Modeling Modern and Historical Data as a Knowledge Graph: A Case Study for Earthquake Data

Iosif Oikonomakis[1,2][0009−0007−3408−7684], Pavlos Fafalios[1,3][0000−0003−2788−526X], Michalis Mountantonakis[1,2][0000−0002−1951−0241] and Yannis Tzitzikas[1,2][0000−0001−8847−2130]

[1] Information Systems Laboratory, FORTH-ICS, Heraklion, Greece
[2] Computer Science Department, University of Crete, Heraklion, Greece
[3] Technical University of Crete, Chania, Greece
{sifisoik, fafalios, mountanton, tzitzik}@ics.forth.gr

Abstract. Historical data are sparse and vague making their collection and analysis difficult. To come up with a methodology for handling such data, in this paper we elaborate on this problem by considering the requirements and challenges for building a Knowledge Graph that contains data about both modern and *historical earthquakes*. We discuss the requirements and challenges, and we report the main observations concerning uncertainties by analyzing the documentation of ancient earthquakes that hit the island of Crete. Afterwards, we analyze representation approaches and related ontologies and then we present a modeling approach that can tackle the requirements. We showcase the feasibility of the approach by implementing it using RDF and we share publicly the result.

Keywords: Historical Data · Modeling Uncertainty · Earthquakes · Knowledge Graphs · CIDOC-CRM family of models

1 Introduction

Historical data are sparse and vague making their representation and analysis difficult. To come up with a methodology for handling such data, in this paper we elaborate on this problem by considering the requirements and challenges for building a Knowledge Graph that contains data about both modern and *historical earthquakes* for the region of Crete. As regards earthquake data, there is a detailed list and characterization of roughly all earthquakes that happened after 1900 in various data repositories[4]. However, for the earthquakes that happened earlier, the information is vague and in many cases it is described in books, like [7, 8]. Figure 1 shows indicative textual descriptions that are found in such books, translated to English.

[4] E.g. at https://earthquake.usgs.gov/earthquakes/search/ and http://www.geophysics.geol.uoa.gr/stations/gmapv3_db/index.php?lang=en

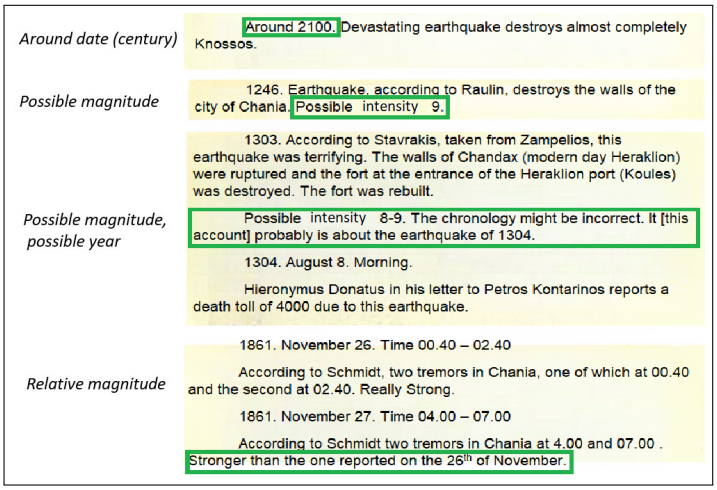

Fig. 1. Indicative descriptions of ancient earthquakes from [8] translated to English.

We can observe information expressed at various levels of detail and precision: dating at century level, statements with magnitude intervals, relative magnitude, statements about the provenance of information, etc. There is no principled and straightforward method to express this information as a KG (Knowledge Graph) and exploit it at the query/analytics level. Consequently there is a need for a methodology for constructing KGs with *both precise* and *imprecise* data for the benefit of analysis and historical research. The main research question is: *what kind of modeling facilitates representation, reasoning and analytics over KGs that contain precise and imprecise data?*
Regarding our contribution, we i) analyze our main observations about the needed requirements by analyzing the contents of books that document ancient earthquakes, ii) discuss possible representation approaches, iii) present our proposed representation based on CIDOC-CRM (ISO 21127)[5] [3] and its extensions, and iv) showcase the adequacy of the model by showing how it can be used to model ancient earthquakes and to query such data.

The rest of this paper is organized as follows: Section 2 describes the motivating scenario and the methodology that we followed. Section 3 discusses the various representation approaches. Section 4 describes and proposes the modeling approach. Section 5 provides indicative examples, Section 6 discusses implementation, and finally, Section 7 concludes the paper.

2 Motivating Scenario and Methodology

We adopted the following methodology: (a) collect the list of questions that we would like to be answerable (for both ancient and modern earthquakes), (b) find the available sources (both digital and printed books), (c) inspect the data they have, and (d) evaluate various representation and modeling approaches. Below we provide information for (a)-(c), while (d) is described in Section 3.

[5] https://www.cidoc-crm.org/

2.1 Initial competency questions

We started by forming a list of competency questions that we would like to be answerable, for being able to analyze earthquake data and building such applications. Below we list the more indicative ones.

Basic-Info Analytic Queries

• Earthquakes with intensity ≥ 6 before 1800, whose focus fall into a given area/s. For each of their key properties (location, magnitude, depth) provide their precision, and the provenance of the information

• Sum and maximum magnitude of earthquakes with focus (epicenter) in Crete by month.

Provenance-based Analytic Queries

• Earthquakes mentioned in Source X, that come from Source Y.

• Earthquakes mentioned in Source X that are wrong (wrong date/ location/-magnitude) according to Book Y.

• Sources that contain information about the earthquake X, organized chronologically, and accompanied by the corresponding citation network.

2.2 Data Sources and Data Collection

We collected data from several data sources and for many earthquakes, as described below.

Data Sources. Earthquake data from 1900 and onward was mainly gathered from two sources, from the earthquake catalogues of the United States Geological Survey (USGS)[6] and from the earthquake catalogues of the University of Athens (UoA)[7]. Finding data about earthquakes before 1900 was quite challenging. We managed to collect data about earthquakes before 1900 (for the island of Crete) mainly from two sources, a book about Cretan (ancient and modern) earthquakes [7] and an article from 1950 published in a Cretan journal [8].

Data Collection Process. For the needs of our analysis we decided to download from USGS all earthquakes of magnitude 3 or above, according to the rectangle drawn on a map from latitude 34.666 N - 35.934 N and longitude 23.138 E - 26.676 E, from all the datasets contained on those sites. The coordinates were chosen such that they contain the island of Crete and the sea near it, up to at least 30 kilometers from Crete's coast.

From the book and the article, we had to manually collect information about all earthquakes listed and their citations.

Size of Collected Data. From USGS we collected 733 earthquakes, from [7] 219 earthquakes, while from [8] 271 earthquakes. In total, we have 1223 descriptions of earthquakes.

2.3 Observations

Here, we describe the different levels of detail and precision that we have noticed by reading the books for the ancient earthquakes that hit the island Crete.

[6] https://earthquake.usgs.gov/earthquakes/search/

[7] http://www.geophysics.geol.uoa.gr/stations/gmapv3_db/index.php?lang=en

• **A. Dating Uncertainty**: In such a case, there is no evidence about the exact year or even century of the earthquake.

• **B. Magnitude/Depth Uncertainty**: In such cases, the sources do not provide the exact magnitude or/and depth of an earthquake, but a magnitude interval (e.g., 8-9 Richter). Moreover, in some cases there can be a comparison with other earthquakes that can aid us to estimate the magnitude/depth (e.g. "stronger than the one on the 26th of November").

• **C. Location Uncertainty**: For a given earthquake, there can be an uncertainty for the exact location of the earthquake, e.g. "Southeast of Santorini".

• **D. Provenance of Information**: For the same earthquake one or more sources can either agree or disagree with one or more key characteristics of the earthquake (location, magnitude, depth, etc.). This becomes even harder for ancient earthquakes.

The key requirement is to model these different levels of granularity, precision, as well as the provenance, for aiding the query formulation process and the presentation and visualization of results (of plain or analytic queries).

3 Representation and Modeling Approaches

3.1 Knowledge Graphs Limitations

Current KGs are based on the W3C standard RDF (Resource Description Framework) and OWL, which essentially offer a labeled directed graph data model. However, the expressiveness of this data model is limited: it does not allow quantitative statements, i.e., it cannot host data associated with degrees that express various kinds of uncertainty (probability, trust, etc). This limits the applicability, and thus the value of KGs, since we cannot host in a disciplined manner information that has been produced by information extraction and data mining techniques. Existing solutions (like reification) have not been widely adopted due to several limitations, e.g. they do not allow inferences on (reified) triples (important for producing new knowledge) and require an addition of a large number of triples to annotate a single statement (enlarging the already large dataset). Also, current technology for KGs does not offer any intrinsic mechanism for hosting (and properly managing) knowledge expressed in different levels of granularity and nested provenance levels (i.e. provenance of the provenance). There are only a few works that elaborate on fuzzy or probabilistic KGs (a review of the basic concept and definitions of knowledge reasoning and the methods for reasoning over KGs, including fuzzy extensions, is given by Chen et al, 2020 [1]), however they have not reached a level of maturity to enable their application in real datasets. There is also a recent W3C proposal, under the name RDF-Star[8], that extends RDF with a way to make statements about other statements, and this can be exploited for associating an RDF edge (triple) with scores, weights, temporal aspects and provenance.

However the general question that arises is: how to select what approach to use for modeling various levels of granularity/precision and provenance? The one

[8] https://www.w3.org/2022/08/rdf-star-wg-charter/

extreme is to tackle all such aspects at the conceptual modeling level (i.e. at the level of the ontology); the other extreme is to tackle them at the data model level (i.e. adopt a data model more expressive than RDF). A deeper understanding is required for deciding (based on the characteristics of the application context) what approach (or what kind of hybrid approach) is beneficial to adopt. In this paper, we investigate the conceptual modeling level.

3.2 Ontologies Related to Earthquakes

As regards ontology-based modeling of earthquakes, there are some works for modeling various perspectives of earthquakes. This includes a) General Emergency Management ontologies, like SOFERS emergency response system [5] and the Emergency Case Ontology Model [12], b) Earthquake Emergency Management Ontologies, like EEM, an ontology for modeling sensors and actuators (for evaluating both the earthquake entity and to respond to the emergency) and EDER, a disaster response geo-ontology [9], c) Earthquake Engineering ontologies for Research [4] and for building seismic risk assessment [11], d) Earthquake Emergency Training ontologies, for example a drill script design ontology [2]. The aforementioned ontologies focus on particular aspects, they cannot document all information about an earthquake, do not meet our requirements (as listed in §2.1), and cannot tackle the difficulties (as listed in §2.3).

In addition, there exists e) a data warehousing ontology [6], created for the purpose of storing earthquake data and predicting earthquakes. We found one existing attempt at creating an f) Earthquake LOD, using recent datasets from the USA and Japan, with the purpose of improving the discovery and retrieval of earthquake data [10]. While these ontologies serve the purpose of storing earthquake data from different datasets, they are still not extensive enough for our case, as there is no planning for earthquakes before 1900, ambiguity in data and conflict between observatories and sources. The data warehousing ontology does not even account for earthquake data sources. There exists also QuakeML[9], an XML format for exchanging seismological data, but it has not been modeled in RDF.

It is also worth mentioning CRMsci, the "Scientific Observation Model"[10], which is a formal ontology, that uses and extends the CIDOC-CRM, intended to be used as a global schema for integrating metadata about scientific observation, measurements and processed data in descriptive and empirical sciences such as biodiversity, geology, geography, archaeology, cultural heritage conservation and others in research IT environments and research data libraries.

Last, CRMinf (the "Argumentation Model")[11] is another official extension of CIDOC-CRM that is closely related to CRMsci. It is intended to be used as a global schema for integrating metadata about argumentation, inference making, belief adoption and provenance assessment. In this paper we propose the *combined use* and *light extension* of CIDOC-CRM, CRMsci and CRMinf.

[9] https://quake.ethz.ch/quakeml/
[10] https://www.cidoc-crm.org/crmsci/
[11] https://www.cidoc-crm.org/crminf/

4 The Proposed Modeling

Based on the requirements and the competency questions, we decided to use CIDOC-CRM and its compatible extensions CRMsci and CRMinf. CIDOC-CRM is a well developed and recognized ontology providing us an ontological base for our model (we represent earthquakes as a subclass of E5 Event). CRMsci is necessary, since a class for measurements of temporal entities is required, while CRMinf is used for representation of provenance and uncertainty. We shall call our model **CRM-EQ**. The main concepts that it captures are shown in Figure 2. In brief, we mainly use classes and properties from the aforementioned models, plus some original classes and properties that extend CIDOC-CRM and CRMsci to facilitate the needs of our own model. A more detailed overview of CRM-EQ (also annotated with examples of instances for each class) is shown in Figure 3. Below, we provide details on how we model each type of information.

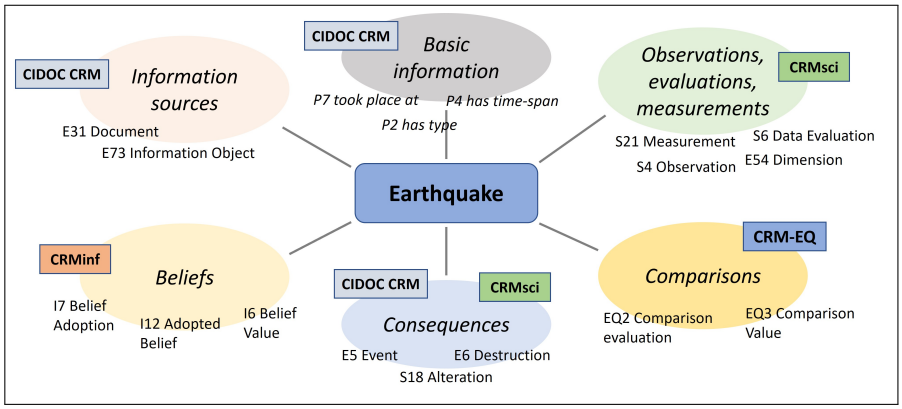

Fig. 2. Overview of modeled information.

Modeling basic earthquake information. First, for representing an instance of an earthquake, we define a new class with the name `EQ1 Earthquake` as a subclass of the CIDOC-CRM class `E7 Event`.[12] Then, we make use of the CIDOC-CRM properties *P7 took place at*, *P4 has time-span* and *P2 has type* (inherited from super-classes of E7 Event) for representing the general location which suffered damages, e.g. Crete (the island) or Arkalochori (town in the Heraklion regional unit of Crete)[13], its timestamp in any granularity level (e.g. year, year and month, date, etc.), and one or more types (e.g. aftershock, tectonic, etc.). We can also use the property *P183 ends before the start of* for connecting an earthquake to aftershocks, or the property *P9 consists of* if we consider an earthquake as a set of distinct shocks/earthquakes.

[12] We prefer defining a new class instead of directly using E7 Event together with the property *P2 has type*, as it makes querying more efficient and because we reserve the P2 property for defining the type of an earthquake (e.g. aftershock).

[13] Notice that this location might be different from the epicenter of the earthquake.

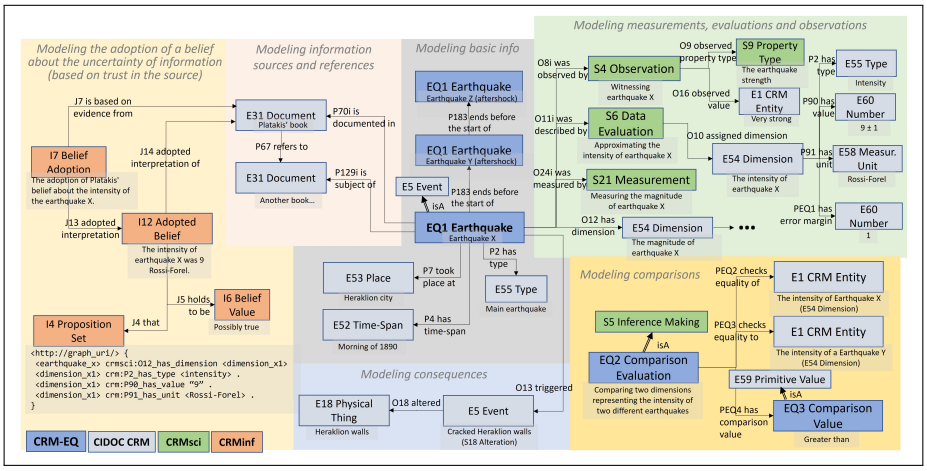

Fig. 3. The CRM-EQ model.

Modeling measurements, evaluations and observations. We make use of CRMsci which offers classes and properties for describing measurements for events (`S21 Measurement`), evaluations based on established hypotheses and inference rules (`S6 Data Evaluation`) as well as general observations providing additional information about an earthquake (`S4 Observation`). These constructs allow us to represent characteristics of the earthquake together with metadata information, such its magnitude (and which observatory measured it), its intensity, etc. Here, the new property *PEQ1 has error margin* can be used to assign an error margin to a recorded dimension.

Modeling information sources. For historical earthquakes that are documented in archival sources and books (known to have occurred prior to the early 20th century), we need to retain the data provenance information. This is important for the verification and the long-term validity of research findings that have made use of this data. To capture such information, we can make use of constructs of the core CIDOC-CRM model. In particular, we can link an earthquake with a document (`E31 Document`) that provides information about it using the property *P70 documents (is documented in)*, or using the property *P129 is about (is subject of)* if the earthquake is the primary subject of the document. Also, we can link a document with other documents it refers to (or other referred types of entities, in general) using the property *P67 refers to (is referred to by)*. Additional (metadata) information can be included using existing CIDOC-CRM constructs, such as who wrote the book, when, etc.

Modeling beliefs on earthquake characteristics. We have seen that historical information about earthquakes, as documented in books, is expressed in various levels of detail and precision. To maintain the full provenance chain of information, we use the CRMinf model which supports the structured description of beliefs/propositions of variable certainty as documented in books or other information sources (using the classes `I7 Belief Adoption`, `I12 Adopted`

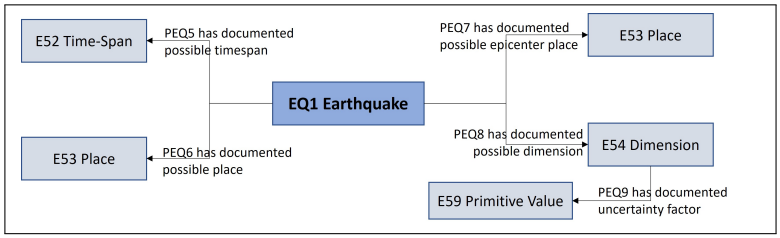

Fig. 4. Shortcuts of CRM-EQ model.

`Belief`, `I4 Proposition Set`, and `I6 Belief Value`). This implies a trust to the considered information source(s).

Modeling consequences. For representing the consequences of an earthquake, e.g., damages and destructions, we can use the property *O13 triggered* of CRMsci for linking an event (an earthquake in our case) with other events occurred due to the earthquake and with the physical things that altered by this activity.

Modeling comparisons. Comparing historical earthquakes is quite common due to the absence of precise information about an earthquake characteristic. An example is the witnessing that an earthquake was stronger than another earthquake. We provide new classes and properties for supporting the description of such information, in particular the class `EQ2 Comparison Evaluation` which is linked to the two entities under comparison using the properties *PEQ2 checks equality of* and *PEQ3 checks equality to*, and to an instance of the class `EQ3 Comparison Value` using the property *PEQ4 has comparison value*.

Shortcuts. We provide a set of shortcut properties (see Fig. 4), each one representing a deduction/join of a long data path. These properties are introduced for the cases where there is less detailed knowledge (that would allow for full data path description), while it can also highly accelerate data querying. For example, the property *PEQ8 has documented possible dimension* is a shortcut of the fully developed path: `EQ1 Earthquake` - *P70i is documented in* - `E31 Document` - *J14 adopted interpretation of* - `I12 Adopted Belief` - {*J4 holds to be* = "Possibly true", *J4 that* = <propositions describing an earthquake dimension>}.

5 Examples of Modelling Modern and Ancient Earthquakes

5.1 Example 1: Modeling a relatively modern earthquake

In Figure 5 we can see how we can model the basic information of a modern (20th century) earthquake, where we do not have to deal with any uncertainty. We can see how we model information about the earthquake, like place, epicenter, date and measurements, along with the source of the information regarding the earthquake. In Figure 6 we can see how we model the measurements (intensity, duration, quake movement type, quake direction) regarding the earthquake in a specific location, Malia in our example.

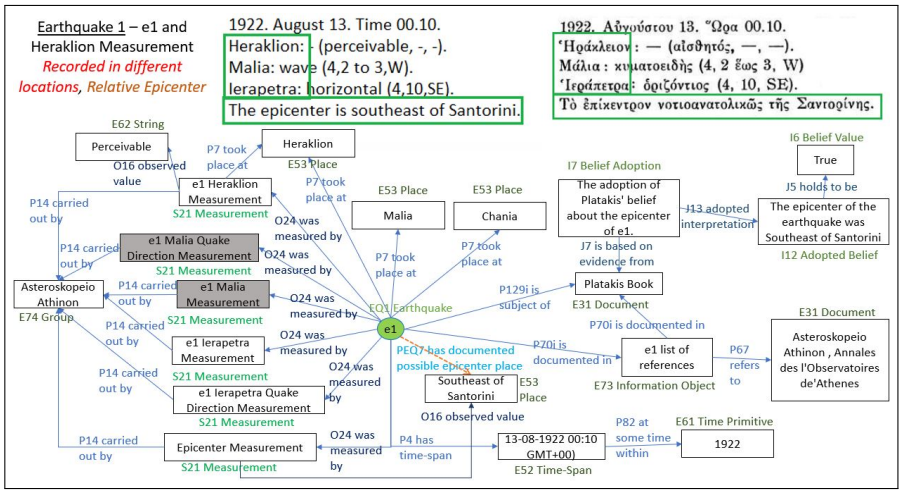

Fig. 5. 20th Century Earthquake - General information and Heraklion measurement

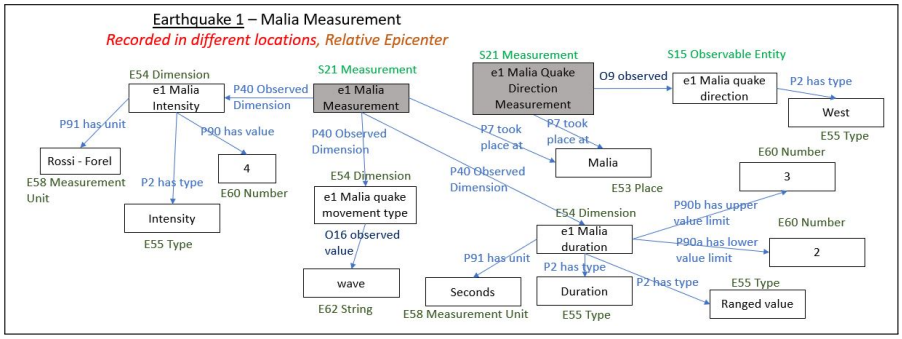

Fig. 6. 20th Century Earthquake - Malia measurement

5.2 Example 2: Modeling uncertain intensity and caused damages

In Figure 7 we can see how we model a statement about the possible intensity of an earthquake. Note how we model uncertainty, i.e. we associate the statement about intensity with a tag "Possible". Classes from CRMinf are used to model uncertainty and shortcut properties, shown in dotted lines, are used to quickly connect uncertainty values with the corresponding entity, intensity in this case. Additionally, we can see how we model the damages caused by this earthquake (the earthquake triggers an alteration/destruction or death event).

5.3 Example 3: Modeling time uncertainty and possible earthquake equivalences

In Figure 8 we can see a more complex example, with various kinds of uncertainty. In particular we show how we model the third book excerpt in Figure 1. Apart from ranged intensity uncertainty, here we can see how we model uncertainty as regards the time point/period, and uncertainty as regards equivalence with another reported earthquake $e4$, shown in Figure 9.

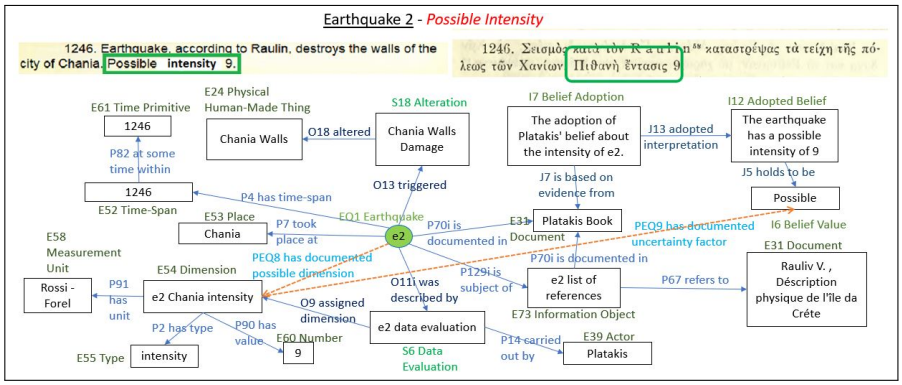

Fig. 7. 1246 earthquake. Modelling uncertainty and damages.

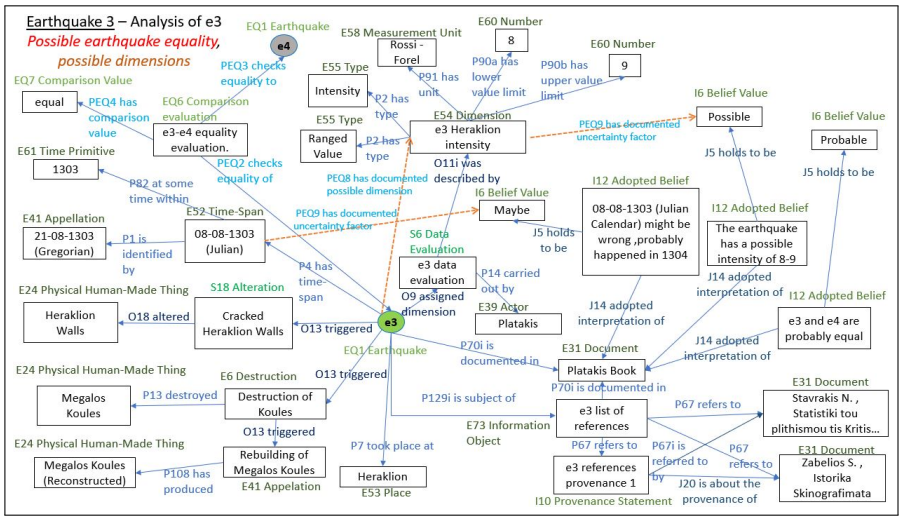

Fig. 8. 1303 earthquake($e3$)

Figure 10 shows an example on how to model the adoption of a documented belief about the wrong date of an earthquake. Indeed, we adopt the belief in the Platakis book that the date of an earthquake that hit Heraklion is not 1303-08-08 as sourced from Stavrakis' book. In the RDF data model, the proposition set can be implemented using named graphs. Similarly, we can model the adoption of a belief about the possible magnitude, intensity or epicenter of an earthquake, the adoption of a belief that two earthquakes mentioned in different sources are probably the same earthquake, etc. This allows tracking (and querying) the full provenance chain of uncertain or false information as documented in sources.

6 Implementation and Application

We have implemented the proposed modeling approach. In particular we have created the ontology described. Using that ontology we have documented the

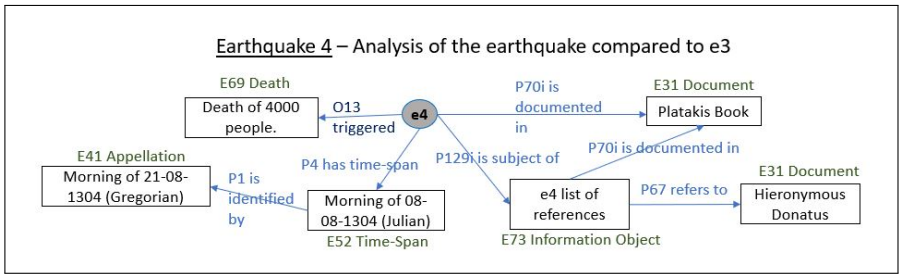

Fig. 9. 1304 earthquake($e4$)

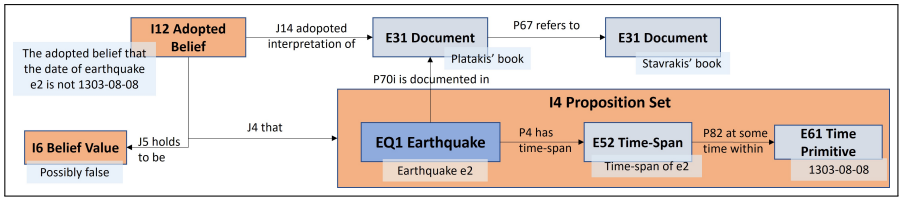

Fig. 10. Modeling the adoption of a belief about the wrong date of an earthquake.

earthquakes described in the previous section. These data are publicly available[14], including a .ttl file that contains the RDF representation of all earthquakes described in §5. We have tested the ontology and the earthquake description by trying to formulate the competency questions described in §2.1. For reasons of space, below we show only the following query: *"Give me all earthquakes with intensity >= 6 before 1800, whose focus falls into Crete. For each earthquake, give the following information: date, place, intensity, uncertainty factor (if any), information sources"* can be expressed in SPARQL as:

```
SELECT ?eq ?date ?place ?intensity ?uncertainty ?source_title
WHERE {
  ?eq a eq:EQ1_Earthquake ; crm:P4_has_time-span ?timespan ; crm:P7_took_place_at ?place .
  ?place crm:P89_falls_within eq:Crete.
  ?timespan crm:P82_at_some_time_within ?date . FILTER (year(?date)<1800) .
  ?eq crmsci:O12_has_dimension ?intDim .
  ?intDim crm:P2_has_type eq:intensity ;
          crm:P90_has_value ?intensity . FILTER(?intensity >= 6) .
  OPTIONAL { ?intDim eq:PEQ9_has_documented_uncertainty_factor ?uncertainty } .
  OPTIONAL { ?source crm:P70_documents ?eq ; crm:P102_has_title ?source_title } }
```

Note here that, for exploiting the full power of RDF and SPARQL and also simplifying query writing, the underlying engine/triplestore should have inference enabled for the *rdfs:subClassOf*, *rdfs:subPropertyOf* and *owl:sameAs* relationships, as well as have the required inclusion relationships between toponyms materialised so that the queries work correctly. For example, the above query will not return earthquakes that took place at Heraklion, if inclusion relationships between toponyms have not been materialized (i.e. the relation that Heraklion falls within Crete). Similarly the query will not return uncertain dimensions that make use of the property *PEQ8 has documented possible dimension* (subproperty of *O12 has dimension*), if *subPropertyOf* inference is not enabled.

[14] https://demos.isl.ics.forth.gr/crm-eq

7 Concluding Remarks

In this paper, we presented an approach for handling historical data which are sparse and vague. In particular, we focused on a case study for earthquake data for the island of Crete. Indeed, we discussed the requirements and challenges for building a Knowledge Graph for the mentioned data, a list of competency questions and we recorded the main observations concerning uncertainties. Afterwards, we analyzed representation approaches and related ontologies and then we presented the proposed representation for making it feasible to provide an answer to the competency questions. We have showcased the feasibility of the approach by implementing it using RDF. As a future work and research, we plan to produce and make publicly available a complete dataset about all earthquakes in Crete (the 1223 earthquakes described in §2.2), and to investigate various analytic and visualization services that are suitable for uncertain data.

References

1. Chen, X., Jia, S., Xiang, Y.: A review: Knowledge reasoning over knowledge graph. Expert Systems with Applications **141**, 112948 (2020)
2. Chou, C.C., Jeng, A.P., Chu, C.P., Chang, C.H., Wang, R.G.: Generation and visualization of earthquake drill scripts for first responders using ontology and serious game platforms. Advanced Engineering Informatics **38**, 538–554 (2018)
3. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine **24**(3), 75–75 (2003)
4. Hasan, R., Farazi, F., Bursi, O.S., Reza, M.S.: A faceted lightweight ontology for earthquake engineering research projects and experiments. In: Experimental Research in Earthquake Engineering: EU-SERIES Concluding Workshop. pp. 11–19. Springer (2015)
5. Liu, Y., Chen, S., Wang, Y.: Sofers: Scenario ontology for emergency response system. Journal of Networks **9**(9), 2529 (2014)
6. Nimmagadda, S.L., Dreher, H.: Ontology based data warehouse modeling and mining of earthquake data: prediction analysis along eurasian-australian continental plates. In: 2007 5th IEEE International Conference on Industrial Informatics. vol. 1, pp. 597–602 (2007). https://doi.org/10.1109/INDIN.2007.4384825
7. Papadopoulos, G.A.: A Seismic History of Crete. The Hellenic Arc and Trench - Earthquakes and Tsunamis: 2000 BC - 2011 AD. Ocelotos, Athens (2011)
8. Platakis, E.: Oi Seismoi tis Kritis apo ton Arxaiotaton mexri ton Kath' Imas Xronon, vol. D. Andreas G. Kalokairinos (1950)
9. Spalazzi, L., Taccari, G., Bernardini, A.: An internet of things ontology for earthquake emergency evaluation and response. In: 2014 International Conference on Collaboration Technologies and Systems (CTS). pp. 528–534. IEEE (2014)
10. Uematsu, H., Takeda, H.: Earthquake ontology and LOD (2023), http://www-kasm.nii.ac.jp/papers/takeda/23/uematsu23ijckg.pdf
11. Xu, M., Zhang, P., Cui, C., Zhao, J.: An ontology-based holistic and probabilistic framework for seismic risk assessment of buildings. Buildings **12**(9), 1391 (2022)
12. Yang, P., Wang, W., Dong, C.: Application of emergency case ontology model in earthquake. In: 2009 International Conference on Management and Service Science. pp. 1–5. IEEE (2009)