XML Representation and Management of Temporal Information on the Web for Cultural Heritage Applications

Fabio Grandi, Maria Rita Scalas

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 extensions 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 has a prominent relevance, with potential positive impacts on Electronic Education (e.g. by publishing on the Web digital libraries, virtual museums and archaeological sites with historical indexing and advanced temporal filtering facilities) and Electronic Science (e.g. by allowing history and archaeology researchers to share on the Web large collections of hypertextual documents, to be used as primary sources, provided with semantically and computationally powerful temporal search engines).

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

I. INTRODUCTION

THE eXtensible Markup Language (XML) [39] is becoming the newly emerging standard for data management and exchange over the Internet [2]. 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* to be automatically used by advanced computer tools, like "intelligent" search engines (towards a Semantic Web [51]).

In this context, the CSITE-CNR database research group [47] has been interested in recent years in the introduction of *temporal* aspects into the Web, by adapting and extending concepts and techniques deriving from a more than decennial experience in temporal database research [36], [25], [52], [37], [13]. 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 [12] and "Summaries of Current Work" in [13]. In such research, we first explored the applicability to the Web of the basic notions of *transaction*

Authors are with C.S.I.TE.-C.N.R. and D.E.I.S., University of Bologna, Italy. E-mail: {fgrandi,mrscalas}@deis.unibo.it

time and *valid time*. According to [23], transaction time is the time 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 some fact is *valid in the real world*. With respect to the Web [21], 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 superposition 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" [49], 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 with the introduction of transaction time. We investigated several techniques for the implementation of a Web site with versioned resources, aimed at reducing data duplication, and extensions to the HTTP protocol for the negotiation of transaction-time versions [7]. We also developed a prototype Web site [8] with temporal navigation facilities along transaction time. As to valid time, we proposed in [21] a first timestamping scheme for the explicit encoding of historical information into Web pages, based on non-standard HTML and custom browser extensions. In the same time, we also investigated "killer applications" for the best exploitation of valid-time temporal semantics on the Web and, as a result, we definitly focused on Cultural Heritage applications with our participation in another national research project ("Cultural Heritage" [48], funded by the Italian National Research Council).

Subsequently, in order to put into production the introduction of the valid time dimension into 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 [14], [17]. 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 (Ie5 [50]) and its successive versions. In order to show its potentialities, the infrastructure has been implemented on a demo prototype [15] (also available on-line [16]) showing, as application example, the functionalities of a temporal fine-arts Web museum [32], that is a virtual environment in which it is possible to cut personalized visit 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, that is a collaboration with the Computer and History group at the University of Florence [28], which has been involved for a couple of years in the study of a new edition, in electronic form, of Emanuele Repetti's historical-geographical dictionary of Tuscany (nineteenth century). In particular, our contribution to the "XML/Repetti" project is an 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 a redesign of the overall system architecture with an efficient organization of the temporal search engine and of the large-sized resulting XML document repository, including optimized search algorithms and temporal indexing facilities [19], [18].

We emphasize the fact that the worldwide accessibilty 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 it frequently happens in medieval archaeology that written sources have the same importance as material evidence. It has already been pointed out how the role of Internet in archaeological investigation is continuously increasing, as wide, fast and easy sharing of information on the Web has a substantial impact on the archaeological methodology [4], [22], which could be ever boosted by the deployment of XML-related technologies (as also evidenced in [28], [29], [31]). The representation and automated management of temporal aspects adds computational search power to simple accessibility, with a leap with respect to the availability of paper printed editions.

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

II. "The Valid Web": a Simple Temporal XML/XSL Infrastructure

The addition of valid time to Web documents we proposed in "The Valid Web" approach [17] is based on the extension of the XML markup language [39] with timestamping tags. The proposed infrastructure (also including an XML schema [44] and an XSL stylesheet [40]) is fully 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 the addition 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 to be used 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 [24]. For instance, the following code:

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

   <!-- 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 $[1280{-}1285] \cup [1295{-}1300]$. The time constants can be specified according to the ISO 8601 format [9] (also corresponding 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 fetched 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-sized 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 [38] 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 reengineering of legacy HTML-based Web sites, since the resulting temporal documents only present a very small space and network transfer overhead with respect to the non-temporal case (basically due to the added timestamping tags, in addition to a bit of XML extra formatting, necessary to make legacy documents *well-formed*).

More in detail, the valid-time selection relies on the XSL stylesheet, named `Valid.xsl`, which can be seen in Fig. 1: the first part consists of a simple identity-transformation

```
<?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 ]">
            <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>
```

Fig. 1. The XSL `Valid.xsl` stylesheet.

template, whereas 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 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* 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 finalized to support 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 act on valid time selection to change the historical period of interest. For instance, we can choose the High Renaissance period, by selecting the validity range 1495–1520. Hence, we may start our virtual visit entering some virtual hall or gallery: only the temporally relevant paintings or sculptures would be present; by changing the validity context we could see some works vanish and some different 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 to 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 [1], with dedicated journals (e.g. *Archives and Museum Informatics*) and conferences (e.g. *Museums and the Web*). The "Web Museum" [32], authored and maintained by Nicholas Pioch, 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 [15], [16]. The pages of the site are organized in two frames (see Fig. 3). A small service frame in the bottom part of the window contains all the required controls to deal with the user interactive 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 almost all the browser 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 an independent choice of the 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 analog 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 [42], which contains two graphic sliders: the former to select the year and the latter to select the day of the year (see Fig. 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.
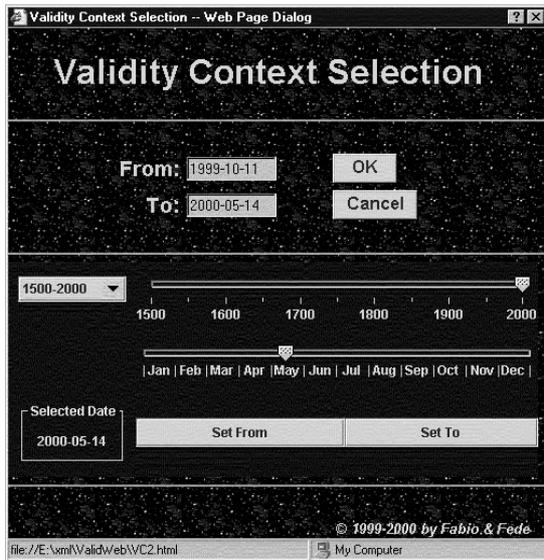
Fig. 2. The Java Applet for the selection of a Validity Context.

The chosen value can then be assigned to the From or To interval boundary by means of the corresponding "Set" *button*. However, also editable input fields for direct typing of a valid time value (in the "YYYY-MM-DD" string format) are 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 [43] supported by the Java Plug-in 1.2.2 [41].

Once the navigation validity context choice is confirmed by the user, the temporal selection over the currently displayed document is automatically re-executed by means of the DOM method calls described above. Furthermore, in order to enable a full-fledged temporal navigation, each time the user changes the displayed document in the usual way (e.g. by following a link), the current validity context is automatically "inherited" by the newly loaded page, if also the new document is a temporal XML one. 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 Fig. 1. The actual stylesheet implements a dynamic callback mechanism by inserting some JavaScript code in the preamble of the processed document. Such a script provokes, on load of the document by the browser, the immediate activation of the temporal selection functionality: as 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 selection condition) and then re-applied to the displayed document.

For example, Fig. 3 shows two snapshots of the navigation of a sample page containing Rembrandt's self-portraits. The full page contains seven pictures, dating from 1629 to 1669. Fig. 3 shows on the left the page when the validity context has been set to [1635–1640] and only two pictures are visible (the third and the fourth one) and on the right 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 Fig. 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 first reference 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 temporal structured or semistructured data, and enabling the utilization of functionalities as developed by temporal database research (e.g. TSQL2-like temporal query languages [35]).

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

Other kinds of Web sites of interest for Cultural Heritage applications which are extremely suitable to benefit form 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.

III. The "XML/Repetti" Project – An Extension of the Temporal XML Encoding Scheme

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 the temporal information contained in the dictionary and the availability of techniques for temporal search support.
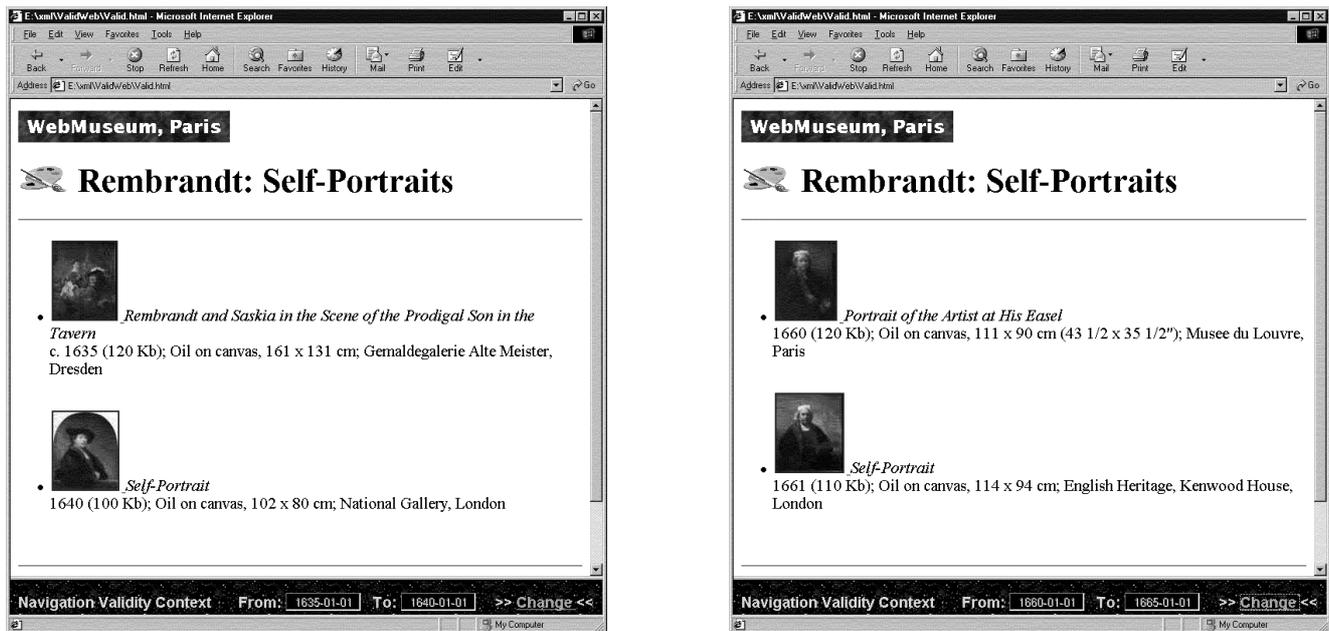
Fig. 3. Temporal navigation of the Web Museum.

However, advanced functionality and efficiency specifications (*Repetti*'s Dictionary is a quite *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 architecture, including an 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 temporal expressions widespread 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 as *Repetti*'s Dictionary, we introduced in [18] a broad classification of the temporal expressions denoting single indeterminate events into four main categories[1]. If we denote by the term Reference Temporal Expression (RTE) the time literal written in text, 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* [11], 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, happened on a certain date located in 1276, rather than to an activity lasting for the whole year. Moreover, we cannot *a priori* ascertain on which day the event actually happened and there is no reason to prefer one date with respect 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 action likely required several years to be completed. However, since it is not known the exact date the works began and ended, the expression denotes

---

[1] 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.

an indeterminate interval, whose boundaries are indeterminate C_1-type dates.

Summing up, every indeterminate temporal expression found in the text and concerning a single event can be reduced to 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 Section. 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 orignal text the suitable markup which enables the exploitation of powerful and fast search engines, which can be used at least to effectively trim down the amount of text which must be examined by experts for a refined analysis. This is very different, for example, from the use of relational databases which has been done 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 lead to a deprecated information loss [20].

A. 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" [11], [35] and the *fuzzy* approach [10]. In particular, Dutta used a *fuzzy set* approach [53] to deal with *generalized temporal events* [10], 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, at a certain extent, non completely specified (i.e. "high" is not a so-called *crisp* predicate). In Dutta's model, a generalized event allows to represent 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 already consider as "high" a body temperature higher that 37.5 Celsius degrees). With the probabilistic approach instead, we prefer to represent an *indeterminate validity* connected to the occurrence of an event, which remains anyway 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, in a greater or lesser way depending on its membership *degree*. The two approaches are deeply different as far as the representation of incomplete temporal information is concerned and, in our case, we prefer the latter approach: an historical event (say the death of a King) must be occurred on a precise date, even if no precise and unique determination can be found in the sources.

Cat.	Prototype expression	Shape	Distrib.
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

TABLE I
DISTRIBUTIONS ASSOCIATED WITH INDETERMINATE DATES.

The probabilistic model has been introduced in the design of the temporal query language TSQL2 [35], which is a consensual proposal for temporal extensions of the standard query language SQL, and further developed by Dyreson and Snodgrass in [11]. In this approach, an indeterminate event t is represented through its *probability distribution* P, different from zero only in an interval of possible occurrence, whose boundaries (t^- and t^+) are said *lower support* and *upper support*:

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

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 ($t_1 \equiv t_2$) iff 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 which is used to define all the other temporal comparison operators [35]. In the indeterminate semantics, the "*Before()*" primitives includes an additional parameter to specify an ordering *plausibility*, whose value can range from 0 to 100 (high plausibility means high precedence probability between the compared instants). Its complete definition becomes thus:

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

where the precedence probability is evaluated as:

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

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 [18]) a small set of predefined distributions which can be assigned to the $C_1 - C_4$ indetermination categories as summarized in Table I. 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 [18] how this choice, which is fairly correct from a semantic viewpoint, allows us to exploit extremely efficient comparion algorithms (with an optimized evaluation of formula (1)) without any storage space overhead, which, on the contrary, would have made unfeasible the direct application of the basic approach in [11]. For the distributions but 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 the *principal interval* and P is one of the availiable 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 exactly 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 with at least one contained 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. To 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 will be provided in Fig. 4) has several attributes, some of which are specific for the indeterminacy support:
• GRANULARITY, which allows to specify the granularity used to express the date value as "DAY" (*default*), "MONTH", "YEAR" or "CENTURY";
• VALUE, which allows to specify the date expression (obviously in a way consistent with the assigned granularity);
• INDETETERMINATE, with values "YES" or "NO" (*default*), which allows to specify whether the date is expressed in indeterminate format or not; in case the attribute value is YES, do the further attributes have meaning:
− DISTRIBUTION, whose value can be one of the supported probability distributions in Tab.I (with their variants);
− DURATION, which expresses (with *default* "1"), as granularity multiples, the amplitude of the *principal interval*

(also corresponding to the width of all the base intervals on which the probability density is constant);
• CALENDAR, which allows to reference a specific calendar, as explained later.

The *principal interval* is expressed in *implicit* way, by means of an interval having as lower boundary the first day of the VALUE temporal expression (e.g. 1456/1/1 for VALUE="1456") and an amplitude which can be evaluated as the specified granularity GRANULARITY, converted to days, multiplied by the DURATION attribute value. Such an interval coincide with the whole interval between the two supports 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 loss of generality, this choice permits to easily exploit the granularity to directly express base intervals with unit duration. 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* support encoding makes it a bit more "transparent" and user-friendly, so that the user (i.e. the history researcher) can best concentrate on the choice of an intuitive "form factor" among a few available alternatives rather that on mathematical details of distributions like the support computation or the variance. Notice also how a uniform encoding choice, in which value and granularity *exactly* correspond to the textual expression used in the document (which we called **rigorous encoding rule** [18], [19]), represents itself *metainformation* (on the original form of the text contents) to be used for advanced searches.

The possibility to express dates at different granularity levels makes it difficult to use them together, as required to analysis and temporal reasoning tools (e.g. the comparison between dates expressed with different granularities becomes a problem). However, this possibility mirrors the richness of forms in which dates are actually recorded in written historical sources; the rigorous encoding rule we introduced provides a tight correspondence and, thus, allows to maintain such a richness also in the XML format for the benefit of the researcher. In TSQL2, it is possible to define and utilize different calendars, each of which can be provided with specific (i.e. calendar-dependent) granularity scales. The TSQL2 calendar and calendric system support is the answer to a demand expressed for the first time in [3] and satisfied for the first time in the MultiCal system implementation [27]. In general, TSQL2 calendars

(predefined or specified by the user or database administrator) have an associated granularity lattice defining the time unit hierarchy and a set of conversion rules between different units to from one granularity level to another.

In our case, the management of different calendars does not present particular difficulties, as all the calendars in use have the day as base granularity and seemingly use the same lattice (the only granularities of interest are always: day, month, year and century). Hence, in the context of the DATE type (i.e. in the `<AT>`, `<FROM>` e `<TO>`) we can simply introduce an attribute CALENDAR to declare the name of the reference calendar for the date specification (e.g. GREGORIAN, which is the default value, JULIAN, ROMAN). The additional use of special calendar styles[2] can be encoded via variants of the calendar (e.g.. GREGORIAN_FLORENCE_STYLE for the "*Florentine*" style of the Gregorian calendar). In order to be processed, all the dates found in a temporal XML document, possibly specified via different calendars, can be automatically converted according to a common reference calendar (e.g. the Gregorian) before successive operations like comparisons. The specification of the calendar as markup attribute permits a correct conversion, and also represent *metainformation* concerning the calendar used in the original document, which can be used for specific investigations too. For instance, the expression "**(anno di R. 469)**" found in *Repetti*'s Dictionary is translated into a determinate time interval whose boundaries are respectively the first of January and the last of December of the 469 according to the Roman calendar (whose year numbering starts from the founding of the city). The complete specification is the following:

```
<INTERVAL>
    <FROM VALUE="0469-01-01" CALENDAR="ROMAN" />
    <TO VALUE="0469-12-31" CALENDAR="ROMAN" />
      (anno di R. 469)
</INTERVAL>
```

or better (according to the rigorous encoding rule):

```
<EVENT>
    <AT VALUE="0469" GRANULARITY="YEAR" CALENDAR="ROMAN" />
      (anno di R. 469)
</EVENT>
```

For reference, we add in Fig. 4 a (simplified) Document Type Definition (DTD [34]) for the XML encoding of temporal expressions in *Repetti*'s Dictionary. Our contribution will eventually be integrated in the global work which is carried on at the University of Florence on *Repetti*'s Dictionary electronic edition, where the activities of other researchers [29] include the development of a complete DTD for the XML markup of *Repetti*'s Dictionary, also adapting previous work on document coding with XML/SGML, based on the TEI [46] or TEI Lite [45] DTDs, which already experienced a positive acceptance by the community

---

[2]The *style* of a calendar concerns the day chosen to start a new year [5], which is January 1st in the *modern* style. Several different styles were in use in Tuscany during the Middle Age (e.g. in Florence's style the year began on the 25th of March, while in Pisa's style it began on the 25th of March of the year before) and, thus, dates found in *Repetti*'s Dictionary may refer to different modern-style dates depending on the local context they were handed on.

```
<!DOCTYPE DICTIONARY[
<!ENTITY % DATE " CDATA #REQUIRED " >
<!ENTITY % DATE_ATTR
  " GRANULARITY (DAY|MONTH|YEAR|CENTURY) 'DAY'
    VALUE CDATA #REQUIRED
    INDETERMINATE (YES|NO) 'NO'
    DISTRIBUTION (DURING|EARLY|VERY_EARLY|LATE|
      VERY_LATE|AROUND|STRICTLY_AROUND|WIDELY_AROUND) #IMPLIED
    DURATION CDATA '1'
    CALENDAR (GREGORIAN|ROMAN|JULIAN) 'GREGORIAN' " >
<!ELEMENT DICTIONARY (ITEM)* >
<!ELEMENT ITEM (#PCDATA|EVENT|INTERVAL)* >
<!ELEMENT EVENT (AT,#PCDATA) >
<!ELEMENT INTERVAL (FROM,TO,#PCDATA) >
<!ELEMENT AT %DATE; >
<!ATTLIST AT %DATE_ATTR; >
<!ELEMENT FROM %DATE; >
<!ATTLIST FROM %DATE_ATTR; >
<!ELEMENT TO %DATE; >
<!ATTLIST TO %DATE_ATTR; >
]>
```

Fig. 4. The temporal DTD for the "XML/Repetti" project.

interested in digital libraries and Humanities. Therefore, the DTD in Fig. 4 is going to be merged into the final DTD which will globally result from "XML/Repetti" project.

B. Implementation

An important part of the "XML/Repetti" project is also the design and implementation of a development tool for the computer-aided encoding of temporal XML documents and of a search engine for their fruition on the Web [19].

Starting form an electronic version of *Repetti*'s Dictionary (in HTML format), the development tool tries to find out, using *regular expressions*, as many contained temporal determinations as possible. After this pre-processing phase, for each individuated temporal expression, the tool assists the user in the encoding process to produce the corresponding markup, by proposing solutions strictly dependent on the temporal expression type itself. Such a semi-automatic process leaves anyway to the user the freedom of choosing whether to accept the encoding prompted by the tool, or to change it on the basis of his/her own interpretation of the text. Moreover, the user will always able to select pieces of text to be encoded as he/she thinks they contain historical information, although the tool did not succeed in pointing them out. The system will be made available on the Internet and usable through a standard Web browser like Netscape or Explorer. Fig. 5 shows a prototype version of the tool and, in particular, a possible organization of the user interface.

Its user interface is divided in two parts: a main area (upper window) where the document under processing can be seen with the temporal expressions evidenced by the pre-processing (e.g. "**nel 1404**" in the figure) and a lower window which allows the user to specify structure and attributes of the XML tags to be inserted via friendly input forms. If the user, when browsing the document in the
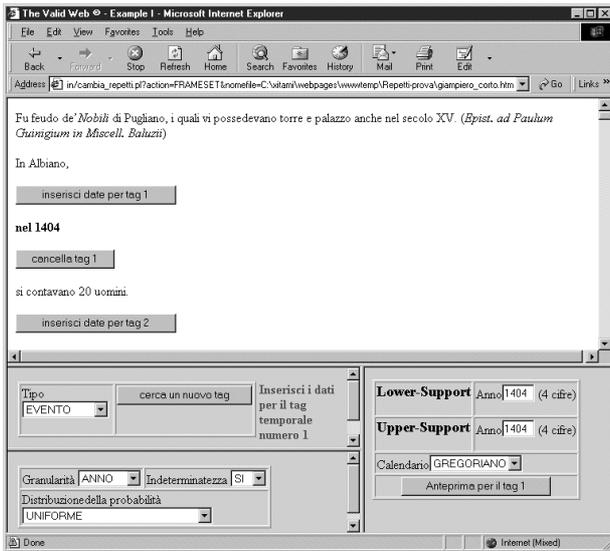
Fig. 5. The graphic interface of the development tool.

main window, finds a temporal expression which has not been automatically spotted by the tool, he/she can proceed with a manual selection of the corresponding text and its markup. The successive design phases of the tool functionalities and its graphic interface will be effected in strict collaboration with the history researches who are working to the "XML/Repetti" project, as they represent the final users of the tool.

As far as the search engine is concerned, its design has been guided by three main requirements:

• **efficiency** the engine must have an optimized implementation to cope with the *Repetti*'s Dictionary dimension and the computational complexity of the retrieval procedures; this will impact on the overall system architecture, on the dictionary storage organization and on the temporal query processing algorithms;

• **powerful semantics** the engine must exploit all the temporal semantic richness supported by the adopted XML-markup scheme, including indeterminacy, multiple calendars and granularities, full TSQL2-like temporal query expressivity; furthermore, also the use of "second-order" metainformation (e.g. based on the rigorous encoding rule) should be supported;

• **availability** the engine must be accessible on the Internet with a standard Web browser, it must have easy-to-understand functionalities and a friendly interface for user interaction.

The design solution we proposed is based on the use of a temporal index structure and optimized management techniques. In particular, the solution is based on the adoption of a MAP21 temporal index [30], which relies on standard "off-the-shelf" technology. In the overall system architecture, the search engine runs on the server side: in our first prototype implementation, it is implemented by means of

PERL scripts, which are activated by the client through a standard CGI mechanism. The retrieval work is effected by compiled C^{++} programs that access the only MAP21 index leaves that may contain qualifying dates to fetch the exact expressions of the (indeterminate) dates which are passed to the function doing probabilistic comparison. For all the dates passing the test at the assigned plausibility level, the pointers found in the index leaves are used to access the dictionary items on disk. The search results to be returned to the Web client are assembled by the PERL script, according to some user's option, either as a single XML/HTML file explicitly containing all the retrieved items, or as a "digest" containing a list of links to the selected items. It should also be noted that the temporal indexing via the MAP21 is a secondary indexing, which does not prevent other indexes to be built (e.g. on names, places, etc.). The client-side functionalities only require the execution of controls for the management of the query formulation, including the full specification of the temporal period of interest, possibly including all indeterminacy, granularity and calendar parameters. In particular, the user-interaction will be based on a friendly interface, so that it could be easily used by non computer-experts, which can be similar to the one designed for the Valid Web approach (Fig. 2).

IV. CONCLUSIONS

In this paper we reviewed the research work of the CSITE-CNR database group in the field of temporal extensions of the World Wide Web. In particular, we described two XML-based technological infrastructures, results of "The Valid Web" and "XML/Repetti" projects, which have been developed for the support of valid time for Cultural Heritage applications dealing with historical information. The implemented demo prototypes include a temporal virtual Web museum and a repository of ancient texts in digital form equipped with a smart temporal search engine. As also shown by those experiences, the applications which could most benefit from the representation and management of historical information and temporal semantics are, in a broad sense, *digital libraries*, with a likely potential positive impact on Electronic Education and Electronic Science (e.g. in the field of Archaeology).
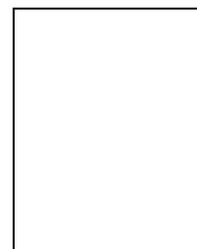
Fabio Grandi is currently an Associate Professor in the Faculty of Engineering of the University of Bologna, Italy. Since 1989 he has worked at the CSITE (formerly CIOC) center of the Italian National Research Council (CNR) in Bologna in the field of neural networks and temporal databases. For this work he was initially supported by a fellowship from the CNR. In 1993 and 1994 he was an Adjunct Professor at the Universities of Ferrara, Italy, and Bologna. He joined his current department (Dept. of Electronics, Computer Science and Systems) as a Research Associate in 1994. His scientific interests include temporal databases, storage structures, access cost models, world wide web extensions. He received a Laurea degree in Electronics Engineering and a PhD in Electronics Engineering and Computer Science from the University of Bologna.

Maria Rita Scalas is currently an Associate Professor in the Faculty of Engineering of the University of Bologna, Italy. From 1975 to 1979 she worked at the Universities of Pisa and Bologna supported by a fellowship from the Italian Ministry of Education. In 1980 she became a Research Assistant in Computer Science at the University of Bologna and a consultant at the CIOC-CNR center of the National Research Council in Bologna. In 1986 she was a visiting scientist at the IBM Scientific Center in Heidelberg, Germany, where she took part in the AIM-P project. In 1987 she became an Associate Professor at the University of Trieste, Italy. She holds a Laurea degree in Physics from the University of Bologna. Her research interests are in the area of database management systems, temporal databases, access structures, optimizers and database design.

REFERENCES

[1] Archives & Museums Informatics, http://www.archimuse.com.

[2] S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[3] T.L. Anderson, "Modeling Time at Conceptual Level", Proc. of *Intl' Conf. on Databases*, Jerusalem, Israel, 1982.

[4] I. Bogdanovic, O. Vicente, J.A. Barcelo, "A Theory of Archaeological Knowledge Building by using Internet: the DIASPORA Project", Proc. of *Intl' Conf. on Computer Applications in Archaeology (CAA'99)*, Dublin, Ireland, 1999.

[5] A. Cappelli, "Chronology, Chronography and Perpetual Calendar", Hoepli, Milan, Italy, 1998 (*in Italian*).

[6] J. Clifford, A. Tuzhilin (eds.), *Recent Advances in Temporal Databases* (Proc. *Intl. Workshop on Temporal Databases*), Springer-Verlag, Berlin, 1995.

[7] C. Cristofori, F. Grandi, F. Mandreoli, M.R.Scalas, "Management of Temporal Versions of World Wide Web Resources along Transaction Time", Tech. Rep. INTERDATA T6-R03, July 1998 (*in Italian*), http://www.difa.unibas.it/users/gmecca/interdata/.

[8] C. Cristofori, F. Grandi, F. Mandreoli, M.R.Scalas, "Web Server with Transaction Time Versioning", Prototype Descr.INTERDATA T6-S09, February 1999 (*in Italian*), http://www.difa.unibas.it/users/gmecca/interdata/

[9] Date and Time Formats, W3C Consortium Note, http://www.w3.org/TR/NOTE-datetime.

[10] S. Dutta, "Generalized Events in Temporal Databases", Proc. of *Intl' Conf. on Data Engineering (ICDE)*, Los Angeles, CA, February 1989.

[11] C.E. Dyreson, R.T. Snodgrass, "Supporting Valid-time Indeterminacy", *ACM Trans. on Database Systems*, Vol. 23, No. 1, 1998.

[12] O. Etzion, S. Jajodia and S. Sripada (eds.), *Temporal Databases*, Dagstuhl-Seminar-Report N. 183 (9726), June 1997.

[13] O. Etzion, S. Jajodia and S. Sripada (eds.), *Temporal Databases - Research and practice*, LNCS N. 1399, Springer-Verlag, 1998.

[14] F. Grandi, F. Mandreoli, "The Valid Web: it's Time to Go...", TIMECENTER TR-46, 1999, http://www.cs.auc.dk/research/DP/tdb/TimeCenter/.

[15] F. Grandi, F. Mandreoli, "The Valid Web ©", Proc. of *Software Demonstrations Track at the EDBT'2000 Intl. Conf.*, Konstanz, Germany, March 2000.

[16] F. Grandi, F. Mandreoli, "The Valid Web ©" Home Page, http://degas.deis.unibo.it/ValidWeb/.

[17] F. Grandi, F. Mandreoli, "The Valid Web: an XML/XSL Infrastructure for Temporal Management of Web Documents", Proc. *Intl' Conf. on Advances in Information Systems (ADVIS'2000)*, Izmir, Turkey, 2000, LNCS N. 1909, Springer-Verlag, 2000.

[18] F. Grandi, F. Mandreoli, "Effective Representation and Efficient Management of Indeterminate Dates", *Proc. Intl' Symposium on Temporal Representation and Reasoning (TIME-01)*, Cividale del Friuli, Italy, June 2001, IEEE Computer Society Press, Los Alamitos, 2001.

[19] F. Grandi, F. Mandreoli, "The "XML/Repetti" Project: XML Encoding and Manipulation of Temporal Information in Historical Text Sources", *Proc. Intl' Cultural Heritage Informatics Meeting (ICHIM'01)*, Milan, Italy, September 2001, Archive and Museum Informatics and Politecnico di Milano, 2001.

[20] F. Grandi, F. Niccolucci, "XML Technologies for the Representation and Management of Spatiotemporal Information in Archaeology", Proc. of *Intl' CODATA Conference - Theme I-6*, Baveno, Italy, 2000, http://www.codata.org/codata/abstracts/TOC.htm.

[21] F. Grandi, M. R. Scalas, "Extending Temporal Database Concepts to the World Wide Web", Proc. of *Natl' Conf. on Advanced Database Systems (SEBD'98)*, Ancona, Italy, 1998.

[22] S. Hermon, F. Niccolucci, "The Impact of Web-shared Knowledge on Archaeological Scientific Research", Proc. of *Intl' Conf. on Current Research on Information Systems (CRIS2000)*, Helsinki, Finland, 2000.

[23] C.S. Jensen, J. Clifford, R. Elmasri, S.K. Gadia, P. Hayes, S. Jajodia (eds.) *et al.*, "A Consensus Glossary of Temporal Database Concepts - February 1998 Version," in [13].

[24] C.S. Jensen, M.D. Soo, R.T. Snodgrass, "Unifying Temporal Data Models via a Conceptual Model," *Information Systems* 19(7), 1994.

[25] N. Kline, "An Update of the Temporal Database Bibliography," *SIGMOD Record* 22(4), 1993.

[26] A. Montanari, E. Maim, E Ciapessoni, E. Ratto, "Dealing with Time Granularity in the Event Calculus", Proc. of *Intl' Conf. on Fifth Generation Computer Systems*, Tokyo, Japan, 1992.

[27] *The* MULTICAL *System - release 1.0*, Dept. of Computer Science, Univ. of Arizona, Tucson, 1993.

[28] F. Niccolucci, A. Zorzi, M. Baldi, F. Carminati, P. Salvatori, T. Zoppi, "Historical Text Encoding: an Experiment with XML on Repetti's Historical Dictionary", Proc. of *Conf. of the Association for History and Computing - UK Branch (AHC-UK'99)*, London, UK, 1999.

[29] A. Benvenuti, F. Niccolucci, S. Baragli, C. Carpini, "Advances in XML Treatment of Historical Documents", in *La Historia en una Nueva Frontera*, AHC, Toledo, Spain, 2000.

[30] M. Nascimento, M.H. Durham, "Indexing Valid Time Databases via B^{+}-trees", *IEEE Trans. on Knowldege and Data Engineering*, 11(8), 1999.

[31] F. Niccolucci (ed.), *From the Sources to the Network – Proc. of Workshop on XML Applications to Historical Source Encoding*, Florence, Italy, 2000.

[32] N. Pioch, "The Web Museum" , http://www.cnam.fr/wm/.

[33] N. Pissinou, R.T. Snodgrass, R. Elmasri, I.S. Mumick, M.T. Öszu, B. Pernici, A. Segev, B. Theodoulidis, U. Dayal, "Towards an Infrastructure for Temporal Databases: Report of an Invitational ARPA/NSF Workshop", *ACM SIGMOD Record*, Vol. 23, N. 1, 1994.

[34] Prolog and Document Type Declaration, in *Extensible Markup Language (XML) 1.0*, W3C Recommendation, http://www.w3.org/TR/REC-xml#sec-prolog-dtd.

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

[36] M.D. Soo, "Bibliography on Temporal Databases," *SIGMOD Record* 20(1), 1991.

[37] A. Tansel, J. Clifford, V. Gadia, S. Jajodia, A. Segev, R.T. Snodgrass (eds.), *Temporal Databases: Theory, Design and Implementation*, Benjamin/Cummings, Redwood City, CA, 1993.

[38] The Document Object Model (DOM) Home Page, W3C Consortium, http://www.w3.org/DOM/.

[39] The Extensible Markup Language (XML) Home Page, W3C Consortium, http://www.w3.org/XML/.

[40] The Extensible Stylesheet Language (XSL) Resource Page, W3C Consortium, http://www.w3.org/Style/XSL/.

[41] The Java Plug-in Home Page, Sun Microsystems, http://java.sun.com/products/plugin/

[42] Java Foundation Classes (JFC), Sun Microsystems, http://java.sun.com/products/jfc/.

[43] LiveConnect, in *JavaScript Guide*, Ch. 5, Netscape Communications, http://developer.netscape.com/docs/manuals/communicator/jsguide4/livecon.htm.

[44] XML-Data, W3C Note, http://www.w3.org/TR/1998/NOTE-XML-data.

[45] Text Encoding for Interchange, TEI Lite Home Page, http://www.uic.edu/orgs/tei/lite/.

[46] Text Encoding Initiative, TEI Consortium Home Page, http://www.tei-c.org.

[47] The CSITE-CNR DataBase and KnowledgeBase Group Home Page, University of Bologna, Italy, `http://www-db.deis.unibo.it/`.

[48] The "Cultural Heritage" Project Home Page, CNR, Italy, `http://www.culturalheritage.cnr.it/`.

[49] The "INTERDATA" Project Home Page, University of Rome III, Italy, `http://www.dia.uniroma3.it/interdata/`.

[50] The Microsoft Internet Explorer Home Page, Microsoft, `http://microsoft.com/windows/Ie/`.

[51] The Semantic Web Agreement Group Home Page, `http://swag.semanticweb.org`.

[52] V.J. Tsostras, A. Kumar, "Temporal Database Bibliography Update," *SIGMOD Record* 25(1), 1996.

[53] L.A. Zadeh, "Fuzzy Logic", *IEEE Computer*, Vol. 21, N. 4, 1988.