


Enhancing CIDOC-CRM Models for GeoSPARQL Processing with MapReduce

Sara Migliorini

Department of Computer Science, University of Verona, Italy

sara.migliorini@univr.it

 <https://orcid.org/0000-0002-1825-0097>

Abstract

Spatial and temporal dimensions are two important characteristics of archaeological data and cultural heritage in general. The ability to perform some sort of reasoning on them is crucial during the analysis and interpretation process performed by domain experts. Many models have been defined in literature in order to properly describe such data and support the following interpretation process; among them, CIDOC CRM is a formal ontology specifically developed to represent cultural heritage information and many extensions have been proposed in recent years in order to enrich such model. In particular, CRM_{geo} tries to bring the gap between the cultural heritage domain and the geo-spatial domain, by providing a link towards GeoSPARQL and by defining the necessary constructs for the representation of spatial data types and relations. Unfortunately, the current support to the process of spatial functions through SPARQL query engine is still limited and many performance problems remain. The aim of this paper is twofold: (i) to evaluate the applicability of CRM_{geo} in representing spatial characteristics and relations of archaeological objects, and (2) to propose a MapReduce procedure able to efficiently derive spatial relations between objects, in order to automatically enhance an RDF model with them and avoid the performance issues derived from the use of GeoSPARQL query engine.

2012 ACM Subject Classification Information systems → Ontologies

Keywords and phrases Archaeological data, CIDOC CRM, GeoSPARQL, RDF, MapReduce

Digital Object Identifier 10.4230/LIPIcs.COARCH.2018.

Acknowledgements This work was partially supported by the Italian National Group for Scientific Computation (GNCS-INDAM). This work has been supported by “Progetto di Eccellenza” of the Computer Science Department, University of Verona, Italy.

1 Introduction

Many models have been proposed in literature in order to represent cultural heritage information [5, 14, 17, 20], all of them share the presence of some constructs for the representation of spatio-temporal aspects of historical events or remains. Among them CIDOC CRM [15] is an ISO Standard which defines a formal ontology for dealing with cultural heritage information, and $CRM_{archaeo}$ [7] is an extension for supporting the archaeological excavation process and all the various entities and activities related to it.

In [13] we introduce a mapping from a spatio-temporal archaeological model, called *Star* [14], to $CRM_{archaeo}$. In particular, we shown how spatial and temporal dimensions of archaeological concepts can be mapped into CIDOC CRM and $CRM_{archaeo}$ classes. In particular, we concentrate on the spatio-temporal aspects of archaeological information, since these aspects play an important rule during the analysis and interpretation process performed by archaeologists. Indeed space and time are able to reveal important information about the object properties and to discover important relations between objects. However, even

Enhancing CIDOC-CRM Models for GeoSPARQL

if CIDOC CRM and its extension $CRM_{archaeo}$ provide some classes for dealing with spatio-temporal aspects, like for instance E53 Place or E4 Period, the support for the representation of spatio-temporal concepts can be considered limited w.r.t. other geospatial standards, like the ones proposed by OGC, as deeply discussed in [10].

For these reasons, an extension of CIDOC CRM has been developed, called CRM_{geo} [11], with the aim to bring the gap between the standards of the geospatial and the cultural heritage community, in particular it provides a link from GeoSPARQL to CIDOC CRM. It enriches cultural heritage data with precise and well identified descriptions of locations and geometry sites of historical events and remains. GeoSPARQL [16] provides not only a set of constructs for representing features, geometries and their relationships, but also a set of spatial functions for use in SPARQL queries. However, despite the potentiality of GeoSPARQL and the definition of some query engine able to process spatial functions, many problems exist which make not feasible their computation [22]. For these reasons many approaches have been developed in literature in order to increase the performance of a GeoSPARQL query engine [23, 1] which try to define additional software modules able to increase its performance during the processing of spatial functions, for instance through the definition of some sort of online indexes. These problems limit the applicability of GeoSPARQL in real situations, and this is particularly true in a distributed heterogeneous context, where different agencies can share their data, thanks to the use of a standard model (like CIDOC CRM), but cannot made any assumption about the particular software implementation adopted by each of them for querying this shared model.

Besides to these technological considerations, archaeological data also pose additional challenges to the application of existing systems for performing spatio-temporal analysis on them, due to their inherent vagueness and incompleteness. As we will see in the next sections, in the archaeological domain the spatial location and/or extent of an object can be defined in different ways: sometimes it is possible to have a correct specification of its geometry, other times the location is described through a place appellation, and finally it is not rare to know only the spatial relation existing between some objects, without knowing anything about its real location. As discussed in [3, 4] it is a common practice in archaeology to represent topological relations between objects without having a realization of their geometries. Therefore, it is clear that testing the existence of a particular topological relation may become a difficult task, not only due the amount of data to be processed, but also because some relations have to be retrieved from geometries, other from the explicit specification of properties, and other from a mix of these two aspects.

The aim of this paper is twofold: (i) to evaluate the applicability of CRM_{geo} for the representation of spatial characteristics of archaeological information, referring to a conceptual model called *Star* (Sect. 3), and (ii) to propose a procedure to be applied before the mapping of a conceptual model to CIDOC-CRM, in order to automatically discover the spatial relations existing between objects and represent them as properties in the classical triple RDF format (Sect. 4). In this way, any SPARQL query engine can be applied to infer new knowledge, without any additional overhead, since all the necessary spatial derivations are performed off-line immediately after the mapping process.

2 Related Work

The problem of processing spatio-temporal data with SPARQL has been widely treated in literature. In [1] the authors provide an overview of the current state of the art in industry and research about geo-spatial data in the Semantic Web, with a focus on GeoSPARQL.

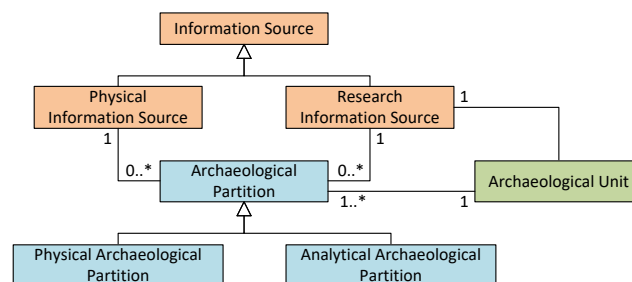
In particular, they analyse the limited support to GeoSPARQL provided by existing triple stores with a reference to the processing of the prescribed spatial function. They point out that sometimes existing systems provide similar custom functionalities or only a subset of them. They also concentrate on the Parliament system in order to provide a fully support to GeoSPARQL. As regards to the efficiency issues in processing spatial functions, in [23] the authors propose an approach based on the creation of spatial indexes on-the-fly to prune the search space and the parallelization of join computations within queries.

A different approach to solve the efficiency problems is based on the observation that queries are usually evaluated completely either on the server or on the client side, without any alternative between them. For these reasons in [19] the author proposed the concept of Triple Pattern Fragment (TPF) which provides a compromise consisting in breaking down queries into simple queries that the server needs to return and the client uses to compute the final result. In [6] the authors propose an extension of such idea to support client-side processing of GeoSPARQL functions, the main intuition is to store the geometries in a simple geospatial database on the client-side and compute the geospatial predicates on it.

Finally, in recent years some attempts have been made in order to use MapReduce environment, such as Hadoop or Spark, for processing SPARQL query [12, 18]. However, these works concentrate only on SPARQL without its spatial extension, so they can be good candidate in processing the RDF produced by our transformation procedure proposed in Sect. 4, but not GeoSPARQL in general.

3 Mapping of the *Star* Model to CIDOC CRM

The *Star* (Spatio-Temporal Archaeological) model has been developed in order to consistently collect, record and process archival documents (reports, plans, drawings, photographs and other materials), excavations processes and other archaeological researches (field surveys, geophysical prospections, etc.), archaeological findings and remains. In particular, information can come from both the archives of an archaeological agency and from data recorded in publications, manuscripts and maps. The kernel of the *Star* model is composed of three main concepts: information source, archaeological partition and archaeological unit, which are all characterized by some spatial and temporal dimensions. These concepts and their relationships are depicted in Fig. 1 through a UML class diagram, where arrows with a white big head denotes hierarchies, while lines represent relations with their cardinality.



■ **Figure 1** Hierarchy of the main classes contained in the *Star* model with their relationships.

This section briefly introduces these concepts and provides an overview of their mapping towards the CIDOC CRM standard and two of its extensions, CRM_{archaeo} and CRM_{geo}, with a particular attention to the representation of their spatial characteristics. Fig. 2 illustrates the hierarchy of classes considered in this paper, where blue boxes are classes from

Enhancing CIDOC-CRM Models for GeoSPARQL

GeoSPARQL, purple boxes are classes from CIDOC CRM, green boxes are class of CRM_{geo}, gray boxes are classes of CRM_{sci} (included also in CRM_{archaeo}) and pink boxes are classes of CRM_{archaeo}. As you can notice, classes of CRM_{geo} represent a link between CIDOC CRM and GeoSPARQL. In particular, the concept of space-time volume has been originally proposed by CRM_{geo} and then included into the set of classes of CIDOC CRM starting from version 6. The main idea behind CRM_{geo} is the distinction between phenomenal and declarative space-time volumes. A Phenomenal Spacetime Volume defines a 4 dimensional fuzzy point set (volume) which a material phenomena (event or physical thing) occupies in space and time. Its spatio-temporal extent is unique but is unknown and unobservable in an exact manner. As regards to the space dimension, it is represented by a phenomenal place which derives its identity from the event or the physical thing it comes from. Since a phenomenal place cannot be exactly observed or determined, it can only be approximated through a declarative place. A declarative place can be derived from a measurement of some points related to a physical feature or as a result of an interpretation of a place on a map.

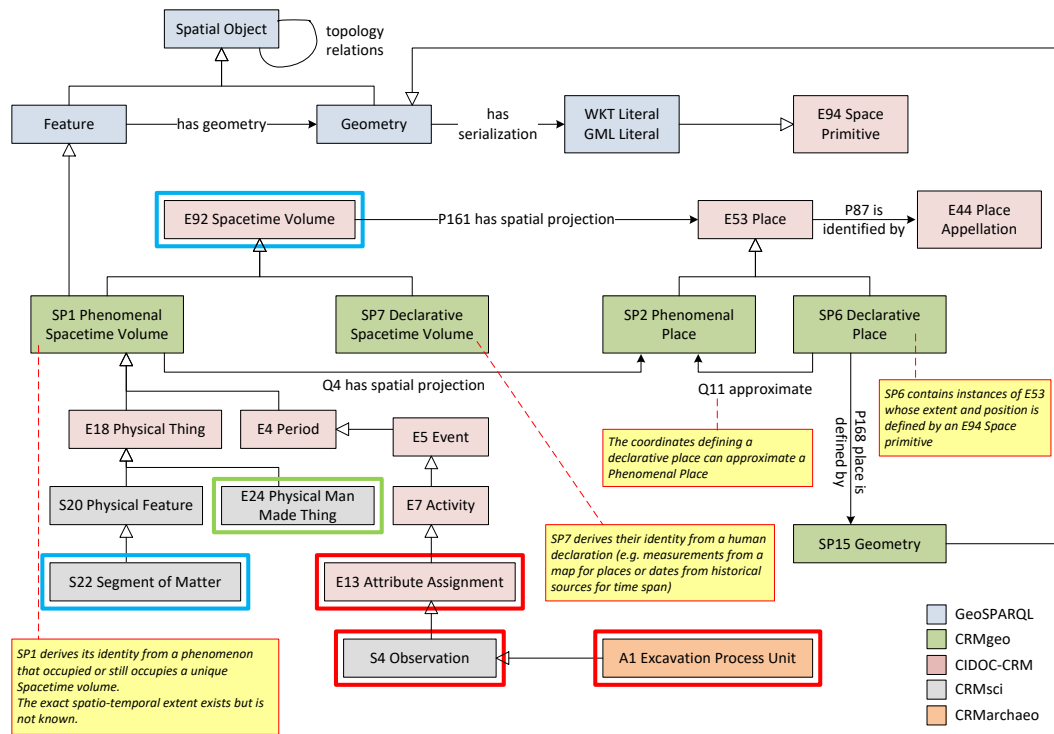


Figure 2 Hierarchy of classes coming from the CIDOC CRM, CRM_{archaeo}, CRM_{sci} and CRM_{geo} considered in this paper.

Referring to Fig. 2, the class E92 Spacetime volume has two sub-classes SP1 Phenomenal Spacetime Volume and SP7 Declarative Spacetime Volume which in particular specialize the property P161, so that an instance of SP1 can only have a spatial projection which is a SP2 Phenomenal Place, while an instance of SP7 can only have a spatial projection which is a SP6 Declarative Place. Moreover, an instance of SP2 does not have a direct representation in terms of geometric primitives, but it can only be approximated (property Q11) by an instance of SP6. Only instances of SP6 can be defined in terms of geometric primitives through the property P168 place is defined by towards an instance of SP15 Geometry. This last class is the link between the hierarchy of CRM_{geo} and GeoSPARQL, since an instance

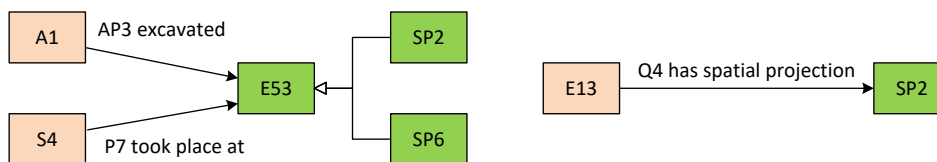
of SP15 is essentially an instance of *Geometry* which can have a serialization (representation) in WKT or GML. Notice that Fig. 2 highlights another property of E53 which is P87 is identified by towards an instance of E44 *Place Appellation*, because in many cases the location of an archaeological object can only be described in terms of its address or historical name, specially when the information came from ancient documents. As we will see in the next sections, all these concepts and the distinction between phenomenal and declarative places will be very important during the mapping of archaeological data.

3.1 Information Source

An *information source* (IS) represents the way used to start collecting information about an archaeological finding or remain. It is a very general class of objects and its instances are classified by the acquisition methodology that characterizes them. In particular, we distinguished between (i) sources, which describe a physical process of data collection and (ii) research studies, which analyse documents and other literary sources, obtaining two sub-classes: *physical information source* and *research information source*, as shown in Fig. 1.

These sub-classes have been mapped towards CRM_{archaeo} depending on their different acquisition methodology value into: A1 *Excavation Process Unit*, E13 *Attribute Assignment* and S4 *Observation*. In particular, instances of A1 identify physical ISs related to some kind of excavation (both extended or sample). A1 is a good candidate for the representation of this kind of ISs since it “comprises activities of excavating in the archaeological sense which are documented as a coherent set of actions of progressively recording and removing matter from a pre-specified location under specific rules”. Indeed, distinct ISs are instantiated for each excavation activity performed on a given area which is homogeneous in space, time and acquisition methodology. Conversely, instances of A9 are used to represent a coordinated set of excavation activities related to a broader area. Instances of S4 have been chosen for all other classes of physical ISs in which some kind of physical survey on the territory has been carried out but it was not an excavation. Finally, instances of E13 describe research ISs, in particular they represent the specific case of monograph study applied/undertaking w.r.t. the unambiguous identification and description of an archaeological monument or complex.

These three classes have been highlighted with a red frame in Fig. 2, and as you can notice, they are all sub-classes of E7 *Activity* which in turn is a sub-class of E92 *Spacetime Volume*, namely they have a spatial and temporal projection. Even if both aspects are very important in discovering relations between archaeological objects and perform some kind of interpretation, this paper concentrates only on the spatial one, even if a similar reasoning can be made also for the temporal counterpart.



■ **Figure 3** Possible mappings of the spatial characteristics of an information source.

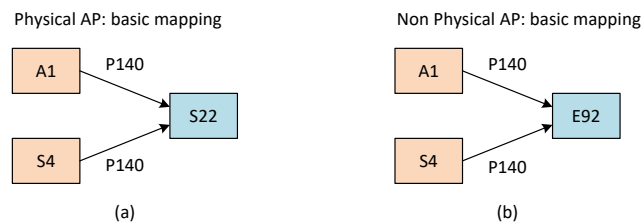
As regards to S4 and A1, they both inherit the property P161 *has spatial projection* from their super-class SP1 which goes towards an instance of SP2 *Phenomenal place*. However, they also have a specialized property towards a spatial attribute (E53 *Place*), namely P7 *took place at* for S4, and AP3 *excavated* for A1, which are considered during the mapping of *Star*, as illustrated in Fig. 3. The nature of this place can be different depending on the nature of

Enhancing CIDOC-CRM Models for GeoSPARQL

the original information, for instance in case of a contemporary excavation, its location and extension can be derived from performed measurements and mapped to a declarative place. Conversely, in other cases, such as excavations retrieved from ancient documents, its location and extension can be described in terms of the occurrence of a particular phenomenon. Finally, as regards to E13, in this case there is not a measured geometry and its spatial projection is represented through property Q4 has spatial projection towards a SP2 instance. Notice that independently from the kind of place used, the location of an information source can be described (also or only) through a place appellation, namely an official address or an historical name commonly used to identify the place in the past.

3.2 Archaeological Partition

An *archaeological partition* (AP) concerns the scientific description of an archaeological finding/remain of structural and non-structural nature, provided that it has an informational value in ancient topography terms, at least also minimal or uncertain at the first recording time. AP represents a very flexible concept used to describe observations at different level of refinement. Therefore, different types of mapping are possible depending on the level of refinement of the represented information. In the general case, an AP is mapped to an instance of S22 Segment of Matter with the only exception of the case an AP does not represent a physical observation, but it is used to describe a hypothesis of reconstruction or other piece of information produced by an interpretation process usually based on some performed analysis; in this case an AP is represented by an instance of E92 Spacetime Volume. The link towards the instances of A1 or S4, representing the related information source, is obtained using the property P140i was attributed by (*i* stands for inverse of the mentioned property), as illustrated in Fig. 4.



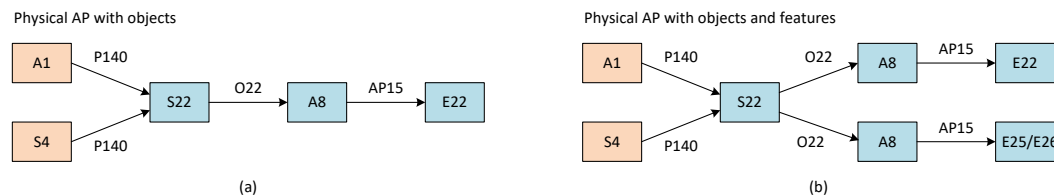
■ **Figure 4** Possible mappings of an archaeological partition and linking towards the related information source.

As regards to the choice of the S22 class, we can notice that the AP semantics finds a very good correspondence with the concept of segment of matter explicitly described in CRM_{archaeo} and CRM_{sci}. In particular, as specified in the standard, S22 comprises “*physical material in a relative stability of form (substance) within a specific spacetime volume (unity, extend)*”: this is a fundamental binomial aspect in AP semantics within the first phase of the identification and aggregation process. Indeed, it is strictly related to (i) the *substance* of an AP, which is normally an aggregate of structures and sometimes also of nonstructural stratigraphic units; (ii) its *space* and *time* boundaries, which are usually well defined and determined through the process identified by its related information source.

In this paper we concentrate only on the spatial aspect of an AP and regarding this, the standard says that the “*spatial extend of a S22 Segment of Matter is defined by humans usually because the constellation is subject to a specific interest for and investigations of the geometric arrangement of physical features or parts of them on or within the specified S22 Segment of*

Matter”: this is another peculiar, methodological aspect of the identification and aggregation of each AP, especially when it is analyzed during a subsequent elaboration process. Finally, relatively to the discovery of an AP, the standard says that an instance of S22 comes “*into existence as being an object of discourse through S4 Observation or declaration*”: exactly as it happens during the process described by the corresponding information source (regardless of the type of S4 Observation, namely the kind of information source).

When an AP contains evidences of man made objects, we add to the instance of S22 the property O22 partially or completely contains an instance of A8 Stratigraphic Unit which is connected to an instance of E22 Man Made Object through the property AP15 is or contains remains of, as illustrated in Fig. 5.a. Moreover in some specific cases, when the partition also contains evidence of features, we add also the property: O22 partially or completely contains an instance of A8 Stratigraphic Unit as before, but in this case it refers to an instance of E25 Man Made Feature (or E26 Physical Feature) again with the property AP15 is or contains remains of, as illustrated in Fig. 5.b.



■ **Figure 5** (a) Mapping of an AP containing an evidence of a man made object, and (b) mapping of an AP containing evidence of both a man made object and a man made/physical feature.

S22 Segment of Matter is a sub-class of S20 Physical Feature which in turn is a sub-class of SP1 Phenomenal Spacetime Volume, as reported in Fig. 2. Therefore, it follows that in general the spatial extent of an AP is represented by an instance of SP2, indeed it is unique but may be unknown or unobservable into an exact manner. This consideration is true also (and mainly) for the other kind of archaeological partitions which represent a reconstructive hypothesis and are represented as instances of E92. For this reason, we decide to represent them always as instances of its sub-class SP1.

Notice that a strictly connection exists between the spatial extent of an AP and of its corresponding IS, namely the spatial extent of an AP has to be contained into the spatial extent of its IS [4]. In some cases, the extent of an AP can be approximated by a declarative place without a geometric realization but with a containment topological relation with the declarative place of its corresponding IS, other times it can have a more restrictive approximation defined by a declarative place with a geometric realization. The same considerations made about the specification of the spatial extent of an IS through an address (E44 Place Appellation) can be made also for APs.

3.3 Archaeological Unit

An archaeological unit (AU) is a class of objects representing any archaeological complex or monument obtained from an interpretation process performed by the responsible officer. Such an interpretation is carried out based on some findings, represented by archaeological partitions, retrieved during an excavation process, or a bibliographical analysis or other investigation process, described by an information source. Conventionally, an archaeological unit is identified by the logical union of many archaeological partitions, which can be analyzed together producing an unambiguous archaeological monumental context (e.g. a

Enhancing CIDOC-CRM Models for GeoSPARQL

specific ancient building).

Instances of AU are mapped to instances of the class E24 Physical Man Made Thing which is a subclass of E18 Physical Thing which in turn is an instance of SP1 Phenomenal Spacetime Volume, as illustrated in Fig. 2. Therefore, the spatial extent of an AU is represented through an instance of SP2 Phenomenal Place. The same considerations made for the spatial extent of an AP and its relation with the extent of its IS also apply to the case of an AU.

4 Archaeological Spatial Analysis

As discussed in [3, 4], analysis and interpretation are two of the main activities performed by archaeologists, and these activity typically involves the spatial and temporal characteristics of archaeological records, consider for instance the use of the Harrix matrix. Indeed, these characteristics may reveal many relevant relations between findings and it is not uncommon that the dating of one remain derives from the dating of (spatially) neighboring objects. Therefore, the ability to perform spatial queries on the collected data is an important activity, and the use of standards, like CIDOC CRM, promotes interoperability and allows to perform such analysis on data coming from different sources and collected by different agencies.

This section briefly introduces GeoSPARQL highlighting some issues that can prevent its use in real world archaeological domains, such as the mapping presented above, and then proposes a solution based on a MapReduce procedure which automatically discovers spatial relations between findings that may be relevant for a subsequent processing.

4.1 GeoSPARQL and Spatial Analysis

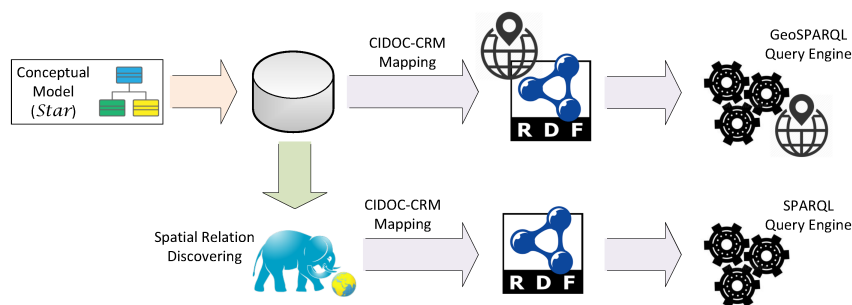
As discussed in Sect. 1, CRM_{geo} is an extension of CIDOC CRM which has been developed with the aim to provide a link towards GeoSPARQL. The OGC GeoSPARQL [16] supports the representation and querying of geospatial data on the Semantic Web. In particular, it defines a vocabulary to represent features, geometries and their relationships in RDF, and it defines an extension of the SPARQL query language for processing geo-spatial data. As illustrated in Fig. 2, GeoSPARQL includes two different ways to represent geometry literals and their associated type hierarchies: WKT and GML which are linked through the properties `geo:asWKT` and `geo:asGML`, respectively.

GeoSPARQL includes a standard way to check the existence of a topological relationship between features. In particular, such relationships are given in form of binary properties between the entities and geo-spatial filter functions. These properties can be used in SPARQL query triple pattern like a normal property and expressed using three distinct vocabularies: the OGC's Simple Feature, Egenhofer's 9-intersection model, and RCC8. Clearly, in order to evaluate such properties, it is necessary that the chosen RDF query engine understand the GeoSPARQL ontology and provide support for its spatial functions. As discussed in Sect. 2, such support is not uniform and each GeoSPARQL query engine can implement only one of three vocabularies mentioned above, or eventually implement custom similar functions. Moreover, as discussed in Sect. 1, spatial data retrieval using Geo-SPARQL can be very inefficient and other additional issues can arise in presence of archaeological data processing.

During the analysis of different remains the main operation to be performed is the computation of the spatial join between different objects in order to identify the topological relation between them, or to select all objects which are in a particular topological relation with another one. This is because many new knowledge can be achieved from neighbour object, for instance as regards to the dating [3, 4] and classification processes.

Relatively to the mapping presented in Sect. 3, the place representing the location and/or the extent of a feature can be defined in different ways: sometimes it is possible to have a correct specification through a geometry expressed using a space primitive, other times the location can be described through a place appellation, or again the location can only be instantiated but without a geometry. As discussed in [3, 4], vagueness and incompleteness are key concepts in archaeology, sometimes only the relations existing between objects are known without any knowledge about their real locations. Therefore, it is a common practice to represent topological relations between objects without having a representation of their geometries. In this case, the topological relation cannot be directly derived from the geometries but has to be explicitly defined [2]. Moreover, sometimes the level of uncertainty is so high that it is necessary to represent a disjunction of topological relations as a set of relations that are equally possible, but within which it is not possible to choose. This is particularly true in case of an information collected starting from ancient documents which can be vague and without a precise geo-localization of objects.

Therefore, it is clear that testing the existence of a particular topological relations can become a difficult task, not only for the amount of data to be processed, but also because some relations have to be retrieved from geometries, other from the explicit specification of properties, and other from a mix of the two aspects. For this reason, in this paper we propose a procedure to be applied before the mapping from a conceptual model, like *Star*, to CIDOC-CRM in order to automatically discover the relations existing between objects and represent them as properties in the classical triple RDF format. In this way, any SPARQL engine can be applied to infer new knowledge, and no performance issues related to spatial processing remains, since all the necessary spatial derivations are performed off-line during the transformation process.



■ **Figure 6** Possible architectures for the analysis of archaeological data using SPARQL, the top row represents the classical approach based on the use of spatially-enhanced SPARQL query engine, while the bottom row represents the alternative approach proposed in this paper.

Fig. 6 shows the different kind of architectures that can be implemented. The top row represents the traditional architecture proposed in literature, where the original data (for instance contained into a relational database) is mapped to CIDOC CRM obtaining an RDF enhanced with geospatial aspects, which has to be processed by a GeoSPAQL query engine. This solution presents the limitations discussed above. Conversely, the bottom row depicts the solution proposed by this paper: in this case the original data is initially processed by a MapReduce procedure which identifies all existing topological relations between objects, instantiating a corresponding property. The obtained RDF file contains some spatial constructs, in particular spatial topological relations, but they can be treated as traditional property and processed by plain SPARQL query engine. The following section discusses in details such MapReduce procedure, called *Spatial Relation Discovery*.

4.2 Spatial Relation Discovering with MapReduce

The MapReduce procedure proposed in this paper has the main aim to discover the topological spatial relations existing among all ISs, APs, AUs, and between any IS with every AP or AU, and between any AP with every AU. Notice that, while the relation existing between an AP/AU and its related IS is always known (it has to be a containment relation), this procedure wants to discover also the relation existing between an IS and any AP, even the ones not related to it.

Since the amount of data to be processed can be very huge, we need to exploit some form or parallelism in order to perform it in an efficient way. For this reason we propose to use a MapReduce environment, whose main idea is to subdivide the data into small chunks on which the same initial operation is performed in parallel (map phase), eventually followed by a final operation (reduce phase) which combines the partial results built so far. In particular, we consider SpatialHadoop [8], a spatial extension of Apache Hadoop [21] which provides support for the representation of geometric data types and the execution of spatial functions. In particular, SpatialHadoop contains several different implementations of the spatial join algorithm [9], each one based on a different use of indexes and repartition techniques. In this paper we can safely abstract from a particular implementation, considering the presence of a general spatial join, which is called Distributed Join (DJ). The main observation to be done is that the existing join operators evaluate only the intersection between spatial objects, and they do not determine the specific topological relations existing among them. Conversely, for the purposes of this paper, we are also interested in identifying the specific topological relation. For this reason, we use a modified version of DJ, called here EDJ (Enhanced Distributed Join), which returns not only the pair of intersecting objects, but also the relation existing between them.

The proposed procedure consists of two MapReduce Job, the first one will transform each place appellation into a geometric representation, while the second one will identify the topological relation existing between each pair of objects. Notice that, several locations can be defined for an object, for instance by using multiple instances of the same property, or because a place is defined both in terms of its appellation and a geometry. In this case, we can derive multiple topological relations between the same pair of objects, which have to be all considered as possible alternatives. The possibility to have multiple topological relations (i.e., RDF properties) between the same pair of objects is a way to represent a certain degree of uncertainty which is typical in the archaeological domain, in particular in presence of partial or incomplete information. Future work will regard the study of a more complete way to represent fuzzyness and uncertainty in the spatial and temporal representation of archaeological properties and how to perform some sort of reasoning on them.

In order to apply these MapReduce jobs, it is necessary to extract from the original database two sets of CSV files that can be processed by SpatialHadoop. The first set will contain the places directly connected to an IS, AP or AU whose location and extent is represented by a geometry (datasets D_{geo}^{is} , D_{geo}^{ap} and D_{geo}^{au}). Notice that only the places directly connected to the three main classes are considered, so in most cases they will be phenomenal places which have an indirectly attached geometry through a declarative place, while only in case of a place connected to an information source (i.e., instances of A1 or S4), we can be in presence of a declarative place. Each of these datasets will contain a record for each place composed of its identifier and its geometric representation in WKT, plus a flag denoting if the geometry approximates a phenomenal place or it directly refers to a declarative place. The second set will contain the instance of places directly connected to an IS, AP or AU whose location is represented by a place appellation (datasets D_{addr}^{is} , D_{addr}^{ap}

S. Migliorini

and D_{addr}^{au}). In this case each record contains the feature identifier, the place appellation, and again a flag denoting if the appellation is defined on a phenomenal or a declarative place. If object has a spatial projection defined by both a geometric representation and a place appellation, it will be contained in both datasets. Moreover, if more than one place is defined for a feature, a record will be created for each of them in the corresponding datasets. Notice that these two set of CSV files do not contain the complete serialization of IS/AP/AU, but only the extraction of their spatial properties.

For each dataset D_{addr}^* , the job ADDRDECODER will determine a representative geometry for such address (i.e., place appellation). Each produced result will be added to the corresponding dataset D_{geo}^* . In particular, in accordance with the MapReduce specification, a mapper receives a chunk of data (split) containing a set of records in the form of pairs $\langle key, value \rangle$. In this case, the *key* is the feature identifier, while *value* is a sequence of attributes containing the place appellation and the mentioned flag (accessed by *isDeclarative()*). For each of these records, the mapper retrieves the place appellation from the value and uses the procedure GEOENCODER to determine a symbolic geometric representation for it. For instance, a street address can be represented by a polygon containing all the street area, while the name of a city can be represented by a polygon covering the whole city. No reducer is needed to combine the result of this work. The partial results produced by the mappers do not need any further elaboration to obtain the final result. We assume that the final result produced by ADDRDECODER for dataset D_{addr}^* is stored into a dataset E_{geo}^* that will be added to the corresponding D_{geo}^* .

Algorithm 1: ADDRDECODER job.

```

1 class MAPPER
2   method MAP( $\langle key, value \rangle$ )
3      $n \leftarrow value.appellation$ 
4      $g \leftarrow GEOENCODER(n)$ 
5     return  $\langle key, \langle g, value.isDeclarative() \rangle \rangle$ 

```

Given the enriched datasets D_{geo}^* produced by combining the original content of D_{geo}^* with E_{geo}^* , we build for each of them a spatial index using the functionalities provided by SpatialHadoop. Even in this case, we can safely abstract from the particular kind of spatial index to build, indeed several different indexes are available such as quadtree, rtree, and so on. Assume only that for each dataset D_{geo}^* , a corresponding index I_{geo}^* is available on which the following EDJ job will work. In particular, given the indexes I_{geo}^{is} , I_{geo}^{ap} and I_{geo}^{au} , the job will be applied for the pairs $I_{geo}^{is} \times I_{geo}^{is}$, $I_{geo}^{is} \times I_{geo}^{ap}$, $I_{geo}^{is} \times I_{geo}^{au}$, $I_{geo}^{ap} \times I_{geo}^{ap}$, $I_{geo}^{ap} \times I_{geo}^{au}$ and $I_{geo}^{au} \times I_{geo}^{au}$. This is again a map-only job and its core is the test which determines the topological relation existing between each pair of geometries.

Notice that this procedure is a modified version of the DJ algorithm provided by SpatialHadoop which returns not only the pairs of intersecting objects, but also the kind of topological relation existing between them. Moreover, since the procedure exploits the use of indexes, only pairs of possible intersecting geometries are compared, it follows that all pair of geometries not contained in the result can be considered disjoint. In particular, in accordance with DJ, each mapper works on a combined split which contains as a key a pair of keys, and as value a pair of values, both coming from the two input datasets I_{geo}^1 and I_{geo}^2 . Through the use of a filter, SpatialHadoop ensures that a combined split is built only between pair of splits with intersecting cells. More specifically, for each index I_{geo}^* , the

Enhancing CIDOC-CRM Models for GeoSPARQL

key is represented by the geometry of the index cell, while the value is presented by a list of feature whose geometry intersects that cell. Therefore, given the input of a mapper in Alg. 2, $\langle\langle key_1, key_2 \rangle, \langle list_1, list_2 \rangle\rangle$, the combined key $\langle key_1, key_2 \rangle$ is a pair of cells such that $key_1 \in I_{geo}^1$, $key_2 \in I_{geo}^2$ and $key_1 \cap key_2 \neq \emptyset$, while $list_1 \subseteq D_{geo}^1$ and $list_2 \subseteq D_{geo}^2$ such that the geometry of each element in $list_i$ intersect key_i , for $i \in \{1, 2\}$.

Algorithm 2: EDJ job.

```

1 class MAPPER
2   method MAP( $\langle\langle key_1, key_2 \rangle, \langle list_1, list_2 \rangle\rangle$ )
3      $k \leftarrow key_1 \cap key_2$ 
4      $l_1, l_2 \leftarrow \emptyset$ 
5     foreach  $f \in list_1$  do
6       if  $intersect(f.geo, k)$  then
7          $l_1 \leftarrow l_1 \cup \{f\}$ 
8     foreach  $f \in list_2$  do
9       if  $intersect(f.geo, k)$  then
10         $l_2 \leftarrow l_2 \cup \{f\}$ 
11     $l_1 \leftarrow SORT(l_1); l_2 \leftarrow SORT(l_2)$ 
12     $i, j \leftarrow 0$ 
13    while  $i < |l_1| \wedge j < |l_2|$  do
14      if  $MBR(l_1[i].geo).x_1 < MBR(l_2[j].geo).x_1$  then
15         $j' \leftarrow j$ 
16        while  $j' < |l_2| \wedge MBR(l_2[j'].geo).x_1 \leq MBR(l_1[i].geo).x_2$  do
17          if  $intersect(MBR(l_1[i]), MBR(l_2[j']))$  then
18            return  $\langle s, \langle l_1[i], l_2[j'] \rangle, topoRel(l_1[i], l_2[j']) \rangle\rangle$ 
19           $j' \leftarrow j' + 1$ 
20         $i \leftarrow i + 1$ 
21      else
22         $i' \leftarrow i$ 
23        while  $i' < |l_1| \wedge MBR(l_1[i'].geo).x \leq MBR(l_2[j].geo).x_2$  do
24          if  $intersect(MBR(l_1[i']), MBR(l_2[j]))$  then
25            return  $\langle s, \langle l_1[i'], l_2[j] \rangle, topoRel(l_1[i'], l_2[j]) \rangle\rangle$ 
26           $i' \leftarrow i' + 1$ 
27         $j \leftarrow j + 1$ 

```

For each compound value, a mapper first computes the intersection between the two keys, namely between the cells of the two indexes (line 3). As stated above, the combined split is built so that they have a not empty intersection. Then, the two lists of features in input are filtered w.r.t. this window k (lines 5-10). The obtained lists l_1 and l_2 are sorted based on the x values of their geometry MBR (line 11). Given such sorted list of feature, a plane-sweep like algorithm is applied (lines 13-27) for efficiently comparing the contained geometries. In particular, $intersect(g_1, g_2)$ is a function that returns true or false, depending on whether the intersection between g_1 and g_2 is not empty or empty, respectively; while $topoRel(g_1, g_2)$ is a function that returns the topological relation existing between g_1 and g_2 . Therefore,

S. Migliorini

given a pair of geometries whose MBRs have a not empty intersection (line 17 or line 24), the topological relations existing between them is computed (line 18 or line 25).

EDJ is again a map-only job, namely the partial results produced by the mappers do not need any further process. Each final result produced by EDJ on a different pair of input datasets is stored into a corresponding dataset D_{topo} , where the key is a dummy serial index (see value s in lines 19 and 25 whose declaration and update have been omitted for not cluttering the algorithm), while the value is composed of a pair of objects and the topological relation existing between their geometries. Notice that, since an object f can be replicated multiple time inside the same input dataset D_{geo}^* , due to its multiple spatial representations, multiple different topological relations can be produced between the same pair of objects. All these relations are considered as possible alternatives equally valid.

Algorithm 3: SPATIALLYENHANCERDF algorithm.

```
Input:  $E_{geo}^*, D_{topo}, G$ 
1 foreach  $t \in E_{geo}^*$  do
2    $dp \leftarrow \perp$ 
3   if  $t.geo.isDeclarative()$  then
4      $dp \leftarrow G.get(t.id)$ 
5   else
6      $dp \leftarrow SP6; pp \leftarrow G.get(t.id)$ 
7      $G.add(\langle dp, Q11, pp \rangle)$ 
8    $G.add(\langle dp, P168, SP15 \rangle)$ 
9    $G.add(\langle SP15, "has serialization", t.geo \rangle)$  ▷ GeoSPARQL property
10 foreach  $r \in D_{topo}$  do
11    $d_1 \leftarrow G.get(r.first.id); d_2 \leftarrow G.get(r.second.id)$ 
12    $G.add(\langle d_1, r.topo, d_2 \rangle)$ 
13 return
```

The results produced by ADDRDECODER and EDJ can now be used to enhance the original model and to obtain a more complete RDF output containing the computed topological relations. Procedure SPATIALLYENHANCERDF in Alg. 3 illustrates how this can be done, it receives as input: dataset E_{geo}^* produced by ADDRDECODER, dataset D_{topo}^* produced by EDJ, and the preliminar RDF graph G obtained from the original datasets. The procedure initially processes the content of E_{geo}^* (lines 1-9). For each record in E_{geo}^* a declarative place will be defined in the following way: if the place appellation is connected to a phenomenal place pp , a declarative place dp is built and connected to pp through property Q11 (lines 6-7); otherwise, the declarative place dp is retrieved (line 4). Given the declarative place dp , it will be connected to an instance of SP15 through property P168 (line 8), this geometry will have a WKT serialization corresponding to the geometric representation computed by procedure ADDRDECODER (line 9). The procedure then processes the content of D_{topo} , for each record $\langle dp_1, r, dp_2 \rangle$ it retrieves the declarative places dp_1 and dp_2 from the graph G and adds a property representing the topological relation r between them (lines 10-12).

5 Conclusion

Space and time are two important characteristics of archaeological data, since they can reveal important properties and relations among objects. Indeed, one of the main activities

Enhancing CIDOC-CRM Models for GeoSPARQL

performed by archaeologists is a deep analysis of the collected data in order to derive new knowledge starting from the available one. It is not uncommon that most of the analysis and interpretation process is based on the study of spatial and temporal relations between objects. Moreover, the archaeological process may greatly benefit from the sharing of information between different agencies. For this reason many efforts have been devoted to the definition of interoperable, standard, spatio-temporal model for archaeology. In this context, CIDOC CRM is one of these standards and several extensions have been defined in order to provide the necessary expressiveness in terms of both the representation of archaeological concepts (i.e., $CRM_{archaeo}$) and spatio-temporal dimensions (i.e., CRM_{geo}).

CRM_{geo} tries to provide a link between the archaeological and cultural heritage domain and the geo-spatial domain, defining a connection between the CIDOC CRM standard and the OGC GeoSPARQL standard. In this way, CIDOC CRM models can be enriched with all spatial types and properties provided by OGC and potentially with a series of functions for spatial analysis. The first contribution of the paper is the discussion of a possible mapping of a conceptual archaeological model, called *Star*, into CIDOC CRM using its geographical extension for the representation of spatial concepts.

However, as discussed in literature, performing spatial analysis on RDF is a cumbersome task for many reasons, and this can become even worse in the archaeological domain, where spatial information can also be incomplete, inaccurate and described in different ways also through appellations. For all these reasons, the second contribution of the paper is the definition of an alternative solution which tries to spatially enhance an RDF schema by automatically discovering all possible topological spatial relations between objects. The resulting RDF will contain all the necessary information and can be processed by any RDF engine in an effective way. Future work will regard the extensive test of such procedure on huge archaeological data in order to evaluate the effective benefits of the approach w.r.t. the traditional approach based on the use of GeoSPARQL query engines. Moreover, additional studies will be performed for including also temporal relations, besides to the spatial ones, and for efficiently reacting to changes in the original datasets, in order to avoid the computation of all relations from the beginning.

References

- 1 R. Battle and D. Kolas. Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. *Semant. web*, 3(4):355–370, 2012.
- 2 A. Belussi and S. Migliorini. A framework for integrating multi-accuracy spatial data in geographical applications. *GeoInformatica*, 16(3):523–561, 2012.
- 3 A. Belussi and S. Migliorini. A Framework for Managing Temporal Dimensions in Archaeological Data. In *Proceedings of the International Symposium on Temporal Representation and Reasoning*, TIME 2014, pages 81–90, 2014.
- 4 A. Belussi and S. Migliorini. A spatio-temporal framework for managing archeological data. *Annals of Mathematics and Artificial Intelligence*, 80(3-4):175–218, 2017.
- 5 B. De Roo, K. Ooms, J. Bourgeois, and P. Maeyer. Bridging archaeology and GIS: Influencing factors for a 4D archaeological GIS. In *Proc. of the 5th Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, pages 186–195, 2014.
- 6 C. Debruyne, E. Clinton, and D. O’Sullivan. Client-side Processing of GeoSPARQL Functions with Triple Pattern Fragments. In *Linked Data on the Web 2017*, pages 1–8, 2017.
- 7 M. Doerr, A. Felicetti, S. Hermon, G. Hiebel, A. Kritsotaki, A. Masur, K. May, P. Ronzino, W. Schmidle, M. Theodoridou, D. Tsifaki, and E. Christaki. Definition of the $CRM_{archaeo}$. An extension of CIDOC CRM to support archaeological excavation

S. Migliorini

- process, version 1.4.5, 2018. <http://www.cidoc-crm.org/crmarchaeo/sites/default/files/CRMarchaeo%20v1.4.5%20%28In%20Progress%29.pdf> Last accessed June 2018.
- 8 A. Eldawy and M. F. Mokbel. SpatialHadoop: A MapReduce framework for spatial data. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1352–1363, 2015.
 - 9 A. Eldawy and M. F. Mokbel. Spatial join with Hadoop. In S. Shekhar, H. Xiong, and X. Zhou, editors, *Encyclopedia of GIS*, pages 2032–2036. Springer, 2017.
 - 10 G. Hiebel, M. Doerr, and Ø. Eide. CRMgeo: A spatiotemporal extension of CIDOC-CRM. *International Journal on Digital Libraries*, 18(4):271–279, 2017.
 - 11 G. Hiebel, M. Doerr, Ø. Eide, and M. Theodoridou. CRMgeo: a spatiotemporal model an extension of CIDOC-CRM to link the CIDOC CRM to GeoSPARQL through a spatiotemporal refinement, version 1.2, 2015. http://new.cidoc-crm.org/crmgeo/sites/default/files/CRMgeo1_2.pdf Last accessed June 2018.
 - 12 W. Li, B. Chen, R. Yao, Y. Li, W. Wen, C. Cheung, and W. Li. SHOE: A SPARQL query engine using MapReduce. In *19th International Conference on Parallel and Distributed Systems*, pages 446–447, 2013.
 - 13 S. Migliorini and P. Grossi. Towards the Extraction of Semantics from Incomplete Archaeological Records. In *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory - Workshop on Computing Techniques for Spatio-Temporal Data in Archaeology And Cultural Heritage*, pages 349–358, 2018.
 - 14 S. Migliorini, P. Grossi, and A. Belussi. An Interoperable Spatio-Temporal Model for Archaeological Data Based on ISO Standard 19100. *ACM Journal on Computing and Cultural Heritage*, 11(1):5:1–5:28, 2017.
 - 15 C. E. Ore, M. Doerr, P. Le Boeuf, and S. Stead. Definition of the CIDOC conceptual reference model, version 6.2.3, 2018. http://www.cidoc-crm.org/sites/default/files/2018-05-16%23CIDOC%20CRM_v6.2.3_esIP%28XDP%29%28XM%29.pdf Last accessed June 2018.
 - 16 M. Perry and J. Herring. OGC GeoSPARQL - A Geographic Query Language for RDF Data, version 1.0, 2012. https://portal.opengeospatial.org/files/?artifact_id=47664 Last accessed June 2018.
 - 17 J. D. Richards, K. Niven, and S. Jeffrey. Preserving our digital heritage: Information systems for data management and preservation. In E. Ch'ng, V. Gaffney, and H. Chapman, editors, *Visual Heritage in the Digital Age*, pages 311–326. Springer, 2013.
 - 18 A. Schätzle, M. Przyjaciół-Zablocki, S. Skilevic, and G. Lausen. S2RDF: RDF querying with SPARQL on spark. *Proc. VLDB Endow.*, 9(10):804–815, 2016.
 - 19 R. Verborgh, M. Vander Sande, O. Hartig, J. Van Herwegen, L. De Vocht, B. De Meester, G. Haesendonck, and P. Colpaert. Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37-38:184–206, 2016.
 - 20 D. Wheatley and M. Gillings. *Spatial technology and archaeology: the archaeological applications of GIS*. Taylor & Francis, 2002.
 - 21 T. White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 4th edition, 2015.
 - 22 W. Zhang, C. Li and T. Zhao. Big geospatial data and the geospatial semantic Web: Current state and future opportunities. In Yulei Wu, Fei Hu, Geyong Min, and Albert Y. Zomaya, editors, *Big Data and Computational Intelligence in Networking*, chapter 3, pages 43–64. CRC Press, 2017.
 - 23 T. Zhao, C. Zhang, L. Anselin, W. Li, and K. Chen. A parallel approach for improving Geo-SPARQL query performance. *Int. Journal of Digital Earth*, 8(5):383–402, 2015.