The "XML/Repetti" Project: Encoding and Manipulation of Temporal Information in Historical Text Sources

Fabio Grandi [(*)] *and Federica Mandreoli* [(#)]

[(*)] C.S.I.TE.-C.N.R. and D.E.I.S., University of Bologna, Italy
E-mail: fgrandi@deis.unibo.it

[(#)] D.S.I., University of Modena and Reggio Emilia, Italy
E-mail: fmandreoli@dsi.unimo.it

ABSTRACT

The paper deals with the deployment of XML-related technologies in Cultural Heritage applications concerning the encoding of temporal semantics in the digital version of historical documents. Since written sources have often the same importance as material evidence in medieval archaeology, our approach can be applied to the development of tools for the support of archaeological research. In previous work, we developed an XML/XSL infrastructure called "The Valid Web" for the definition and management of historical information within Web documents. In this paper we describe the application and extension of such an approach to the realization of the electronic version of Repetti's historical-geographical dictionary of Tuscany. The extension concerns the uniform management of temporal indeterminacy, the use of multiple calendars and granularities and the proposed solutions have been inspired by similar research done for temporal query languages. From the user viewpoint, the proposed XML extensions allow the addition of historical metainformation to the encoded text sources and their "intelligent" temporal navigation via standard Web browsers. The project also involves the definition of optimized search algorithms, storage and temporal indexing of XML-encoded Repetti's Dictionary items, implementation of a prototype. As a byproduct, also a tool for computer-aided temporal XML-encoding of text sources will be developed to be used by Cultural Heritage operators (e.g. archaeology researchers).

KEYWORDS: XML, semantic Web, temporal data management, digital libraries

INTRODUCTION

In the information processing field, XML [21] is becoming a new standard for data management and exchange over the Internet [1]. 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) feature is the capability of easily encoding semantic metainformation in XML documents, to be automatically used by advanced computer tools, like "intelligent" search engines (towards a Semantic Web [23]). In this context, our research interests focused in recent years on 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 [12,14]. To this end, we developed an XML/XSL infrastructure, named "The Valid Web", for the management of temporal documents and data on the Web [6,9]. The proposed techniques enable the explicit encoding of distinguished temporal/historical information within XML (or legacy HTML) documents, whose contents can then be selectively accessed according to their temporal validity with any XML-compliant Web browser (like MS Ie5 [22]). In order to show its potentialities, the infrastructure has been implemented on a demo prototype [7] (also available on-line [8]) showing, as application example, the functionalities of a temporal fine-arts Web museum [19], that is a virtual environment in which it is possible to cut personalized visit routes for a specific epoch of interest. In this paper, we introduce an interesting application and extension of such an approach for the management of historical text sources in digital form. The general framework of this research is a collaboration with the Computer and History group at the University of Florence [18], which has been involved for a couple of years in the study of a new edition, in electronic form, of Repetti's Dictionary. The "*Historical-geographical dictionary of Tuscany*" by Emanuele Repetti is an encyclopedic collection of information about Tuscany published in the XIX century concerning notable places (from large towns to small villages, providing historical, archaeological and artistic information), physical land attributes (viz. mountains, rivers, lakes, wetlands, etc.) and special items (like statistical tables). It has been anastatically reprinted many times and its digital edition under study should incorporate hypertextual links and additional multimedia data and should be made available on the Web to be worldwide accessible through the Internet. In our case, this aspect has a noteworthy relevance from a Cultural Heritage and also scientific point of view: 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 [3,13], which could be ever boosted by the deployment of XML-related technologies (as also evidenced in [2,11,17]).

THE "XML/REPETTI" PROJECT

As a starting point, the straightforward application of "The Valid Web" approach to the "XML/Repetti" project allows the uniform classification and encoding of the temporal information contained in the dictionary and the availability

of techniques for temporal search support. However, advanced functionality and efficiency specifications (Repetti's Dictionary is a quite large collection of text) require an improvement of the basic approach. Such an improvement actually provides for three goals:

1. The extension of the XML/XSL infrastructure to deal with the specificity and 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 (with temporal indexing facilities);

3. The design and implementation of a user-friendly tool for computer-aided temporal encoding and document mark-up, which could save history or archaeology researchers from "manual" editing as much as possible.

In particular, the infrastructure extension requires 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 indeed) involving **indeterminacy** (as in: "around 1456"), **multiple calendars** (e.g. use of the Julian calendar) and **different granularities** (e.g. months versus years). Special attention has been devoted to the indeterminacy problem, which has interesting theoretical implications and requires the most consistent infrastructure extensions.

Advanced XML Encoding of Temporal Information

An analysis of Repetti's Dictionary contents evidences the large use of imprecise temporal expressions that could only be represented with the original "The Valid Web" infrastructure with a unacceptable oversimplification. For instance, the following examples can all be found in the "A letter" section of Repetti's Dictionary (the expression are evidenced in boldface; an English translation is added in square brackets):

1. Un terzo giro fu tracciato con ampio pomerio, profondi fossi e più regolari vie, **circa il 1276** [about in 1276], per ordine del vesc. Guglielmino degli Ubertini,...

2. Quindi, **intorno al 1530** [around 1530], il francese Marcilla dipinse a vetri colorati le belle finestre...

3. Fu Comunità sino **alla fine del secolo XVIII** [at the end of the eighteenth century], compresavi la popolazione della piccola borgata di Petrognola.

4. ...Cosimo III fondò un convento ai frati Minori della riforma di Spagna, mantenuti a spese del R. erario, soppressi **sulla fine del secolo XVIII** [near the end of the eighteenth century].

5. Con simile vocabolo fu **designata nei secoli intorno al mille** [in the centuries around the first century A.D.; the sentence talks about a name given to a church] la pieve di S. Cresci a Maccioli

6. La grandiosa e ricca cappella della Madonna fu fondata **sulla fine del sec. XVIII** [near the end of the eighteenth century] nella parete sett. del Tempio...

7. ...con una porzione di quella Maremma stata già donata **verso l'anno 830** [towards the year 830] dall'imp. Lodovico Pio.

8. ...nell'ultima ricostruzione delle mura aretine ordinata da Cosimo I, che di nuovi baluardi e cortine **fra il 1549 e il 1568** [between 1549 and 1568; the sentence talks about building of city walls] le fortificò.

Similar expressions correspond to the denotation, affected by some imprecision and/or indeterminacy, of the occurrence time of a fact that can be instantaneous (in this case we properly call it an event), as in Ex. 5, or a period with non-null duration, which can be described by a temporal interval, as in Ex. 8. Obviously, the concept of instantaneousness of a fact depends (in addition to the interpretation of the text, which could also be ambiguous) on the choice of the reference temporal granularity (scale unit on the time axis), which in the case of Repetti's Dictionary ranges from the day to the century. For example, the adoption of the day as reference granularity implies that any historical fact whose validity can be located within a given date has to be interpreted as an event (i.e. ideally *instantaneous*, being the reference unit indivisible). On the contrary, the validity of historical facts whose duration can be expressed via granularities coarser than the day can also be expressed via an interval of days in which we surely know the event happened, though we do not know exactly when. The solution we adopt for the representation of indeterminate dates in the "XML/Repetti" project is based on the *probabilistic approach* introduced for the temporal query language TSQL2 [20] and further developed in [5]. In this approach, every indeterminate date t is represented by a tuple (t_L , t_U , P), where t_L and t_U (resp. lower and upper support) are the boundaries of an interval in which t may have occurred and P is the probability distribution for the occurrence of t in such an interval. For query evaluation, two dates can be compared by evaluating their precedence probability:

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

where $P_k(x)$ is the occurrence probability of t_k at time x.

The encoding scheme. In particular, we developed an enhanced encoding scheme for indeterminate dates, based on piecewise-constant probability distributions, which is described in detail [10]. It is based on the classification of textual expressions into four main categories associated to predefined probability distributions according to the correspondence shown in the Table which follows:

Cat.	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`

The (very frequent) C_1 case corresponds to a so-called *granularity mismatch* [5], where a determinate expression with higher granularity is used to denote an indeterminate expression with lower granularity. The probability

distributions used are *piecewise-constant* functions over a small number of equal *base intervals* between the lower and upper support. It is shown in [10] how this choice, which is fairly correct from a semantic point of view, yields extremely efficient comparison algorithms between dates without any storage space overhead, which, on the contrary, would make unfeasible the application to Repetti's Dictionary of the basic approach in [5]. For the distributions but DURING, we also consider variants implying a different number of base intervals (i.e. VERY_EARLY, VERY_LATE, STRICTLY_AROUND and WIDELY_AROUND). In particular, indeterminate dates are represented in our encoding scheme by means of a (*I, P*) pair where *I* is the *principal interval* and *P* is one of the available distributions. The principal interval is the base interval where *P* is maximum and exactly corresponds to the temporal expression originally written in the text (and, thus, is a more intuitive parameter to identify than the lower and upper supports).

XML encoding. In order to implement temporal search facilities for Repetti's Dictionary, the dictionary items are used as target units for the search engine. To this end, the textual contents of every XML-encoded item will be enclosed in a tag pair <ITEM> … </ITEM>, which can be selected on the basis of the encoded temporal references they contain. For example, if we are interested in a particular time period, we have to look for every item containing at least one temporal expression which overlaps such period. The temporal expressions of interest include single dates representing the validity of an historical event and time periods, 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 are able to define the <EVENT> and <INTERVAL> tags. The <EVENT> tag contains the <AT> XML element with DATE type, while the <INTERVAL> tag will contains the <FROM> and <TO> elements, both with DATE type. In this way, events can be represented via structures like:

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

whereas the interval markup will be like:

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

The base type DATE 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 (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 seven supported probability distributions (with their variants);

- DURATION, which expresses (with default "1"), as granularity multiples, the amplitude of the principal interval;

- CALENDAR, which allows to reference a specific calendar, as explained in the next section.

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, value converted to days, multiplied by the DURATION value. In this way, a sample expression *around year 1622* can be simply encoded as:

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

The choice of an implicit 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 call **rigorous encoding rule**), represents itself metainformation, on the original form of the text contents, to be used for advanced searches.

Use of different granularities and calendars. The possibility to express dates at different granularity levels [15] mirrors the richness of forms in which dates are actually recorded in written historical sources, although it could make more difficult their management by temporal reasoning tools. The rigorous encoding rule we introduced allows to maintain such a richness also in the XML format for the benefit of the researcher. For instance, the temporal expressions which follow (evidenced in boldface) can all be found in the A-letter items:

[day] La memoria più remota di Anghiari sino a noi pervenuta spetta a un istrumento del **13 nov. 1083**, in forza del quale Bernardo, soprachiamato Sidonia...

[month] ...sino da quando il C. Guido figlio del C. Alberto, stando nella canonica del pievano di Stia, **nell'aprile del 1054**, donò alla vicina chiesa di Sprugnano...

[year] Il cast. di S. Angelo fu diroccato **nel 1282** dai Senesi, che lo diedero in feudo ai Salimbeni.

[century] Fu feudo de' Nobili di Pugliano, i quali vi possedevano torre e palazzo anche **nel secolo XV**.

It should be noted how there often exist nuances of

meaning, for which the context is determinant, between the presence of indeterminacy and the real use of different granularities. As a matter of fact, in the second example which concerns a deed of gift, it is quite likely that it happened on a certain date located in April 1054, rather than it actually went on as an activity lasting for the whole month. In this case, we are in the presence of *granularity mismatch*, which corresponds to a C_1-type indeterminacy and requires the use of an indeterminate date with distribution `DURING`. The fourth example is indeed different, as it concerns the possession of estates by a noble family, that we could guess it was kept for the whole fifteenth century. In this case, the possession validity is more correctly expressed via a determinate interval. In any case, in order to introduce the most appropriate expression, it necessary a very careful choice (a support tool could possibly help by automatically prompting for the seemingly more likely alternatives). In our approach, 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 in practice always: day, month, year and century). Hence, in the context of the `DATE` type, we can simply use the 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 [4] 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, are automatically converted to a common reference calendar (e.g. the Gregorian) before successive operations like comparisons.

DESIGN OF A DEVELOPMENT TOOL

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. Starting form an electronic version of Repetti's Dictionary (in HTML format), such a tool tries to find out, using regular expressions, as many temporal determinations as possible. After this pre-processing phase, the tool assists the user in producing the XML markup, by proposing solutions strictly dependent on the temporal expressions found. Such a semiautomatic 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 be able to select pieces of text to encode 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. A prototype version of the tool is under advanced development. 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 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 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.

IMPLEMENTATION OF A SEARCH ENGINE

The technological infrastructure developed for the "XML/Repetti" project must include an advanced temporal search engine. The design of the search engine 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 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 expressiveness; 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. Special care was devoted to the fulfillment of the first requirement, as it plays a primary role for the success of the "XML/Repetti" project.

The design solution we propose is based on the use of a temporal index structure and optimized search algorithms on a server-side architecture. In particular, the solution is based on the adoption of a MAP21 temporal index [16], 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 (mainly based on a Java2 applet and Javascript functions) [6,9].

CONCLUSIONS

In this paper we presented the evolution of "The Valid Web" approach for the representation and management of temporal information in XML documents and its application to the "XML/Repetti" project. Such project involves the computer-aided encoding of a large collection of historical text sources, for which a support tool is under development, and the efficient implementation of an "intelligent" temporal search engine available on the Web.

REFERENCES

1. S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann Publishers, San Francisco, CA, 1999.
2. 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.
3. I. Bogdanovic, O. Vicente, J.A. Barcelo, "A Theory of Archaeological Knowledge Building by using Internet: the DIASPORA Project", Proc. of Intl' CAA'99 Conf., Dublin, Ireland, 1999.
4. A. Cappelli, *Chronology, Chronography and Perpetual Calendar*, Hoepli, Milan, 1998 (*in Italian*).
5. C.E. Dyreson, R.T. Snodgrass, "Supporting Valid-time Indeterminacy", *ACM Trans. on Database Systems*, 23(1), 1998.
6. F. Grandi, F. Mandreoli, "The Valid Web: it's Time to Go...", TIMECENTER Technical Report TR-46, 1999, *www.cs.auc.dk/research/DP/tdb/TimeCenter/*
7. F. Grandi, F. Mandreoli, "The Valid Web©", Proc. of Software Demonstrations Track at the EDBT'2000 Intl' Conf., Konstanz, Germany, 2000.
8. F. Grandi, F. Mandreoli, "The Valid Web©" Home Page, *degas.deis.unibo.it/ValidWeb*/
9. F. Grandi, F. Mandreoli, "The Valid Web: an XML/XSL Infrastructure for Temporal Management of Web Documents", Proc. of Intl' ADVIS'2000 Conf., Izmir, Turkey, 2000.
10. F. Grandi, F. Mandreoli, "Effective Representation and Efficient Management of Indeterminate Dates", Proc. of Intl' TIME'01 Symposium, Cividale del Friuli, Italy, 2001.
11. F. Grandi, F. Niccolucci, "XML Technologies for the Representation and Management of Spatiotemporal Information in Archaeology", Proc. of Intl' CODATA Conf., Baveno, Italy, 2000.
12. F. Grandi, M. R. Scalas, "Extending Temporal Database Concepts to the World Wide Web", Proc. of Natl' SEBD'98 Conf., Ancona, Italy, 1998.
13. S. Hermon, F. Niccolucci, "The Impact of Web-shared Knowledge on Archaeological Scientific Research", Proc. of Intl' CRIS 2000 Conf., Helsinki, Finland, 2000.
14. 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 *Temporal Databases – Research and Practice*, LNCS 1399, Springer-Verlag, Berlin, 1998.
15. 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.
16. M. Nascimento, M.H. Durham, "Indexing Valid Time Databases via B+Trees", *IEEE Trans. on Knowldege and Data Engineering*, 11(8), 1999.
17. F. Niccolucci (ed.), *From the Sources to the Network*, Proc. of Workshop on XML Applications to Historical Source Encoding, Florence, Italy, 2000.
18. 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 Intl' AHC-UK'99 Conf., London, UK, 1999.
19. N. Pioch, "The Web Museum", *www.cnam.fr/wm/*
20. R.T. Snodgrass (ed.) et al., *The TSQL2 Temporal Query Language*, Kluwer Academic Publishers, Boston, MA, 1995.
21. The eXtensible Markup Language (XML) Home Page, W3C Consortium, *www.w3.org/XML/*
22. The Microsoft Internet Explorer Home Page, *microsoft.com/windows/Ie/*
23. The Semantic Web Agreement Group Home Page, *swag.semanticweb.org*

ABOUT THE AUTHORS

Fabio Grandi is Associate Professor of Information Systems at the University of Bologna. His main scientific interests include temporal databases, Web extensions and schema versioning. E-mail *fgrandi@deis.unibo.it*

Federica Mandreoli is Research Assistant at the University of Modena and Reggio Emilia. Her interests include information retrieval, Web extensions and schema versioning. E-mail *fmandreoli@dsi.unimo.it*