web browser, like Microsoft Internet Explorer 5 (Microsoft, 2002) and its successive versions. In order to show its potential, the infrastructure has been implemented on a demo prototype (Grandi & Mandreoli, 2000a) —also available on-line (Grandi & Mandreoli, 2000b)— showing, as an application example, the functionalities of a temporal fine-arts Web museum (Pioch, 1996), that is a virtual environment in which it is possible to carve out personalized visitor routes for a specific epoch of interest.

Furthermore, our research group was also recently involved in the development of an interesting application and extension of "The Valid Web" approach for the management of ancient text sources in digital form. The general framework of such work is the "XML/Repetti" project, a collaboration with the Computer and History group at the University of Florence (Niccolucci, Zorzi, Baldi, Carminati, Salvatori & Zoppi, 1999). They have been involved in the study of a new edition, in electronic form, of Emanuele Repetti's historical-geographical dictionary of Tuscany (nineteenth century) for a couple of years. In particular, our contribution to the "XML/Repetti" project is the extension of "The Valid Web" infrastructure to deal with the specificity and semantic richness of the temporal information stored in *Repetti*'s Dictionary and similar textual sources (involving vague and imprecise expressions with the use of multiple granularities and calendars) and the redesign of the overall system architecture with efficient organization of the temporal search engine and of the large resulting XML document repository, including optimized search algorithms and temporal indexing facilities (Grandi & Mandreoli, 2001a; Grandi & Mandreoli, 2001b).

We emphasize the fact that the worldwide accessibility of *Repetti*'s Dictionary and analogous ancient text sources, which will be enabled by Web publishing, has a noteworthy relevance from a Cultural Heritage and also a scientific point of view, as frequently written sources have the same importance as material evidence in medieval archaeology. The increasing role of the Internet in archaeological investigation has already been pointed out, as widespread, fast and easy sharing of information on the Web has a substantial impact on the archaeological methodology (Bogdanovic, Vicente & Barcelo, 1999; Hermon & Niccolucci, 2000), which could be boosted by the deployment of XML-related technologies, as also evidenced in (Niccolucci et al., 1999; Benvenuti, Niccolucci, Baragli & Carpini, 2000; Niccolucci, 2002). The representation and automated management of temporal aspects adds computational search power to simple accessibility, enabling a radical change with respect to the availability of paper printed editions of sources.

To sum up, potential uses of the developed infrastructures include a broad range of applications in the Cultural Heritage domain, where the historical perspective is relevant, with a potential positive impact on Electronic Education (e.g. by publishing digital libraries, virtual museums and archaeological sites with historical indexing and advanced temporal filtering facilities on the Web) and Electronic Science (e.g. by allowing history and archaeology researchers to share primary sources as large collections of hypertextual documents provided with semantically and computationally powerful temporal search engines on the Web).

The rest of the paper is organized as follows. Section 2 reviews "The Valid Web" approach, whereas Section 3 describes its developments in the "XML/Repetti" project. Conclusions can be found in Section 4.

2 "THE VALID WEB": A SIMPLE TEMPORAL XML/XSL INFRASTRUCTURE

The addition of valid time to Web documents proposed in "The Valid Web" approach (Grandi & Mandreoli, 2000c) is based on the extension of the XML markup language (W3C Consortium, 1997) with timestamping tags. The proposed infrastructure —which also includes an XML schema (W3C Consortium, 2000) and an XSL stylesheet (W3C Consortium, 2002)— is based on current Web technology and only requires browsers supporting XML (like the latest versions of Microsoft Internet Explorer). In particular, the extension consists of a new XML tag, `<valid>`, to define a *validity context*. The validity context is used to assign a specific time pertinence to a piece of a multimedia document for temporally selective manipulation. Simple timestamps can be specified in a validity context by means of `<validity>` tags, which allow the definition of a temporal interval through its boundaries (i.e. the values of the `from` and `to` attributes of the `<validity>` XML element). In general, multiple intervals can be used: in this case, the timestamp is defined as the union of all the validity intervals specified; formally, the timestamp is a *temporal element* as defined in the BCDM temporal data model (Jensen, Soo & Snodgrass, 1994). For instance, the following code:

```
<!-- definition of a validity context -->
<valid>

    <!-- valid-time timestamping -->
    <validity from="1280-01-01" to="1285-12-31" />
    <validity from="1295-01-01" to="1300-12-31" />

        This is text <b>valid from 1280 to 1285</b>
            but also <b>valid from 1295 to 1300</b>...

</valid>
```

defines a validity context whose validity is $[1280\text{–}1285] \cup [1295\text{–}1300]$. The time constants can be specified according to the ISO 8601 format (Wolf & Wicksteed, 1997), which also corresponds to the XML "`date`" data type.

From a system architecture viewpoint, "The Valid Web" is simply based on client-side document processing on top of standard Web technology (with XML-enabled browsers). In particular, the temporally selective navigation of downloaded documents is based on a client-side filtering of the document contents to be displayed by means of a provided XSL stylesheet. In other words, "The Valid Web" approach requires the temporal document to be retrieved from the Web server and then processed by the XSL stylesheet in the main memory space managed by the browser. This solution is optimal, as it minimizes network traffic, in the presence of small temporal documents and, in particular, when users often change the temporal context during the navigation of a page. In such a case, the XSL stylesheet is updated on-the-fly to reflect the new temporal selection condition and then dynamically re-applied to the document through DOM method calls (W3C Consortium, 2001) invoked by JavaScript control functions (being the stylesheet loaded as a document object in the main memory managed by the browser). For this reason, the approach is best suited for the temporal re-engineering of legacy HTML-based Web sites, since the resulting temporal documents only have a very small space and network transfer overhead compared to the non-temporal case (this is basically due to the added timestamping tags, in addition to a bit of extra XML formatting, needed to make legacy documents *well-formed*).

```
<?xml version="1.0" ?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/TR/WD-xsl">

  <!-- identity transformation template -->
  <xsl:template>
    <xsl:copy>
      <xsl:apply-templates
            select="@*|*|comment()|pi()|text()" />
    </xsl:copy>
  </xsl:template>

  <!-- recursive valid-time selection template -->
  <xsl:template match="valid">
    <xsl:choose>
      <xsl:when test="validity[condition on from and to values]">
        <xsl:copy>
          <xsl:apply-templates
                select="@*|*|comment()|pi()|text()" />
        </xsl:copy>
      </xsl:when>
      <xsl:otherwise>
        <xsl:apply-templates select="*//valid" />
      </xsl:otherwise>
    </xsl:choose>
  </xsl:template>

</xsl:stylesheet>
```

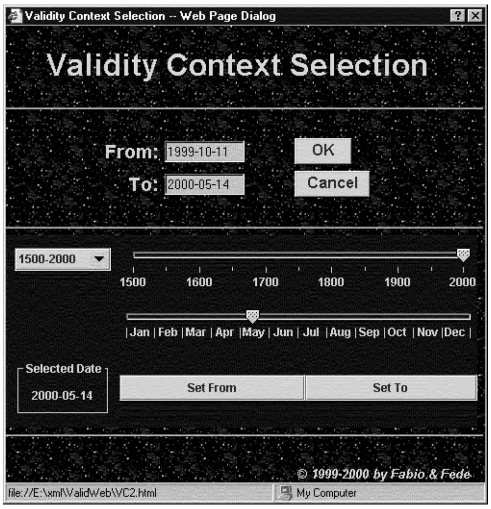
Figure 1: The XSL `Valid.xsl` stylesheet

In more detail, valid-time selection relies on the XSL stylesheet, named `Valid.xsl`, which can be seen in Figure 1: the first part consists of a simple identity-transformation template, while the second part is devoted to the temporal selection of the contents of valid contexts. The processing of the new XML `<valid>` element causes the output of the element's contents when a validity selection condition (involving the `<validity>` timestamps) is verified. For instance, if the condition has the form:

```
@from[.$le$ ’1500-12-31’ ] and @to[.$ge$ ’1500-01-01’ ] ,
```

each `<valid>` element whose validity *overlaps* the year 1500 is included in the stylesheet output: the selection condition matches any `<validity>` element where the `from` attribute value is $\leq$1500/12/31 and the `to` attribute value is $\geq$1500/1/1. The particular structure of the selection template causes the execution of the overlap test with the navigation context on all the `<validity>` timestamps found in the current `<valid>` element. The conditional processing uses the `xsl:choose` instruction which provides for an `xsl:otherwise` case (not supported by the `xsl:if` XSL element), in order to recursively look for nested validity contexts. The `xsl:when` instruction is activated if at least one of the intervals (corresponding to a `<validity>` element) belonging to the timestamp satisfies the selection condition. The `xsl:otherwise` instruction is activated only when none of the timestamps of the current `<valid>` environment satisfies the selection condition.

The Valid Web approach is aimed at supporting temporal navigation in virtual environments which are sources of historical information. An extremely appropriate example of such an environment is a Web museum, where temporal selective browsing allows the definition of personalized visit paths through centuries and artistic or historical periods within the museum collections. In order to plan a visit, we can use valid time selection to change the historical period of interest. For instance, we can choose the High Renaissance period, by selecting the validity



Validity Context Selection -- Web Page Dialog

Validity Context Selection

From: 1999-10-11

To: 2000-05-14

1500-2000

Selected Date 2000-05-14

file://E:\xml\ValidWeb\VC2.html

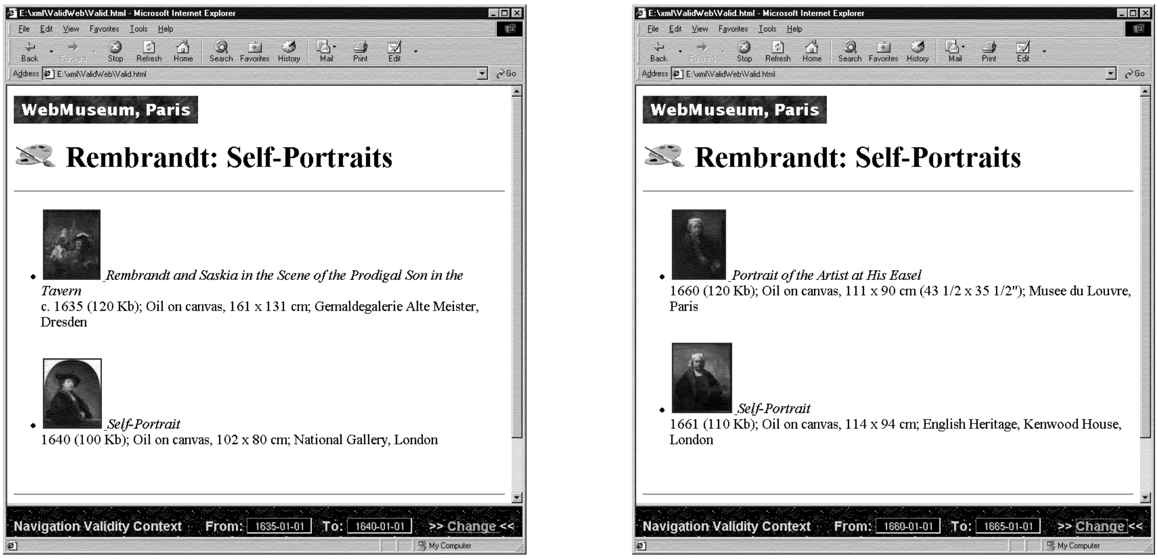
OK
Cancel
1500
1600
1700
1800
1900
2000
|Jan |Feb |Mar |Apr |May |Jun |Jul |Aug |Sep |Oct |Nov |Dec |
Set From
Set To
© 1999-2000 by Fabio & Fede
My Computer

Figure 2: The Java Applet for the selection of a Validity Context

range 1495–1520. Hence, we may start our virtual visit entering some virtual hall or gallery with only the temporally relevant paintings or sculptures present; by changing the validity context we could see some works vanish and other works materialize. For example, in a hall dedicated to the Italian High Renaissance, we could view the evolution of the painting styles of Leonardo da Vinci, Raphael, Michelangelo and Titian and, say, have a look at works contemporaneous to the Mona Lisa picture.

The interest for museum applications on the Internet is constantly growing. This is shown by the increasing number of available museum sites and by the development of a specific discipline (Archives & Museum Informatics, 2001), with dedicated journals (e.g. *Archives and Museum Informatics*) and conferences (e.g. *Museums and the Web*). The "Web Museum", authored and maintained by Nicholas Pioch (1996), was one of the very first to open and is probably the most popular virtual museum on-line. It is basically a collection of image data representing famous paintings, heterogeneous as to their origin, which can accessed, for example, via an artist or a theme index. In order to test our proposal, we realized a temporal version of a subset of the Web Museum pages and developed a Web environment for the temporal browsing of its collections (Grandi & Mandreoli, 2000a; Grandi & Mandreoli, 2000b). The pages of the site are organized into two frames (see Figure 3). A small service frame in the bottom part of the window contains all the required controls to deal with the user's specification of the validity context to be used for temporal navigation, including the visualization of the current validity context. All the controls are implemented as JavaScript functions. A larger frame, occupying nearly all the browser's window space, is used to display temporal documents, that is the results of the temporally selective filtering effected by the `Valid.xsl` stylesheet on timestamped XML documents. The results of such filtering is a plain HTML document which is then rendered by the browser as usual.

In general, the valid-time selection implies the choice of an interval. This can be done by independently choosing two time points representing the interval boundaries. The selection of each interval boundary can be based, for instance, on a graphic *scrollbar* or *slider* for the fine selection of a time-point (at a given granularity level). In our prototype implementation, time-points are dates (i.e. the granularity level is the day) and the selection of an interval can be effected by means of a Java JFC/Swing applet (Sun Microsystems, 2002b), which contains two graphic sliders: the former to select the year and the latter to select the day of the year (see Figure 2). The former slider, for user's convenience, has a 500-year range, which can be changed (from 0–500 to 2000–2500) by means of a *multiple-choice menu* available next to the year slider. Assume we have to fix a date, say 1596/3/7. We can start by choosing the year 1596 with the former slider (with the default range 1500–2000 set) and then choose the March, 7 date with the latter slider. The chosen value can then be assigned to the From or To interval boundary by means of the corresponding "Set" *button*. However, editable input fields for directly typing in a valid time value (in the "`YYYY-MM-DD`" string format) are also always available in the dialog window containing the running applet. The communication between the applet and the JavaScript control functions in the calling service frame (e.g. to return the selected validity context) is managed by means of the LiveConnect package (LiveConnect, 2002)



WebMuseum, Paris

Rembrandt: Self-Portraits

- *Rembrandt and Saskia in the Scene of the Prodigal Son in the Tavern*
  c. 1635 (120 Kb); Oil on canvas, 161 x 131 cm; Gemaldegalerie Alte Meister, Dresden
- *Self-Portrait*
  1640 (100 Kb); Oil on canvas, 102 x 80 cm; National Gallery, London

Navigation Validity Context From: 1635-01-01 To: 1640-01-01 >> Change <<

WebMuseum, Paris

Rembrandt: Self-Portraits

- *Portrait of the Artist at His Easel*
  1660 (120 Kb); Oil on canvas, 111 x 90 cm (43 1/2 x 35 1/2"); Musee du Louvre, Paris
- *Self-Portrait*
  1661 (110 Kb); Oil on canvas, 114 x 94 cm; English Heritage, Kenwood House, London

Navigation Validity Context From: 1660-01-01 To: 1665-01-01 >> Change <<

Figure 3: Temporal navigation of the Web Museum

supported by the Java Plug-in 1.2.2 (Sun Microsystems, 2002a).

Once the navigation validity context choice is confirmed by the user, the temporal filtering of the currently displayed document is automatically re-executed by means of the DOM method calls described above. Furthermore, in order to enable fully fledged temporal navigation, the current validity context is automatically "inherited" by the newly loaded page, if the new document is also a temporal XML one, each time the user changes the displayed document in the usual way (e.g. by following a link). This behaviour is forced in our prototype thanks to the dynamic HTML facilities supported by Ie5. In fact, we used a slightly modified `Valid.xsl` stylesheet with respect to Figure 1. The actual stylesheet implements a dynamic call-back mechanism by inserting some JavaScript code in the preamble of the processed document. Such a script provokes the immediate activation of the temporal selection functionality when the document is loaded by the browser. When the validity context is changed, the `Valid.xsl` filter is updated on the fly (to include the overlap with the current validity context as a selection condition) and then re-applied to the displayed document.

For example, Figure 3 shows two snapshots of navigating through a sample page containing Rembrandt self-portraits. The full page contains seven pictures, dating from 1629 to 1669. The left hand side of Figure 3 shows the page when the validity context has been set to [1635–1640] and only two pictures are visible (the third and the fourth one) while the right hand side shows the same page with the validity context changed to [1660–1665] and two other pictures have been displayed (the fifth and the sixth one). The current navigation context is always visible in the bottom service frame of the window. The "Change" command on the right is a link that activates the applet of Figure 2.

We emphasize that the temporal browsing and navigation enabled by the deployment of the Valid Web infrastructure is really based on the temporal semantics of the contents of Web documents, as encoded by their creators. This is very different from the use of a "traditional" Web search engine (like Google or Altavista) to retrieve Web pages *containing text explicitly matching a search string representing dates or time periods which must supplied by the user*.

In general, the purpose of our proposed timestamping scheme is (at least) twofold:

- it can be used to make temporal "traditional" Web sites (featuring HTML multimedia documents), in order to support the representation of historical information, and enabling a temporally selective navigation with respect to information validity; our Web Museum application corresponds to such an approach;
- it can be used to design new XML-based Web sites and applications, in order to support the management of structured or semi-structured temporal data, and enabling the utilization of functionalities developed by

temporal database research (e.g. TSQL2-like temporal query languages (Snodgrass, Ahn, Ariav, Batory, Clifford, Dyreson et al., 1995)).

Referring to our reference application, the legacy HTML pages making up the Web Museum were first converted into XML documents (their original HTML markup also needed some checking and correcting for conversion into *well-formed* HTML code). This phase of Web site re-engineering can largely be automated. Human intervention and expertise is required during the second re-engineering phase for the addition of timestamps, as the pieces of information to be enclosed in `<valid>` environments have to be carefully identified and appropriate time values have to be assigned to `<validity>` timestamps.

Other kinds of Web sites of interest for Cultural Heritage applications likely to benefit from the encoding of temporal semantics include virtual archaeological sites, historical digital libraries and any other collection of multimedia data and (hyper)textual information where the time dimension may be helpful for selective browsing and navigation.

3 EXTENSION OF THE ENCODING SCHEME: THE “XML/REPETTI” PROJECT

As a starting point, the straightforward application of “The Valid Web” approach to the “XML/Repetti” project would allow the uniform classification and encoding of temporal information contained in the dictionary and make availabile techniques for temporal search support. However, the advanced functionality and efficiency specifications (*Repetti*’s Dictionary is quite a *large* collection of text), in addition to the specificity of the dictionary contents, required an improvement of the basic approach. Such an improvement was aimed at fulfilling three goals:

1. The extension of the XML/XSL infrastructure to deal with the semantic richness of the temporal information stored in *Repetti*’s Dictionary and similar textual sources;
2. The redesign of the overall system’s architecture, including the efficient organization of the temporal search engine (with optimized search algorithms) and the XML document repository (provided with temporal indexing facilities);
3. Last but not least, the design and implementation of a user-friendly tool for computer-aided temporal encoding and document markup, which could save history or archaeology researchers from “manual” intervention and editing as much as possible.

In particular, the infrastructure extension required an enhancement of the markup scheme and search mechanism in order to be be able to capture the semantics of widespread temporal expressions in *Repetti*’s Dictionary involving:

- **indeterminacy** (as in: “towards the end of 1653”);
- **multiple calendars** (e.g. use of the Julian calendar);
- **different granularities** (e.g. months *versus* years).

In particular, special attention was devoted to the indeterminacy problem, which has interesting theoretical implications and required the most consistent infrastructure extensions. Starting form the analysis of a large *corpus* of historical sources such as *Repetti*’s Dictionary, we introduced in (Grandi & Mandreoli, 2001a) a broad classification of the temporal expressions denoting single indeterminate events into four main categories (actually, our classification is based on the analysis of texts written in Italian, but we think it can be applied to texts written in other languages as well). If we denote the literal time written in the text by the term Reference Temporal Expression (RTE), the four categories correspond to the use of temporal expressions with the form: “in RTE” (to reference a validity shorter than the RTE duration) for category C_1, “at the beginning (end) of RTE” for C_2 (C_3), “around RTE” for C_4, as in the following examples:

- The abbey was consecrated to St. Martin **in 1276** (C_1).

- The third circuit of the city walls was added **at the beginning of the fourteenth century** (C_2).
- The famous painter died from the plague **near the end of March 1532** (C_3).
- The delegation of the Emperor arrived in Rome **around Christmas 1467** (C_4).

Notice that in the (actually very frequent) C_1 case, we are in the presence of a so-called *granularity mismatch* (Dyreson & Snodgrass, 1998), where a determinate expression with higher granularity is used to denote an indeterminate expression with lower granularity. As a matter of fact, it is quite likely that the example refers to an event, that happened on a certain date in 1276, rather than to an activity lasting for the whole year. Moreover, we cannot ascertain *a priori* on which day the event actually happened and there is no reason to prefer one date to another. On the other hand, the example "The castle was restored after the fire **between 1549 and 1553**" concerns a real interval since the restoration probably required several years to complete. However, since the exact date the work began and ended is not known, the expression denotes an indeterminate interval, whose boundaries are indeterminate C_1-type dates.

Summing up, every indeterminate temporal expression found in the text concerning a single event can be reduced to an indeterminate date or to an interval whose boundaries are indeterminate dates falling in one of the categories above, whose representation is addressed in the next Subsection. We want to stress that the adopted encoding scheme, although based on four coarse categories and bound to possibly arbitrary interpretations, does not lead to an "impoverishment" of the source representation, as the original text is always maintained in its entirety, and is always visible by the scholar, historian or archaeologist browsing the sources. We only *add* (as metadata) to the original text a suitable markup that enables the exploitation of powerful and fast search engines, which can at least be used to effectively trim down the amount of text that must be examined by experts for a refined analysis. This is very different, for example, from relational databases, which have been used in recent years for the storage and management of archaeological records, where the required normalization and standardization (e.g. wrt dictionaries) of data extracted from antiquarian reports or excavation diaries leads to information loss (Grandi & Niccolucci, 2000).

3.1 Representation of indeterminate dates

In the field of temporal databases, there are basically two mainstream approaches to the management of temporal indeterminacy: the *probabilistic* approach "*à la* TSQL2" (Dyreson & Snodgrass, 1998; Snodgrass et al., 1995) and the *fuzzy* approach (Dutta, 1989). In particular, Dutta (1989) used a *fuzzy set* approach (Zadeh, 1988) to deal with *generalized temporal events*, that is individual events with multiple occurrences. For instance, the event "Tom has high fever" can occur at different time instants, according to the fluctuations of Tom's body temperature. This happens because the meaning of "high" is, to a certain extent, not completely specified (i.e. "high" is not a so-called *crisp* predicate). In Dutta's model, a generalized event allows the representation of all the possibilities for "high", from which the user can select a subset on the basis of his/her judgment and interpretation (e.g. Ann could consider as "high" a body temperature more than 37.5 Celsius degrees). Instead, with the probabilistic approach, we prefer to represent an *indeterminate validity* associated with the occurrence of an event, which remains conceptually single, even if we only know its probability distribution. Hence, an indeterminate instant becomes a set of possible alternatives, only one of which represents the actual validity, with an associated probability. On the other hand, in a *fuzzy* set, each element always belongs to the set, to a greater or lesser extent depending on its membership *degree*. The two approaches are very different as far as the representation of incomplete temporal information is concerned and, in our case, we prefer the latter approach: a historical event (say the death of a King) must occur on a precise date, even if no precise and unique determination can be found in the sources.

The probabilistic model has been introduced into the design of the temporal query language TSQL2 (Snodgrass et al., 1995), which is a consensual proposal for temporal extensions of the standard query language SQL. The TSQL2 probabilistic model has been further developed by Dyreson and Snodgrass in (1998). In this approach, an indeterminate event t is represented through its *probability distribution* P, which is not null on an interval of possible occurrence, whose boundaries (t^- and t^+) are termed the *lower support* and *upper support*:

$$t = (t^- \sim t^+, P)$$

Table 1: Distributions associated with indeterminate dates

Category	Prototype expression	Shape	Distribution
C_1	*about in...*	Flat	**DURING**
C_2	*at the beginning of...*	Decreasing	**EARLY**
C_3	*at the end of...*	Increasing	**LATE**
C_4	*around...*	Bell-shaped	**AROUND**

where $P(i) = \Pr[t = i]$ with $\sum_{i=t^-}^{t^+} P(i) = 1$ and $P(i) = 0$ if $i < t^-$ or $i > t^+$. For query evaluation, two indeterminate instants are considered equivalent ($t_1 \equiv t_2$) if and only if they have exactly the same supports and distributions. Moreover, TSQL2 introduces a suitable extension of the temporal order relation, that is a new definition of the "*Before()*" primitive that is used to define all the other temporal comparison operators (Snodgrass et al., 1995). To deal with indeterminate dates, the "*Before()*" primitive includes an additional parameter to specify an ordering *plausibility*, whose value can range from 0 to 100 (high plausibility means a high precedence probability between the compared instants). Its complete definition thus becomes:

$$Before(p, t_1, t_2) := \neg(t_1 \equiv t_2) \wedge \Pr[t_1 < t_2] \geq p/100$$

where the precedence probability is evaluated as:

$$\Pr[t_1 < t_2] = \sum_{i<j} P_1(i)P_2(j) \qquad (1)$$

where $P_k(x)$ is the occurrence probability of t_k at the instant x. As far as the possible probability distributions are concerned, we adopted —as described in detail in (Grandi & Mandreoli, 2001a)— a small set of predefined distributions which can be assigned to the C_1–C_4 indeterminacy categories summarized in Table 1. The discrete (on a one-day basis) probability densities of the predefined distributions are *piecewise-constant* functions over a small number of equal *base intervals* between the lower and upper supports. It has been shown in (Grandi & Mandreoli, 2001a) how this choice, which is fairly correct from a semantic viewpoint, allows us to exploit extremely efficient comparison algorithms —with an optimized evaluation of formula (1)— without any storage space overhead, which, on the contrary, would have made the direct application of the basic approach in (Dyreson & Snodgrass, 1998) unfeasible. For all the distributions except the uniform, we also considered variants consisting in a greater or lesser accumulation around the mean value (namely VERY_EARLY, VERY_LATE, STRICTLY_AROUND and WIDELY_AROUND), which will correspondingly imply a different number of base intervals. Hence, indeterminate dates can then be represented by a pair (I, P), where I is the *principal interval* and P is one of the available distributions. The principal interval is the base interval where P takes its maximum and exactly corresponds, in the C_1 and C_4 cases (which are by far the most frequent), to the RTE originally written in the text. In any case, it is a more intuitive parameter to identify than the lower and upper supports.

In order to implement temporal search facilities for *Repetti*'s Dictionary, the *validity context* of interest is the dictionary *item*, which becomes the target unit for the search engine. To this end, the textual contents of every item have to be enclosed in a tag pair <ITEM> ... </ITEM>, which can then be selected on the basis of the encoded temporal expressions they contain. For example, if we are interested in a particular time period, we have to look for every item which contains at least one expression that overlaps the period. Temporal expressions of interest include single dates and time intervals, which can be specified through their beginning and end dates. For this purpose, we introduced a "basic type" DATE, to be used alone or in pairs to represent events or intervals, respectively.

By means of the DATE type, we will be able to define the <EVENT> and <INTERVAL> tags. The <EVENT> tag will contain the <AT> XML element with DATE type, while the <INTERVAL> tag will contain the <FROM> and <TO> elements, both with DATE type. In this way, events can be represented via structures like:

```
<EVENT>
     <AT ...  />
         text of the temporal expression (event)
</EVENT>
```

whereas the interval markup will be like:

```
<INTERVAL>
     <FROM ...  />
     <TO ...  />
         text of the temporal expression (interval)
</INTERVAL>
```

The base type `DATE` (which is actually represented as a "macro" `ENTITY` in the DTD that is provided in Figure 4) has several attributes, some of which are specific for supporting indeterminacy:

- `GRANULARITY`, which allows the specification of the granularity used to express the date value as `"DAY"` (default), `"MONTH"`, `"YEAR"` or `"CENTURY"`;
- `VALUE`, which allows the specification of the date expression (obviously in a way consistent with the assigned granularity);
- `INDETERMINATE`, with values `"YES"` or `"NO"` (default), which specifies whether the date is expressed in the indeterminate format or not; when the attribute value is `YES`, the following attributes have meaning:
  - `DISTRIBUTION`, whose value can be one of the supported probability distributions in Table 1 (with their variants);
  - `DURATION`, which expresses (with a default value of `"1"`), as granularity multiples, the amplitude of the *principal interval* (also corresponds to the width of all the base intervals on which the probability density is constant);
- `CALENDAR`, which allows a specific calendar to be referenced, as will be explained later.

The *principal interval* is expressed in an *implicit* way, by means of an interval having as a lower boundary the first day of the `VALUE`'s temporal expression (e.g. 1456/1/1 for `VALUE="1456"`) and an amplitude which can be evaluated as the specified granularity `GRANULARITY`, converted into days, multiplied by the value of the `DURATION` attribute. Such an interval coincides with the whole interval between the lower and upper support in the case of uniform distribution (`DURING`), and with the interval in which the probability is maximal in the other cases (e.g. it is the central interval in the case of `AROUND`, the initial one in the case of `EARLY`). Without losing generality, this choice permits the easy exploitation of granularity to directly express base intervals with unit durations. In this way, the sample expression "around year 1622" can be encoded as:

```
<EVENT>
     <AT GRANULARITY="YEAR" VALUE="1622"
         INDETERMINATE="YES" DISTRIBUTION="STRICTLY_AROUND" />
              around year 1622
</EVENT>
```

or, if we prefer the (default) day granularity, as:

```
<EVENT>
     <AT VALUE="1622-01-01" INDETERMINATE="YES"
         DISTRIBUTION="STRICTLY_AROUND" DURATION="365" />
              around year 1622
</EVENT>
```

The choice of an *implicit* encoding of supports makes it a bit more "transparent" and user-friendly, so that the user (i.e. the history researcher) can concentrate on choosing an intuitive "form factor" among a few available alternatives rather than on the mathematical details of distributions like the support computation or the variance.

Pissinou, N., Snodgrass, R.T., Elmasri, R., Mumick, I.S., Öszu, M.T., Pernici, B., Segev, A., Theodoulidis, B. & Dayal, U. (1994) Towards an Infrastructure for Temporal Databases: Report of an Invitational ARPA/NSF Workshop, *ACM SIGMOD Record 23*(1), 35–51.

Project Cultural Heritage (2001) Homepage of Project Safeguard of Cultural Heritage. Retrieved January 10, 2002 from CNR, Italy: http://www.culturalheritage.cnr.it/.

Project InterData (1997) Homepage of Project InterData, Retrieved January 10, 2002 from University of Rome III, Italy: http://www.dia.uniroma3.it/interdata/.

Semantic Web Agreement Group (2001) Homepage of the Semantic Web Agreement Group. Available from: http://purl.org/swag/.

Snodgrass, R.T. (Ed.), Ahn, I., Ariav, G., Batory, D., Clifford, J., Dyreson, C.E. Elmasri, R., Grandi, F., Jensen, C.S., Käfer, W., Kline, N., Kulkarni, K., Cliff Leung, T.Y., Lorentzos, N., Ramakrishnan, R. Roddick, J.F., Segev, A., Soo, M.D. & Sripada, S.M. (1995) *The TSQL2 Temporal Query Language*, Boston, MA: Kluwer Academic Publishers.

Soo, M.D. (1991) Bibliography on Temporal Databases, *SIGMOD Record 20*(1), 14–23.

Sun Microsystems (2002a) *Homepage of the Java Plug-in*, Retrieved January 10, 2002 from the World Wide Web: http://java.sun.com/products/plugin/.

Sun Microsystems (2002b) *Java Foundation Classes*, Retrieved January 10, 2002 from the World Wide Web: http://java.sun.com/products/jfc/.

Tansel, A.U., Clifford, J., Gadia, S., Jajodia, S., Segev, A. & Snodgrass R.T. (Eds.) (1993) *Temporal Databases: Theory, Design and Implementation*, Redwood City, CA: Benjamin/Cummings.

TEI Consortium (2001) Homepage of the Text Encoding Initiative, Available from: http://www.tei-c.org/.

Tsotras, V.J., & Kumar, A. (1996) Temporal Database Bibliography Update, *SIGMOD Record 25*(1), 41–51.

W3C Consortium (1997) *Homepage of the Extensible Markup Language.* Retrieved January 10, 2002 from W3C Consortium Web site: http://www.w3.org/XML/.

W3C Consortium (2000) *The XML Schema Resource Page.* Retrieved January 10, 2002 from W3C Consortium Web site: http://www.w3.org/XML/Schema.

W3C Consortium (2001) *Homepage of the Document Object Model.* Retrieved January 10, 2002 from W3C Consortium Web site: http://www.w3.org/DOM/.

W3C Consortium (2002) *The Extensible Stylesheet Language Resource Page.* Retrieved January 10, 2002 from W3C Consortium Web site: http://www.w3.org/Style/XSL/.

Wolf M. & Wicksteed C. (1997) Date and Time Formats, *W3C Consortium Note*, Retrieved January 10, 2002 from W3C Consortium Web site: http://www.w3.org/TR/NOTE-datetime.

Zadeh, L.A. (1988) Fuzzy Logic, *IEEE Computer 21*(4), 83–93.