Managing Valid Time Semantics for Semistructured Multimedia Clinical Data

Carlo Combi and Barbara Oliboni

Department of Computer Science — University of Verona
Ca' Vignal 2 — Strada le Grazie 15 — 37134 Verona (Italy)
{carlo.combi,barbara.oliboni}@univr.it

Abstract. In this paper we propose an approach to manage in a correct way valid time semantics for semistructured temporal clinical information. In particular, we use a graph-based data model to represent radiological clinical data, focusing on the patient model of the well known DICOM standard, and define the set of (graphical) constraints needed to guarantee that the history of the given application domain is consistent.

1 Introduction

In the clinical context the amount of multimedia temporal data is growing up. Recently, there has been an increasing attention on semistructured multimedia clinical data motivated also by the growing usage of XML (eXtensible Markup Language) [14] for exchanging medical data and knowledge [6, 12]. Semistructured data have some structure, that may be irregular or incomplete and does not necessarily conform to a schema fixed in advance [1]. In the semistructured data context, the same information can be structured in different ways within the same document, and documents about the same topics can be structured in different ways, thus an important issue is related to the integration of semistructured (XML) data. At this aim, in this paper we use a graphical data model to represent in a simple and intuitive way semistructured documents coming from different sources and (possibly) structured in different ways.

A further aspect which has gained increasing attention also for clinical data is the management of the temporal dimension of information [6]. The time dimension usually considered for clinical data is *valid time* (VT), which is user-provided, and represents the time when a fact is true in the considered domain [7]. Both for conceptual models and for the relational model, several proposals deal with the issue of formalizing and managing valid time semantics of related data [9]: for example, we could require that the database system must be able to constrain the valid time of a visit, i.e. the time during which the visit happened, to be *during* the valid time of the patient, i.e. the time during which a person was hospitalized; on the other hand, we could require that the valid time of the diagnosis starts after the beginning of the symptoms, the diagnosis is based on.

The main issue we consider in this paper is related to the management of valid time semantics for temporal semistructured multimedia clinical data. To

represent semistructured multimedia temporal information we use the Multimedia Temporal Graphical Model (MTGM) [6], which is a general data model for representing semistructured data, having multimedia and temporal features.

We propose a simple approach to manage valid time semantics in the semistructured data context allowing the definition of the set of constraints needed to manage in a correct way the valid time dimension of information represented by means of MTGM. Being MTGM a graph-based data model, we use a graphical approach also to represent the constraints. A constraint will be composed by a *graph*, used to identify the portions of the semistructured data where the constraint has to be applied, and a set of *formulae*, representing restrictions to impose on those information.

In particular, we apply our approach to clinical data, considering DICOM [2] data. DICOM (Digital Imaging and Communications in Medicine) is a standard method to encode and transfer images and related information between heterogeneous sources. DICOM is considered one of the most important standards in radiology, and allows the physicians to store documents containing a diagnostic report related to radiology images. Each report contains the interpretation and the impressions of the radiologist. This kind of document is a typical example of a semistructured document. It has a structure (defined by the DICOM standard) that could be irregular or incomplete: for example the interpretation of the radiologist could be missing in a particular report.

The structure of the paper is as follows: in Section 2 we briefly describe some research directions on multimedia temporal data and then introduce MTGM through general clinical examples. Then, in Section 3, we describe how MTGM allows us to model DICOM data, and in Section 4 we define the constraints needed to manage in a correct way the valid time dimension. In Section 5 we describe a prototype, developed by using the Java technology, managing multimedia temporal clinical data, together with the defined constraints related to the valid time semantics. Finally, in Section 6 we sketch some conclusions.

2 Managing multimedia temporal data

In the context of multimedia database systems, it is possible to distinguish two main research directions: the first one focuses on data modeling and querying issues [4, 5], while the second one is addressed to model multimedia presentations [3, 11]. As examples of the first research direction, in [4] the authors present a unified data model for multimedia types, such as images, sounds, videos and long text data, while in [5] the author faces the problem of modeling temporal aspects (at different granularities and with indeterminacy) of visual and related textual entities in multimedia databases. As for the second research direction, in [3] the authors deal with the mechanisms for the specification of synchronization constraints between different media objects, while in [11] the authors present a methodological approach for checking the temporal integrity of interactive multimedia documents.

To this regard, MTGM deals with both the above research issues, in the management of clinical data [6]. MTGM can be considered as a logical model for semistructured and multimedia data: the main advantages of this model are related to (i) the chance of representing in a flexible and common way information having different structures, and (ii) the possibility of representing, in a seamless way, both temporal aspects of multimedia data and their temporal presentation requirements. In this paper, we do not focus on the graphical description of multimedia presentations, composed by semistructured and multimedia information represented by means of MTGM, even though in the clinical context, the possibility of composing presentations starting from the stored information is an important feature. Details on multimedia presentations with MTGM can be found in [6]. As for data modeling, multimedia temporal data are represented through graphs: an MTGM graph is a directed, labeled graph, with a single root. Figure 1 shows a portion of an MTGM graph representing information about a pregnant patient and her ultrasound exam.

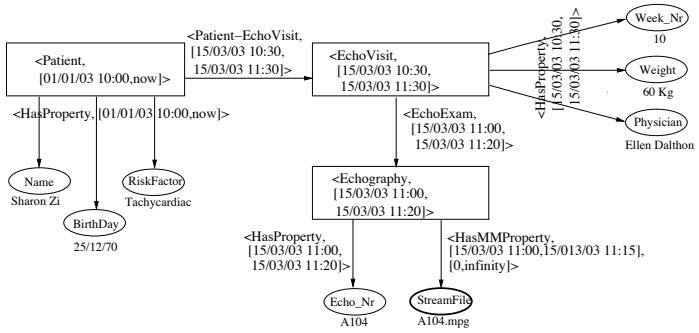


Fig. 1. A portion of an MTGM graph.

MTGM has *complex*, *atomic* and *stream* nodes: complex nodes represent abstract entities and are depicted as rectangles, atomic nodes represent primitive values and are depicted as ovals, and stream nodes (which are a particular kind of atomic node) contain multimedia information and are depicted as thick ovals. In the example of Figure 1, there are three complex nodes, *Patient*, *Echo Visit* and *Echography*, and seven atomic nodes as, for example, *Name*, *Birthday* and *RiskFactor*. The atomic node *StreamFile* (child of *Echography*) is a stream node and contains the string "A104.mpg", which refers to the file that encodes the movie of the echography.

Nodes are connected through direct labeled edges (*relational edges*). Relational edges between complex nodes and their atomic nodes have labels with name "*HasProperty*", while relational edges between complex nodes and their stream nodes have labels with name "*HasMMProperty*".

Valid time is explicitly managed by MTGM both for nodes and edges. The valid time[1] of a complex node is represented in its label. The valid time of an atomic (stream) node is represented in the label of the edge between the atomic node and its parent. For example, in Figure 1, the valid time of the node *Patient* is [01/01/03 10:00, *now*] where "*now*" indicates that the represented fact is currently true (i.e., Sharon Zi is still a patient), while the valid time of the simple node *Physician* related to the complex node *Echo_visit* is [15/03/03 10:30, 15/03/03 11:30] (i.e., the visit has been executed by Ellen Dalthon during the specified interval). The label of a relational edge is composed by the name of the relationship and its valid time. The label of the edge relating a complex node to a stream node contains also the specific subpart of the stream object the complex node is related to. As shown in Figure 1, the label of the edge between *Echography* and *StreamNode* is ⟨*HasMMProperty*, [15/03/03 11:10, 15/03/03 11:15], [0, *infinity*]⟩ and represents the time during which the movie of the echography has been recorded. In particular, the interval [0, *infinity*] describes the fact that all the frames of the movie are related to the echography.

3 Modeling DICOM data by the Multimedia Temporal Graphical Model

DICOM is a standard method to represent images and related information, and in particular it is considered as a standard in radiology. DICOM allows one to store diagnostic reports containing radiology images and the interpretation and the impressions of the radiologist [2].

The DICOM standard aims at allowing interoperability of medical image equipment and considers several aspects: from network communication, to syntax and semantics of exchanged information, to media storage services and file format, to requirements for verifying standard conformance. In particular, in this work, we consider the *DICOM information model* [2], which is an ER-based schema of the domain considered by DICOM. Let us now focus on the representation of (part of) the DICOM model through MTGM [6].

In Figure 2 we show an example of an MTGM graph representing information about a patient and her radiological data, according to the DICOM information model. In Figure 2 *Patient* and *Study* are complex nodes, while nodes *Name* of *Patient* and *Physician's_Name* of *Study* are simple nodes. The node *Raw_Data* of *Image* is a stream node. As for the temporal dimension, the valid time of the *Patient* is [10/01/05, *now*] where "*now*" indicates that the considered fact is currently true. The valid time of an atomic (stream) node is represented in the label of the edge connecting the atomic node and its parent. The label of a relational edge is composed by the already mentioned name of the relationship and by its valid time: for example, the label of the edge between *Patient* and *Study* in Figure 2 is $\langle References, [10/01/05, now]\rangle$. The label of the edge relating a complex node to a stream node contains also the specific subpart of the

[1] In this work dates are represented in the format DD/MM/YY. Format and granularity of timestamps can be chosen with respect to the considered domain.

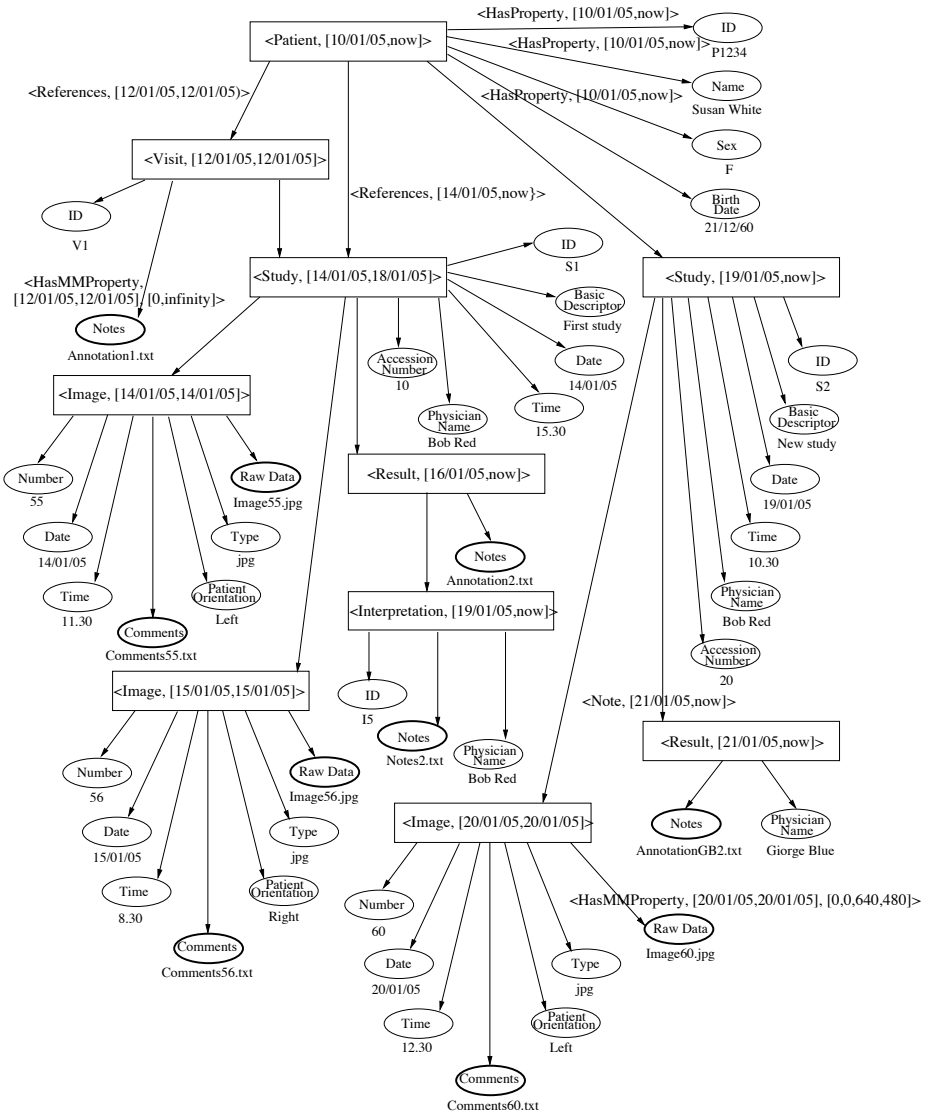


Fig. 2. An example of an MTGM graph.

stream object the complex node is related to. In case of images the subpart is specified by $[x, y, width, height]$ that represents the portion of the image with origin in *(x,y)* and dimensions *width* and *height*: as an example in Figure 2, the label of the edge between the *Image* (with *Number* equal to 60) and *Raw Data* is $\langle HasMMProperty, [20/01/05, now], [0, 0, 640, 480]\rangle$. For readability reasons, in Figure 2 we do not report all the edge labels.

4 Expressing valid time semantics

Constraints on valid times must be able to guarantee that the history of the given application domain is consistent. As an example, at a specific time instant, between two nodes it cannot exist more than one edge representing the same relation.

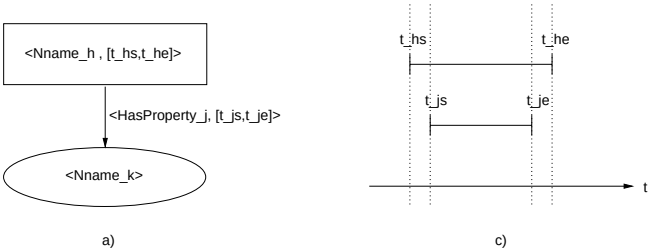
The graphical formalism we use in the following constraints has been described in [8, 13]: a constraint is composed by a *graph*, which is used to identify the subgraphs (i.e., the portions of a semistructured database) where the constraint has to be applied, and a set of *formulae*, which represent restrictions imposed on those subgraphs.

We distinguish two different categories of constraints for valid time values of nodes and edges: *basic constraints* must be satisfied by every MTGM graph; *domain-dependent constraints* are further constraints, which can be defined either for some specific nodes and edges or for the whole graph for a specific clinical domain. In the following, part a) of Figures from 3 to 9 identifies the subgraphs where the constraint has to be applied, and part b) the set of *formulae* representing restrictions imposed on those subgraphs. Part c) shows an example of intervals satisfying the related constraint.

4.1 Basic constraints

In an MTGM graph, we identified the following basic constraints:

1. The time interval of an edge between a complex node and a simple node must be related to the time interval of the complex node (Figure 3). Intuitively, the relation between a complex node and a simple node cannot survive the complex node; thus, the time interval of the edge cannot start before and cannot end after the valid time of the complex node. This is due to the fact that we suppose that a complex node is related to its properties (simple nodes) while it is valid.
2. A complex node cannot have, at a given time, different values for the same property, and thus at a specific time instant, a complex node can be related to at most one simple node with a particular name.
3. At a specific time instant, between two complex nodes it cannot exist more than one edge with the same name. Intuitively, an edge represents a relationship between two complex nodes, thus it makes no sense representing with two edges the same relationship.



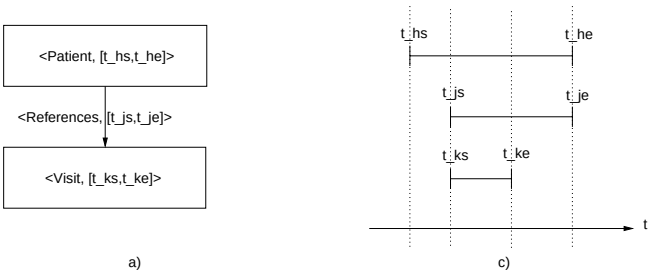
b) $t_{js} \geq t_{hs} \wedge t_{je} \leq t_{he}$

Fig. 3. The VT constraint on the time interval of edges pointing a simple node.

4.2 Domain dependent constraints

Other possible optional constraints that can be considered are the ones for imposing restrictions on the time interval of an edge connecting two complex nodes. These constraints are strictly related to the semantics of the represented objects and relationships. In this work, we consider radiological clinical data represented by means of DICOM, and thus we define the set of constraints needed to manage this kind of information in a consistent way. In Figure 2 we represented information about the DICOM patient information model, related to radiology images, by means of MTGM, and now we define the related set of domain dependent constraints.

1. A patient can be examined only in the period of time in which she is valid (i.e., she is alive), thus the valid time of the visit must be contained in the valid time of the patient. Moreover the valid time of the edge must start at the same time of the visit (Figure 4).

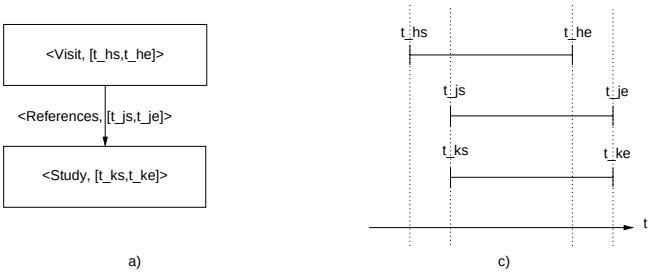


b) $t_{ks} \geq t_{hs} \wedge t_{ke} \leq t_{he} \wedge t_{js} = t_{ks}$

Fig. 4. The VT constraint on the relation between a patient and a visit.

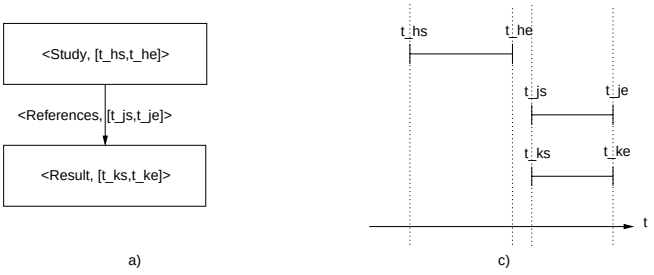
In this constraint we do not consider the end time of the edge, because we do not fix a rigid relation between its value and the valid times of the patient and the visit. Modifying the constraint, it is possible to require that the start time of the edge is related to the start time of the visit, and that the end time of the edge is related to the end time of the patient. In this case, for each subgraph satisfying the structure shown in part a) of Figure 4 we require that $t_{ks} \geq t_{hs} \wedge t_{ke} \leq t_{he} \wedge t_{js} = t_{ks} \wedge t_{je} = t_{he}$. Intervals reported in part c) of Figure 4 satisfy also this version of the constraint.

2. The situation of the patient with respect to the visit can be studied during the visit or after the visit; thus the start time of the study must start at the same time or after the valid time of the visit (Figure 5).



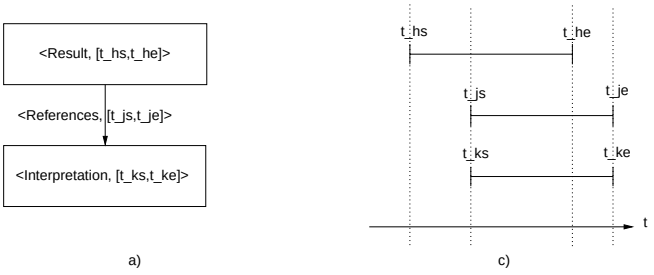
b) $t_{ks} \geq t_{hs}$

Fig. 5. The VT constraint on the relation between a visit and a study.



b) $t_{ks} > t_{hs}$

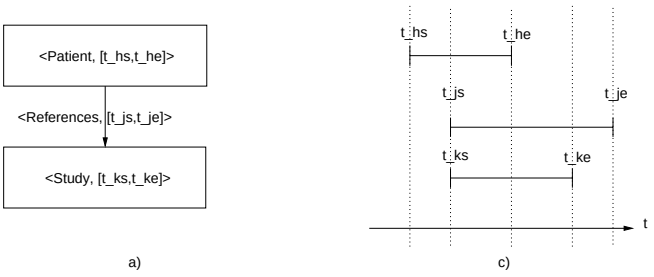
Fig. 6. The VT constraint on the relation between a study and a result.



b) $t_{ks} > t_{hs} \wedge t_{ke} \geq t_{he}$

Fig. 7. The VT constraint on the relation between a result and an interpretation.

3. The result of the study can be defined during or after the study, thus the valid time of the result must start after the start of the valid time of the study (Figure 6).
4. The result can be interpreted after the creation of the result itself, thus the valid time of the interpretation must start after the valid time of the result (Figure 7).
5. A patient can be considered for a study after the moment in which she becomes a patient, thus the valid time of the study must start after the start of the valid time of the patient (Figure 8).

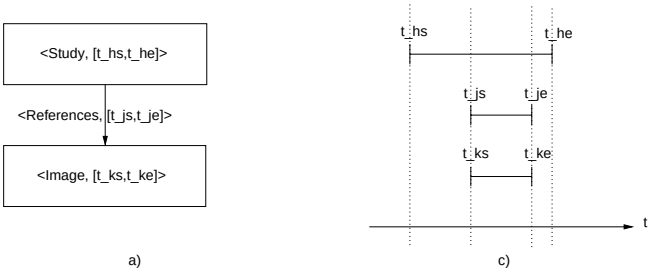


b) $t_{ks} > t_{hs}$

Fig. 8. The VT constraint on the relation between a patient and a study.

6. An image must be related to a study, thus the valid time of the image must be contained in the valid time of the study (Figure 9).

Further, and more complex constraints could be defined: for example we could require that the valid times of two studies related to the same patient do not overlap.



b) $t_{ks} \geq t_{hs} \wedge t_{ke} \leq t_{he}$

Fig. 9. The VT constraint on the relation between a study and an image.

4.3 Translating MTGM graphs into XML documents

MTGM can be seen as a logical model for semistructured data; an MTGM graph can be physically realized through the emerging XML technology. Among the main differences between MTGM and the data model underlying XML, we have to consider (i) the fact that MTGM considers labeled graphs, while XML mainly deals with trees, and (ii) that, while in MTGM both nodes and edges are labeled, in XML only labeled nodes are allowed. Another important difference is that XML node labels are in some way atomic, while in MTGM we deal with compound labels.

The overall, main ideas underlying the designed translation technique can be summarized as follows:

1. Complex nodes are translated into complex elements (i.e., elements which contain other elements); in particular, they have a (nested) element for the corresponding valid time and an element for each outgoing edge.
2. Atomic (stream) nodes are translated into mixed elements (i.e., elements containing both string values and other elements); in particular, they contain the string representing their values and an element for their valid times (which are contained, in the related MTGM graph, in the label of the ingoing relational edge).
3. Edges between complex nodes are represented through complex elements nested into the element corresponding to the complex node, which the edge originates from. The element corresponding to the node the edge points to, is referred through a suitable attribute in the element representing the edge.
4. Edges between a complex node and an atomic (stream) one are not translated (see point 2.).
5. Compound labels are managed by introducing suitable sub-elements (i.e., nested elements), as for representing valid times of nodes and edges.
6. Elements corresponding to MTGM nodes have an attribute (of type ID), which allows one to refer to them in an unambiguous way.

5 Managing temporal clinical data by XML native database systems

Semistructured temporal DICOM data are managed by means of a system prototype developed by the Java technology, and based on the native XML database system eXist [10]. The system prototype has been designed according to the architecture described in [6] and allows us to store clinical radiological data (represented by means of MTGM as shown in Figure 2) in an XML database, as reported in Figure 10.

In the left part of Figure 10 it is possible to see the description of a portion of the MTGM graph reported in Figure 2. The prototype allows us to create and modify an MTGM graph by adding nodes and edges. The description of them can be inserted by means of suitable windows (see the right part of Figure 10). At the end of these operations, the MTGM graph is represented through an XML document and stored in the eXist database.

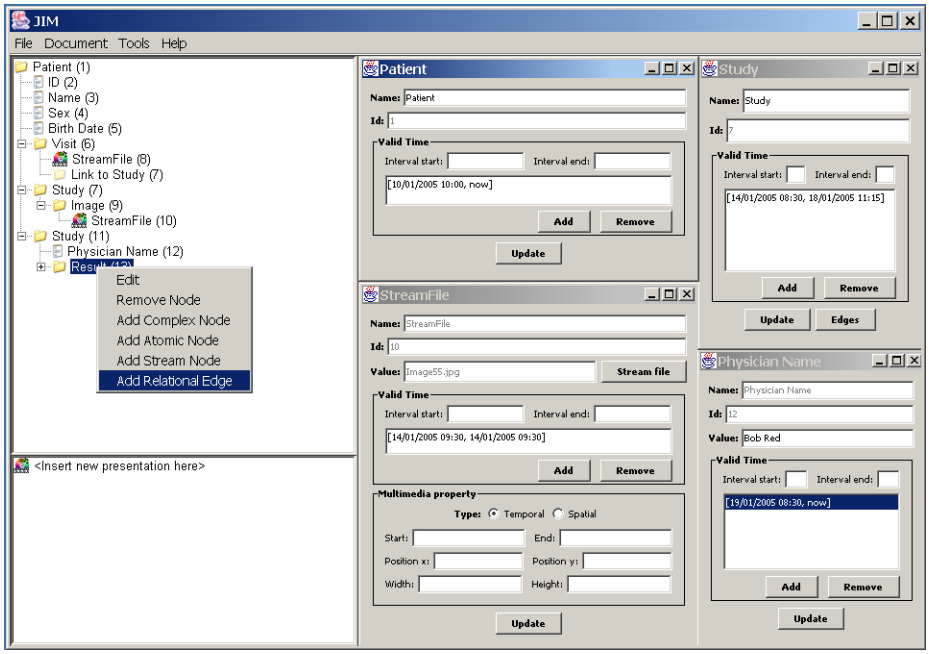


Fig. 10. A screen shot of the described prototype.

The constraints described in Section 4 can be managed by the prototype, which allows us to define a set of constraints on a graph-based representation of information. The defined constraints are verified at each operation on the graph: the graph can be modified only if the constraints are verified.

6 Conclusions

In this work we introduced a graph-based approach to manage in a correct way valid time semantics for semistructured clinical information. In particular we considered the patient data model proposed by the DICOM standard and showed how to represent it by the semistructured temporal data model MTGM; moreover we modeled valid time semantics of DICOM data through graphical constraints. Finally, we showed how to use our system prototype to manage radiological data. We plan to study and define more complex constraints related for example to nodes connected by complex paths involving several edges, and to further extend the implemented system to manage them.

References

1. S. Abiteboul. Querying Semi-Structured Data. In *Proceedings of the International Conference on Database Theory*, volume 1186 of *Lecture Notes in Computer Science*, pages 262–275, 1997.
2. National Electrical Manufactures Association. Digital Imaging and Communications in Medicine (DICOM). 2003. `http://medical.nema.org/dicom/2004.html`.
3. E. Bertino and E. Ferrari. Temporal synchronization models for multimedia data. *IEEE TKDE*, 10(2):612–631, 1998.
4. A. F. Cárdenas and J. D. N. Dionisio. A Unified Data Model for Representing Multimedia, Timeline and Simulation Data. *IEEE TKDE*, 10(5):746–767, 1998.
5. C. Combi. Modeling Temporal Aspects of Visual and Textual Objects in Multimedia Databases. In *TIME 2000*, pages 59–68. IEEE Computer Society Press, 2000.
6. C. Combi, B. Oliboni, and R. Rossato. Merging multimedia presentations and semistructured temporal data: a graph-based model and its application to clinical information. *International Journal Artificial Intelligence in Medicine*, 34(2):89–112, 2005.
7. Carlo Combi and Angelo Montanari. Data models with multiple temporal dimensions: Completing the picture. In *CAiSE*, pages 187–202, 2001.
8. E. Damiani, B. Oliboni, E. Quintarelli, and L. Tanca. Modeling semistructured data by using graph-based constraints. In *OTM Workshops Proceedings*, Lecture Notes in Computer Science, pages 20–21. Springer-Verlag, Berlin, 2003.
9. C. S. Jensen and R. Snodgrass. Temporal data management. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):36–44, 1999.
10. W. Meier. An Open Source Native XML Database. In *Web, Web-Services, and Database Systems*, volume 2593 of *Lecture Notes in Computer Science*, pages 169–183. Springer, 2003.
11. I. Mirbel, B. Pernici, T.K. Sellis, S. Tserkezoglou, and M. Vazirgiannis. Checking the Temporal Integrity of Interactive Multimedia Documents. *VLDB Journal*, 9(2):111–130, 2000.
12. R. Noumeir. Dicom structured report document type definition. *IEEE Transactions on Information Technology in Biomedicine*, 7(4):318–328, 2003.
13. B. Oliboni. *Blind queries and constraints: representing flexibility and time in semistructured data*. PhD thesis, Politecnico di Milano, 2003.
14. World Wide Web Consortium. Extensible Markup Language (XML) 1.0, 1998. `http://www.w3C.org/TR/REC-xml/`.