

Tracking Data Provenance of Archaeological Temporal Information in Presence of Uncertainty

SARA MIGLIORINI, ELISA QUINTARELLI, and ALBERTO BELUSSI, Dept. of Computer Science, University of Verona

The interpretation process is one of the main tasks performed by archaeologists who, starting from ground data about evidences and findings, incrementally derive knowledge about ancient objects or events. Very often more than one archaeologist contributes in different time instants to discover details about the same finding and thus, it is important to keep track of history and provenance of the overall knowledge discovery process. To this aim we propose a model and a set of derivation rules for tracking and refining data provenance during the archaeological interpretation process. In particular, among all the possible interpretation activities, we concentrate on the one concerning the dating that archaeologists perform to assign one or more time intervals to a finding in order to define its lifespan on the temporal axis. In this context we propose a framework to represent and derive updated provenance data about temporal information after the mentioned derivation process. Archaeological data, and in particular their temporal dimension, are typically vague, since many different interpretations can coexist, thus we will use Fuzzy Logic to assign a degree of confidence to values and Fuzzy Temporal Constraint Networks to model relationships between dating of different findings represented as a graph-based dataset. The derivation rules used to infer more precise temporal intervals are enriched to manage also provenance information and their following updates after a derivation step. A MapReduce version of the path consistency algorithm is also proposed to improve the efficiency of the refining process on big graph-based datasets.

CCS Concepts: • **Information systems** → *Data management systems*.

Additional Key Words and Phrases: provenance, temporal constraints, information discovery

ACM Reference Format:

Sara Migliorini, Elisa Quintarelli, and Alberto Belussi. 2021. Tracking Data Provenance of Archaeological Temporal Information in Presence of Uncertainty. *ACM J. Comput. Cult. Herit.* 1, 1, Article 1 (January 2021), 34 pages. <https://doi.org/10.1145/3480956>

1 INTRODUCTION

In the archaeological scenario, the interpretation process allows one to derive and discover new knowledge about ancient findings on the basis of direct and indirect observations performed by domain experts (archaeologists), that are very often combined with previous interpretations performed by themselves or other colleagues. In this context spatial and temporal dimensions are usually relevant for archaeological research, because they contribute to derive new important relationships between findings, to assign chronological and location values, in particular as concern to stratigraphic analysis. In our previous works [9, 11, 12] we suggest the possibility to apply existing automatic reasoning techniques to the dates manually assigned by archaeologists in order to automatically derive more precise temporal

Authors' address: Sara Migliorini, sara.migliorini@univr.it; Elisa Quintarelli, elisa.quintarelli@univr.it; Alberto Belussi, alberto.belussi@univr.it, Dept. of Computer Science, University of Verona, Strada Le Grazie, 15, Verona, Italy, 37134.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

knowledge, find out inconsistencies, or validate existing interpretations, exploiting both the available spatial and temporal information. This paper starts from such previous contributions, which present a very detailed spatial and temporal analysis of archaeological information, and enriches them by adding data provenance knowledge.

Among all possible interpretation and derivation activities performed by archaeologists, we concentrate on the dating process, since it is one of the most challenging and interesting task from an automatic reasoning point of view. Archaeological data, and more specifically their temporal dimensions, are typically uncertain and thus many different interpretations can coexist, each one with its own degree of confidence and consequently several different global conclusions can be inferred from them. Each interpretation is typically associated with its author and the confidence greatly depends on the archaeologist's reputation in the field. For these reasons, during the interpretation process, it is necessary not only to refine the acquired knowledge, e.g. to infer more precise time intervals, but also to keep track of the provenance information that has affected such inference. More specifically, it is necessary to consider from which pieces of information (past interpretations) the current and updated knowledge has been originated, together with their authorship and their degree of influence.

In computer science, the term *provenance* means data lineage and is the ability to record the history of data together with its place of origin, and is useful to determine the chronology of the ownership, custody or location of any object and to provide a critical foundation for assessing authenticity and enabling trust. As highlighted in [16], *data provenance* is separable from other forms of provenance. In our specific archaeological scenario, the term provenance comes originally from the art world and it has been applied in archaeology and paleontology as well, where it refers to having trace of all the steps involved in producing a scientific result, such as a finding, from experiment design through acquisition of raw data, and all the subsequent steps of data selection, analysis and visualization. Such information is necessary for the reproduction of a given result, it can be useful to establish precedence (in case of patents, Nobel prizes, etc.) [31] and its meaning is different from that of provenience.

In our previous work [38] we introduced a graph-based model and the idea of a set of derivation rules that are able to track the data provenance regarding temporal information during the archaeological interpretation process. More specifically, even if many specific temporal dimensions can be described in the archaeological context, for keeping the discussion simple we concentrated on the dating process used by archaeologists to assign one or more lifespans to an archaeological object as formalized in [9, 11], with the purpose of checking the temporal data consistency, or reducing the vagueness with the use of Fuzzy Temporal Constraint Networks (FTCN) [4, 50]. The extension of the proposed approach to a more articulated and complex temporal model is straightforward, as it will be clear in the following sections. Moreover, in [38] we made a step forward by exploring how to manage and infer new knowledge including approximate provenance of data and complex inferences. Since the dating process, and more generally any interpretation process performed by an archaeologist, greatly depends on previous interpretations and discoveries performed by some colleagues, it is of paramount importance the ability to track data provenance in order to correctly recognize the authorship of the information and provide a correct history of the interpretation process.

The main aim of this paper is to extend the work in [38] by (i) providing a complete formalization of the core *Star* model used for provenance tracking, (ii) defining a set of translation rules towards a FTCN enhanced with provenance information, called here Provenance-Aware Fuzzy Temporal Constraint Network (PA-FTCN), (iii) proposing a MapReduce [18] implementation of the path-consistency algorithm, which exploits the characteristics of the model at hand in order to apply some derivations in parallel and improve the overall efficiency of the derivation procedure.

The basic idea behind the proposed MapReduce algorithm is that archaeological information is naturally partitioned into a set of highly connected objects among which the temporal relations are very dense, while the number of relations

between different groups are typically very few. More specifically, during the interpretation process, archaeologists start from elementary objects (called *archaeological partitions* in the *Star* model) and use their characteristics in order to derive more complex objects (called *archaeological units* in the *Star* model). This group determines a subdivision of the information into quite independent chunks, which can be efficiently processed in parallel.

The remainder of the paper is organized as follows: Sect. 2 provides an overview of the literature about the derivation of new temporal knowledge and the representation of provenance information. Sect. 3 introduces the *Star* model, an archaeological spatio-temporal model which allows also to model detailed provenance information for archaeological and temporal data. Sect. 4 discusses the proposed framework by starting with the notion of Provenance-Aware Fuzzy Temporal Constraint Network (Sect. 4.1) and presenting a set of rules for translating a *Star* into a PA-FTCN (Sect. 4.2), then it concentrates on the path consistency algorithm by defining extended derivation operations that take care of provenance information (Sect. 4.3) and finally it introduces an efficient MapReduce version of the path consistency algorithm (Sect. 4.4). In Sect. 5 we illustrate a case study taken from a real world scenario concerning the archaeological data of Verona. Finally, Sect. 6 concludes the work.

2 RELATED WORK

Archaeological Temporal Information. Archaeological data are not traditionally stored inside databases, but they are usually collected and maintained inside reports, drawings and other kinds of documents. However, in recent years many efforts have been devoted to define suitable models for representing such kind of data and to store them inside databases, with the main aim to improve interoperability, information exchange and querying between researchers, professionals and also the public. As discussed in [24], we can distinguish two different ways to describe the passing of time in conceptual models: (a) by representing time as a cross-cutting aspect and (b) by using an explicit representation of time in relevant classes. Even if (a) can be considered the most powerful and expressive approach in most cases, the explicit approach (b) is most suitable when objects have a strong temporal semantics. In this paper, we choose to discuss the application of approach (b) for those objects which represent the main archaeological concepts of interest, since for them the temporal characterization is of paramount importance.

Usually the archaeological context requires to represent additional and specific time dimensions, which are not typically considered in the temporal database research. In [30] the authors identify six potential time categories for archaeological finds which includes: excavation time, database time, stratigraphic time, archaeological time, site phase time and absolute time. While database time corresponds to the traditional transaction time [47], the other temporal characteristics can be seen as a specialization of the valid time [48], each one with a particular meaning that can influence each other producing new knowledge. The *Star* model proposed in [11] and extended in this paper includes many of these time categories, in particular: the excavation time, the stratigraphic time (in terms of relative temporal positions between findings), the archaeological time (e.g. Roman Time or Middle Age), the site phase time (i.e. the distinction of different phases during an object life), and the absolute time. Moreover, it introduces a way to represent both uncertain temporal values by means of fuzzy set theory and *qualitative* temporal relations through a topological approach which allow to represent precedence relations without any realization as time instants. Indeed, as we will see in the following section, one of the characteristics provided by the *Star* model is the possibility to represent also topological temporal relations between archaeological objects. This allows one to represent classical temporal relations between objects, like the Allen's temporal relations [2], as well as stratigraphic relations derivable from the Harris matrix [25], without any actual instant or period characterization. In [11] we provide a deep description of this representation together with a way to derive new temporal knowledge from the combination of stratigraphic,

temporal but also 2D spatial topological relations [21]. In recent years, other works have investigated the possibility to build graph structures in which the information of an Harris Matrix can be combined together with other temporal information like the Allen's Algebra [6], in order to both derive new ordering of data [14] or perform semantically enriched deductions which fundamentally link archaeological data together [35]. Even if this is an important research topic, it is out the scope of this paper and will be not considered in the remainder. Moreover, in this paper, for not cluttering the notation and keeping the discussion simple, we consider a subset of the richer *Star* model presented in [11] and in particular we take into account only a subset of time dimensions. The extension to a richer model can be easily obtained by extending the general approach described here.

The inherent uncertainty of spatio-temporal archaeological data is widely recognized in literature. In [20] the author confirms that even if actual time and place of past events is not known, it lies within the boundaries given by dating and localization of the evidence. Therefore, they propose the construction of a probabilistic model in order to reduce such uncertainty and derive new precise knowledge. Fuzzy set theory and probability theory are two related but different ways for modeling uncertainty. Probability statements are about the likelihoods of outcomes: an event either occurs or does not, and you can choose on it. This theory is typically used to make predictions and is characterized by only two outcomes: true and false. Conversely, fuzzy set theory was introduced as a mean to model the uncertainty of natural language and is extended to handle the concept of partial truth (or degree of truth). It cannot say clearly whether an event occurs or not and is usually applied for describing happened events. For these reasons, a fuzzy representation of time has been adopted in the definition of the *Star* model, as deeply discussed in [7].

The joint statistical analysis of spatial and temporal information is widely used in archeology with reference to many other interpretation contexts which are not covered in this paper, such as the identification and explanation of trends or patterns [17, 36] and the summarization of main characteristics of data [23]. Clearly, also this kinds of derivations can benefit from the use of some techniques for tracking and updating provenance information and represent an interesting future point of extension for our work.

Provenance. In recent years, together with the rapid growing of data to manage, different formal models for provenance storage, maintenance, and querying has been proposed. As motivated in [46], metadata describing data related to a considered scenario is essential to enrich, make sense and enable the reuse of data. Among important metadata, data provenance, also called lineage, is defined as any information about entities, activities, and people involved in producing a piece of data [42]. Indeed, it keeps track of the derivation history of a data object from its original sources and it is in general relevant to protect intellectual property but also to determine the veracity and quality of any information. In GIS, provenance is the information describing materials or findings and transformations applied to derive the data [32].

Due to the importance of provenance information, the W3C has proposed PROV, which is the recommendation for provenance data model and language [1], extended in the GIS context by the metamodel in [44] for capturing the complete history of lineage information provided by e-science experiments. Data provenance [15] differs from other forms of meta-data because it often is based on relationships among objects [26]. Indeed, the ancestry relationships used in provenance for correlated objects form a directed graph that can be represented though semistructured data models. In [41] the authors have encoded provenance graphs into Datalog and expressed inference rules and constraints with the same declarative language, in order to determine inconsistencies with respect to temporal constraints or provenance information (e.g. inconsistent cycles).

Temporal reasoning. A Temporal Constraint Network (TCN) [19] is a formalism for representing temporal knowledge based on metric temporal constraints. It supports the representation of temporal relations and is provided with efficient

algorithms based on CSP (Constraint Satisfaction Problem) techniques. An extension of the traditional TCN, called Fuzzy Temporal Constraint Network (FTCN), which is able to deal with fuzzy sets, have been proposed in [50] and further developed in [3, 4] with the aim to represent vague and imprecise temporal knowledge. In this paper we propose an extension of the FTCN, called Provenance-Aware FTCN, to model and reason on time and provenance concepts together.

TCN belongs to the research area known as *temporal reasoning*, which analyzes existing data in order to determine their consistency, answer queries about scenarios satisfying all constraints, and derive missing information. These techniques are particularly useful in the archaeological context, where incomplete temporal data with some constraints are typically available. Conversely, another temporal research area, known as *temporal data mining*, analyzes large amount of temporal information to discover existing patterns. Many approaches exist for temporal data mining which are based on various data model and are suitable for different applications. In [43] the authors provide a unified view of such concepts and a guideline for selecting the appropriate method and data model based on the specific purpose. The use of data mining technique can be useful also in the archaeological context in order to discover other kinds of relations between objects that are out the scope of this paper. In particular, the use of fuzzy clustering techniques for archeological data analysis is discussed in [8].

In general, the use of computational intelligence techniques in archaeology has been discussed in [5], where the author analyses if it is possible to automate the archaeological knowledge production, coining the term *computable archaeology*. The main conclusion is that artificial intelligence could provide an invaluable tool for supporting archaeologists in their work, particularly for dealing with the structure and growth of scientific knowledge.

The constraint propagation performed on a TCN is typically done by applying the classical path-consistency algorithm [34]. However, this algorithm presents a high computational complexity and sooner or later becomes unusable for treating real-world problems. For this reason, several efficient versions of the path consistency algorithm, which try to reduce its computational complexity [33], have been implemented. In this paper we propose a different approach which uses the MapReduce programming paradigm by exploiting the semantical characteristics of the network at hand.

3 THE SPATIO-TEMPORAL ARCHAEOLOGICAL MODEL

This paper refers to the graph-based Spatio-Temporal ARchaeological model (*Star*) presented in [9, 11]. More specifically, for not cluttering the notation and keeping the discussion simple, we consider a core version of the model which is sufficiently expressive for describing the potentiality of the proposed approach. This section describes the main concepts contained in this core version of the model together with necessary extensions needed for properly representing and managing data provenance.

In the *Star* model three main objects of interest can be recognized: `ST_InformationSource`, `ST_ArchaeoPart` and `ST_ArchaeoUnit`, which are depicted in yellow in Fig. 1. An `ST_ArchaeoUnit` is a complex archaeological entity obtained from an interpretation process performed by the responsible officer. Such an interpretation is done based on some findings (represented by `ST_ArchaeoPart` instances) retrieved during an excavation process, a bibliographical analysis or other investigation processes (represented by `ST_InformationSource` instances). In other words, each `ST_ArchaeoUnit` is connected to one or more constituent `ST_ArchaeoParts`, which have been collected during an investigation activity described by an `ST_InformationSource`. The concept of `ST_ArchaeoPart` is more articulated than the mere representation of a simple finding. As discussed in [13], an `ST_ArchaeoPart` can be a structure, in a more or less complete form, a movable element, a stratigraphic or geological substrate which has particular importance

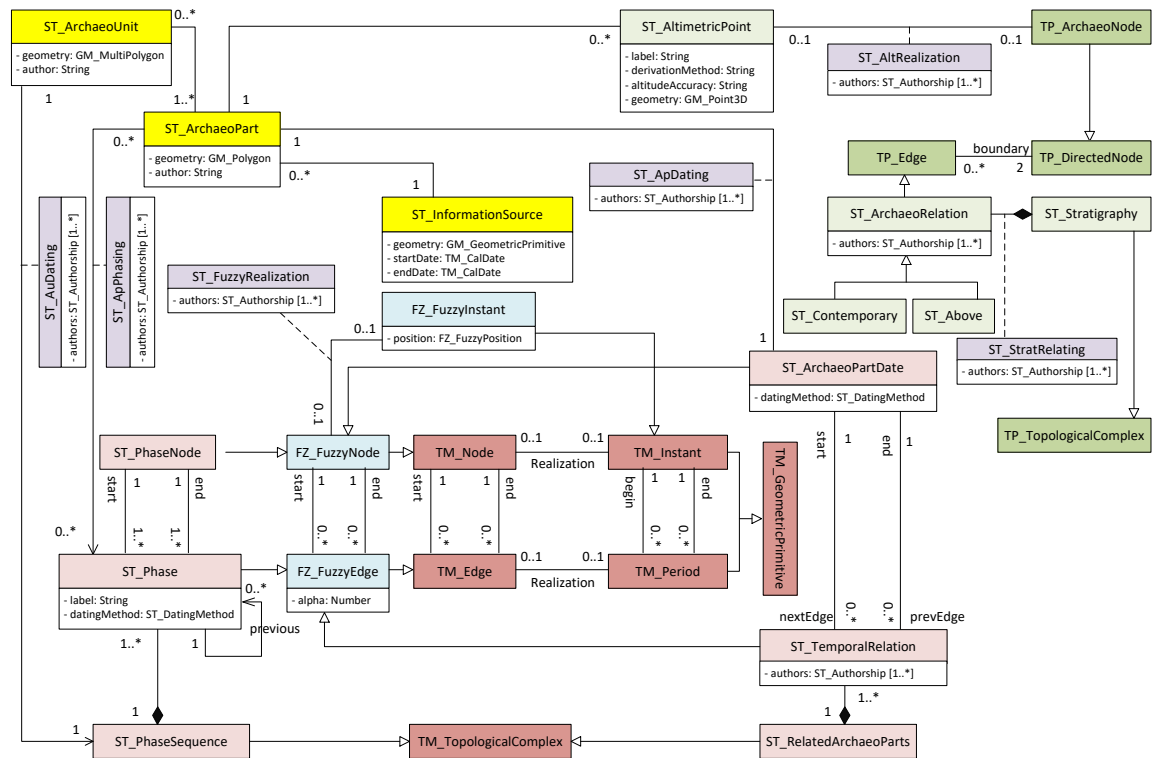


Fig. 1. UML Class diagram of the main *Star* classes. The prefixes GM_ and TP_ are used to denote the classes coming from the Standard ISO 19107 for the definition of geometric primitive types and geometric topological types, respectively. Similarly, prefix TM_ is used to denote the classes coming from the Standard ISO 19108 for the definition of temporal concepts. Finally, prefixes ST_ and FZ_ denote the classes introduced and defined in the *Star* model. More specifically, the former is used for the definition of archaeological objects, while the latter for the fuzzy extension of temporal data types, as further described in Fig. 2. The diagram uses the standard UML notation, where arcs represent simple relations between classes, while big arrows denote hierarchical relation, and diamond arrows are used for composition relations.

in the historical perspective. It can be used to represent both well recognized findings, as well as raw observations or high level reconstructions. Indeed, several specializations of this base class have been defined in [40]. Similarly, an ST_InformationSource can represent many different kinds of investigations and several specializations of the base class have been defined in [40] to properly represent each of them. Anyway, for the purposes of this paper, we can concentrate on the most general classes, since provenance information can be directly related to them.

Fig. 1 summarizes the main structure of the core *Star* model considered in this paper. In particular, the new aspects related to data provenance tracking have been highlighted (i.e., additional attributes author or authors added to some classes and association classes). As you can notice, the three main archaeological objects are all characterized by both spatial and temporal properties and, in accordance with Standard ISO 19107 [29] and 19108 [28], spatial and temporal properties can be described in a geometrical or topological way. While a geometric representation is suitable for defining quantitative positions in space or time, for instance a polygon shape or a temporal instant, a topological representation allows one to model qualitative relationships between objects. More specifically, topological structures can be built by connecting nodes (i.e., objects) through edges (i.e., relations). Inside a topological structure, some nodes can be *realized*,

namely they can have an associated quantitative characterization (i.e., an associated spatial geometry or time instant value), while others can be defined only qualitatively by means of (spatial or temporal) relations with other nodes (i.e., represented as dummy nodes connected to other nodes).

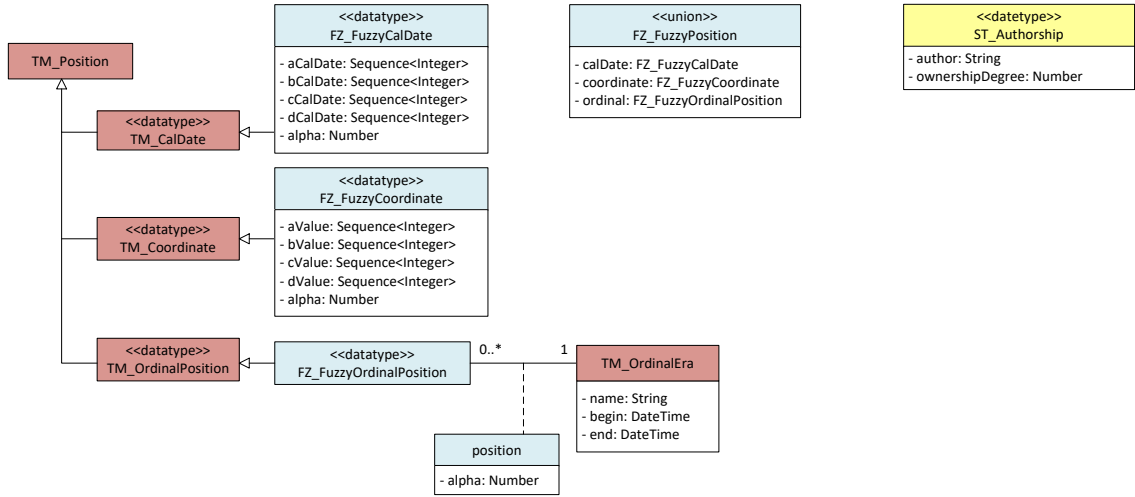
As far as temporal aspects are concerned, the *Star* model provides a fuzzy representation of temporal instants, as illustrated in Fig. 2. Indeed, in archaeology, temporal knowledge is usually characterized by a certain level of vagueness and dates are therefore expressed as periods of great confidence together with a safety additional interval. For instance, the construction date of a building can be expressed as: between 1830-1850 with more confidence, plus or minus 10 years of safety. In particular, given a fuzzy set F , the term *support* denotes the set of elements with a possibility greater than zero, while the term *core* denotes the set of elements with a possibility equal to 1. As discussed in [11], this kind of knowledge can be properly captured by a fuzzy model, where each fuzzy set is represented by means of a trapeze $T_k = \langle a_k, b_k, c_k, d_k \rangle [\alpha_k]$, where the characteristics 4-tuple is enriched with a degree of consistency α_k representing its height (see Sect. 4.1). Datatypes FZ_FuzzyCalDate and FZ_FuzzyCoordinate of Fig. 2 represent a calendar date or a coordinate using the trapeze representation of a fuzzy set, respectively.

An ST_ArchaeoUnit is temporally characterized by a sequence of *phases* of its evolution, such as, foundation, usage, disposal, renovation, re-use, and so on. This sequence is represented in Fig. 1 by the topological complex called ST_PhaseSequence which is obtained from the composition of a set of ST_Phases instances. The association between an ST_ArchaeoUnit and its related ST_PhaseSequence is represented by the association class ST_AuDating, which has an attribute of kind ST_Authorship with multiplicity greater than one. The structure of the datatype ST_Authorship is depicted in Fig. 2, it has two attributes useful for representing a provenance statement, namely an author name (or label) and a degree of ownership.

The dating of an ST_ArchaeoPart can be specified in three different ways. The first one is represented by the specification of an ST_ArchaeoPartDate, namely a topological node which can eventually be realized as an FZ_FuzzyInstant. The association between an ST_ArchaeoPart and its ST_ArchaeoPartDate is represented by the association class ST_ApDating, which specifies a set of authors, namely a set of domain experts who contributed to such dating process, together with a degree of ownership for each of them. Each ST_ArchaeoPartDate can be realized or not, namely the domain expert can use them to position the archaeological partition on time axis (*quantitative* specification), or to define some precedence relations between archaeological partitions (*qualitative* specification). Indeed, an ST_ArchaeoPartDate is a topological node which could be connected to other nodes through topological edges, called ST_TemporalRelations. The set of connected topological nodes and topological edges of this kind forms a topological complex (see the Standard ISO 19107) called ST_RelatedArchaeoParts, through which we can represent any given subset of the Allen's temporal relations.

The following example illustrates a case of topological complex representing the temporal relations existing between archaeological partitions.

Example 3.1. Let us consider four archaeological findings labeled as f_1 , f_2 , f_3 and f_4 which are coarsely dated as follows: f_1 , f_2 have been located in the 19th century by archaeologist a_1 , while f_3 has been dated 1850 by a_2 and f_4 has been dated 1820 by a_3 . Besides these geometrical values, the following temporal relations have been detected: f_1 before f_2 and f_3 by a_4 , while f_2 before f_3 and after f_4 by a_5 . This knowledge can be represented by the topological complex in Fig. 3. Dates associated to nodes f_3 and f_4 are realized as the years 1850 and 1820, respectively. Conversely, dates related to nodes f_1 and f_2 are not realized, but they are located between two dummy nodes representing the years 1800 and 1899. Notice that both nodes and arcs can have an additional label representing the archaeologists that define

Fig. 2. Data types used in the *Star* model.

such quantitative or qualitative temporal information. Given such topological relations, some automatic reasoning techniques can be applied in order to specialize some coarse-grained dates and realize the dummy nodes. For instance, as regards to this example, the geometric temporal value associated to f_2 can be restricted from 1800-1899 to 1820-1850, and consequently the dating of f_1 can be restricted from 1800-1899 to 1800-1820. When considering the provenance propagation, we can observe that the new dating of f_2 is determined by archaeologist a_2 who generally locates it in the 19th century, but also more specifically by a_5 who defines the relations with f_3 , f_4 and by a_2 and a_3 who give a precise date to f_3 and f_4 . Similar considerations can be done also for the new dating of f_1 . \square

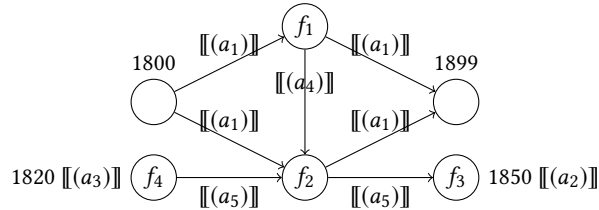


Fig. 3. Example of topological complex representing temporal relations between archaeological partition.

The second kind of dating for an *ST_ArchaeoPart* is given by its assignment to an *ST_Phase* of a related *ST_ArchaeoUnit*. This association is represented by an instance of the association class *ST_ApPhasing*, which also specifies the authors of such specific dating. As in the previous case, an *ST_ArchaeoPartDate* can be simply a qualitative node, or it can be realized as a *FZ_FuzzyInstant*. Clearly, a consistency constraint exists between the *ST_ArchaeoPartDate* and the *ST_ArchaeoPhases* associated to the same archaeological partition. Namely, the specified date has to be contained inside the period represented by the phase of the archaeological unit. Notice that an archaeological partition can be associated to multiple archaeological units, each one resulting from a different interpretation process performed in

different periods or by different domain experts. From this, it results that the same archaeological partition can be assigned to different phases, one for each associated archaeological unit.

The realization of both types of fuzzy nodes (i.e., `ST_PhaseNode` and `ST_ArchaeoPartDate`) may be given by an instance of the association class `ST_FuzzyRealization` which collects also its authors, namely the specification of who has assigned a quantitative temporal value to a topological node. The class `FZ_FuzzyInstant` has only one attribute, called `position`, of type `FZ_FuzzyPosition`. The structure of this data type is reported in Fig. 2 where we can notice that a temporal position may be given in three different ways: as a fuzzy calendar date, as a fuzzy position inside a temporal reference system or as a fuzzy ordinal position inside a temporal Era. This plurality of representation techniques is compliant with the Standard ISO 19107 and is very useful in the archaeological context, where temporal knowledge could come from different sources and in different formats.

The last possible kind of dating information associated to an `ST_ArchaeoPart` is a temporal precedence relation defined by means of a spatial stratigraphic relation. Each `ST_ArchaeoPart` can have some associated altimetric points that have been collected during the investigation process. From the principles of stratigraphy, we know that an object o , which is located below another one p , can be considered more ancient than p . Therefore, an `ST_AltimetricPoint` can be considered as a realization of a topological node and connected to other related objects through the edge called `ST_ArchaeoRelation`. The set of these nodes and edges forms the `ST_Stratigraphy` topological complex which allows to represent the information derivable from a Harris's matrix. Notice that two kinds of archaeological relations can be modelled, i.e. the `ST_Contemporary` and the `ST_Above`. The definition of `ST_ArchaeoRelations` is again characterized by the specification of the authors of such information.

We can observe that the *Star* model allows us to represent a wide range of information. More specifically, the availability of temporal topological complexes allows us to represent not only precise temporal values, but also relations of various kinds, from Allen's to stratigraphic ones. Furthermore, the origin of such temporal information could be manifold: from C14-like dating to the interpretation performed by archaeologists starting from stratigraphic relations, material analysis, and so on.

As regards to the `ST_InformationSource` class, since it can be used to describe an excavation process, or a bibliographic analysis, or other investigation processes performed in the current days and so clearly documented, we can safely assume that in this case the temporal information is known and could be represented through classical temporal data types. Therefore, we do not consider them in the remainder of the paper. Anyway, for further details about this concept and of its spatio-temporal characterization, the reader can refer to [11, 40].

In Fig. 1 we report for completeness also the geographical characterization of the objects. Since only the main classes of the hierarchy have been reported, the types of these objects is also the most generic one: `GM_GeometricPrimitive`. Indeed, each specialization of these classes can be characterized by a different geometric type, as described in [11, 40], for instance for an excavation it can be a polygon while for a survey or a bibliographic analysis it can be a point.

From *Star* to CIDOC-CRM

The presented *Star* model has been defined and developed in the context of a National project called SITAR regarding the collection of archaeological data about the city of Rome. It has been subsequently extended and adapted to collect also data about the city of Verona and some other small towns in Northern Italy [7]. Currently only the Verona database contains about 330 information sources (referring mainly to excavation of the last 50 years, but also to ancient documents about very old excavation campaigns) and 720 archaeological evidences that describe raw archaeological findings. In the last years, a pilot activity has been performed [13], in cooperation with the ARIADNE European infrastructure, for

defining a mapping from the *Star* model to the CIDOC-CRM model and its CIDOC-CRM_{archaeo} extension [37, 39]. The CIDOC Conceptual Reference Model (CRM) [27] is an ontology specifically developed for the information integration in the field of cultural heritage. Conversely, its CIDOC-CRM_{archaeo} extension [22] has been developed with the purpose to support the representation of the archaeological excavation process and it takes advantages from the concepts defined in another extension called CRM_{sci}.

With reference to the classes in Fig. 1, the main primitives describing temporal features are ST_Phase, ST_PhaseNode and ST_FuzzyInstant, which can be mapped into the CIDOC-CRM entities as briefly described below and illustrated in Fig. 4. For improving readability, Fig. 4 reports only the code of the CIDOC classes, while the corresponding mapped *Star* classes have been reported in light gray.

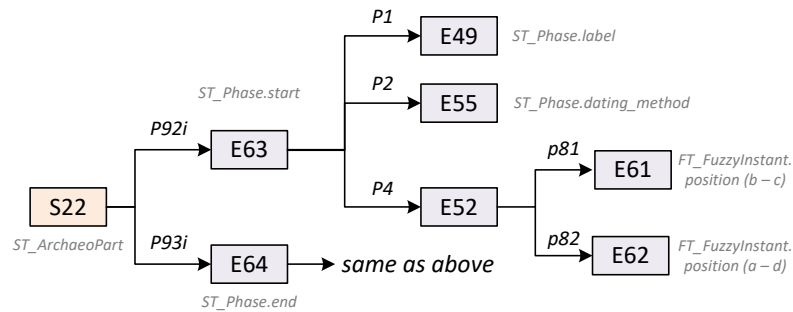


Fig. 4. Representation of the temporal properties of an ST_ArchaeologicalPartition through the CIDOC-CRM ontology.

First of all, the core concept ST_ArchaeoPart has been translated into the entity S22 Segment of Matter, namely a physical material in a relative stability of form (substance) within a specific space-time volume. This is the most appropriate definition for an archaeological partition, which can be a structure, in a more or less complete form, a movable element, a stratigraphic or a geological substrate, which has a particular importance in an historical perspective. The boundaries of each phase (i.e., ST_Phase) are represented as ST_PhaseNode instances and define the range of existence of a phase, they can be represented through an instance of E63 Beginning of Existence and E64 End of Existence which are connected to S22 through the properties P92i was brought into existence by and P93i was taken out of existence by, respectively.

Each ST_Phase is then characterized by a datingMethod property, which can be represented by an instance of E55 Type connected to an instance of E63 through the property P2 has type. Similarly, the label associated to a phase can be represented by an instance of E49 Time appellation and connected to E63 by the property P1 is identified by. Finally, the realization of a ST_PhaseNode as a ST_FuzzyInstant can be obtained by an instance of E52 TimeSpan connected through the property P4 has time-span. Indeed, each time span is characterized by two properties: P81 ongoing throughout and 82 at some time within, each one defining a connection towards E61 Time Primitive instances. As stated in the CIDOC-CRM documentation, since time spans may not have precisely known temporal extents, the CRM supports statements about their minimum and maximum temporal extents. More specifically, property P81 defines the minimum temporal extent (i.e. its inner boundary), which can be considered the core of a fuzzy instant, while property 82 can be used to define the maximum temporal extent, namely the support of a fuzzy instant.

As regards to the representation of the authorship and provenance of the dating information, a convenient mapping can be defined between some classes of the *Star* model and the CRM_{inf} extension [49] which has been specifically

developed to support argumentation. In particular, each relation in Fig. 1 specifying the authors of a given temporal assignment can be represented with an instance of I10 Provenance Statement which comprises statements about the provenance of an instance of E73 Information Object. An E73 Information Object is connected to a generic E1 CRM Entity through the property P129 is about.

From these considerations it follows that the *Star* model complies with both the ISO Standards and the CIDOC-CRM Standard. Therefore, the application of the reasoning technique proposed in this paper can be adapted to other models in a straightforward manner by simply defining translation rules similar to the ones presented in Sect. 4.2.

4 PROPOSED SOLUTION

This section illustrates how new temporal knowledge can be extracted and how provenance information can be tracked in a *Star* model. More specifically, we start by introducing the concept of Provenance-Aware Fuzzy Temporal Constraint Network (PA-FTCN), a fuzzy extension of a temporal constraint network in which the constraints contain, not only fuzzy temporal information, but also provenance statements (Sect. 4.1). Then, we illustrate how a *Star* model can be translated into a PA-FTCN in order to apply some reasoning on it (Sect. 4.2). Finally, we propose an efficient MapReduce implementation of the path-consistency algorithm that allows to produce new temporal knowledge and provenance information (Sect. 4.4).

4.1 Provenance-Aware Fuzzy Temporal Constraint Network (PA-FTCN)

A Temporal Constraint Network (TCN) [19] is a formalism for representing temporal knowledge based on *metric* constraints among pairs of time-points. In literature, TCNs have been applied in several areas in order to derive new temporal knowledge or check the satisfaction of temporal constraints, such as scheduling, planning, temporal databases and common sense reasoning [45]. A TCN can be represented by a directed graph, where each node is associated with a variable and each arc corresponds to the constraint between the connected variables.

Definition 4.1 (Temporal Constraining Network). A Temporal Constraint Network (TCN) \mathcal{N} is a tuple $\langle \mathcal{X}, \mathcal{K} \rangle$, where \mathcal{X} is a set of variables representing time points and \mathcal{K} is a set of binary constraints on those variables. Variables take values on \mathbb{R} , while each constraint $k_{ij} \in \mathcal{K}$ restricts the duration of the time elapsed between two temporal variables $x_i, x_j \in \mathcal{X}$.

As discussed in the previous section, temporal knowledge in archaeology is characterized by a certain level of vagueness which can be properly captured by a fuzzy representation. For this reason, we consider an extension of TCN, which is called Fuzzy Temporal Constraint Network (FTCN) [50]. A fuzzy temporal constraint network (FTCN) is a generalization of TCN where a *degree of possibility* is associated with each possible value of a temporal constraint. In other words, given a pair of time points, a constraint between them represents a possibility distribution over temporal distances.

Definition 4.2 (fuzzy temporal constraint). Given two temporal variables x_i and x_j , a *fuzzy temporal constraint* k_{ij} between them is represented as a *possibility distribution function* $\pi_{ij} : \mathbb{R} \rightarrow [0, 1]$ that constrains the possible values for the temporal distance $x_j - x_i$. □

In other words, $\pi_{ij}(d)$ is the possibility degree for the distance $x_j - x_i$ to take the value d under the constraint k_{ij} . As done in our previous works [9, 11, 38], this paper considers only trapezoidal distributions, which are on one side expressive enough in practical contexts, on the other side computationally less expensive during the reasoning

process. They can be represented as a 4-tuple $T = \langle a, b, c, d \rangle$, where the intervals $[b, c]$ and $[a, d]$ represent the core and the support of the fuzzy set, respectively. Each tuple representation $\langle a, b, c, d \rangle$ of a trapeze T is enriched with an additional value α , called *degree of consistency*, which denotes the height of the trapeze and allows the representation of non-normalized distributions. More specifically, given a trapeze $T = \langle a, b, c, d \rangle[\alpha]$, the support of π_T is defined as $\text{supp}(\pi_T) = \{x : \pi_T(x) > 0\} = [a, d]$, while the core as $\text{core}(\pi_T) = \{x : \pi_T(x) = \alpha\} = [b, c]$. The components of a trapeze T takes values as follows: $a, b \in \mathbb{R} \cup \{-\infty\}$ and $c, d \in \mathbb{R} \cup \{+\infty\}$. As already discussed in our previous work, even if α is commonly set equal to 1, its specification becomes necessary during the reasoning, since the conjunction of some constraints can produce trapezes with a height less than one.

Definition 4.3 (trapezoid possibility distribution function). The *possibility distribution function* of a generic trapeze $T = \langle a, b, c, d \rangle[\alpha]$ can be written as:

$$\pi_T(x) = \begin{cases} 0 & \text{if } x < a \vee x > d \\ \alpha \cdot ((x - a)/(b - a)) & \text{if } a \leq x < b \\ \alpha \cdot ((d - x)/(d - c)) & \text{if } c < x \leq d \\ \alpha & \text{otherwise} \end{cases}$$

□

Moreover, since we consider only *well-formed* trapezes, we always have that $a \leq b \leq c \leq d$.

In case of a PA-FTCN, we enrich each temporal constraint with some provenance (authorship) information represented as a set of provenance statements $\Omega = \{(o_1, d_1), \dots, (o_n, d_n)\}$. Each *provenance statement* $\omega_i = (o_i, d_i) \in \mathcal{A} \times [0, 1]$ is composed by a label identifying the data owner (or author) $o_i \in \mathcal{A}$, where \mathcal{A} is a set of labels representing known authors, and a number $d_i \in [0, 1]$ representing the degree of ownership.

Definition 4.4 (provenance-aware fuzzy trapezoidal constraint). Given two variables x_i and x_j , a provenance-aware fuzzy trapezoidal temporal constraint $C_{ij} = \{T_1, \dots, T_m\}$ is a disjunction of trapezoidal distributions $\pi_{T_k} : \mathbb{R} \rightarrow [0, 1]$ for $k = 1, \dots, m$, each one denoted by a *well-formed* trapeze $T_k = \langle a, b, c, d \rangle[\alpha][\Omega]$, where the characteristic tuple is enriched with a set of provenance statements $\Omega = \{(o_1, d_1), \dots, (o_n, d_n)\}$. □

To make the notation more readable, in the following we will omit the specification of the degree of consistency α , when it is not strictly necessary.

As discussed in the previous sections, in archaeology it is quite common to know only qualitative temporal information between objects, namely temporal precedence relations. This has been captured by the extensive usage of topological complexes in the *Star* model. From the point of view of a PA-FTCN, we need to translate a qualitative temporal information into a quantitative one; this step can be performed as follows.

Definition 4.5 (provenance-aware fuzzy qualitative constraint). Given two variables x_i and x_j , a provenance-aware fuzzy qualitative constraint C_{ij} between them is defined as: $C_{ij} = \{\text{before}[\alpha_1][\Omega_1], \text{equal}[\alpha_2][\Omega_2], \text{after}[\alpha_3][\Omega_3]\}$ where *before*, *equal* and *after* are the possibile qualitative relations between two time points, $\alpha_k \in [0, 1]$ is the degree of confidence of such relation, and Ω_k is the set of provenance statements identifying the authorship of such relations.

A provenance-aware fuzzy qualitative constraint $C_{ij} = \{\text{before}[\alpha_1][\Omega_1], \text{equal}[\alpha_2][\Omega_2], \text{after}[\alpha_3][\Omega_3]\}$ can be translated into the corresponding quantitative constraint as follows:

$$\begin{cases} \text{if } \alpha_1 > 0 & \text{then } \langle 0, 0, +\infty, +\infty \rangle[\alpha_1][\Omega_1] \\ \text{if } \alpha_2 > 0 & \text{then } \langle 0, 0, 0, 0 \rangle[\alpha_2][\Omega_2] \\ \text{if } \alpha_3 > 0 & \text{then } \langle -\infty, -\infty, 0, 0 \rangle[\alpha_3][\Omega_3] \end{cases} \quad (1)$$

Example 4.6. Let us consider a simple PA-FTCN \mathcal{N} composed of three temporal variables $\mathcal{X} = \{s, e, f\}$ representing the following temporal knowledge: (i) the lifespan of an ancient building b , which has been determined by archaeologist a_1 in approximately 100-200 years, plus or minus approximately 10 years (the start of the lifespan is represented by node s and its end by the node e), and (ii) the dating of a finding f which has been discovered in the same area by archaeologist a_2 , who has also determined that f has been added to b after 5-10 years from its foundation with an approximation of ± 1 year, and certainly before its destruction.

The resulting set of constraints becomes $\mathcal{K} = \{s \xrightarrow{\langle 90, 100, 200, 210 \rangle[a_1]} e, s \xrightarrow{\langle 4, 5, 10, 11 \rangle[a_2]} f, f \xrightarrow{\langle 0, 0, \infty, \infty \rangle[a_2]} e\}$, while the network can be graphically represented as in Fig. 5.

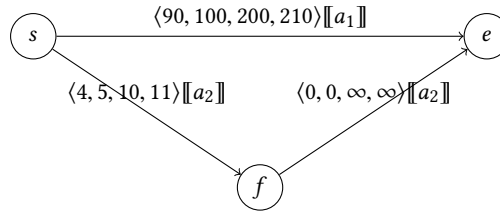


Fig. 5. Graphical representation of the FT CN described in Example 4.6.

Let us consider the edge connecting the nodes s and e on which the trapeze $\langle 90, 100, 200, 210 \rangle[a_1]$ is defined. In this case the possibility distribution function π_{se} says that the temporal distance between the nodes s and e will take a value x with the following possibilities:

$$\pi_{se}(x) = \begin{cases} 0 & \text{if } x < 90 \vee x > 210 \\ \left(\frac{x - 90}{100 - 90} \right) & \text{if } 90 \leq x < 100 \\ \left(\frac{210 - x}{210 - 200} \right) & \text{if } 200 < x \leq 210 \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

In other words, while such distance cannot be less than 90 or greater than 210, its possibility to take a value between 100 and 200 is 1 (or better α), and to take a value between 90 and 100 or between 200 and 210 is greater than 0 but less than α .

In the following Sect. 4.2 we will discuss how a *Star* model can be translated into a PA-FTCN, while in Sect. 4.3 we will extend the operations on trapezoidal distributions, that will be used during the derivation process, for taking into account also the presence of provenance statements. Finally, a MapReduce version of the path consistency algorithm is proposed in Sect. 4.4 with the aim to effectively apply the derivation process.

4.2 Translate a *Star* into a PA-FTCN

This section introduces a set of rules for translating a *Star* model into a PA-FTCN. As mentioned in the previous sections, archaeological temporal information can be expressed using different coordinate reference systems. Therefore, in order to ease the required comparisons and operations, it is necessary to preliminary establish a common reference system and translate all dates into real numbers. The origin of such reference system will become the start node of the PA-FTCN. In the remainder of this section, we assume that a common temporal reference system has been chosen and all the dates have been properly expressed w.r.t. to it.

The construction of the PA-FTCN starts from the topological complexes described in the *Star* model at hand. More specifically, in a *Star* model three kinds of topological complex can be recognized: ST_PhaseSequence, ST_RelatedArchaeoParts and ST_Stratigraphy. In particular, the first two extend the TM_TopologicalComplex class and are used for defining the dating of an ST_ArchaeoUnit as a sequence of phases, while the second one establishes temporal relationships among ST_ArchaeoParts. Both kinds of topological complex have nodes and edges represented by specializations of FZ_FuzzyNode and FZ_FuzzyEdge, respectively. Finally, each fuzzy node may have a realization represented by an instance of FZ_FuzzyInstant. Conversely, the third one extends the TP_TopologicalComplex class and is used for adding additional constraints between the dating of archaeological partitions.

Before presenting the translation of these constructs into components of a PA-FTCN, we explain how a provenance statement can be generated starting from instances of the ST_Authorship class.

RULE 1 (ST_Authorship). *Given an instance of ST_Authorship containing the attributes author and ownershipDegree, it can be translated into a provenance statement $\omega = (o, d)$ where o is the value of author, namely a label that identifies the data owner, while d is the value of ownershipDegree and quantifies the degree of ownership associated to the author o . A list L of ST_Authorship instances can be translated into a sequence of provenance statements $\Omega = \{\omega_1, \dots, \omega_{|L|}\}$, one for each element in L .*

Given this rule, we start with the topological complex ST_PhaseSequence, which is used for dating an ST_ArchaeologicalUnit through the relation ST_AuDating. As a topological complex, it is composed of a set of nodes (i.e., ST_PhaseNodes) connected through edges (i.e., ST_Phases).

RULE 2 (NOT-REALIZED ST_PhaseNode). *Each not-realized ST_PhaseNode included into an ST_Phase p is translated into a node x of the PA-FTCN and connected by an incoming arc with the network start node s , as illustrated in Fig. 6a. In this case, the arc label becomes $\langle 0, 0, +\infty, +\infty \rangle [1][\Omega]$, since we only know that x occurs after the start node s . As regards to the provenance statements Ω , these are built starting from the ST_Authorship instances contained in related ST_AuDating relation r , as explained in Rule 1.*

RULE 3 (ST_Phase). *Each ST_Phase p from a ST_PhaseNode n_x to a ST_PhaseNode n_y is translated into an edge from x to y labelled with the constraint $\langle 0, 0, +\infty, +\infty \rangle [1][\Omega]$, as illustrated in Fig. 6a, where x and y are the representation of n_x and n_y in the PA-FTCN, respectively, while Ω is built starting from the ST_Authorship instances contained in related ST_AuDating r , as explained in Rule 1.*

RULE 4 (ST_Phase ORDERING). *Given an ST_Phase p_1 for which the relation previous has been specified towards another phase p_2 , this relation is translated as an arc from node $n_{p_{1e}}$ to node $n_{p_{2s}}$, where nodes $n_{p_{1e}}$ and $n_{p_{2s}}$ denote the end of phase p_1 and the start of phase p_2 , respectively. The arc is labelled with the constraint $\langle 0, 0, +\infty, +\infty \rangle [1][\Omega]$, where Ω is built starting from the ST_Authorship instances contained in related ST_AuDating r , as explained in Rule 1.*

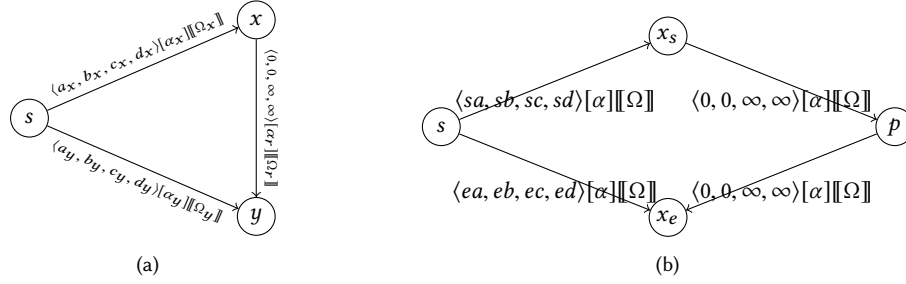


Fig. 6. (a) Translation of a FZ_FuzzyNode. (b) Translation of a FZ_FuzzyOrdinalPosition p inside a FZ_FuzzyOrdinalEra.

In case a phase node has a realization as a FZ_FuzzyInstant, its translation takes a different label edge which depends on the specific kind of fuzzy position. Notice that since it is necessary that all the temporal values in the network are defined as multiple of a common chosen granularity, we eventually need to preliminary transform all dates in the model w.r.t. a minimum common granularity.

RULE 5 (MINIMUM GRANULARITY). Let g the minimum common granularity in the considered model (i.e., day, month or year). Any fuzzy date $x = \langle a, b, c, d \rangle [\alpha]$ whose components have a granularity smaller than g , will be transformed into a date with granularity g in the following way.

- If g is day and the granularity of x is month: components a and b become the first day of the given month, while c and d become the last day of the given month.
- If g is day and the granularity of x is year: a and b become the first day of the first month of the given year, while c and d become the last day of the last month of the given year.
- If g is month and the granularity of x is year: a and b become the first month of the given year, c and d become the last month of the given year.
- All the other combinations do not require any transformation.

This transformation is useful only for reasoning purposes, but it does not affect the granularity of the represented knowledge which will be finally represented using their original granularity. Given such rule, in the following all the other transformations assume that temporal values are represented with the same granularity.

RULE 6 (ST_PhaseNode REALIZED AS FZ_CalDate). Each ST_PhaseNode p realized as a FZ_CalDate $x = \langle aCalDate, bCalDate, cCalDate, dCalDate \rangle [\alpha]$ through the relation ST_FuzzyRealization, is translated as illustrated in Fig. 6a and similarly to what has been done in Rule 2. However, in this case the tuple $\langle a, b, c, d \rangle$ where $a, b, c, d \in \mathbb{R}$ is the representation of aCalDate, bCalDate, cCalDate and dCalDate in the chosen coordinate reference system, respectively. Moreover, the authorship statement Ω is obtained starting from the content of the authors attribute of relation ST_FuzzyRealization.

RULE 7 (ST_PhaseNode REALIZED AS TM_FuzzyCoordinate). Each ST_PhaseNode p realized as a TM_FuzzyCoordinate $x = \langle aValue, bValue, cValue, dValue \rangle [\alpha]$ is translated as in Fig. 6a where the label $\langle a, b, c, d \rangle [\alpha]$ is obtained by transforming each component position w.r.t. the coordinate reference system chosen for the PA-FTCN, while Ω is obtained from the translation of the authors attribute contained in the relation ST_FuzzyRealization. If the fuzzy coordinate x and the network are already expressed in the same reference system, no transformation is required and simply $a = aValue$, $b = bValue$, $c = cValue$ and $d = dValue$.

The last kind of position is represented by a fuzzy ordinal position inside an ordinal era. In this case, the representation is quite different, because we need to represent both the era and the ordinal position inside it.

RULE 8 (ST_PhaseNode REALIZED AS FZ_FuzzyOrdinalPosition). *Each FZ_FuzzyOrdinalEra x spanning a period of time (start, end), where start = $\langle \text{saCalDate}, \text{sbCalDate}, \text{scCalDate}, \text{sdCalDate} \rangle[\alpha]$ and end = $\langle \text{eaCalDate}, \text{ebCalDate}, \text{ecCalDate}, \text{edCalDate} \rangle[\alpha]$, is translated into two nodes x_s and x_e , connected to the start node s as explained in Rule 6. Finally, each TM_FuzzyOrdinalPosition inside x is represented by a node p connected to the node x_s by an incoming edge labeled $\langle 0, 0, \infty, \infty \rangle[\alpha][\Omega]$, and to the node x_e by an outgoing edge labeled $\langle 0, 0, \infty, \infty \rangle[\alpha][\Omega]$, representing together the inclusion of the position inside the era. The authorship statement Ω is obtained from the translation of the authors attribute contained in the relation ST_FuzzyRealization. The overall presentation of an ordinal position is depicted in Fig. 6b.*

An ST_ArchaeoPart can be dated in two ways: the first one through the definition of an ST_ArchaeoPartDate instance and the other one through the assignment to an ST_Phase. An ST_ArchaeoPartDate is again a topological node which can have or not a realization as a FZ_FuzzyInstant.

RULE 9 (ST_ArchaeoPartDate). *Each ST_ArchaeoPartDate defining the dating of an ST_ArchaeoPart is translated similarly to what described in Rule 2 or Rule 6-8, depending on the fact that the node is realized or not. The only difference regards the construction of the provenance statement Ω that in case of not-realized nodes will come from the ST_Authorship instances specified in the relation ST_ApDating.*

If the archaeological partition is also associated to a phase, we represent the containment constraint between its ST_ArchaeoPartDate and the connected ST_Phase, as explained by the following rule.

RULE 10 (ST_ApDating AND ST_ApPhasing CONTAINMENT). *For each ST_ArchaeoPart which has associated both an ST_ArchaeoPartDate d and an ST_Phase p , two additional edges are added to the PA-FTCN: the first one from node n_{p_s} to node n_d and the second one from node n_d to node n_{p_e} , where n_{p_s} and n_{p_e} are the nodes representing the phase boundaries, while n_d is the node representing the date d , respectively. Both edges are labeled as $\langle 0, 0, \infty, \infty \rangle[1][\Omega]$ for representing the temporal precedence relation, while Ω is built starting from the ST_Authorship instances defined in the relation ST_ApPhasing.*

Some relations can be represented between the dating of different partitions, leading to the definition of another topological complex called ST_RelatedArchaeoParts. Namely, topological nodes of type ST_ArchaeoPartDate can be connected through topological edges of type ST_TemporalRelation.

RULE 11 (ST_TemporalRelation). *Each ST_TemporalRelation can be translated similarly to what has been done in Rule 3 for the ST_Phase, but in this case the provenance statement Ω is obtained starting from its ST_Authorship attributes specified in ST_TemporalRelation.*

Finally, as regards to the spatial topological complex, in this case the translation is slightly different, since the relation specified through a TP_Edge between two TP_DirectNode is translated as an edge between the dating of the involved ST_ArchaeoPart, as specified by the following rule.

RULE 12 (ST_ArchaeoRelation). *Each ST_ArchaeoRelation from an ST_AltimePoint x_{ap_1} belonging to an archaeological partition ap_1 and an ST_AltimePoint x_{ap_2} belonging to an archaeological partition ap_2 , is translated into an edge between the nodes d_{ap_1} and d_{ap_2} , representing the dating of ap_1 and ap_2 , respectively. The edge will be labelled as $\langle 0, 0, 0, 0 \rangle[\alpha]$ or $\langle 0, 0, \infty, \infty \rangle[\alpha]$ depending on the fact that the relation is ST_Contemporary or ST_Above.*

4.3 Constraint Propagation and Knowledge Discovery

Given a PA-FTCN containing several constraints between time variables, the common way to extract new knowledge from it is to apply the *path consistency algorithm* [34]. The idea behind this algorithm is very simple, given three time variables x_i , x_j and x_k , such that there exists a constraint C_{ij} between x_i and x_j , a constraint C_{ik} between x_i and x_k and a constraint C_{kj} between x_k and x_j that completes the triangle, a new constraint can be derived between x_i and x_j by properly combining them. A complete picture of the path-consistency algorithm is presented in Alg. 1, while the main operation performed on each triangle for the constraint propagation is explained in the following definition.

Definition 4.7 (path-consistency algorithm). Given three variables x_i , x_k and x_j of a PA-FTCN \mathcal{N} , a new constraint between x_i and x_j can be induced from pre-existing constraints by the path consistency algorithm as follows:

$$C_{ij} \otimes (C_{ik} \circ C_{kj}) \quad (3)$$

where C_{ij} is the constraint existing between x_i and x_j , $C_{ik} \circ C_{kj}$ is the composition (addition between fuzzy sets) of two constraints and $C_{ij} \otimes C$ is the conjunction (intersection between fuzzy sets). \square

In order to determine the result of the operation in Eq. 3, it is necessary to define the semantics of the required operations. More specifically, it is necessary to specialize some operations on fuzzy sets to operations on trapezes with provenance statement. In particular, the specialization of the inversion (T_k^{-1}), composition ($T_1 \circ T_2$), conjunction ($T_1 \otimes T_2$) and disjunction ($T_1 \oplus T_2$) contained in [10] have to be specialized in order to take care also of the provenance information. In particular, our aim is from one side to propagate provenance labels, but also to provide a degree of ownership to each author, thus, we need to define the concept of similarity between two trapezes.

Definition 4.8 (trapeze similarity). Given two trapezes $T_1 = \langle a_1, b_1, c_1, d_1 \rangle [\alpha_1] [\llbracket \Omega_1 \rrbracket]$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle [\alpha_2] [\llbracket \Omega_2 \rrbracket]$, the degree of similarity $sim(T_1, T_2) \in [0, 1]$ between them is defined as:

$$sim(T_1, T_2) = \frac{area(T_1 \cap T_2)}{area(T_1 \cup T_2)} \quad (4)$$

In other words the similarity is maximum (equal to 1) when the two trapezes coincide, while it is minimum (equal to 0) when the two trapezes are completely disjoint (the intersection is empty), otherwise it is proportional to the degree of overlap between them. Notice that there can be two cases where the degree of similarity is equal to 0: i) when the intersection is empty, and ii) when the union of the two trapezes generates an infinite trapeze. This second case is possible, for instance, when at least one of the trapezes represents a qualitative precedence constraint. In order to distinguish these two situations, we use the symbol 0 when the intersection is empty (no similarity at all), and the symbol \perp when the union is infinite (very low similarity).

Example 4.9. Let us consider two trapezes $T_1 = \langle 190, 200, 250, 260 \rangle [1] [\llbracket \Omega_1 \rrbracket]$ and $T_2 = \langle 210, 220, 280, 290 \rangle [1] [\llbracket \Omega_2 \rrbracket]$, the similarity between them is given by

$$sim(T_1, T_2) = \frac{area(T_1 \cap T_2)}{area(T_1 \cup T_2)} = \frac{area(\langle 210, 220, 250, 260 \rangle)}{area(\langle 190, 210, 280, 290 \rangle)} = \frac{35}{85} = 0.412 \quad (5)$$

In this case the computed value indicates that the similarity between the two trapezes T_1 and T_2 is about 41%.

During the various derivations performed by the path consistency algorithm, the degree of ownership assigned to each author for a constraint C_{ij} is computed on the basis of the starting degree of ownership and the similarity between the original constraint and the new obtained one.

Given such observation about the degree of similarity, we now discuss the extension of the four basic operations: inversion, composition, conjunction and disjunction which are used during the path consistency algorithm.

Definition 4.10 (inversion). Given a constraint $C_{ij} = \{T_1, \dots, T_m\}$ between two variables x_i and x_j , the constraint C_{ij}^{-1} represents the equivalent constraint holding between x_j and x_i . Such constraint can be obtained by making the inversion of each constituent trapezes $T_k = \langle a_k, b_k, c_k, d_k \rangle [\alpha_k] [\Omega]$ contained in C_{ij} , as follows:

$$T_k^{-1} = \langle -d_k, -c_k, -b_k, -a_k \rangle [\alpha_k] [\Omega] \quad (6)$$

□

Notice that the provenance statements of the original constraint are not affected by the inversion, the rationale is that this operation does not introduce changes in the original constraint, it is only useful for obtaining a constraint in the opposite direction w.r.t. the original one.

We have a different situation in case of a composition between two constraints C_1 and C_2 : the composition is essentially the addition or union of the two original constraints.

Definition 4.11 (composition \circ). Given two constraints C_1 and C_2 , the composition of two generic trapezes $T_1 = \langle a_1, b_1, c_1, d_1 \rangle [\alpha_1] [\Omega_1] \in C_1$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle [\alpha_2] [\Omega_2] \in C_2$ is defined as:

$$T_1 \circ T_2 = \langle a_1 + a_2, b'_1 + b_2, c'_1 + c_2, d_1 + d_2 \rangle [\min\{\alpha_1, \alpha_2\}] [\Omega_1 \cup \Omega_2] \quad (7)$$

where, assuming that $\alpha_1 \geq \alpha_2$:

$$\begin{aligned} b'_1 &= a_1 + (\alpha_2/\alpha_1)(b_1 - a_1) \\ c'_1 &= d_1 - (\alpha_2/\alpha_1)(d_1 - c_1) \\ \llbracket \Omega_1 \cup \Omega_2 \rrbracket &= \{ \omega_i = (o_i, d_i) \mid (\omega_i \in \Omega_1 \wedge \omega_i \notin \Omega_2) \vee \\ &\quad (\omega_i \in \Omega_2 \wedge \omega_i \notin \Omega_1) \vee \\ &\quad (\exists (o_i, d_{i_1}) \in \Omega_1 \wedge \exists (o_i, d_{i_2}) \in \Omega_2 \wedge d_i = \max(d_{i_1}, d_{i_2})) \} \end{aligned} \quad (8)$$

□

The composition of two constraints produces a trapeze with a bigger extension in terms of its main base, namely it indicates that the support becomes bigger. Indeed, in this case the provenance information becomes the union of the input ones. More specifically, all original authors will be considered with a degree of ownership which is the maximum degree for that author in the original trapezes.

The conjunction of two generic fuzzy possibility distribution functions π_1 and π_2 is defined as: $\forall d \in \mathbb{R} (\pi_1 \otimes \pi_2)(d) = \min\{\pi_1, \pi_2\}$. When applied to two generic trapezes T_1 and T_2 , this operation essentially computes an intersection between them. Unfortunately, the intersection of two trapezes does not always produce a trapeze, see for instance the situations reported in Fig. 7.a and Fig. 7.c. Therefore, some sort of approximation is necessary in order to always obtain a valid trapeze as a result of a conjunction operation, as illustrated in Fig. 7.b and Fig. 7.d.

The rationale behind the defined conjunction approximation \otimes is that: (i) the core has to correspond to the core produced by the exact conjunction (i.e., \otimes_e): $core(\pi_T) = core(\pi_{T_1} \otimes_e \pi_{T_2})$ and $\alpha(\pi_T) = \alpha(\pi_{T_1} \otimes_e \pi_{T_2})$, while the support should be slightly modified in order to produce a valid trapeze but ensuring that $supp(\pi_T) \subseteq supp(\pi_{T_1} \otimes_e \pi_{T_2})$. This operation is formalized as follows.

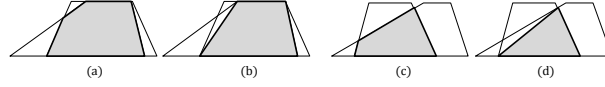


Fig. 7. Two examples of approximated conjunction operation \otimes between trapezoids: in (a) and (c) the result of the classical conjunction operation between fuzzy possibility distribution functions, and in (b) and (d) the corresponding approximation which produces a trapeze.

Definition 4.12 (conjunction \otimes). Given two constraints C_1 and C_2 , the conjunction between two trapezes $T_1 = \langle a_1, b_1, c_1, d_1 \rangle[\alpha_1][\Omega_1] \in C_1$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle[\alpha_2][\Omega_2] \in C_2$ can be computed starting from their intersection:

$$T_1 \otimes T_2 = (\max\{a_1, a_2\}, b', c', \min\{d_1, d_2\})[\min\{\alpha'_b, \alpha'_c\}][\Omega_1 \cup \Omega_2] \quad (9)$$

where b', c', α'_b and α'_c are computed as in Tab. 1 on the basis of the intersection points between the two trapezes, while:

$$\begin{aligned} \llbracket \Omega_1 \cup \Omega_2 \rrbracket = \{ \omega_i = (o_i, d'_i) \mid & ((o_i, d_i) \in \Omega_1 \wedge o_i \notin \mathcal{A}_2 \wedge d'_i = d_i \cdot \text{sim}(T_1, T)) \vee \\ & ((o_i, d_i) \in \Omega_2 \wedge o_i \notin \mathcal{A}_1 \wedge d'_i = d_i \cdot \text{sim}(T_2, T)) \vee \\ & ((o_i, d'_i) \in \Omega_1 \wedge (o_i, d'_i) \in \Omega_2 \wedge d'_i = \max(d_i^1 \cdot \text{sim}(T_1, T), d_i^2 \cdot \text{sim}(T_2, T))) \} \end{aligned} \quad (10)$$

In order to compute the intersection points b' and c' , we need to consider: (i) the possible ways the side $[a_1, b_1]$ of T_1 can intersect T_2 (or the side $[a_2, b_2]$ of T_2 can intersect T_1), producing the intersection point b' , and (ii) the possible ways the side $[c_1, d_1]$ of T_1 can intersect T_2 (or the side $[c_2, d_2]$ of T_2 can intersect T_1) producing the new intersection point c' . The formulas in Tab. 1 are obtained from the equation of the two possible intersecting sides (one for each trapeze) and by combining them to obtain the intersection point. Given the intersection point, its x coordinate corresponds to the value of b' (or c'), while the y coordinate is the value of α' .

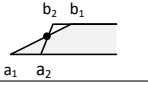
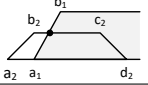
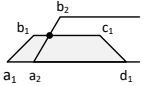
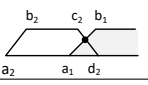
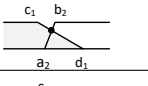
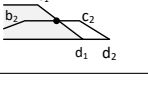
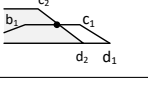
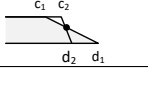
From Def. 4.12 it is clear that the height of the new trapeze produced by the conjunction operation can become less than one, hence the specification of the degree of consistency α becomes strictly necessary. As regards to the provenance, in this case we keep track of all authors who contribute to the trapeze conjunction, but we update the degree of ownership on the basis of the similarity between the original information and the obtained one. Notice that, when the same author o_i is present in both the two trapezes T_1 and T_2 , we will compute its degree of ownership as $d_i = \max(\text{sim}(T_1, T), \text{sim}(T_2, T))$. Moreover, $\max(\perp, \text{sim}(T_i, T)) = \text{sim}(T_i, T)$. In other words the conjunction operator produces a more stringent constraint, i.e. the resulting trapeze is narrower; thus, the degree of provenance (i.e. the author contribution in the result) is proportional to the similarity between the original trapeze and the resulting one.

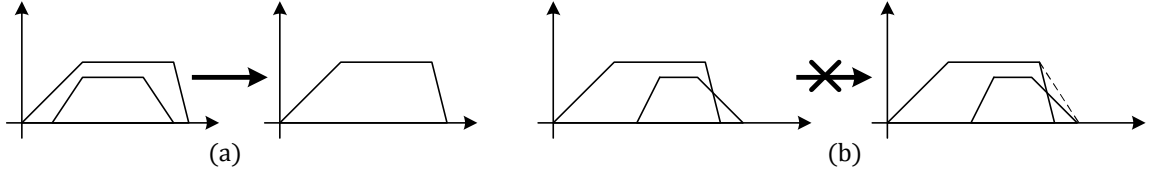
Finally, the last operation we will consider is the disjunction between two constraints. It is not required by the path consistency algorithm, but it can be applied for eliminating redundant trapezes introduced during the constraint propagation. Therefore, it is an operation useful for compressing available information. In other words, the disjunction of two general fuzzy distribution functions π_1 and π_2 is defined as $\forall d \in \mathbb{R} : \pi_1 \oplus \pi_2(d) = \max\{\pi_1(d), \pi_2(d)\}$. However, like conjunction, disjunction is not closed in the algebra of trapezes. Therefore, the idea is to compute also in this case a tentative trapeze and then check whether it corresponds to the disjunction of the involved constraints (i.e., correspond of one of the two involved trapezes), otherwise the constraints will be maintained separated.

Definition 4.13 (disjunction \oplus). Given two constraints C_1 and C_2 , the disjunction between two trapezes $T_1 = \langle a_1, b_1, c_1, d_1 \rangle[\alpha_1][\Omega_1] \in C_1$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle[\alpha_2][\Omega_2] \in C_2$ is defined as follows:

$$T_1 \oplus T_2 = \langle a, b, c, d \rangle[\max\{\alpha_1, \alpha_2\}][\Omega_1 \cup \Omega_2] \quad (11)$$

Table 1. Computation of the intersection point between two trapezes.

Condition	b'	c'	α'_b/α'_c	
$\text{int}([a_1, b_1], [a_2, b_2])$	$\frac{\alpha_1 a_1 (b_2 - a_2) - \alpha_2 a_2 (b_1 - a_1)}{\alpha_1 (b_2 - a_2) - \alpha_2 (b_1 - a_1)}$		$\alpha_1 \frac{(b' - a_1)}{(b_1 - a_1)}$	
$\text{int}([a_1, b_1], [b_2, c_2])$ $\wedge \alpha_1 \geq \alpha_2$	$\frac{\alpha_2}{\alpha_1} (b_1 - a_1) + a_1$		α_2	
$\text{int}([a_1, b_1], [b_2, c_2])$ $\wedge \alpha_1 < \alpha_2$	$\frac{\alpha_1}{\alpha_2} (b_2 - a_2) + a_2$		α_1	
$\text{int}([a_1, b_1], [c_2, d_2])$	$\frac{\alpha_1 a_1 (d_2 - c_2) + \alpha_2 d_2 (b_1 - a_1)}{\alpha_1 (d_2 - c_2) + \alpha_2 (b_1 - a_1)}$		$\alpha_1 \frac{(b' - a_1)}{(b_1 - a_1)}$	
$\text{int}([c_1, d_1], [a_2, b_2])$		$\frac{\alpha_1 d_1 (b_2 - a_2) - \alpha_2 a_2 (d_1 - c_1)}{\alpha_1 (b_2 - a_2) - \alpha_2 (d_1 - c_1)}$	$-\alpha_1 \frac{(c' - d_1)}{(d_1 - c_1)}$	
$\text{int}([c_1, d_1], [b_2, c_2])$ $\wedge \alpha_1 \geq \alpha_2$		$-\frac{\alpha_2}{\alpha_1} (d_1 - c_1) + d_1$	α_2	
$\text{int}([c_1, d_1], [b_2, c_2])$ $\wedge \alpha_1 < \alpha_2$		$-\frac{\alpha_1}{\alpha_2} (d_2 - c_2) + d_2$	α_1	
$\text{int}([c_1, d_1], [c_2, d_2])$		$\frac{-\alpha_1 d_1 (d_2 - c_2) + \alpha_2 d_2 (d_1 - c_1)}{-\alpha_1 (d_2 - c_2) + \alpha_2 (d_1 - c_1)}$	$-\alpha_1 \frac{(c' - d_1)}{(d_1 - c_1)}$	

Fig. 8. Two examples of approximated disjunction operation \oplus between trapezoids: in (a) the operation can be performed, while in (b) the operation cannot be performed.

where

$$\begin{aligned}
 a &= \min\{a_1, a_2\} & d &= \max\{d_1, d_2\} \\
 b &= \begin{cases} b_1 & \text{if } \alpha_1 > \alpha_2 \\ b_2 & \text{if } \alpha_2 > \alpha_1 \\ \min\{b_1, b_2\} & \text{otherwise} \end{cases} & c &= \begin{cases} c_1 & \text{if } \alpha_1 > \alpha_2 \\ c_2 & \text{if } \alpha_2 > \alpha_1 \\ \min\{c_1, c_2\} & \text{otherwise} \end{cases}
 \end{aligned} \tag{12}$$

$$\begin{aligned}
 \llbracket \Omega_1 \cup \Omega_2 \rrbracket &= \{ \omega_i = (o_i, d_i) \mid (o_i \in \mathcal{A}_1 \wedge o_i \notin \mathcal{A}_2 \wedge d_i = \text{sim}(T_1, T)) \vee \\
 &\quad (o_i \in \mathcal{A}_2 \wedge o_i \notin \mathcal{A}_1 \wedge d_i = \text{sim}(T_2, T)) \vee \\
 &\quad (o_i \in \mathcal{A}_1 \wedge o_i \in \mathcal{A}_2 \wedge d_i = \max(\text{sim}(T_1, T), \text{sim}(T_2, T))) \}
 \end{aligned}$$

Fig. 8.a illustrates a case where the disjunction is executed, while Fig. 8.b illustrates a case where it cannot be executed. As regards to the provenance statement, in case the disjunction is able to eliminate a redundant constraint, the resulting one will take the maximum degree of ownership for each author. Therefore the disjunction has the same behaviour of the conjunction operation as regards to the provenance statements.

Algorithm 1: Application of the path-consistency algorithm for the propagation of the constraints in a PA-FTCN.

```

1 function ConstraintPropagation( $\mathcal{N} = \langle \mathcal{X}, \mathcal{K} \rangle$ )
2    $Q \leftarrow \text{triangles}(\mathcal{N})$ 
3   while  $Q \neq \emptyset$  do
4      $\langle x_i, x_k, x_j \rangle \leftarrow \text{dequeue}(Q)$ 
5      $C'_{ij} \leftarrow C_{ij} \otimes (C_{ik} \circ C_{kj})$ 
6     if  $C'_{ij} \neq C_{ij}$  then
7        $Q \leftarrow Q \cup \{ \langle x_i, x_k, x_j \rangle \}$ 
8        $\text{replace}(\mathcal{N}, C_{ij}, C'_{ij})$ 
9     end
10  end
11  return  $\mathcal{N}$ 
12 function triangles( $\mathcal{N} = \langle \mathcal{X}, \mathcal{K} \rangle$ )
13   $T \leftarrow \emptyset$ 
14  for  $1 \leq i < j \leq |\mathcal{X}|, 1 \leq k \leq |\mathcal{X}| \wedge k \neq i \wedge k \neq j$  do
15    if  $\exists C_{ij} \in \mathcal{K} \wedge (\exists C_{ik} \in \mathcal{K} \vee \exists C_{ik}^{-1} \in \mathcal{K}) \wedge (\exists C_{kj} \in \mathcal{K} \vee \exists C_{kj}^{-1} \in \mathcal{K})$  then
16       $T \leftarrow T \cup \{ \langle x_i, x_k, x_j \rangle \}$ 
17    end
18  end
19  return  $T$ 

```

Given the semantics of the operations provided in Def. 4.10-4.13, we can apply the path-consistency algorithm presented in Alg. 1. However, as already studied in literature, this algorithm has a theoretical complexity equal to $O(n^3k^5)$ where n is the number of nodes and k is the number of arcs. Clearly, this complexity made the algorithm unusable in many real-world situations where the number of nodes and arcs increases. Many optimized versions of the algorithm have been proposed in literature in order to reduce such complexity [33]. In this paper, we propose a different approach which is based on the observation that, during the processing of a *Star* model, the obtained graph is naturally partitioned into independent sub-graphs, each one related to the semantical subdivision given by the notion of archaeological unit. More specifically, the interpretation process through which archaeological units are determined, is essentially the result of the discovering of related archaeological partitions and their grouping into a unique entity. From the network point of view, each archaeological unit contained in a *Star* model originates a sub-graph whose nodes are highly connected, while very few connections are established between nodes belonging to different archaeological units. Starting from this observation we develop a MapReduce version of the algorithm in Alg. 1 which is discussed in the following section and allows to overcome the scalability problems of the original algorithm.

4.4 A MapReduce Version of the Path Consistency Algorithm

As discussed at the end of the previous section, in the *Star* model archaeological information can be semantically repartitioned around the concept of archaeological unit. Each archaeological unit is constituted by a set of archaeological

partitions that represent raw data on which several interpretations and reconstruction hypothesis can be made by archaeologists. Following this idea, the overall network built by using the rules in Sect. 4.2 can be subdivided into a set of sub-networks which can be considered quite independent from each other, in the sense that the connections inside them are very dense while the connections among them are quite sparse. As a general idea, the definition of such sub-networks is done by starting from each archaeological unit and the set of archaeological partitions connected to it, and then by considering the temporal connections among these objects. These connections can originate by both instances of `ST_PhaseSequences` and `ST_RelatedArchaeoParts` (both subclasses of `TM_TopologicalComplex`) and instances of `ST_Stratigraphy` (subclass of `TP_TopologicalComplex`).

From an operative point of view, the translation of a *Star* model into a PA-FTCN starts from Rules 2-8, which regard the topological complex `ST_PhaseSequence` that is used for dating an `ST_ArchaeologicalUnit`. For each of these topological complexes, the rules generate a distinct highly connected sub-graph. Starting from that, the sub-graphs could be enriched by Rules 9-10, which regard the dating of archaeological partitions and their containment relations with the assigned phases. Since an archaeological partition can belong to different archaeological units, the same elements produced by Rule 9 can be contained in different sub-graphs. Moreover, different instances of `ST_ArchaeoPartDate` could be connected together through the application of Rule 11. This rule can generate connections not only internal to the same sub-graph, but also external ones. Another way to generate connections between different sub-graphs is through the third considered topological complex, the `ST_Stratigraphy`, which could again generate relations among archaeological partition dates and they can both belong to the same or to different sub-graphs. We can observe that the connections inside each sub-graph may be very dense, while the connections among different sub-graphs are relatively sparse. This is also true for stratigraphic relations, indeed objects involved in the definition of the same archaeological unit are typically nearby to each other and on them some stratigraphic relations can have been defined, conversely objects regarding different archaeological units can be very far from each other and in this case they can be hardly involved in the construction of the same `ST_Stratigraphy` complex.

Example 4.14. Let us consider the network depicted in Fig. 9 which has been obtained from a simple *Star* model composed of two archaeological units A_1 and A_2 . Both archaeological units are characterized by a single phase in their life, whose boundaries are represented by nodes $A1_s - A1_e$ and $A2_s - A2_e$, respectively. Moreover, archaeologists have assigned to A_1 an archaeological partition P_1 , whose dating is represented by node $P100$ that has to be contained inside phase $A1$. Conversely, two archaeological partitions have been attached to A_2 whose datings are represented by nodes P_{200} and P_{201} , both belonging to the phase $A2$. Finally, the following temporal relation has been determined: “ P_{200} before P_{201} ”, and the following stratigraphic relation has been discovered: “ P_{200} below (before) P_{100} ”. As you can notice, the network in Fig. 9 is partitioned inside two sub-graphs, each one corresponding to a different archaeological unit. Moreover, while the connections inside each sub-graph are very dense, only one connection is established between the two sub-graphs, due to the relation between archaeological partitions P_{200} and P_{100} , which belongs to different archaeological units. In the network of Fig. 9 we omitted the edge labels for not cluttering the diagram.

From these considerations a MapReduce implementation of the algorithm in Alg. 1 can be easily obtained; the general idea is to start by considering each sub-graph corresponding to a different archaeological unit alone (during a map phase) and then combine the obtained results in order to process also the connections between different sub-graphs (during a reduce phase). Notice that each sub-graph processed by a map task will contain a copy of the network start node and of any other shared node (i.e., archaeological partition date). The reduce phase may work on a compact version of the complete graph containing only the nodes with a connection outside its topological complex, together

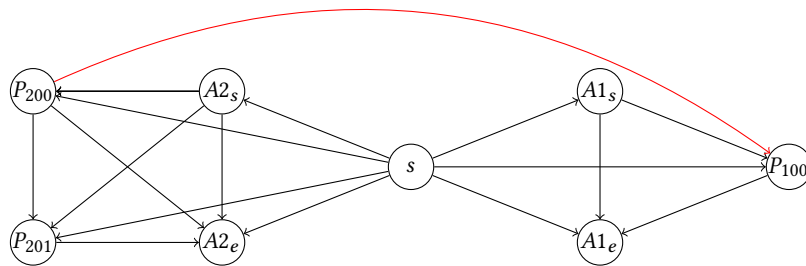


Fig. 9. Simple example of PA-FTCN built by using the rules in Sect. 4.2.

with the start node and its outgoing connections towards the other nodes in the compact graph. If the reduce phase produces a modification on the edges of the compact graph, another iteration of the overall job is necessary in order to propagate such changes also inside the sub-graphs.

Example 4.15. Let us consider again the network introduced in Fig. 9, in this case two sub-graphs can be obtained, each one corresponding to a different archaeological unit, they are depicted in Fig. 10 (a) and (b). Each of these sub-graphs will be processed by a distinct map task that will eventually derive new constraints for the edges in each of them. Conversely, the reducer will work on the compact network depicted in Fig. 10 (c), which contains the nodes P_{200} and P_{100} that are involved in the external connection, together with the start node s and the edges connecting s to P_{200} and P_{100} , respectively.

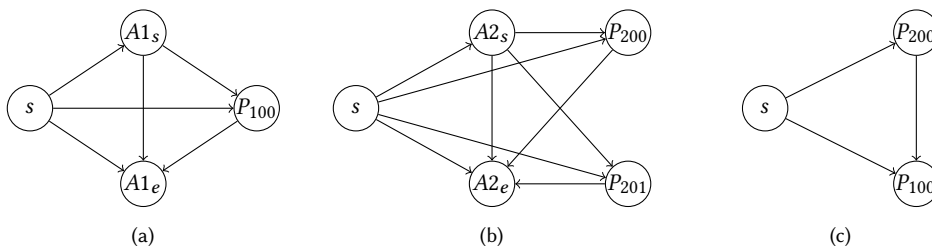


Fig. 10. (a) Sub-graph corresponding to A_1 , (b) sub-graph corresponding to A_2 and (c) compact network obtained from Fig. 9.

Alg. 2 illustrates how the inputs for the MapReduce job are prepared. In particular, starting from a *Star* model \mathcal{S} , the algorithm produces: (i) the overall network \mathcal{N} , (ii) a list of sub-graphs L_n each one corresponding to an archaeological unit, and (iii) a compact network \mathcal{N}_t containing only the nodes with external connections and the related edges.

The algorithm starts by processing the topological complexes of type *ST_PhaseSequence* contained in the model (lines 5-16). For each of them, it generates a new sub-graph \mathcal{N}_i which is populated with the start node s and by using Rules 2-8. Moreover, for each archaeological partition ap associated to the current archaeological unit, we apply Rule 9 and Rules 6-8 for generating the nodes corresponding to its dating. Finally, we apply Rules 10-12 for generating the additional edges inside the sub-graph. Indeed, while Rule 9 is applied starting from the archaeological partition ap , Rules 10-12 are applied w.r.t. to both ap and the current topological complex t . In other words, Rules 10-12 on the dating of the current archaeological partition and the dating of phases reachable from elements belonging to the same

Algorithm 2: Function responsible for building the global network \mathcal{N} , the list of sub-graphs L_n processed by the Mappers and the compact network \mathcal{N}_t processed by the Reducer.

```

1 function BuildNetwork( $\mathcal{S}$ )
2    $\mathcal{N} \leftarrow \langle \{s\}, \emptyset \rangle$ 
3    $L_n \leftarrow \{\}$ 
4    $\mathcal{N}_t \leftarrow \langle \{s\}, \emptyset \rangle$ 
5   foreach  $t \in \mathcal{S}.ST\_PhaseSequence$  do
6      $\mathcal{N}_i \leftarrow \langle \{s\}, \emptyset \rangle$ 
7      $\mathcal{N}_i \leftarrow \text{transf}(t, \text{Rule 2}, \text{Rule 3}, \text{Rule 4}, \text{Rule 6}, \text{Rule 7}, \text{Rule 8})$ 
8     foreach  $ap \in t.\text{archaeoUnit}.\text{archaeoParts}$  do
9        $\mathcal{N}_i \leftarrow \text{transf}(ap, \text{Rule 9}, \text{Rule 6}, \text{Rule 7}, \text{Rule 8})$ 
10       $\mathcal{N}_i \leftarrow \text{transf}(t, ap, \text{Rule 10})$ 
11       $\mathcal{N}_i \leftarrow \text{transf}(t, ap, \text{Rule 11})$ 
12       $\mathcal{N}_i \leftarrow \text{transf}(t, ap, \text{Rule 12})$ 
13     end
14      $L_n \leftarrow L_n \cup \{\mathcal{N}_i\}$ 
15      $\mathcal{N} \leftarrow \text{expand}(\mathcal{N}, \mathcal{N}_i)$ 
16   end
17   foreach  $t \in \mathcal{S}.ST\_RelatedArchaeoParts$  do
18     foreach  $(d_s, d_e) \in t.ST\_TemporalRelation$  do
19       if  $d_s.\text{archaeoPart}.\text{archaeoUnits} \cap d_e.\text{archaeoPart}.\text{archaeoUnits} = \emptyset$  then
20          $\mathcal{N} \leftarrow \langle \mathcal{X}, \mathcal{K} \cup \{(n_{d_s}, n_{d_e})\} \rangle$ 
21          $\mathcal{N}_t \leftarrow \langle \mathcal{X} \cup \{(n_{d_s}, n_{d_e})\}, \mathcal{K} \cup \{(s, n_{d_s}), (s, n_{d_e}), (n_{d_s}, n_{d_e})\} \rangle$ 
22       end
23     end
24   end
25   foreach  $t \in \mathcal{S}.ST\_Stratigraphy$  do
26     foreach  $(p_s, p_e) \in \mathcal{S}.ST\_ArchaeoRelation$  do
27       if  $p_s.\text{archaeoPart}.\text{archaeoUnits} \cap p_e.\text{archaeoPart}.\text{archaeoUnits} = \emptyset$  then
28          $\mathcal{N} \leftarrow \langle \mathcal{X}, \mathcal{K} \cup \{(n_{ap_s}, n_{ap_e})\} \rangle$ 
29          $\mathcal{N}_t \leftarrow \langle \mathcal{X} \cup \{(n_{ap_s}, n_{ap_e})\}, \mathcal{K} \cup \{(s, n_{ap_s}), (s, n_{ap_e}), (n_{ap_s}, n_{ap_e})\} \rangle$ 
30       end
31     end
32   end
33   storeDfs( $L_n$ )
34   store( $\mathcal{N}_t$ )
35   return  $\mathcal{N}$ 

```

archaeological unit. The function `transf` is responsible for properly applying the transformation rules described in Sect. 4.2. At the end of this inner cycle, the obtained sub-graph \mathcal{N}_i is added to the list L_n (line 14) and the overall network \mathcal{N} is also accordingly expanded (line 15).

The subsequent two main cycles are used to build the connections among different sub-graphs. In particular, starting from a topological complex `ST_RelatedArchaeoParts` (lines 17-24), we identify the set of relations involving dates associated to archaeological partitions which do not belong to the same archaeological unit. Indeed, the condition in line 19 checks the presence of at least one archaeological unit shared between the two involved partitions, if this

object exists it means that the current relation has been already processed as a inner connection inside some topological complex; otherwise, the connection should be considered external. For each external connection, we add it to the global network (line 20) and we expand the compact network (line 21) by adding the nodes involved in the external connection, their relation with the start node s and the connection itself. With the label n_{d_s} and n_{d_e} we denote the nodes representing the dates d_s and d_e in the relation, respectively.

Algorithm 3: Mapper class implementing the operations to be performed in parallel on each sub-graph $\mathcal{N}_i \in L_n$.

```

1 class Mapper
2   method setup()
3     |  $M \leftarrow \emptyset$ 
4   method map( $id, \langle x_i, x_j, C \rangle$ )
5     |  $M \leftarrow M.put(\langle x_i, \langle x_j, C \rangle \rangle)$ 
6     |  $M \leftarrow M.put(\langle x_j, \langle x_i, C^{-1} \rangle \rangle)$ 
7   method cleanup()
8     |  $Q \leftarrow triangles(\mathcal{N}_i, M)$ 
9     | while  $Q \neq \emptyset$  do
10      |  $\langle x_i, x_k, x_j \rangle \leftarrow dequeue(Q)$ 
11      |  $C'_{ij} \leftarrow C_{ij} \otimes (C_{ik} \circ C_{kj})$ 
12      | if  $C'_{ij} \neq C_{ij}$  then
13      |   |  $Q \leftarrow Q \cup \{\langle x_i, x_k, x_j \rangle\}$ 
14      |   | replace( $\mathcal{N}_i, C_{ij}, C'_{ij}$ )
15      |   | replace( $\mathcal{N}, C_{ij}, C'_{ij}$ )
16      |   end
17      | end
18   method triangles( $\mathcal{N}, M$ )
19     |  $Q \leftarrow \emptyset$ 
20     | foreach  $\langle x_i, x_j, C_{ij} \rangle \in \mathcal{N}$  do
21     |   | foreach  $x_k \in M.get(x_i).keys()$  do
22     |   |   | if  $x_j \in M.get(x_k).keys()$  then
23     |   |   |   |  $Q \leftarrow \langle x_i, x_k, x_j \rangle$ 
24     |   |   |   | return  $Q$ 
25     |   |   |   end
26     |   |   end
27     |   end

```

Similar operations are performed for each topological complex `ST_Stratigraphy` (line 26-32), in this case starting from the relation between two altimetric points p_s and p_e , we go back to the corresponding archaeological partitions, if they do not share a common archaeological unit, the relation between their dates is added to both the global network \mathcal{N} and the condensed network \mathcal{N}_t . Notice that we add to \mathcal{N}_t also the edges from the start node s to the considered dating nodes. In the pseudo-code, we use n_{ap_s} and n_{ap_e} to denote the nodes representing the dating of the archaeological partitions associated to p_s and p_e , respectively.

Finally, the algorithm stores in a distributed way the partial sub-graphs (line 33), namely each sub-graph will correspond to a different split (using the MapReduce terminology) and will be processed by a different map task. On the contrary, the compact network is stored separately for being processed by a single reduce task. Each graph is stored as a sequence of tuples $\langle x_i, x_j, C_{ij} \rangle$, where x_i and x_j are two nodes connected by an edge with label (constraint) C_{ij} .

Given the inputs produced by Alg. 2, the MapReduce job responsible for performing the constraint propagation can be applied (see Alg. 5). More specifically, each mapper works on a different subgraph $\mathcal{N}_i \in L_n$ by performing the operations illustrated in Alg. 3. The setup method is responsible for initializing the necessary auxiliary data structure, in particular, to ease the discovery of the existing triangles in the network, we maintain a map M which summarizes the node connections. More specifically, the keys of M will be the nodes in \mathcal{N} , for each node $x_i \in \mathcal{N}$ the corresponding value is again a map M_i whose keys are the nodes $x_j \in \mathcal{N}$ which are reachable from x_i through an edge, and the value is the constraint characterizing such edge. The variable M is populated by the map method (lines 4-6) which processes one record of type $\langle x_i, x_j, C_{ij} \rangle$ at time. As you can notice, we include inside M both the direct edges and their inverse. The real constraint propagation activity is performed by the cleanup method which initially builds a list Q of triangles contained inside the network \mathcal{N} by exploiting the content of M .

The identification of the triangles in \mathcal{N} is done by the method `triangles`, for each edge $\langle x_i, x_j, C_{ij} \rangle \in \mathcal{N}$ it searches the presence of a node x_k such that there exists: (i) a direct or indirect edge from x_i to x_k and (ii) a direct or indirect edge from x_k to x_j . The identification of both direct and indirect edges may be easily done because both edges have been added to M by the map method.

Given a triangle $\langle x_i, x_k, x_j \rangle$, the constraint propagation formula is applied (line 11). In case a different constraint C'_{ij} is derived, the triangle is added again to Q and both the subgraph network \mathcal{N}_i and the global network \mathcal{N} are updated.

Algorithm 4: Reducer class implementing the operations to be performed on the condensed network \mathcal{N}_t .

```

1 class Reducer
2   method setup()
3      $M \leftarrow \emptyset$ 
4   method reduce( $id, \langle x_i, x_j, C \rangle$ )
5      $M \leftarrow M.put(\langle x_i, \langle x_j, C \rangle \rangle)$ 
6      $M \leftarrow M.put(\langle x_j, \langle x_i, C^{-1} \rangle \rangle)$ 
7   method cleanup()
8      $Q \leftarrow triangles(\mathcal{N}_t)$ 
9      $c \leftarrow false$ 
10    while  $Q \neq \emptyset$  do
11       $\langle x_i, x_k, x_j \rangle \leftarrow dequeue(Q)$ 
12       $C'_{ij} \leftarrow C_{ij} \otimes (C_{ik} \circ C_{kj})$ 
13      if  $C'_{ij} \neq C_{ij}$  then
14         $Q \leftarrow Q \cup \{\langle x_i, x_k, x_j \rangle\}$ 
15        replace( $\mathcal{N}_t, C_{ij}, C'_{ij}$ )
16        replace( $\mathcal{N}, C_{ij}, C'_{ij}$ )
17         $c \leftarrow true$ 
18      end
19    end
20  return  $c$ 

```

The reducer performs the same operations done by the mappers, with the exception that it works on the compact network \mathcal{N}_t . In the reduce method, the reducer is responsible for updating the compact network \mathcal{N}_t with the new constraints determined during the map phase. More specifically, for each constraint C contained in the compact network, if C has been modified by a mapper, the constraint is accordingly modified in \mathcal{N}_t . This constraint can regard the

connection between the dating of archaeological partition d and the start node s . Since an archaeological partition can be associated to multiple archaeological units, more than one constraint could be generate for the pair (s, d) . The function update is responsible for combining them, eventually by applying the disjunction operation in Def. 4.13.

Algorithm 5: Job providing a MapReduce implementation of the ConstraintPropagation function in Alg. 1.

```

1 function ConstraintPropagationMR( $S$ )
2    $\mathcal{N} \leftarrow \text{BuildNetwork}(S)$ 
3    $job.setMapper(\text{Mapper}, L_n, \mathcal{N})$ 
4    $job.setReducer(\text{Reducer}, \mathcal{N}_t, \mathcal{N})$ 
5    $continue \leftarrow \text{true}$ 
6   while  $continue$  do
7      $continue \leftarrow job.run()$            // run an iteration of the Mappers and the Reducer
8   end

```

A boolean variable c , which keeps track of the eventual changes done to the constraints in \mathcal{N}_t , is maintained. In case a constraint in the compact network \mathcal{N}_t has been modified during the Reducer cleanup method, another iteration of the overall MapReduce job in Alg. 5 is necessary.

The theoretical complexity of each mapper and reducer is again $O(n^3k^5)$, since all of them apply the classical path consistency algorithm presented in Alg. 1. However, in case of the mappers, the value of n is not the global number of nodes in \mathcal{N} , but the average number n_i of nodes in each sub-graph, while the value of k becomes the average number of connections inside a topological complex. Similarly, in case of the reducers, the value of n becomes the number n_r of nodes having a connection outside its topological complex and k is the value k_r of external connections. Clearly, we can observe that $n_r \ll n_i \ll n$ and $k_r \ll k_i \ll k$. As regards to the number of times the overall job in Alg. 5 is executed, this is at most equal to the number of constraints contained in the compact network, since in the worst case we are able to remove only one external constraint at time.

5 CASE STUDY

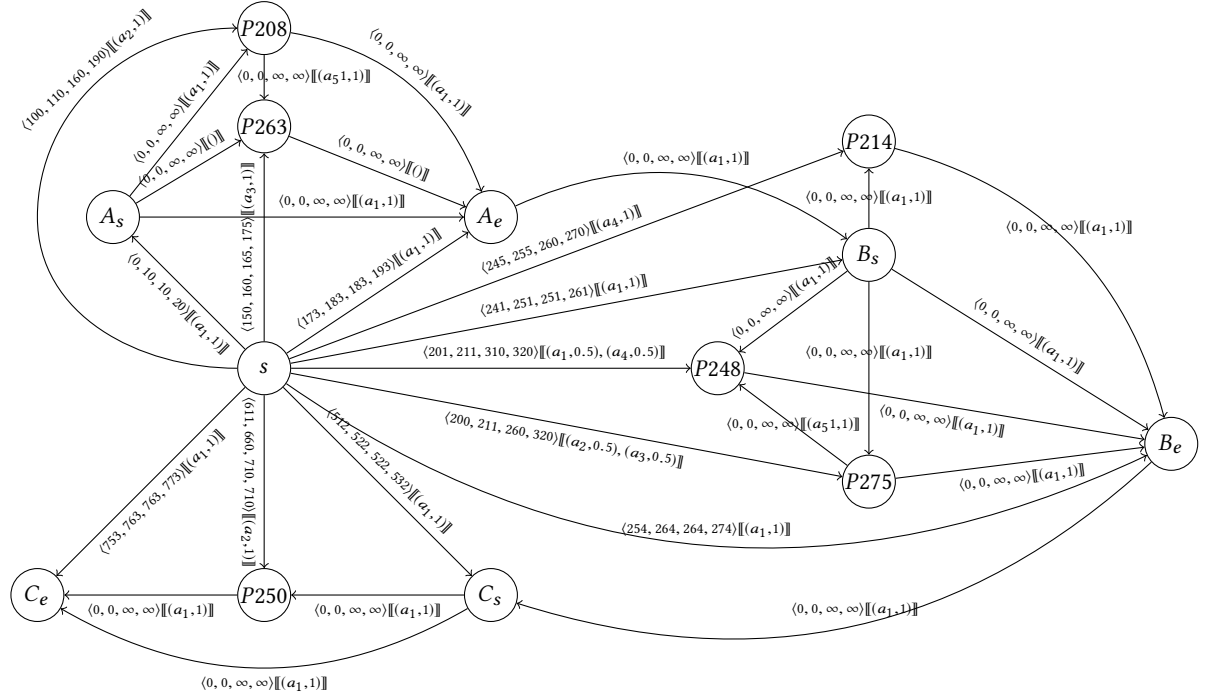
This section illustrates an example of reasoning performed on archaeological data that allows the identification of some new temporal and data provenance knowledge. It regards an archaeological object called *Porta Borsari* which is an ancient Roman gate in Verona and a historical building adjacent to it. The two objects have been modelled as two ST_ArchaeoUnits by archaeologist a_1 who also identifies some distinct phases in their lives. In the following, we will refer to the *Porta Borsari* as au_1 and the other historical building as au_2 .

As regards to au_1 , archaeologist a_1 identifies three distinct phases into its life:

- Phase A – first foundation as *Porta Iovia* during the Late Republican Time, which spans from 200 B.C. to 27 B.C.;
- Phase B – reconstruction during the Claudian Time, which spans from 41 A.C. to 54 A.C.;
- Phase C – Teodorician changes during the Middle-Age, which spans from 312 A.C. to 553 A.C.

The same archaeologist decides to add ± 10 years of safety to the temporal boundaries of each phase.

Subsequently, other archeologists have identified some findings as archaeological partitions belonging to this archaeological unit. Table 2 reports some information about them together with the associated dating. As regards to the dating, we assume that the first archaeologist who found an archaeological partition simply assigns it to one of the identified phases, while later the same or other authors will restrict such dating as soon as new information becomes

Fig. 11. Sub-graph for archaeological unit au_1 .

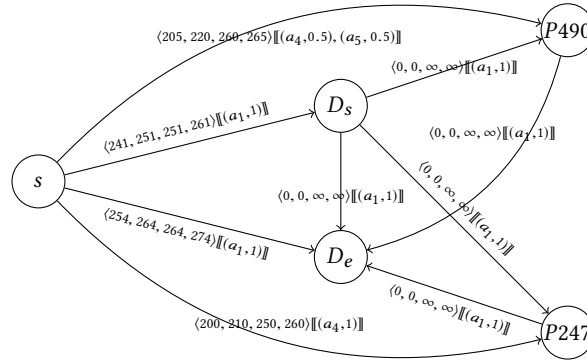
available. The author responsible for the identification of the phase membership is reported in column **Ph** inside round brackets together with the phase name, while the author(s) responsible for the fine-grained dating is (are) reported in column **Dating**. Notice that in order to not cluttering the notation, we have omitted to report the original unitary

Table 2. Dating and associated phase for the archaeological partition belonging to au_1 . Negative values indicate B.C. years.

Archaeo. Partition	Ph	Dating
P208 Foundation and North Tower	A (a_1)	$\langle -110, -100, -50, -20 \rangle \llbracket (a_2, 1) \rrbracket$
P263 Structures of eastern facade	A (a_1)	$\langle -60, -50, -45, -35 \rangle \llbracket (a_3, 1) \rrbracket$
P214 Front of the external facade	B (a_1)	$\langle 35, 45, 50, 60 \rangle \llbracket (a_4, 1) \rrbracket$
P248 External Foundations	B (a_1)	$\langle -9, 1, 100, 110 \rangle \llbracket (a_1, 0.5), (a_4, 0.5) \rrbracket$
P275 Internal Foundations	B (a_1)	$\langle -10, 1, 50, 100 \rangle \llbracket (a_2, 0.5), (a_3, 0.5) \rrbracket$
P250 Defensive structures	C (a_1)	$\langle 401, 450, 500, 500 \rangle \llbracket (a_2, 1) \rrbracket$

height of the trapeze (namely [1]). Moreover, since the table reports initial information, we assume that when more than one author is present in Tab. 2, the contribution provided by each author is equal, i.e. the reporting date is the result of a joint work.

Finally, author a_5 identifies the following temporal relations between partitions: P208 terminates before P263 starts, and P275 terminates before P248 starts. These precedence relations have to be modeled with an arc $\langle 0, 0, \infty, \infty \rangle \llbracket (a_5, 1) \rrbracket$. The sub-graph for archaeological unit au_1 is reported in Fig. 11.

Fig. 12. Sub-graph for archaeological unit au_2 .

Conversely, as regards to au_2 , archaeologist a_1 identifies a unique phase of its life: D , which is traceable back to the Claudian Time from 41 A.C. to 54 A.C., and authors add a safety range of ± 10 years to it. Moreover, two archaeological partitions have been identified and assigned to this phase, as reported in Table 3. The sub-graph for the archaeological unit au_2 is reported in Fig. 12.

Table 3. Dating and associated phase for the archaeological partition belonging to au_2 . Negative values indicate B.C. years.

Archaeo. Partition	Ph	Dating
P247 External Foundations	D (a_1)	$\langle -10, 0, 60, 70 \rangle \parallel (a_1, 0.5), (a_4, 0.5) \parallel$
P490 Internal Foundations	D (a_1)	$\langle -5, 10, 50, 60 \rangle \parallel (a_4, 0.5), (a_5, 0.5) \parallel$

Finally, author a_5 determines a precedence relation between the archaeological partitions P248 and P247, which belong to two different archaeological units but are adjacent from a spatial point of view. This is an external relation which has to be represented in the compact network \mathcal{N}_t , as illustrated in Fig. 13. The compact network contains the external relation together with the edges connecting the involved nodes and the start node s .

Notice that, according to the transformation rules of the previous section, the first operation to perform is the definition of a common coordinate reference system. The origin of such system is set to 210 B.C. (i.e., 200 B.C. minus 10 years of safety), since it is the earliest date in the model, while the granularity is the year, since for all dates the minimum granularity is at least a year.

Following the algorithm presented in Sec. 4.4, the two sub-graphs in Fig. 11 and 12 are processed by two distinct mappers, while the compact graph in Fig. 13 is processed by a reducer, which also takes care of new constraints between s and P247, and between s and P248 produced during the map phase.

Let us consider for instance the triangle in Fig. 11 composed of nodes $i = s$, $j = P248$ and $k = B_s$, by applying the formula in Def. 4.7: $\pi'_{ij}(x) = \pi_{ij} \otimes_a (\pi_{ik} \circ \pi_{kj}(x))$, we obtain the following new constraints C_{ij} between s and P248:

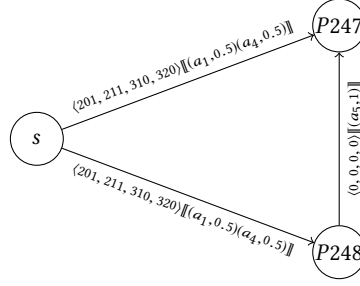


Fig. 13. Compact network \mathcal{N}_f containing the external relations involving both au_1 and au_2 .

$$\begin{aligned}
 \pi'_{s,P248} &= \pi_{s,P248} \otimes (\pi_{s,B_s} \circ \pi_{B_s,P248}) \\
 &= \langle 201, 211, 310, 320 \rangle \llbracket (a_1, 0.5), (a_4, 0.5) \rrbracket \otimes \\
 &\quad (\langle 241, 251, 251, 261 \rangle \llbracket (a_1, 1) \rrbracket \circ \langle 0, 0, \infty, \infty \rangle \llbracket (a_1, 1) \rrbracket) \\
 &= \langle 201, 211, 310, 320 \rangle \llbracket (a_1, 0.5), (a_4, 0.5) \rrbracket \otimes \langle 241, 251, \infty, \infty \rangle \llbracket (a_2, 1) \rrbracket \\
 &= \langle 201, 251, 310, 320 \rangle \llbracket (a_1, 0.4), (a_4, 0.4), (a_5, \perp) \rrbracket
 \end{aligned}$$

This derivation produces a restriction of the distribution core while the support remains unchanged, it represents in any case a restriction of the uncertainty or better the refinement of the information regarding the dates of archaeological partition P248, which is now more precise. Relatively to the ownership of the information, we observe that authors a_1 and a_4 have a smaller degree of ownership w.r.t. the original information, because the final result depends also on the observations done by a_5 . Author a_5 also appears in the final result however with a very low degree of ownership, since she only provides a specification of an undirect relation.

Similarly, if we consider the second sub-graph in Fig. 12 and the triangle $i = s, j = P247$ and $k = D_e$, we obtain the following derivation:

$$\begin{aligned}
 \pi'_{s,P247} &= \pi_{s,P247} \otimes (\pi_{s,D_e} \circ \pi_{D_e,s}) \\
 &= \langle 200, 210, 270, 280 \rangle \llbracket (a_4, 1) \rrbracket \otimes \\
 &\quad (\langle 254, 264, 264, 274 \rangle \llbracket (a_1, 1) \rrbracket \circ \langle 0, 0, \infty, \infty \rangle^{-1} \llbracket (a_1, 1) \rrbracket) \\
 &= \langle 200, 210, 270, 280 \rangle \llbracket (a_4, 1) \rrbracket \otimes \\
 &\quad (\langle 254, 264, 264, 274 \rangle \llbracket (a_1, 1) \rrbracket \circ \langle -\infty, -\infty, 0, 0 \rangle \llbracket (a_1, 1) \rrbracket) \\
 &= \langle 200, 210, 270, 280 \rangle \llbracket (a_4, 1) \rrbracket \otimes \langle -\infty, -\infty, 264, 274 \rangle \llbracket (a_1, 1) \rrbracket \\
 &= \langle 200, 210, 264, 274 \rangle \llbracket (a_4, 0.9), (a_1, \perp) \rrbracket
 \end{aligned}$$

In this case we obtain a restriction of both the core and the support of the distribution. As regards to the degree of ownership, we register the small contribution provided by author a_1 which collocates P247 inside phase D and a consequent reduction of the ownership associated to a_1 due to the produced change.

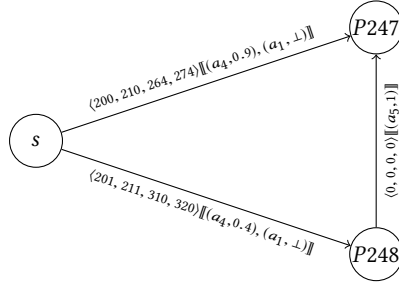


Fig. 14. Compact network \mathcal{N}_t containing the external relations involving both au_1 and au_2 , after the map phase.

In light of the new obtained constraints, the compact network \mathcal{N}_t in Fig. 13 changes as illustrated in Fig. 14 and can be processed by the reducer. In this case the application of the constraint propagation rule can be useful for obtaining a more precise precedence relation between P248 and P247.

$$\begin{aligned}
 \pi'_{P248, P247} &= \pi_{P248, P247} \otimes (\pi_{P248, s} \circ \pi_{s, P247}) \\
 &= \langle 0, 0, \infty, \infty \rangle \llbracket (a_4, 1) \rrbracket \otimes \\
 &\quad \langle \langle 200, 210, 264, 274 \rangle^{-1} \llbracket (a_4, 0.9), (a_1, \perp) \rrbracket \circ \langle 201, 211, 310, 320 \rangle \llbracket (a_4, 0.4), (a_1, \perp) \rrbracket \rangle \\
 &= \langle 0, 0, \infty, \infty \rangle \llbracket (a_4, 1) \rrbracket \otimes \\
 &\quad \langle \langle 274, 264, 210, 200 \rangle \llbracket (a_4, 0.9), (a_1, \perp) \rrbracket \circ \langle 201, 211, 310, 320 \rangle \llbracket (a_4, 0.4), (a_1, \perp) \rrbracket \rangle \\
 &= \langle 0, 0, \infty, \infty \rangle \llbracket (a_4, 1) \rrbracket \otimes \langle 475, 475, 520, 520 \rangle \llbracket (a_4, 0.4), (a_1, \perp) \rrbracket \\
 &= \langle 475, 475, 520, 520 \rangle \llbracket (a_4, 1), (a_1, \perp) \rrbracket
 \end{aligned}$$

In this case, we obtain that the interval between the two archaeological partitions is about 45 years thanks to the major contribution of author a_4 , which defines both the dating of the two partitions and the qualitative temporal relation between them, and in a minor part to author a_1 .

Clearly, these are only examples of the derivations that can be obtained by executing the path-consistency algorithm on the overall network and considering all the triangles. However, these examples make clear the utility of applying existing temporal reasoning techniques on archaeological data.

6 CONCLUSION

The dating process is one of the main activities performed by archaeologists who, during their interpretation of the founded objects, try to give them some location in time. Time in archaeology is typically characterized by a certain level of uncertainty and dates are usually provided as an interval of great confidence with a certain range of safety added by domain experts. In this context the use of reasoning techniques could be very useful to reduce the level of uncertainty and increase the temporal knowledge about the objects at hand [9, 11]. During these derivations, it is essential to keep track of the interpretation provenance, namely the set of authors that are involved in the definition of new temporal information.

In this paper we propose an extension of the model called *Star* which is able to represent archaeological information characterized by uncertain temporal properties and relations together with data provenance attributes [9, 11, 40]. We

also present a way to translate this model into a PA-FTCN, on which some reasoning technique can be applied in order to produce new temporal knowledge or reducing the uncertainty of the available one. At this regard, the traditional operations have been extended in order to deal with both uncertain temporal information and data provenance. More specifically, each derivation operation produces new degree of ownership values for the derived information.

In order to increase the efficiency of the constraint propagation procedure, we exploit the semantic characteristics of the model to propose a MapReduce implementation of the path consistency algorithm: it processes independent subgraphs in parallel and then combine the partial results in order to obtain the final one. Finally, we illustrate an example of application of the proposed framework by considering a real-world case scenario regarding some archaeological data of Verona. This application reveals the utility and potentiality of the proposed approach and encourages future investigations in this direction.

As future work we plan to extend the methodology proposed in this paper for tracking data provenance information in other interpretation activities performed by archaeologists, not only the dating one. More specifically, despite the need to define proper extensions of the *Star* conceptual model and of the translation rules, the general approach presented in this paper can be applied any time we have provenance information attached to objects with some mutual connections. Indeed, the overall mechanism of constraint propagation is based on the existence of some redundant (direct and indirect) connections between objects on which some attribute can be measured or evaluated. This can be for instance the case of an interpretation based on the analysis of material, or a compositional analysis, or any pattern analysis, in which the existence of the same or similar pattern can suggest some properties of the objects.

Another interesting extension that can be easily made to the already articulated *Star* temporal model is the possibility to express relative values together with absolute ones. This can be already partially obtained through the use of topological temporal constructions, through which relative temporal relations between objects can be expressed. Anyway, the topological model can be enriched with the possibility to express some quantitative measure to these qualitative precedences. This can be considered a simple change, but it is of invaluable importance in the archaeological application domain since many dates are expressed in relative form rather than in an absolute way.

ACKNOWLEDGMENTS

This work was partially supported by the Italian National Group for Scientific Computation (GNCS-INDAM) and by “Progetto di Eccellenza” of the Computer Science Dept., Univ. of Verona, Italy.

REFERENCES

- [1] 2013. World Wide Web Consortium - PROV-DM: The PROV Data Model. <https://www.w3.org/TR/prov-dm/>.
- [2] J. F. Allen. 1983. Maintaining Knowledge About Temporal Intervals. *Communications of the ACM* 26, 11 (1983), 832–843.
- [3] S. Badaloni, M. Falda, and M. Giacomini. 2004. Integrating Quantitative and Qualitative Fuzzy Temporal Constraints. *AI Communications* 17, 4 (2004), 187–200.
- [4] S. Badaloni and M. Giacomini. 2006. The Algebra IA^{fuz} : A Framework for Qualitative Fuzzy Temporal Reasoning. *Artificial Intelligence* 170, 10 (2006), 872–908.
- [5] J. A. Barceló. 2010. Computational Intelligence in Archaeology. State of the Art. In *Computer Applications & Qualitative Methods in Archaeology (CAA), Proceedings of the 37th International Conference*. 11–21.
- [6] Juan Antonio Barceló. 2019. Computing Archaeological Stratigraphies. A State-of-the-Art. In *Book of Abstracts of the 47th Computer Applications and Quantitative Methods in Archaeology (CAA 2019)*. 178. https://2019.caaconference.org/wp-content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf
- [7] P. Basso, P. Grossi, B. Bruno, A. Belussi, and S. Migliorini. 2017. From Rome, to Verona, to the Agro areas: Roundtrip. An experimentation of interoperability between SITAR, SITAVR and SITAIS. *Archeologia e Calcolatori* 2017 (2017), 157–170.
- [8] M. J. Baxter. 2009. Archaeological Data Analysis and Fuzzy Clustering. *Archaeometry* 51, 6 (2009), 1035–1054.

- [9] A. Belussi and S. Migliorini. 2014. A Framework for Managing Temporal Dimensions in Archaeological Data. In *Proceedings of 21st International Symposium on Temporal Representation and Reasoning (TIME)*. 81–90. <https://doi.org/10.1109/TIME.2014.15>
- [10] A. Belussi and S. Migliorini. 2014. *Modeling Time in Archaeological Data: the Verona Case Study*. Technical Report RR 93/2014. Department of Computer Science, University of Verona. <http://www.di.univr.it/report>
- [11] A. Belussi and S. Migliorini. 2017. A spatio-temporal framework for managing archeological data. *Annals of Mathematics and Artificial Intelligence* 80, 3 (Aug 2017), 175–218. <https://doi.org/10.1007/s10472-017-9535-0>
- [12] Alberto Belussi, Sara Migliorini, and Piergiorgiana Grossi. 2015. Managing Time Dimension in the Archaeological Urban Information System of the Historical Heritage of Rome and Verona. In *Proceedings of the 21st Century Archaeology: Concepts, methods and tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology (Paris, France) (CAA 2014)*. 235–244. <https://caa2014.sciencesconf.org/45964/document>
- [13] A. Belussi, S. Migliorini, and P. Grossi. 2016. Mapping of the SITAR and NIOBE data models towards the standard CIDOC-CRM_{archaeo} of the ARIADNE project with data transformation into RDF format. <https://www.di.univr.it/?ent=progetto&id=4578&lang=en> Last accessed May. 2020.
- [14] Igor Bogdanović, Vasiliki Andreaki, Joan A. Barceló, Raquel Piqué, Xavier Terradas, Antoni Palomo, Núria Morera, and Oriol López. 2019. Discovering the time of La Draga. In *Book of Abstracts of the 47th Computer Applications and Quantitative Methods in Archaeology (CAA 2019)*. 182. https://2019.caaconference.org/wp-content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf
- [15] P. Buneman, S. Khanna, and W. C. Tan. 2001. Why and Where: A Characterization of Data Provenance. In *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*. 316–330. https://doi.org/10.1007/3-540-44503-X_20
- [16] P. Buneman and W. C. Tan. 2018. Data Provenance: What next? *SIGMOD Record* 47, 3 (2018), 5–16. <https://doi.org/10.1145/3316416.3316418>
- [17] Enrico R. Crema and Anne Kandler. 2019. An R package for inferring patterns of social learning from archaeological frequency data. In *Book of Abstracts of the 47th Computer Applications and Quantitative Methods in Archaeology (Kraków, Poland) (CAA 2019)*. 70. https://2019.caaconference.org/wp-content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf
- [18] Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*. San Francisco, CA, 137–150.
- [19] R. Dechter, I. Meiri, and J. Pearl. 1991. Temporal Constraint Networks. *Artificial Intelligence* 49, 1-3 (1991), 61–95.
- [20] Peter Demján. 2019. Analysing settlement dynamics using statistics based on archaeological theory. In *Book of Abstracts of the 47th Computer Applications and Quantitative Methods in Archaeology (CAA 2019)*. 75. https://2019.caaconference.org/wp-content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf
- [21] M. J. Egenhofer and R. Franzosa. 1991. Point-set topological spatial relations. *International Journal of Geographic Information Systems* 2, 5 (1991), 161–174.
- [22] A. Felicetti, A. Masur, A. Kritsotaki, G. Hiebel, K. May, M. Theodoridou, M. Doerrand, P. Ronzino, S. Hermon, and W. Schmidle. 2016. Definition of the CRM_{archaeo}. An extension of CIDOC CRM to support archaeological excavation process. <http://www.cidoc-crm.org/crmarchaeo/ModelVersion/version-1.4.1> Last accessed May. 2020.
- [23] Gabriele Gattiglia, Nevio Dubbini, and Francesca Anichini. 2019. Spatio-temporal network analysis applied to Roman Terra Sigillata data. In *Book of Abstracts of the 47th Computer Applications and Quantitative Methods in Archaeology (CAA 2019)*. 77–78. https://2019.caaconference.org/wp-content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf
- [24] Cesar Gonzalez-Perez. 2018. *Temporality*. Springer International Publishing, 143–155. https://doi.org/10.1007/978-3-319-72652-6_15
- [25] Edward C. Harris. 1989. *Principles of archaeological stratigraphy, 2nd ed.* Academic Press.
- [26] David Holland, Uri Braun, Diana Maclean, Kiran-Kumar Muniswamy-Reddy, and Margo Seltzer. 2008. Choosing a Data Model and Query Language for Provenance. In *Proceedings of the 2nd International Provenance and Annotation Workshop (IPAW '08)*. Springer.
- [27] ICOM/CIDOC CRM Special Interest Group. 2020. Definition of the CIDOC Conceptual Reference Model, version 6.2.2. <http://www.cidoc-crm.org/>
- [28] ISO 2002. *ISO 19108 Geographic Information – Temporal Schema*. ISO. <https://www.iso.org/standard/26013.html>
- [29] ISO 2019. *ISO 19107 Geographic Information – Spatial Schema*. ISO. <https://www.iso.org/standard/66175.html>
- [30] M. Katsianis, S. Tsipidis, K. Kotsakis, and A. Kousoulakou. 2008. A 3D Digital Workflow for Archaeological Intra-Site Research using GIS. *Journal of Archaeological Science* 35, 3 (2008), 655–667.
- [31] C. C. Kolb. 2014. *Provenance Studies in Archaeology*. Springer New York, New York, NY, 6172–6181. https://doi.org/10.1007/978-1-4419-0465-2_327
- [32] David P. Lanter. 1991. Design of a Lineage-Based Meta-Data Base for GIS. *Cartography and Geographic Information Systems* 18, 4 (1991), 255–261. <https://doi.org/10.1559/152304091783786718>
- [33] Zhiguo Long, Michael Sioutis, and Sanjiang Li. 2016. Efficient Path Consistency Algorithm for Large Qualitative Constraint Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (New York, New York, USA) (IJCAI'16)*. AAAI Press, 1202–1208.
- [34] Alan K. Mackworth. 1977. Consistency in networks of relations. *Artificial Intelligence* 8, 1 (1977), 99–118. [https://doi.org/10.1016/0004-3702\(77\)90007-8](https://doi.org/10.1016/0004-3702(77)90007-8)
- [35] Keith May, James Stuart Taylor, and Steve Roskams. 2019. When Harris met Allen in The Matrix: How can the conceptual modelling of stratigraphic relationships facilitate deeper understanding of archaeological space and time?. In *Book of Abstracts of the 47th Computer Applications and Quantitative Methods in Archaeology (CAA 2019)*. 182. https://2019.caaconference.org/wp-content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf
- [36] Adam Mertel and David Zbiral. 2019. Early Christian Baptisteries: Geocoding, Exploring and Analysing a Spatiotemporal Dataset. In *Book of Abstracts of the 47th Computer Applications and Quantitative Methods in Archaeology (Kraków, Poland) (CAA 2019)*. 49–50. https://2019.caaconference.org/wp-content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf

- content/uploads/sites/25/2019/04/CAA2019_programabstracts_v20190423.pdf
- [37] S. Migliorini. 2019. Enhancing CIDOC-CRM Models for GeoSPARQL Processing with MapReduce. In *Proceedings of 2nd Workshop On Computing Techniques For Spatio-Temporal Data in Archaeology And Cultural Heritage*. CEUR-WS, 51–65. <http://ceur-ws.org/Vol-2230/>
 - [38] Sara Migliorini, Alberto Belussi, and Elisa Quintarelli. 2020. Promoting Data Provenance Tracking in the Archaeological Interpretation Process. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference (CEUR Workshop Proceedings, Vol. 2578)*. CEUR-WS.org. <http://ceur-ws.org/Vol-2578/PIE5.pdf>
 - [39] S. Migliorini and P. Grossi. 2018. Towards the Extraction of Semantics from Incomplete Archaeological Records. In *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*. Springer International Publishing, 349–358. https://doi.org/10.1007/978-3-319-63946-8_52
 - [40] Sara Migliorini, Piergiorganna Grossi, and Alberto Belussi. 2017. An Interoperable Spatio-Temporal Model for Archaeological Data Based on ISO Standard 19100. *J. Comput. Cult. Herit.* 11, 1, Article 5 (2017), 28 pages. <https://doi.org/10.1145/3057929>
 - [41] P. Missier and K. Belhajjame. 2012. A PROV Encoding for Provenance Analysis Using Deductive Rules. In *Provenance and Annotation of Data and Processes - 4th International Provenance and Annotation Workshop, IPAW*. 67–81. https://doi.org/10.1007/978-3-642-34222-6_6
 - [42] Paolo Missier, Khalid Belhajjame, and James Cheney. 2013. The W3C PROV family of specifications for modelling provenance metadata. In *Joint 2013 EDBT/ICDT Conferences, EDBT '13 Proceedings, Genoa, Italy, March 18-22, 2013*, Giovanna Guerrini and Norman W. Paton (Eds.). ACM, 773–776. <https://doi.org/10.1145/2452376.2452478>
 - [43] F. Mörchen. 2007. Unsupervised Pattern Mining from Symbolic Temporal Data. *SIGKDD Explor. Newsl.* 9, 1 (2007), 41–55.
 - [44] Guillem Santos, Joan Masó, Alaitz Zabala Torres, Lluís Pesquer, and Xavier Pons. 2019. A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: Conceptual model and implementation. *Transactions in GIS* 23 (07 2019). <https://doi.org/10.1111/tgis.12555>
 - [45] Eddie Schwalb and Lluís Vila. 1998. Temporal Constraints: A Survey. *Constraints An Int. J.* 3, 2/3 (1998), 129–149. <https://doi.org/10.1023/A:1009717525330>
 - [46] Yogesh Simmhan, Beth Plale, and Dennis Gannon. 2005. A survey of data provenance in e-science. *SIGMOD Rec.* 34, 3 (2005), 31–36. <https://doi.org/10.1145/1084805.1084812>
 - [47] Richard T. Snodgrass, Michael H. Boehlen, Christian S. Jensen, and Andreas Steiner. 1996. Adding Transaction Time to SQL/Temporal. change proposal, ANSI X3H2-96-502r2, ISO/IEC JTC1/SC21/ WG3 DBL MAD-147r2.
 - [48] Richard T. Snodgrass, Michael H. Boehlen, Christian S. Jensen, and Andreas Steiner. 1996. Adding Valid Time to SQL/Temporal. change proposal, ANSI X3H2-96-501r2, ISO/IEC JTC1/SC21/ WG3 DBL MAD-146r2.
 - [49] Stephen Stead, Martin Doerr, Christian-Emil Ore, Athina Kritsotaki, et al. 2019. CRM_{inf}: the Argumentation Model. An Extension of CIDOC-CRM to support argumentation. <http://www.cidoc-crm.org/crminf/ModelVersion/version-10.1> Last accessed May. 2020.
 - [50] Lluís V. and Lluís G. 1994. On Fuzzy Temporal Constraint Networks. *Mathware and Soft Computing* 3 (1994), 315–334.