To Automate or not to Automate the Transcription of Ancient Earthquakes: Toward a Global Knowledge Graph about Ancient Earthquakes

Sophia Sideri[1,2]*, Emmanouil Patronakis[1,2]*, Evangelos Nikiforos[2]*, Isidoros Chatzichrysos[2], Apostolos Baniotis[1,2], Valantis Zervos[1,2], Stavros Tzormpatzakis[1,2], Michalis Saridakis[1,2], Iosif Oikonomakis[1], Michalis Mountantonakis[1,2], Pavlos Fafalios[1,3], and Yannis Tzitzikas[1,2]

[1] Institute of Computer Science, FORTH
{sophisid, patronakis, baniotis, vzervos, stzorba, saridakism, sifisoik, mountant, fafalios, tzitzik}@ics.forth.gr,
[2] Computer Science Department, University of Crete, Heraklion, Greece
{csd4966, csd4338}@csd.uoc.gr
[3] School of Production Engineering and Management, Technical University of Crete, Greece

Abstract. The transcription of ancient earthquakes remains a significant challenge due to the ambiguity and uncertainty of historical descriptions. The vision is to document this information, found in books and manuscripts in various languages, in a structured, semantically rich, and interoperable manner, enabling a complete picture of past seismic activity. While ontologies like CRM-EQ offer semantically expressive models for capturing seismic events, the manual transcription of such information from historical sources remains time-consuming and requires expertise. In this paper we propose and evaluate a hybrid and collaborative workflow for transcribing earthquakes from historical texts. We compare five different methodologies from manual curation to LLM-based pipelines, assessing them in terms of effort, accuracy, and scalability. Our findings show that semi-automated approaches are practical, balancing effort and accuracy. We captured 248 earthquake events and produced over 24,500 RDF triples, demonstrating the feasibility of semi-automated transcription for historical seismic data.

Keywords: Knowledge Graph Construction · Historical Data · Earthquakes · Ontology-based Data Extraction · Large Language Models (LLMs)

1 Introduction

Recently, the use of semantic technologies and knowledge graphs to represent historical and uncertain data has gained significant traction, especially in the

* These authors contributed equally to this work.

domain of cultural heritage, history, and natural disasters [28,7,13]. Many descriptions of ancient earthquakes can be found in various books [35,39] which contain 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, uncertainties and beliefs. Such textual representation does not facilitate the machine-readable data necessary for researchers and the public to assess the frequency of catastrophic earthquakes [37].

Particularly, it would be very useful if all data that are now described in various books or manuscripts could be recorded in machine readable formats along with their semantics. That would provide a more comprehensive understanding of what happened in the past and would help seismologists discover patterns in the frequency of the occurred earthquakes. The vision would be to construct a global knowledge graph about all ancient earthquakes. However, the manual transcription of such historical data is laborious and time-consuming.

To the best of our knowledge, no prior work has systematically evaluated the range of solutions from entirely manual to fully automated methods for this problem. To address this, in this paper we explore a hybrid and collaborative workflow that combines Large Language Model (LLM) services with human transcription, aiming to reduce human effort while ensuring accuracy and semantic consistency.

The key contributions of this paper are: (a) the design and empirical evaluation of a collaborative workflow for the transcribing of historical earthquakes according to CRM-EQ [30] (b) a comparative analysis of the methodologies used (c) the enrichment and instance matching of the transcribed data with authorities, and (d) the development of a user-interface for semantic exploration of the resulting Knowledge Graph (KG). The transcription was based on the book "Oi Seismoi tis Kritis" by Platakis [39], which describes earthquakes in Crete (Greece). This work was designed following key principles for semantic content management systems, introduced by Hyvönen et al. [22], including consistent URI naming, modular data structures, validation steps and usability publishing practices.

Outline. The rest of the paper is organized as follows. In Section 2, we cover some background on semantic technologies and KGs from archival data. In Section 3, we present the collaborative methodology and approaches that we followed. In Section 4, we present our insights of each method and step. Finally, Section 5 concludes the paper and presents directions for future work.

2 Background & Related Work

Recently, there has been a growing interest in using semantic technologies and knowledge graphs for representing historical and uncertain data, such as in the domain of cultural heritage [11,28,17], biodiversity [1], Virtual Research Environments (VREs) [26,40], and natural disasters [27,30,38]. For our use case, we utilised the CRM-EQ [30] ontology, to represent the uncertainty in earthquake data. It is based on two CIDOC-CRM [8] extensions, CRMsci [9] and CRMinf [10],

and supports the representation of provenance, approximate dates, multiple sources, and inconsistent or incomplete information.

Uematsu and Takeda [45,46] suggest the creation of an Earthquake Linked Open Data (LOD) dataset utilizing Ontology Oriented Design Patterns. A part of this approach was the development of a new Earthquake Ontology that represents specialized domain concepts including hypocenters, seismic motion, and the seismic intensity scale of the Japan Meteorological Agency, reusing existing ontologies such as SOSA [23] for sensor data.

Several works have explored workflows for transcribing and curating historical data. Fafalios et al. [15] proposed a provenance-aware workflow model where transcription and curation is separated, to preserve original data. Their workflow was applied to the SeaLiT project [6,14] demonstrating how semantic technologies can be used in large-scale historical research. A similar work in the field of maritime history developed a collaborative platform for historians to curate archival sources like ship logbooks [16,36]. While these approaches focus mainly on manual transcription by domain experts, our goal is to explore whether this transcription task can be automated, or at least partially. A relevant automated approach from Piantanida et al. [38] provides a pipeline for extracting geological knowledge from reports using annotators and machine learning. Similarly, Mastoras et al. [27] built a KG for modern earthquakes using Natural Language Processing (NLP) modules and satellite data.

Recent works have shown the promising usage of LLMs to automate the process of digitizing historical records. In geosciences, Ma et al. [25] suggested a workflow that uses models offered in their public Llama 2 [44]. Wei et al. [49] propose a chain-of-thought strategy to derive location and depth information from archives. Similarly, a proof-of-concept study in the social sciences led by Nicole Schwitter [41] applied Cohere Labs Command R+ model [4] to enhance the efficiency and accuracy of extracting information from 19th-century newspapers.

In conclusion, most related approaches focus either on structured modelling or on partial automation of historical data transcription. No prior work evaluates the different approaches in terms of quality and time. This constitutes the central focus of our approach. Our work examines a hybrid workflow that combines manual assertions, automated information extraction from books using LLM, and rule-based post-processing steps to ensure data consistency.

3 The Steps of Transcribing Ancient Earthquakes

3.1 Collaborative Methodology

We investigated a collaborative approach, involving 23 persons with different backgrounds: 10 postgraduate students, 9 undergraduates, and 4 professionals with expertise in semantic technologies. We formed teams with particular roles and asked all members to record the steps and the time they spent on each task. In this workflow, also shown in Figure 1, we identified four main roles: **R1** for producing RDF triples from the book (14 people), **R2** for organizing the

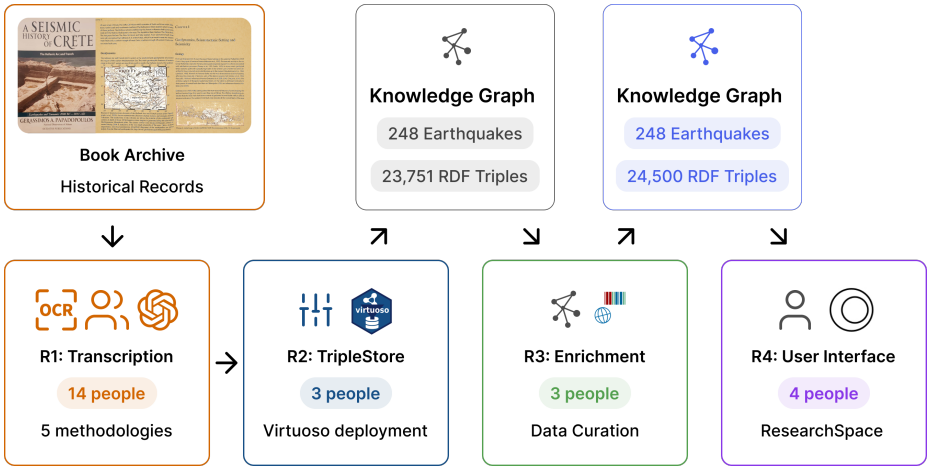

Fig. 1: Project Workflow.

produced data (3 people), **R3** for quality checking and enrichment (3 people), and **R4** for deploying various data access services (4 people).

Transcription Methodologies. Converting historical earthquake descriptions into structured RDF representations aligned with the CRM-EQ ontology is a foundational element of our knowledge graph construction. Given the collaborative nature of this project, we explored various methodologies, ranging from fully manual transcription to automated pipelines using the latest LLMs. Through analysis of the approaches adopted by the 14 participants in the R1 group, we identified five distinct methodological patterns for RDF production:

1. Manual Transcription using Protégé: In this approach we manually created RDF triples using the Protégé editor [29]. This required deep understanding of the ontology but provided the highest accuracy. We noted that after an initial learning curve of 1-2 hours of transcription, the process became more efficient.

2. OCR-Assisted Manual Transcription: We employed Google OCR (Optical Character Recognition) [19] to digitize the scanned pages, then manually authored the RDF/TTL using Protégé [29]. This method reduced the time needed for text extraction while maintaining human oversight for semantic accuracy.

3. LLM-Supported RDF Generation: We utilized LLMs, such as ChatGPT-4 [32], GPT-o1 [33], Gemini 2.0 [5], and DeepSeek [3] to generate RDF triples from OCR-extracted text. To help guide the models, we provided the ontology along with example triples. However, all participants who used this method found that the output needed manual corrections due to various errors in the LLM output.

4. Direct Vision-Language Model Processing: Another approach involved using multimodal LLMs (GPT-4o) to directly process page screenshots, bypassing traditional OCR. This method needed approximately one minute per page for text extraction. However, manual verification remained essential.

5. Custom Pipeline with Intermediate Representation: We developed a custom pipeline to improve accuracy and consistency. As shown in Figure 2, the process involves: (a) multimodal LLM to extract text directly from page screenshots (b)

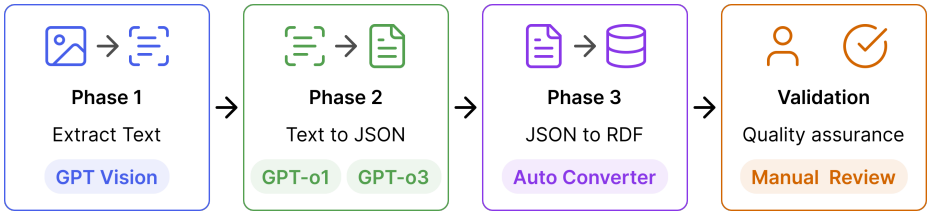

Fig. 2: Custom Pipeline with Intermediate Representation.

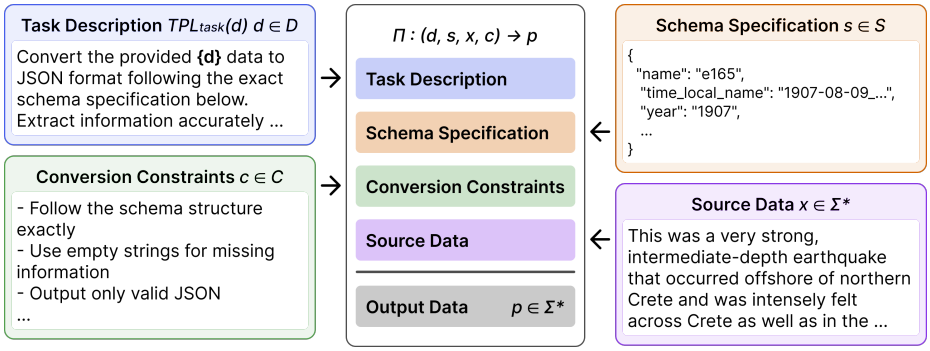

Fig. 3: Formal Prompt Construction.

applying reasoning models (GPT-o1 [33], GPT-o3 [34]) with specific instructions to transform text into standardized JSON schemas containing earthquake information such as dates, locations, seismic data, references and related phenomena, and (c) a dedicated tool, the CRM-EQ JSON Converter[1], to transform the JSON into CRM-EQ compliant RDF triples.

Analysis of the Custom Pipeline. Among the aforementioned methods, the custom pipeline stood out due to its modular architecture and its potential for scalable RDF generation. We introduce an intermediate representation (IR) as a JSON that mediates between historical text (images or OCR output) and RDF triples. By constraining the extraction to a schema whose fields have defined semantics, the pipeline reduces the ambiguity and errors of the transformation.

The IR organizes earthquake information into groups. We have spatiotemporal information, seismological details, source data, associated phenomena, and uncertainties or beliefs indicating when something is approximate or disputed, and we also keep the exact original phrase. Writing things this way keeps the original meaning, but removes random wording that would confuse the mapping into RDF. To ensure accurate extraction, we used schema-based prompting. Each prompt instance is composed of four XML-labeled templates plus a final emission directive. Instead of treating this prompt in an ad hoc manner, we formalize it as a higher-order function that constructs a well-formed extraction request and specifies an expected contract for the model output.

[1] `https://github.com/EmmanouilPatronakis/crm-eq-json-converter`

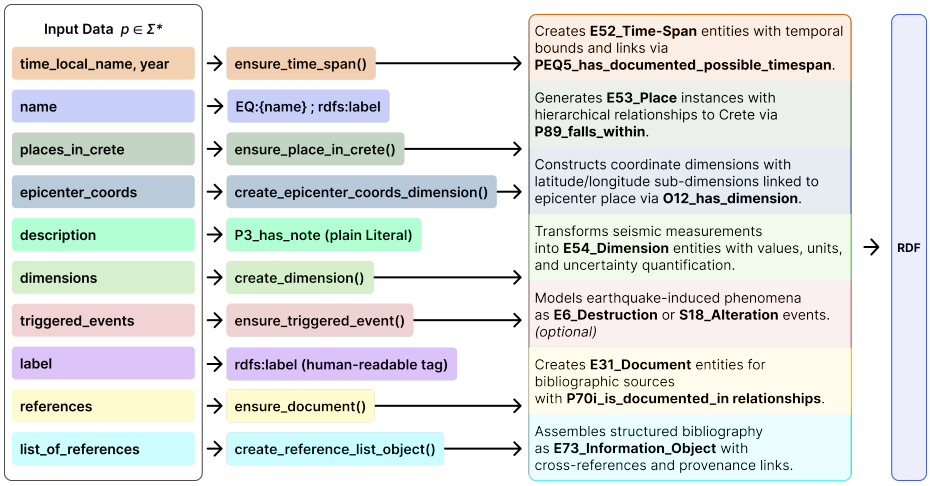

Fig. 4: Transformation to RDF.

Formal Prompt Construction. Figure 3 illustrates the formulation of the prompt used as input to the LLM. It shows how main components are combined into a structured prompt string $p \in \Sigma^*$. Let $\mathcal{D}$ denote the set of supported domain labels (e.g., `EARTHQUAKE_HISTORY`). Let $\mathcal{S}$ be the set of JSON Schema specifications for the IR, and Σ^* the set of Unicode strings (source transcriptions). We define the function:

$$\Pi : (d, s, x, c) \mapsto p \quad \text{with} \quad d \in \mathcal{D},\ s \in \mathcal{S},\ x \in \Sigma^*,\ c \in \mathcal{C}$$

where $\mathcal{C}$ is the set of finite lists of conversion constraints. The element $p \in \Sigma^*$ is the prompt string supplied to the reasoning LLM. Specifically:

$$\begin{aligned}\Pi(d, s, x, c) = {} & \texttt{"<task\_desc>"} + \mathrm{TPL}_{\mathrm{task}}(d) + \texttt{"</task\_desc>"} \\ & + \texttt{"<schema\_specs>"} + s + \texttt{"</schema\_specs>"} \\ & + \texttt{"<constraints>"} + c + \texttt{"</constraints>"} \\ & + \texttt{"<source>"} + x + \texttt{"</source>"} \\ & + \texttt{"Output the converted JSON"} \end{aligned} \tag{1}$$

The task $\mathrm{TPL}_{\mathrm{task}}(d)$ is parametrized as, *"Convert the provided"* + d + *"data to JSON format following the exact schema specification below. Extract information faithfully; do not infer or fabricate."* The constraint list c typically includes (but is not limited to): *Follow the schema exactly; Use empty strings for missing atomic values; Preserve numeric precision; No comments; Output only valid JSON.* We define $\mathcal{J} \subseteq \Sigma^*$ as the set of strings that parse (under a strict JSON parser) to an object instance conforming to schema. The model is considered successful on input tuple (d, s, x, c) iff its emission is valid JSON conforming to the schema. Any deviation, like, malformed JSON, schema violations, needs human intervention.

Transformation to RDF. Figure 4 illustrates each field mapping during transformation from valid JSON input data $p \in \Sigma^*$ to RDF. Each input

field is processed by a dedicated mapping function (e.g., `ensure_time_span`, `create_dimension`) which creates the triples, contributing to the final RDF output. Let $f : \mathcal{J} \to \mathcal{R}$ map valid JSON IR documents to the set $\mathcal{R}$ of CRM-EQ compliant RDF triple sets. We decompose f as a composition of field-level deterministic mappings $f = g_n, g_{n-1}, \ldots, g_1$. Each g_i consumes one semantic entity (e.g. temporal, spatial, magnitudes) and emits a new set of triples.

For a JSON document j, we write $f(j) = (T, E)$, where T is the (multi)set of triples and E a (possibly empty) list of validation exceptions (e.g. missing mandatory path `event.time.begin`). A document is *ontology-valid* if $E = \varnothing$. Each mapping employs URI templates of the form, $\text{URI}(\text{entity}, \alpha) = b/\text{entity}/h(\alpha)$, with base IRI b and hash function h over a canonical serialization ensuring trivial regeneration of the RDF entities.

Common Transcription Challenges. Participants faced several challenges in all methodologies. The most significant were *LLM hallucinations and errors*, including missing properties, syntax issues, and incomplete date representations. Furthermore, the *learning curve of the CRM-EQ ontology* delayed the process. The uncertainty structure and conflicting sources demanded a long learning time period. Most overcame this through collaborating and sharing strategies, gradually learning the ontology. *Ambiguous historical data* required a careful interpretation. Uncertain dates, conflicting locations, and different magnitude scales were modelled using CRM-EQ's belief entity to retain the original source that had the claims and their uncertainties for each earthquake. Finally, *OCR inaccuracies* remained problematic despite modern tools. Historical typography and notation, often led to recognition errors. Participants spent considerable time on manual correction, highlighting the need for human oversight.

3.2 Uploading the Files To the TripleStore

Following the transcription phase, the team R2, created a collaborative workflow, to upload the Terse RDF Triple Language (TTL) files into a Virtuoso triplestore [12]. More specifically, each member of R1 uploaded their TTL files to a shared cloud repository which were then imported into Virtuoso to formulate the RDF graphs. Each graph was assigned a URI, using the pages of the book assigned to each person, and the student's registration number, making it unique and consistent based on the following format:

`http://csd.uoc.gr/<starting page>-<ending page>-<student regist. number>`

For instance, the TTL file *http://csd.uoc.gr/022-051-1234* represents content from pages 22 to 51 of the book and was curated by the student with registration number 1234. Because the KG was part of a group project, we used the unique IDs of each student, for ease of use, and to ensure unique URIs within the triplestore. Following upload, a validation step was performed to ensure that the triples were syntactically correct, and aligned to the ontology. Once the RDF graphs were successfully integrated, the SPARQL endpoint of the Virtuoso Triplestore was shared with Teams R3 and R4.

3.3 Enrichment and Instance Matching

The next step, implemented by the R3 team, was associated with improving the data and linking it with external authoritative resources such as Wikidata [47] and Geonames [18]. This enables interoperability within the Linked Open Data [2] ecosystem and follows the FAIR principles [50]. We used the standard OWL [20] relation, `owl:sameAs`, to indicate entities as identical. Since OWL does not provide a relation for approximate matches we introduced a custom property `custom:closeMatch` to capture such semantics. It is important to note that we do not incorporate all external information into our knowledge graph. Instead, we keep basic information, and query the rest on demand. In our approach, we start by enriching the data with the external resources and use them to match local data.

Enrichment. We perform different enrichment approaches on each entity type, querying external sources to retrieve additional metadata. The **Places** (`E5_Place`) were enriched using the GeoNames API [18], retrieving coordinates, administrative divisions, and alternative labels. We identified candidate matches by their geographic coordinates, if available, otherwise, we used the label to perform a keyword search. We cached GeoNames responses to avoid redundant API requests, as geographic information is stable. **Persons** (`E21_Person`) were enriched using Wikidata, incorporating fields such as occupations, birth and death years, and external identifiers. However, due to incomplete data, especially missing birth or death dates, exact matching was often not feasible. To improve the ranking of Wikidata candidates, we introduced a weighted scoring prioritizing relevant occupations (e.g., geologists, seismologists, historians) and temporal consistency

$$S(e) = \sum_{i=1}^{n} w(o_i) + \delta_{\text{birth}}(e) \cdot w_b + \delta_{\text{death}}(e) \cdot w_d \tag{2}$$

where $S(e)$ is the total score of candidate entity e, $o_1, \ldots, o_n$ are the occupations of e, $w(o_i) \in [0, 5]$ is the weight assigned to occupation o_i based on relevance, with $w(o_i) = 0$ if o_i is irrelevant, $\delta_{\text{birth}}(e) = 1$ if the birth year matches the reference and $\delta_{\text{death}}(e) = 1$ if the death year matches the reference (both 0 otherwise), and $w_b = w_d = 5$ are the weights assigned for birth and death year matches. All weights in the scoring function, including those for occupations ($w(o_i)$), birth year (w_b), and death year (w_d), were empirically chosen. Occupations considered highly relevant to seismic data, such as geologist, seismologist, and historian were assigned higher weights[2].

In many cases, names were incomplete (e.g., only surname was available), and associated dates were ambiguous or found along with their names (e.g., Troulinos (1900)), possibly representing birth or death year or unrelated facts. We utilized these dates as possible birth/death dates for the purpose of scoring. To address incomplete name labels, we searched for all individuals who shared

[2] The full list of weights is available at: `https://github.com/sophisid/instance_matching_EQ/blob/master/occupations_weights.json`

the same family name and used the scoring system to identify the most relevant. As a result, person entities were typically linked using the custom relation `custom:closeMatch`.

Instance Matching. Following the enrichment, we performed instance matching to identify duplicate or equivalent entities across both the corpus and external linked data. We preserve all local data and use `owl:sameAs` or `custom:closeMatch` to record equivalence, depending on the confidence level.

Places. We define two places Pl_1, Pl_2 as equivalent if they share the same GeoURI, or if they have highly similar labels and are located near each other:

$$\begin{aligned} Pl_1 \equiv Pl_2 \iff & \ \text{GeoURI}(Pl_1) = \text{GeoURI}(Pl_2) \\ & \vee (\text{sim}(L_1, L_2) \geq \tau \ \wedge \ d(C_1, C_2) \leq \delta) \end{aligned} \tag{3}$$

where, $\text{sim}(L_1, L_2)$: label similarity between names L_1 and L_2, computed using Levenshtein distance [24], $\tau = 0.95$: similarity threshold, $d(C_1, C_2)$: Haversine distance [42] between coordinates C_1 and C_2, $\delta = 0.2\,\text{km}$: spatial proximity threshold. When these criteria are not satisfied but close matches are observed (e.g., $\tau \geq 0.8$, $\delta \leq 5\,\text{km}$), we assign `custom:closeMatch`.

Earthquakes. To identify equivalent earthquake events, we rely on a combination of label similarity, temporal overlap, and spatial proximity. The equivalence conditions are defined as:

$$\begin{aligned} E_1 \equiv E_2 \iff & \quad (\text{sim}(L_1, L_2) \geq \tau \ \wedge (\Delta T(B_1, B_2) \leq \theta \ \vee \ \Delta T(E_1, E_2) \leq \theta)) \\ & \vee \ (\text{sim}(L_1, L_2) \geq \tau \ \wedge \ d(C_1, C_2) \leq \delta) \\ & \vee \ ((\Delta T(B_1, B_2) \leq \theta \ \vee \ \Delta T(E_1, E_2) \leq \theta) \ \wedge \ d(C_1, C_2) \leq \delta) \end{aligned} \tag{4}$$

where, E_1, E_2 two earthquakes, $\text{sim}(L_1, L_2)$: label similarity (e.g., Levenshtein distance) between L_1 and L_2, $\tau = 0.95$: similarity threshold, $\Delta T(B_1, B_2)$, $\Delta T(E_1, E_2)$: time difference between begin and end dates, $\theta = 1\,\text{day}$: temporal threshold, $d(C_1, C_2)$: Haversine distance between coordinates, and $\delta = 50\,\text{km}$: spatial threshold. If none of the strict equivalence conditions are met, but weaker ones apply (e.g., $\tau \geq 0.80$), we assign a `custom:closeMatch`.

Persons. We define two persons P_1, P_2 as equivalent if they share a Wikidata URI through the `owl:sameAs` relation, or if they have very similar labels and matching years of birth and death.

$$\begin{aligned} P_1 \equiv P_2 \iff & \quad \text{WikidataURI}(P_1) = \text{WikidataURI}(P_2) \\ & \vee \ (\text{sim}(L_1, L_2) \geq \tau) \\ & \vee \ (\text{YOB}_1 = \text{YOB}_2 \ \wedge \ \text{YOD}_1 = \text{YOD}_2) \end{aligned} \tag{5}$$

where, $\text{WikidataURI}(P)$: external identifier of person P in Wikidata, $\text{sim}(L_1, L_2)$: Levenshtein similarity between labels L_1 and L_2, $\tau = 0.95$: label similarity threshold, YOB: year of birth, YOD: year of death. If $\text{sim}(L_1, L_2) \in [0.80, 0.95)$ or if one label is a strict subset of the other, we assign a weaker equivalence relation, `custom:closeMatch`, instead of `owl:sameAs`.

Enriching the data not only enhances its semantic expressiveness, but also facilitates integration with external data sources. Upon completion of this phase, the graph had transformed from a basic transcription archive to a rich semantic knowledge graph. Ideally, we would like to extend instance matching to cover `E5_Event` and `E7_Activity`, by comparing actors, time spans, and locations. Another direction is to explore how to connect beliefs (`I2_Belief`) together by using LLMs to match uncertain information.

3.4 Access Services

To ensure the accessibility of the transcribed and enriched data to both technical and non-technical users, we developed a user interface built on top of the ResearchSpace (RS) framework [31], and a SPARQL endpoint.

SPARQL Endpoint. The underlying knowledge graph is provided through a SPARQL endpoint[3]. For domain experts, the SPARQL endpoint allows direct querying and interaction with the raw RDF data, facilitating tasks such as validation and enrichment.

User Interface. For non-expert users we also created a public interface[4], to navigate the KG through interactive tools, including a structured search component that helps build complex queries by selecting entities and relations from a predefined catalogue, a keyword search, and a map navigation. RS was selected for its native support for semantic data exploration. The UI has been customized around the main entities of our dataset: Earthquakes, Places, Persons, and Documents. Each entity page includes a dynamic RDF catalogue showing incoming and outgoing relations. In addition, it queries information from the external authorities linked in the enrichment process.

4 Evaluation

To ensure data quality, we implemented several validation strategies. To verify the RDF triples' syntax we used the W3C RDF Validation Service [48]. To ensure ontology compliance, we checked property domains and ranges using Protégé's [29] reasoning capabilities and OntoGraf [43] for visual consistency checking. Additionally, we did a peer review process, where we cross-checked the triples to identify inconsistencies or errors. Finally, we tested the correctness of the generated RDF data by executing predefined SPARQL queries derived from a group of competency questions.

Competency Questions. Team R1 contributed a set of natural language competency questions, which were manually translated into SPARQL queries by R2. More specifically, the set included questions such as:

- How many / Which earthquakes occur between a specific time period?
- How many / Which earthquakes occur in a specific place?

[3] http://demos.isl.ics.forth.gr/earthquakes-rs/sparql
[4] http://demos.isl.ics.forth.gr/earthquakes-rs/

– How many / Which earthquakes had more than a specific Richter number?
– Which places have the most earthquakes?

We formulated queries based on the previous questions, and used them during the early stages of development to test the integrity of the graphs and validate the endpoint's functionality.

Empirical Observations. Our findings reveal that while automated approaches using LLMs are scalable, they require human oversight to achieve acceptable quality levels. Participants who started with manual methods reported improved efficiency as they became familiar with the ontology, often achieving comparable or better throughput than those using automated methods when factoring in correction time. The most effective approach appeared to be a hybrid methodology that combines automated text extraction with human-guided RDF generation.

The time investment in the transcription process varied across methods and participants, ranging from 7 to 36 hours for completing the assigned pages (∼22 pages per person). Table 1 summarizes the key characteristics observed across different methodologies, including average time per page, error rate, scalability, and the level of ontology expertise required. As shown, manual transcription has the lowest error rate but is the least scalable and demands expertise. In contrast, automated or semi-automated methods are faster and scalable, with moderate error rates, emphasizing the trade-offs between efficiency and accuracy.

Method	**Time/Page**	**Error Rate**	**Scalability**	**Ont. Expertise**
Manual (Protégé)	30-45 min	Very Low	Low	High
OCR + Manual	20-30 min	Low	Medium	High
OCR + LLM	15-25 min	Medium	High	Medium
Custom Pipeline	10-20 min	Low-Medium	High	Medium
Vision-LLM	5-10 min	Medium	Very High	Low

Table 1: Comparison of different RDF production methods.

Effort Spent. The collaborative transcription effort involved 14 participants in the role of R1: 1 expert supervisor, 9 post-graduate students, and 4 undergraduate students. The total time investment reached 8,644 minutes (144 hours), with each participant spending an average of 11 hours on their assigned sections. This time included the initial learning curve for learning the CRM-EQ ontology, familiarizing with the book content, formalizing effective LLM prompts, and the actual transcription work.

The R2 team, comprising one post-graduate, one undergraduate, and an expert supervisor, spent approximately one week setting up the Virtuoso triple store. Similarly, the R3 team, including a post-graduate, an undergraduate, and an expert supervisor, devoted two weeks to formulate the scoring system and implement the corresponding code. Finally, R4 with the same composition, developed the user interface in approximately two weeks.

Post Processing. As shown in Table 2, the post-processing phase contributed a total of 749 additional triples across three entity types. The number of triples increased from 23,751 to 24,500, which cover all entities, incoming and outgoing

Entity	**Named Graphs**	**Triples**	**Source**
EQ1 Earthquake	custom:earthquake	337	Manuscript
E53 Place	custom:places, custom:geonames	342	Manuscript & Geonames
E21 Person	custom:people, custom:wikidata	70	Manuscripts & Wikidata

Table 2: Triples per entity and source added after the enrichment.

Type	Value
Triples	24,500
Triples with Object URIs	14,262
Triples with Object Literals	10,130
Distinct URIs	3,085
Distinct Literals	5,895
Distinct Subjects	3,040
Distinct Objects	8,545

Table 3: Statistics about the KG.

Class	URIs	Triples (T)	avg T/URI
E54 Dimension	673	4,235	6.29
EQ1 Earthquake	248	3,695	14.89
E52 Time Span	205	1,188	5.79
E53 Place	146	6,234	42.69
E73 Information Object	104	1,200	11.53
I12 Adopted Belief	83	728	8.77

Table 4: Statistics of some key KG Classes.

relations. In total, the project led to the documentation of 248 ancient earthquakes, fully described according to the CRM-EQ ontology, covering diverse locations and time periods as described in the book.

Knowledge Graph Statistics. Table 3 shows some key statistics for the resulted KG, such as the number of distinct triples, literals, URIs, and others. Table 4 offers a more detailed view of some of the main classes in the KG. For instance, the **E53 Place** class has the most triples (6,234), reflecting the large number of spatial references in the corpus. Earthquake instances (**EQ1 Earthquake**) are associated with an average of 14.89 triples each. Other classes, such as **E54 Dimension** and **E73 Information Object**, also show many measurements and textual information material. This highlights the semantic richness and expressivity of the knowledge graph.

Reproducibility. To ensure reproducibility, we provide open access to all relevant resources. The CRM-EQ JSON configuration and conversion code is available at `https://github.com/EmmanouilPatronakis/crm-eq-json-converter`, accompanied by working examples. The enrichment and instance matching components are available at `https://github.com/sophisid/instance_matching_EQ`, and the source code of the ResearchSpace application is accessible at `https://github.com/sophisid/EarthquakeRS`.

5 Concluding Remarks

This paper presented a collaborative workflow for transcribing historical earthquake information into a knowledge graph. The process resulted in 248 earthquake entities, producing over 24,500 RDF triples, and showed that semi-automated pipelines reduced time by over 50% compared to manual methods.

Lessons Learned. The collaborative process revealed several important insights that can inform future efforts in similar domains:

- **Hybrid approaches are most effective:** Combining automated extraction using LLMs and OCR with structured intermediate representations and manual validation achieves high-quality and scalable results.
- **Ontology learning curve matters:** Although the ontology learning was initially challenging, the use of CRM-EQ became more efficient over time.
- **Intermediate schemas help control quality:** Using intermediate representations restricted LLM hallucination outputs and improved the pipeline.
- **Collaboration fosters consistency:** Well-defined roles, peer feedback and validation, significantly improved data quality and participant engagement.

The results of this work can be useful for enabling the community to use semi-automated workflows, aiming to create a knowledge graph that contains as complete information as possible about all ancient earthquakes. Moreover, this approach can be easily applied in other domains. Future work can explore how LLMs and OCR can accelerate the transcription process, with strict validation and corrections of the data, to ensure semantic accuracy, or, try other ontologies like OMIT [21]. In addition, more advanced prompt engineering strategies could further reduce human effort by improving the reliability and consistency of LLM outputs.

Acknowledgments. Many thanks to all students who participated in this project, namely: Dimitris Andronikou, Pavlos Ziotas, Georgios Panagiotis Mellios, Christos Konstantinos Panorios, Vasileia-Eirini Pitikaki, Emmanouil Troulis, Petros Tsalikis, Torbjorn Fries, Evaggelos Mixelioudakis, Ilias Kapsis, Maria Foukaki and Despoina Athanasiou.

References

1. Basset, A., Los, W.: Biodiversity e-science: Lifewatch, the european infrastructure on biodiversity and ecosystem research. Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology **146**(4), 780–782 (2012)
2. Bauer, F., Kaltenböck, M.: Linked open data: The essentials. Edition mono/-monochrom, Vienna **710**(21) (2011)
3. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)
4. Cohere: Command r+. `https://cohere.com/blog/command-r-plus` (2024)
5. DeepMind, G.: Gemini 1 technical report. `https://deepmind.google/technologies/gemini/` (2023)
6. Delis, A.: Seafaring lives at the crossroads of mediterranean maritime history. International Journal of Maritime History **32**(2), 464–478 (2020)
7. Dirgahayu, T., Setiaji, H., et al.: Semantic web in disaster management: a systematic literature review. In: IOP Conference series: Materials science and engineering. vol. 803, p. 012043. IOP Publishing (2020)
8. Doerr, M.: The CIDOC Conceptual Reference Model: an ontological approach to semantic interoperability of metadata. AI magazine **24**(3), 75–75 (2003)

9. Doerr, M., Hiebel, G., Kritsotaki, A., Rousakis, Y., Schmidle, W., Theodoridou, M., Velios, A., et al.: Definition of the CRMsci - An Extension of CIDOC CRM to support scientific observation. Tech. rep., https://cidoc-crm.org/crmsci
10. Doerr, M., Ore, C.E., Fafalios, P., Kritsotaki, A., Stead, S., et al.: Definition of the CRMinf - An Extension of CIDOC-CRM to support argumentation (2023), https://cidoc-crm.org/crminf
11. Doerr, M., Theodoridou, M.: Mapping archaeological databases to CIDOC-CRM. In: CAA (2011), https://publications.ics.forth.gr/_publications/DOERR_MappingCIDOC_1_wFig.pdf
12. Erling, O.: Virtuoso, a Hybrid RDBMS/Graph column store. IEEE Data Eng. Bull. **35**(1), 3–8 (2012)
13. Fafalios, P., Konsolaki, K., Charami, L., Petrakis, K., Paterakis, M., Angelakis, D., Tzitzikas, Y., Bekiari, C., Doerr, M.: Towards semantic interoperability inhistorical research: Documenting research data and knowledge with synthesis. In: ISWC. pp. 682–698 (2021)
14. Fafalios, P., Kritsotaki, A., Doerr, M.: The SeaLiT ontology–an extension of CIDOC-CRM for the modeling and integration of maritime history information. ACM Journal on Computing and Cultural Heritage **16**(3), 1–21 (2023)
15. Fafalios, P., Marketakis, Y., Axaridou, A., Tzitzikas, Y., Doerr, M.: A workflow model for holistic data management and semantic interoperability in quantitative archival research. Digital Scholarship in the Humanities **38**(3), 1049–1066 (2023). https://doi.org/10.1093/llc/fqad018
16. Fafalios, P., Petrakis, K., Samaritakis, G., Doerr, K., Kritsotaki, A., Tzitzikas, Y., Doerr, M.: Fast cat: collaborative data entry and curation for semantic interoperability in digital humanities. Journal on Computing and Cultural Heritage (JOCCH) **14**(4), 1–20 (2021)
17. Felicetti, A., Gerth, P., Meghini, C., Theodoridou, M.: Integrating heterogeneous coin datasets in the context of archaeological research. In: Proceedings of the Workshop on Extending, Mapping and Focusing the CRM co-located with 19th International Conference on Theory and Practice of Digital Libraries (2015), Poznań, Poland, September 17, 2015. CEUR Workshop Proceedings, vol. 1656, pp. 13–27 (2015), https://ceur-ws.org/Vol-1656/paper2.pdf
18. GeoNames Team: Geonames. https://www.geonames.org (2024)
19. Google Cloud: Cloud Vision API: Optical Character Recognition (OCR). https://cloud.google.com/vision/docs/ocr (2024)
20. Group, W.O.W.: Owl 2 web ontology language document overview (second edition). https://www.w3.org/TR/owl2-overview/
21. Huang, J., Dang, J., Borchert, G.M., Eilbeck, K., Zhang, H., Xiong, M., Jiang, W., Wu, H., Blake, J.A., Natale, D.A., et al.: Omit: dynamic, semi-automated ontology development for the microrna domain. PLoS One **9**(7), e100855 (2014)
22. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked data finland: A 7-star model and platform for publishing and re-using linked datasets. In: ESWC. pp. 226–230 (2014)
23. Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D., Lefrançois, M.: Sosa: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics **56**, 1–10 (2019)
24. Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710. Soviet Union (1966)
25. Ma, Z., E. Santos, J., Lackey, G., Viswanathan, H., O'Malley, D.: Information extraction from historical well records using a large language model. Sci. Rep. **14** (2024). https://doi.org/10.1038/s41598-024-81846-5

26. Martin, P., Remy, L., Theodoridou, M., Jeffery, K.G., Zhao, Z.: Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment. Future Gener. Comput. Syst. **101**, 1–13 (2019). https://doi.org/10.1016/J.FUTURE.2019.05.076
27. Mastoras, V., Vassiliades, A., Rousi, M., Diplaris, S., Mavropoulos, T., Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I.: Towards a framework for seismic data. In: Iberoamerican Knowledge Graphs and Semantic Web Conference. pp. 106–119 (2023)
28. Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., Van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web **6**(6), 539–564 (2015)
29. Musen, M.A.: The protégé project: a look back and a look forward. AI matters **1**(4), 4–12 (2015)
30. Oikonomakis, I., Fafalios, P., Mountantonakis, M., Tzitzikas, Y.: Modeling modern and historical data as a knowledge graph: A case study for earthquake data. In: Proceedings of the 18th Metadata and Semantics Research Conference (MTSR) (2024)
31. Oldman, D., Tanase, D.: Reshaping the knowledge graph by connecting researchers, data and practices in researchspace. In: ISWC. pp. 325–340 (2018)
32. OpenAI: Gpt-4 (large language model). https://openai.com/research/gpt-4
33. OpenAI: Openai o1. https://openai.com/o1/ (2024)
34. OpenAI: Openai o3. https://openai.com/index/introducing-o3-and-o4-mini/ (2025)
35. Papadopoulos, G.A.: A Seismic History of Crete. The Hellenic Arc and Trench - Earthquakes and Tsunamis: 2000 BC - 2011 AD (2011)
36. Petrakis, K., Samaritakis, G., Kalesios, T., Domingo, E., Delis, A., Tzitzikas, Y., Doerr, M., Fafalios, P.: Digitizing, curating and visualizing archival sources of maritime history: the case of ship logbooks of the nineteenth and twentieth centuries. Drassana pp. 60–87 (2021). https://doi.org/10.51829/Drassana.28.649
37. Philips, J.P., Tabrizi, N.: Historical document processing: A survey of techniques, tools, and trends (2020), https://arxiv.org/abs/2002.06300
38. Piantanida, M., Bonamini, E., Dolfi, M., Auer, C., Caborni, C., Bergero, F., Staar, P.: Using knowledge graphs to navigate through geological concepts extracted from documents (2021)
39. Platakis, E.: Oi seismoi tis kritis. Kritika Chronika **4**, 494–496 (1950)
40. Remy, L., Ivanovic, D., Theodoridou, M., Kritsotaki, A., Martin, P., Bailo, D., Sbarra, M., Zhao, Z., Jeffery, K.G.: Building an integrated enhanced virtual research environment metadata catalogue. Electron. Libr. **37**(6), 929–951 (2019)
41. Schwitter, N.: Using large language models for preprocessing and information extraction from unstructured text: A proof-of-concept application in the social sciences. Methodological Innovations **18**(1), 61–65 (2025). https://doi.org/10.1177/20597991251313876
42. Sinnott, R.W.: Virtues of the haversine. Sky and telescope **68**(2), 158 (1984)
43. Stanford Center for Biomedical Informatics Research: Ontograf - protégé wiki. https://protegewiki.stanford.edu/wiki/OntoGraf (2025)
44. Touvron, H., Martin, L., Stone, K., et al.: Llama 2: Open foundation and fine-tuned chat models (2023), https://arxiv.org/abs/2307.09288
45. Uematsu, H., Takeda, H.: Earthquake lod: Seismic dataset construction with ontology oriented design patterns. In: 12th International Joint Conference on Knowledge Graphs (IJCKG 2023) (2023)

46. Uematsu, H., Takeda, H.: Earthquake ontology and lod. In: ISWC (Posters/Demos/Industry) (2023)
47. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
48. W3C: W3C RDF validation service. `https://www.w3.org/RDF/Validator/` (2025)
49. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), `https://arxiv.org/abs/2201.11903`
50. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)