Musical Meetups Knowledge Graph (MMKG): a collection of evidence for historical social network analysis

Alba Morales Tirado[1][0000−0001−6984−5122] Jason Carvalho[1][0000−0001−6557−3190]
Marco Ratta[1][0000−0003−3788−6442] Chukwudi Uwasomba[1][0000−0002−6572−2835]
Paul Mulholland[1][0000−0001−6598−0757] Helen Barlow[1]Trevor
Herbert[1][0000−0002−0436−3314] Enrico Daga[1][0000−0002−3184−5407]

The Open University, Milton Keynes, United Kingdom {alba.morales-tirado, jason.carvalho, marco.ratta, chukwudi.uwasomba, helen.barlow, trevor.herbert, paul.mulholland, enrico.daga}@open.ac.uk

Abstract. Knowledge Graphs (KGs) have emerged as a valuable tool for supporting humanities scholars and cultural heritage organisations. In this resource paper, we present the Musical Meetups Knowledge Graph (MMKG), a collection of evidence of historical collaborations between personalities relevant to the music history domain. We illustrate how we built the KG with a hybrid methodology that, combining knowledge engineering with natural language processing, including the use of Large Language Models (LLM), machine learning, and other techniques, identifies the constituent elements of a *historical meetup*. MMKG is a network of historical meetups extracted from ∼33k biographies collected from Wikipedia focused on European musical culture between 1800 and 1945. We discuss how, by providing a structured representation of social interactions, MMKG supports digital humanities applications and music historians' research, teaching, and learning.

Keywords: Knowledge Graph · historical encounters · MEETUPS · Historical Social Network Analysis.

Resource Type: Knowledge Graph

DOI: 10.5281/zenodo.7924618
URL: https://github.com/polifonia-project/meetups-knowledge-graph
Endpoint: https://polifonia.kmi.open.ac.uk/meetups/sparql/

1 Introduction

The Semantic Web community has been very active in generating knowledge from unstructured sources in domains such as scientific knowledge, social media, and digital humanities. In this regard, Knowledge Graphs (KGs) have emerged

as a valuable tool for supporting humanities scholars and cultural heritage organisations [3,1,26,5]. In the EU project Polifonia[1], we study how knowledge graphs can support scholarship in music history. In this paper, we present the Musical Meetups Knowledge Graph (MMKG), a collection of evidence of historical collaborations between personalities relevant to the music history domain. Our resource aims at representing documentary evidence of social interactions in the music history domain, to support the needs of humanities scholars. This includes capturing the evidence text and decorating it with semantic entities, such as type of events, participants, temporal expressions and spatial instances, enabling the exploration of complex *historical meetups*. We build on previous work focused on conceptualising the domain of interest in the Meetups Ontology [21]. Here, we report on the work undertaken to generate a KG extracted from 33,309 biographies collected from Wikipedia focused on European musical culture between 1800 and 1945. Our work provides a structured representation of historical exchanges to enable the discovery of new insights and support music historians' research, teaching, and learning. Therefore, in this paper, we contribute 1. Musical Meetups Knowledge Graph (MMKG), a Knowledge Graph of documentary evidence of musical collaborations from ∼33k biographies from Wikipedia 2. A KG generation pipeline, combining Natural Language Processing (NLP), Large Language Models (LLM), Knowledge Engineering, and Knowledge Graph Construction (KGC) techniques. The rest of the paper is structured as follows. We discuss the background and motivation for our work in Section 2. In Section 3 we illustrate the Meetups Ontology. Section 4 describes the hybrid knowledge graph generation pipeline. We evaluate the knowledge graph in Section 5 for its ability to answer key competency questions and via a survey with domain experts. Next, we discuss feedback on the utility and usability of MMKG from two types of stakeholders: domain experts of a Music Department and application developers of Digital Humanities tools (Section 6). We report on relevant related work in Section 8, before discussing conclusions and future work (Section 9).

2 Motivation

Music historians and those involved in the arts and humanities research process rely heavily on information and knowledge contained within historical manuscripts and the biographies of figures from history. A common method used for historical investigation is narrative inquiry [23] in which historical evidence is first organised into a chronicle including evidence of events in temporal order. The evidence is then filtered from the chronicle comprising exchanges with common features of interest, generating, for example, storylines of encounters of a particular location (e.g. London), purpose (e.g. music making) or participant (e.g. Elgar). Such storylines can then be used to investigate comparisons and temporal shifts, for example, why music-making is more common in certain locations during particular periods.

[1] http://polifonia-project.eu

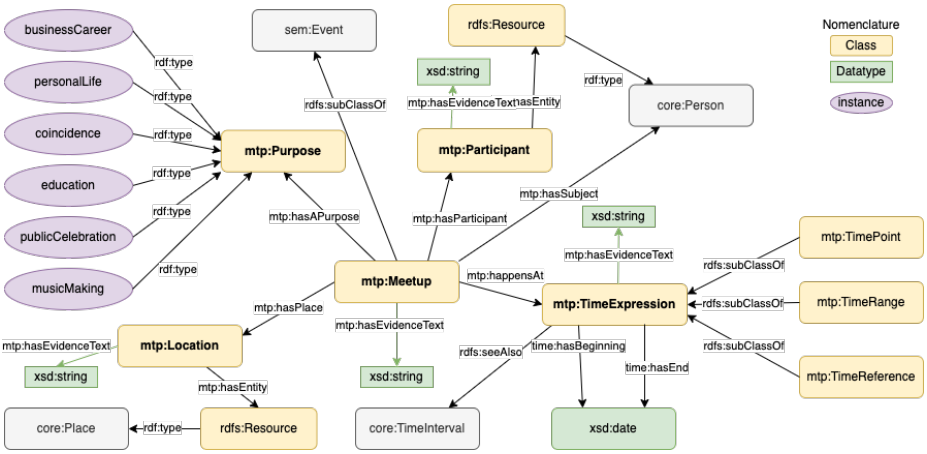

Fig. 1: Meetups Ontology.

Such forms of inquiry are highly resource-intensive in terms of extracting the chronicle of social interactions from source documents and organising them into storylines for further analysis. The development of a knowledge graph comprising the evidence of historical interactions extracted from source documents that can then be queried to create storylines (e.g. the music creation events of a composer or those occurring in a geographic location) would not only enhance the depth of historical analysis but also facilitate a more thorough and nuanced interpretation of the dynamic interplay between music and societal evolution.

A knowledge graph would also open up the possibility for Exploratory Data Analysis (EDA) [25] of storylines to reveal patterns and behaviours that might evade detection during examinations at a smaller scale. By leveraging these techniques, historians can pinpoint temporal and spatial trends, discern correlations between genres and historical periods, and unveil unexpected connections, thus enriching their grasp of the cultural and social contexts surrounding musical encounters.

3 Meetups Ontology

We use as a guiding framework the Meetups Ontology[2] (See Figure 1) that models the elements of a historical meetup. As presented in previous work [21], the scope of the ontology is the analysis of historical encounters and collaborations of people in the musical world. These knowledge requirements are formalised as Competency Questions (listed in the ontology repository), following good practices in ontology engineering. The ontology considered commonly used vocabu-

[2] Meetups Ontology repository: https://github.com/polifonia-project/meetups-ontology

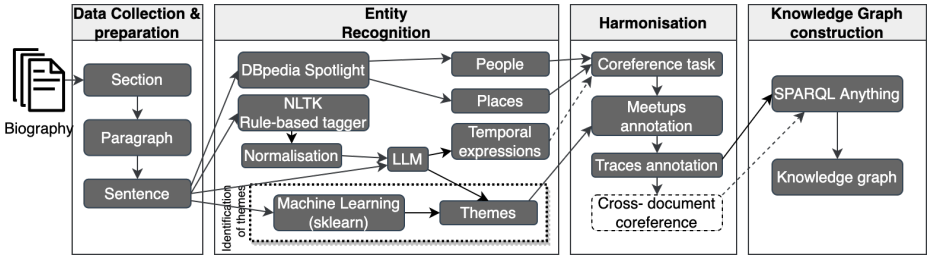

Fig. 2: MMKG construction pipeline.

laries such as Time Ontology, PROV Ontology, SEM and the Polifonia CORE Ontology.

Here we give a brief description of the main Meetups Ontology classes. A historical meetup – `mtp:Meetup`, is derived from evidence within a biography – `mtp:hasEvidenceText`. Mentions of at least one or more participants and places are represented by the `mtp:Participant` and `mtp:Location` class, respectively. Each mention is an entity (`mtp:hasEntity`) extracted and linked to DBpedia or Wikidata (Section 4 gives details on the extraction process). To represent the time when the meetup took place, we use the `mtp:TimeExpression` class. It is composed of start `time:hasBeginning` and end `time:hasEnd` dates as well as the text from where it was compiled. Lastly, the purpose of the encounter is defined by (`mtp:Purpose`) and six different subclasses specified by domain experts. These classes are adjacent but not disjoint and each meetup can be annotated with more than one purpose, namely: `mtp:BusinessCareer`, `mtp:PersonalLife`, `mtp:Coincidence`, `mtp:Education`, `mtp:PublicCelebration` or `mtp:MusicMaking`.

4 Knowledge Graph generation pipeline

This section describes the knowledge extraction pipeline developed to build MMKG (see Figure 2). Our approach focuses on extracting evidence that describes historical meetups according to Meetups Ontology. We apply knowledge extraction techniques and methods for text processing to recognise, classify and link the entities that are part of a historical meetup, particularly: people, places, time expressions and themes. In what follows, we describe in detail the steps taken towards the construction of the KG.

4.1 Data collection & preparation

The first step of the pipeline is dedicated to the collection of data. We rely on Wikipedia and its database of open-access biographies. To obtain the list of music personalities, we queried the DBpedia SPARQL endpoint, obtaining all the entities of type `dbo:MusicalArtist`. The biographies of 33,309 personalities

were collected[3] in text format[4]. Next, we prepare the text by removing empty spaces, redundant line breaks and special characters. We organise the text by assigning indexes to each sentence, grouped by paragraphs and sections. The output is a corpus of indexed sentences for each biography.

4.2 Entity recognition

Identification of people and place entities. People and places correspond to two of the main elements that characterise a historical meetup; such entities indicate who was involved and where encounters happened.

Therefore, we use the DBpedia Spotlight[5] tool to automatically annotate mentions of people and places linking them to DBpedia resources. Once entities are extracted, we select people and places. For places, we filter the following types: `wikidata:Q41176, wikidata:Q486972, dbo:Place, dbo:Location` or `wikidata:Q6256`. For people, types should be one of `dbo:Person, wikidata:Q215627` or `dbo:MusicalArtist`. To solve a data quality issue of DBpedia Spotlight (e.g., errors due to name variation or entity ambiguity [22]), we perform an additional evaluation to verify that the entity is an instance of class `wd:Q5` and has a date of birth on Wikidata. We also filter out people whose date of birth is posterior to the subjects' date of death or whose date of death is before the subject's dob (meaning that they could not have actually met them).

The final output is a bag of entities containing mentions of person and place, grouped by sentence and biography.

Identification and normalisation of temporal expressions. This task is divided into two parts: **(a) identification of temporal expressions and (b) normalisation**. To **(a)** extract temporal expressions from text, we use a rule-base tagger, based on the research by [27] and their implementation of SynTime[6]. The tool is a three-layer system that recognises time expressions using syntactic token types (part-of-speech POS tags) and general heuristic rules. Unlike SynTime, our implementation exploits the NLTK Toolkit and was developed in Python. Importantly, the heuristics rules were revised and expanded adapting them to the Meetups corpus. Furthermore, we classify each expression according to the type of dates they represent, we use the syntactic token types for this classification: *time range* (e.g., from 1959 to 1970), *time point* (e.g., exact date, 23/03/1294), and *time reference* (e.g., usually incomplete dates (19 April), later this year). The result is a list of temporal expressions (a short piece of text), for instance, "2 June 1857" (POS tag "CD NNP CD") for each sentence.

To allow temporal analysis of meetups with SPARQL, we **(b)** normalise temporal expressions into a XSD date time compliant format. We consider time

[3] Data was collected on January 2022
[4] Meetups corpus: https://github.com/polifonia-project/meetups_corpus_collection
[5] DBpedia-Spotlight: https://www.dbpedia-spotlight.org/
[6] SynTime software: https://github.com/zhongxiaoshi/syntime

```
You are a tool that extracts time references and returns results in ISO8601 date format. I
↪  will provide a sentence and the target text. 1. You should decide if the target text
↪  represents a time reference. 2. If the target text is not a specific time reference
↪  return "NO". 3. If the target text is a specific time reference then estimate the
↪  approximate dates in the context of the sentence. 4. If the information is not enough
↪  to calculate a date then use {}'s biography information. 5. Check the results are in
↪  ISO8601 format. 6. Return the results in JSON format using two keys: start_date and
↪  end_date. Estimate dates after point 4. Don't return fake dates. Sentence {text},
↪  temporal expression {}
```

Listing 1: Prompt to normalise temporal expressions

as ranges, with a start and end time point. We follow the ISO8601[7] format (`YYYY-MM-DD`) using two Python libraries (`dateutil`[8] and `approx-dates`[9]). On average 65% of temporal expressions (by biography) are normalised automatically. The left 35% corresponds to expressions such as *"the next seven years"* or *"of the twentieth century"*. To cope with these cases, we make use of the LLM tool ChatGPT[10], providing as input context the temporal expression, the sentence where the expression was identified, and the subject of the biography (see listing 1). We perform dedicated experiments tailoring the prompt to return the best results regarding quality (accurate to the expression) and coverage (number of temporal expressions normalised). By including LLMs in the normalisation process we increase the number of temporal expressions parsed to 82% on average.

The final output is a bag of temporal expressions represented by the textual evidence, start and end date and additional information such as POS tags and the way it was normalised (python library or LLM).

Improving identification of meeting purpose using LLM tools. The main element of a historical meetup is the reason for the encounter, a type of meetup, which is named as *Purpose* according to the MEETUPS Ontology. As detailed in previous work [21] our pipeline included a Machine Learning approach that follows a semi-supervised classification process, annotating each sentence in the corpus and assigning them one of the meetup types. In [21], we identified that 74% of predictions were correct, and the 26% left were either Partially Correct or Incorrect. Therefore, we explored the use of the LLM tool ChatGPT to increase the accuracy of the automatic classification. We followed a zero-shot learning approach meaning we provided as input the piece of text to analyse, and the list of classes. We ask the tool to return the two classes that better describe the meetup type. The prompt (See listing 2) was designed having three main elements. First the context of the task (line 1); the expected format output, a JSON response (line 2); and finally instructing the tool on the classification task according to the set list of classes (lines 3 to 6).

[7] ISO8601 International Standard: https://en.wikipedia.org/wiki/ISO_8601
[8] dateutil library: https://dateutil.readthedocs.io/en/stable/index.html
[9] approx-dates library: https://pypi.org/project/approx-dates/
[10] ChatGPT: https://openai.com/product

For each sentence, we gather the two most probable topics of the text. For instance, the text: *'His father, William Henry Elgar (1821–1906), was raised in Dover and had been apprenticed to a London music publisher.'* will be classified as follows:

1. "Music making"; the explanation being "The sentence mentions a music publisher, indicating a connection to music making"; and
2. "Personal life"; the given explanation is "The sentence also provides information about the personal life of Elgar's father".

```
You are a knowledge classification system that annotates sentences according to their main topic.
Respond in json format using the following keys: thm_type_1, thm_type_2, thm_explanation_1 and thm_explanation_2.
The value for thm_type_1 is the first most probable topic, use only one of the following keys: ['Music making', 'Business meeting', 'Personal life', 'Coincidence', 'Public celebration', 'Education'].
The value for thm_type_2 is the second most probable topic, use only one of the following keys: ['Music making', 'Business meeting', 'Personal life', 'Coincidence', 'Public celebration', 'Education'].
The value for thm_explanation_1 should be a short explanation for the topic in thm_type_1. Less than 100 characters.
The value for thm_explanation_2 should be a short explanation for the topic in thm_type_2. Less than 100 characters.
```

Listing 2: Prompt to classify sentences according to the meetup type (purpose)

We sampled 83 sentences and asked three annotators to verify the accuracy of the *purpose* generated. We compared these manual annotations with the results of the automatic extraction using the Machine Learning (ML) approach, and the Large Language Model (LLM). Table 1 displays the results in terms of Precision. Using LLM tools leverages the classification of text, increasing its precision to 85% (on average).

		Machine Learning (ML) prediction		**LLM prediction**	
	# Sentences	**Precision @ 1**	**Precision @ 2**	**Precision @ 1**	**Precision @ 2**
Edward Elgar	56	0.34	0.64	0.45	0.73
Yehudi Menuhin	11	0.36	0.45	0.55	0.82
Clara Butt	16	0.69	0.75	0.88	1
	83	0.46	0.62	*0.62*	*0.85*

Table 1: Classification results. Comparison ML and LLM Precision @ 1 and 2

The output of this task is a list of sentences grouped by paragraph, each with two of the most probable meetup types. Since we use LLM tools and the API to obtain the results. We decided to keep results from the ML approach when it was not possible to query the tool and obtain an LLM response; this is reflected in the ontology using the `mtp:hasSourcePurpose` attribute.

4.3 Harmonisation

Coreferences While the entity identification task identifies named entities, there are also mentions of people and places that are not always automatically identified by DBpedia Spotlight, this happens when such mentions are implicit, in the form of noun phrases or pronouns. For example, people referenced in the text as he or she. To maximise the identification of entities we perform a Coreference Resolution (CR) task, finding entities and their coreference mentions [16] and linking them to DBpedia or Wikipedia resources.

We use the spaCy library *coreferee*[11]. The library receives as input a paragraph text, then it identifies the entities' mentions (person or a place) and groups them into chains of mentions. We use the coreference chains to verify that an entity is part of the bag of entities in a sentence. New entities are added when implicit mentions are listed in the chain. The final output is a dataset of sentences, each with an extended bag of entities that now includes coreferent mentions.

Identification of historical meetups At this stage, we have a dataset of sentences, each of them including zero or more entity types (people, place, temporal expressions and meetup type). However, a *historical meetup* can be described in consecutive sentences, having complementary information. In this step, we harmonise the data by joining adjacent sentences representing the same social interaction.

To identify historical meetups, we built an algorithm that traverses adjacent sentences, incrementally (see listing 3) and applies a set of heuristics. The method checks the sentence being evaluated (A) and the previous sentence (B). The algorithm starts comparing if A has all elements of a meetup: time, place and person (all sentences have a purpose annotation), and then applies the following rules:

- If sentence A does not have time but its place is the same as B's place then sentence A inherits B's time (line 6).
- On the contrary, if sentence A does not have a place but its time is the same as B's time then sentence A inherits B's place (line 7).
- If sentence A does not have a person (participants) but its time is the same as B's, and A's place is the same as B's, then sentence A inherits B's people (line 8).
- Finally, if A does not have time and place, but its person is the same as B's, then A inherits B's entities (line 9).

In any case, the algorithm verifies that whenever an entity type is missing, it can be complemented by the previous sentence given that it complies with having the same participants, place or time accordingly. If A has all the elements, whether complemented by B or not, then it is considered a meetup (line 10).

The second part of the algorithm checks whether sentence B complements sentence A:

[11] Coreferee a spaCy library: https://spacy.io/universe/project/coreferee/

- If all elements of A are the same as B, sentences describe the same evidence and, consequently, can be considered the same meetup (line 11).
- Other cases include when A and B have the same entity, and B lacks the other two: 1. Line 13 describes the case of equal A's and B's time but B lacking person and place, 2. line 14 same person, but B having no time or place, and finally, 3. line 15 the same place but B having no person or time.

For instance, the evidence *"His only formal musical training beyond piano and violin lessons from local teachers consisted of more advanced violin studies with Adolf Pollitzer, during brief visits to London in 1877–78."* has all entity types. People participating (Elgar and Adolf), a location (London) and a date (1877-18), meeting conditions at line 10. We can mention an example of adding context thanks to the coreference task. Sentences (B) *"Dessay had collaborated frequently with Michel Legrand in concerts."* and (A) *"In May 2009, she dedicated two concerts of songs written by him in Toulouse."* represent a meetup. Sentence A satisfies conditions in line 11 (having all the elements of a meetup). And in line 13, B lists the same people but no time or place entities. Therefore, sentences A and B can be joined in a single meetup.

Importantly, when one of the entity types is not present we annotate the evidence as *historical trace*. This type of evidence can be useful to complement social analysis. For example, the sentence *"While in France, he visited his fellow composer Frederick Delius at his house at Grez-sur-Loing."* describes how Edward Elgar and Frederick Delius met in France. However, it does not indicate the date when it took place. This information is still important if we want to answer questions such as the places he visited or the people he met throughout his life.

```
1  A = aSentence() # the current sentence being analysed
2  B = previousSentence() # the previous sentence
3  # place -> all place entities
4  # person -> all people entities
5  # time -> all temporal expressions
6  if !A.time & A.place == B.place -> A.time = B.time
7  if !A.place & A.time == B.time -> A.place = B.place
8  if !A.person & A.time == B.time & A.place == B.place -> A.person = B.person
9  if !A.time & !A.place & A.person == B.person -> A.time = B.time & A.place = B.place
10 if A.time & A.place & A.person -> A = meetup
11 if A.time = B.time & A.place = B.place & A.person = B.person
12     -> meetup = A + B # sentence A and B have the same entities, can be considered a meetup
13 if A.time = B.time & !B.person & !B.place or
14         A.person = B.person & !B.time & !B.place or
15         A.place = B.place & !B.person & !B.time
16     ->  meetup = A + B # A + B have complementary entities and can be considered a meetup
```

Listing 3: Historical meetups identification

The output is a dataset that contains the text (typically a sentence or a set of sentences), and the list of entities that account for a meetup. The results are stored in CSV files, grouped by biographies. This dataset is ready to be transformed into the MMKG in the following step.

4.4 Knowledge Graph construction

The KG was constructed using the CSV files resulting from the process described so far and applying the Meetups ontology described in Section 3. We use SPARQL Anything [2] and design CONSTRUCT mappings, to create triples from each biography. MMKG contains data from 33,309 artists' biographies, 16,748 of which have at least one historical meetup. The KG describes a total of 45,812 historical meetups. The meetups mention 49,170 people involved in different encounters. So far, the historical meetups gathered around 7,107 places and 51,120 time expressions. The KG is currently published in Turtle and N-quads RDF format and available in the MMKG GitHub repository.

5 MMKG evaluation

We evaluate MMKG by implementing queries to answer the competency questions of the Meetups Ontology and by means of a survey with domain experts.

5.1 Answering CQs

The knowledge requirements, which are the foundation of the Meetups ontology design, were formalised as a list of Competency Questions (CQs) in [21]. In this section, we take as guidelines these CQs (Table 2[12]) and design a series of SPARQL queries[13] to evaluate that the MMKG data meets the knowledge requirements.

Table 2: List of Competency Questions (CQs).

#	Competency Questions	Entity focus
1	What places did musician Z visit in his/her career?	Place
2	Where did musician X and performer Y meet?	where?
3	Why did musician X and performer Y meet?	Purpose /
4	What is the nature of the event (a celebration, a festival, a private event, a performance, accidental)?	Meetup type why?
5	When did musician X and performer Y meet?	Temporal
6	Did musician X and performer Y ever meet?	when?
7	Who other musicians were working at the same time?	Participants
8	What was the composer's network?	who?

CQs focus on place dimension. One of the main requirements is about the places where people met or visited. We take the example of the German pianist Clara Schumann and build a query (Figures 3a and 3b) that retrieves the list of places she visited during her life (16 places in total). To answer question 2 we build a query that lists all the places she and Joseph Joachim shared, evidence shows that they met in Germany and the UK.

[12] Due to space constraints, Table 2 displays a summary of the CQs

[13] Queries and results obtained available in the MMKG repository - queries folder https://github.com/polifonia-project/meetups-knowledge-graph/

```
SELECT DISTINCT ?resource ?placeLabel
WHERE {
  VALUES ?subject { <http://dbpedia.org/resource/Clara_Schumann> }
  ?s mtp:hasSubject ?subject ; mtp:hasType ?type .
  FILTER (regex ( str (?type), str ("HM") ) ) .
  ?s  mtp:hasPlace ?aPlaceIRI .
  ?aPlaceIRI mtp:hasEntity ?resource .
  ?resource rdfs:label ?placeLabel . }
```

resource	placeLabel
http://dbpedia.org/resource/Leipzig	Leipzig
http://dbpedia.org/resource/Berlin	Berlin
http://dbpedia.org/resource/Dresden	Dresden
http://dbpedia.org/resource/London	London

(a) Query: Places visited by an artist (b) Results: Places visited by an artist

Fig. 3: Illustrative example for CQs focused on place entities

CQs focus on purpose. Following the previous example, we can expand the queries to include the meetup's purpose and answer questions 3 and 4. Evidence shows that Clara and Joachim's meetings mainly related to "Music Making". For instance, *"In October–November 1857, Schumann and Joachim went on a recital tour to Dresden and Leipzig."*.

CQs focus on temporal expressions. In the previous example, we have textual evidence that indicates the time Clara and Joachim met in 1857. This information is available in the MMKG and can be queried in the form of start and end dates (Figures 4a and 4b).

```
SELECT DISTINCT ?text ?nrmlsdValueStart ?nrmlsdValueEnd ?timeEvidenceText
WHERE {
  VALUES ?subject { <http://dbpedia.org/resource/Clara_Schumann> }
  VALUES ?participant { <http://dbpedia.org/resource/Joseph_Joachim>
    <https://www.wikidata.org/wiki/Q159976> }
  ?s mtp:hasSubject ?subject ; mtp:hasType ?type .
  FILTER (regex ( str (?type), str ("HM") ) ) .
  ?s mtp:happensAt ?timeExpression_IRI .
  ?timeExpression_IRI a mtp:TimeExpression ;
  OPTIONAL {  ?timeExpression_IRI  time:hasBeginning ?nrmlsdValueStart ;
    time:hasEnd ?nrmlsdValueEnd ; mtp:hasEvidenceText ?timeEvidenceText }.
```

```
<result>
    <binding name='text'>
        <literal>In October–November 1857, Schumann and Joachim went
            on a recital tour to Dresden and Leipzig.</literal>
    </binding>
    <binding name='nrmlsdValueStart'>
        <literal datatype='http://www.w3.org/2001/XMLSchema#date'>
            1857-01-01</literal>
    </binding>
    <binding name='nrmlsdValueEnd'>
        <literal datatype='http://www.w3.org/2001/XMLSchema#date'>
            1857-12-31</literal>
    </binding>
    <binding name='timeEvidenceText'>
        <literal>1857</literal>
    </binding>
</result>
```

(a) Query: Date of the meetup (b) Results: Date of the meetup

Fig. 4: Illustrative example for CQs focused on temporal expressions

CQs focus on person dimension. We follow the previous example and build queries that return the list of people Clara met (17 people in total) (Figure 5a). Importantly, we can expand this query to include all the artists Clara met (a total of 50 people), even if the evidence does not provide data on the time or place they met (a historical trace); results are shown in Figure 5b.

<http://dbpedia.org/resource/Carlo_Alfredo_Piatti>	Carlo Alfredo Piatti
<http://dbpedia.org/resource/Clara_Schumann>	Clara Schumann
<http://dbpedia.org/resource/Franz_Grillparzer>	Franz Grillparzer
<http://dbpedia.org/resource/Franz_Liszt>	Franz Liszt
<http://dbpedia.org/resource/Frederick_Augustus_II_of_Saxony>	Frederick Augustus II of Sax
<http://dbpedia.org/resource/Friedrich_Wieck>	Friedrich Wieck
<http://dbpedia.org/resource/Greifswald>	Greifswald
<http://dbpedia.org/resource/Johannes_Brahms>	Johannes Brahms
<http://dbpedia.org/resource/Joseph_Joachim>	Joseph Joachim

<http://dbpedia.org/resource/Franz_Grillparzer>	Franz Grillparzer
<http://dbpedia.org/resource/Franz_Liszt>	Franz Liszt
<http://dbpedia.org/resource/Franz_Schubert>	Franz Schubert
<http://dbpedia.org/resource/Frederick_Augustus_II_of_Saxony>	Frederick Augustus II of Saxony
<http://dbpedia.org/resource/Friedrich_Kalkbrenner>	Friedrich Kalkbrenner
<http://dbpedia.org/resource/Friedrich_Wieck>	Friedrich Wieck
<http://dbpedia.org/resource/Frédéric_Chopin>	Frédéric Chopin
<http://dbpedia.org/resource/George_Bernard_Shaw>	George Bernard Shaw
<http://dbpedia.org/resource/George_Frideric_Handel>	George Frideric Handel

(a) Query: Meetups and participants (b) Results: All people she met

Fig. 5: Illustrative example for CQs focused on people

Importantly, we produced SPARQL queries to extract statistics about the top visited places, people met, years of activity and purpose of meetings for each artist. These queries can be found in the MMKG repository[14]

5.2 Feedback questionnaire from domain experts.

In a survey with 12 domain experts from the Music Department of our University, we evaluated the value to users of MMKG[15]. Participants were either/or researchers (75%), musicians (33%), educators (33%), and historians (25%), with a significant 91.7% reporting daily engagement with music-related content. All the respondents agreed on the value of documenting musical history encounters, considering them either important (50%) or very important (50%). All respondents reported not being aware of any tool/database to store and organise historical music encounters. We asked about the value of specific elements of the ontology. All respondents rated the importance of documenting the people involved in musical encounters as "very important"(Fig 6). The place and time of the encounter were rated very important and important by 83.33% and 16.67%, respectively. Responses were varied for the purpose of encounters: 58.34% rated it as "Very Important," 33.33% as "Important," and 8.33% as "Moderately" important.

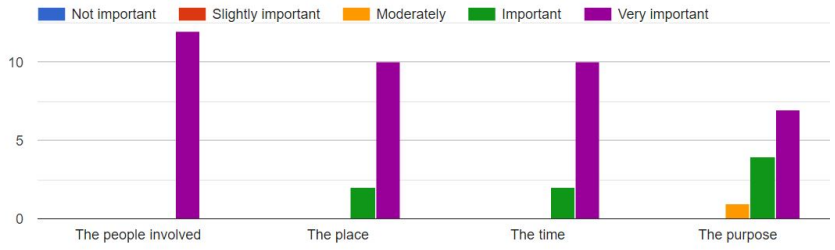

Fig. 6: Dimensions in musical encounters.

Business meetings and music-making have been highly valued throughout music history, which received over 83% of the rating as "very important" and "important, respectively". The survey also revealed that 75% of the respondents respectively rated personal life and public celebration as "very important" and "important" purpose for musical encounters. While education and coincidence were rated the same by 66.67%.

In addition, all participants agree on the utility of exploring the geographical, temporal, and thematic proximity of subjects, even outside direct evidence of documented encounters. Finally, the usefulness of knowledge graphs in teaching music history is largely (75%) acknowledged by experts.

[14] Stats queries: https://github.com/polifonia-project/meetups-knowledge-graph/blob/main/queries/top-entities.sparql

[15] Complete results can be found at this address: https://docs.google.com/document/d/1PGknESmJm4f_-QwSdebTuiHIZ3bMaBrtcs4-9AKLPcc/edit?usp=sharing

6 Using the MMKG

In this section, we discuss the value of the MMKG from the perspective of both domain experts in their research and exploration and developer communities leveraging the data to build novel applications and exploration tools.

6.1 Music historians and domain experts

The MMKG, as described in Section 2, offers an exploration of historical information at a macro-scale using EDA principles, in ways not possible when simply studying the biographies of individual musical figures.

This is especially significant for the period before globalisation through broadcasting and recording technologies, when it is often much less obvious how musical ideas and influences were disseminated. The MMKG opens the possibility of identifying points of cultural and musical exchange that have been unsuspected till now, revealing patterns of travel and contact, and intersections of musical figures identified by place and time. For example, from the biography of a composer, the researcher might generate a visualisation that reveals meetings with another musician with whom they were not previously known to have been in contact. This may shed new light on music the composer subsequently wrote, or on a performer's subsequent repertoire choices. Similarly, a previously unknown meeting of a musician with an instrument maker might turn out to be related to the development of a new or modified instrument, or the composition of a new piece for a particular type of instrument. The knowledge graph may also reveal a cluster of meetings in a particular place, prompting the researcher to develop new research questions exploring that place as musically or culturally significant.

Traditional humanities research methods are based on linear investigations in which evidential sources are aligned to research questions and interrogated within predetermined contexts. In providing a non-linear vision of accumulated data capable of stimulating new and often unconsidered ideas, MMKG both complements and expands this process. It offers a different set of perspectives on which to base the evaluation and ordering of data, to shape research questions and to determine the direction and priorities of a research project.

6.2 Developer communities

In Section 2 we give an overview of some of the principles of Exploratory Data Analysis, within the context of the MMKG project. These are facilitated by the use of carefully selected and formatted attributes and annotations that allow developers to leverage and integrate the data into data-driven visualisations and applications. These annotations can be used across various mapping and time-display technologies, libraries, and software. Geolocation data, in particular, enables seamless integration with mapping libraries, Geographical Information System (GIS) software, and advanced geoprocessing routines. Common

JavaScript web mapping libraries such as Leaflet[16] or Openlayers[17] can be used to render points or clusters on maps, making use of additional annotations for nuanced visualisations such as colour-coding by theme. The data is also adaptable for integration into GIS software, allowing intricate geoprocessing and geoanalysis tasks such as creating thematic heatmaps, plotting paths, implementing advanced clustering algorithms, and executing task-specific querying and analysis processes. Temporal data offers similar visualisation techniques through the use of timeline displays, frequency distribution analysis, and visual time-based search tools. Again, GIS software can also leverage this temporal information to provide advanced temporospatial geoprocessing.

Beyond academic research, the MMKG offers developers the chance to create applications in other domains such as education, archives, and exploration. MMKG's data can power interactive learning platforms, browsable archive repositories, immersive museum exhibits and exploration tools. This flexibility caters to diverse needs, showcasing KG's potential to transcend disciplinary boundaries.

Accessible through an open SPARQL endpoint, the KG seamlessly integrates into numerous platforms such as mobile apps and desktop applications. Developers can integrate straightforward HTTP calls into their application development and also build intermediary APIs, facilitating efficient data retrieval and reuse of common SPARQL queries. This accessible approach, often using standardised formats like JSON, not only streamlines integration but also promotes collaboration within the developer community, encouraging collective utilisation.

7 Resource Availability, Reusability, Sustainability

Our work on this project has been completed whilst adhering to the FAIR[18] principles. We have built the resource in a number of specific ways to ensure that the outputs are in compliance with the FAIR Guiding Principles for scientific data management and stewardship.

Resource availability for the MMKG is ensured through its presence on GitHub, providing both source data and documentation. Additionally, a permanent SPARQL endpoint is openly accessible on the web, facilitating ease of use and integration into various applications. Furthermore, developers can encapsulate these results as API calls, promoting sharing and collaboration within the developer community. These features contribute to the sustainability of the MMKG, ensuring ongoing accessibility, usability, and collaborative engagement over time.

8 Related work

Below we consider related work on musical social relationships, the publishing of biographical and prosopographical data, event-based KGs and their application.

[16] https://leafletjs.com/
[17] https://openlayers.org/
[18] https://www.go-fair.org/fair-principles/

Within the musical domain, [6] presents an analysis of musical influence networks for sample-based music. In contrast, [15] analyses the MusicBrainz dataset relationship metadata to uncover how music artists influence one another.

For biographical/prosopographical data, [26] provide a linked data model to integrate biographical and cultural heritage data while [5] presents a KG of biographical information of German academics from the 16th to 18th century. [17] performs similar research but with short textual biographies about Finnish and Swedish academics. [20] extracts information from text about the historical movements of people, focusing on biographies from the first half of the 20th century and [4] presents a relational database of biographies spanning 5,500 years. [13] [11] [24] and [12] argue towards the adoption of Linked Data practices for the development of databases, applications and Artificial Intelligence systems based on biographies and/or prosopographies.

An Event KG is a graph where the knowledge representation is centred upon dynamic events happening in time rather than entities and relations. While [10] provides a general survey of Event KGs, [19] and [18] research related technical aspects involved in their construction KGs, such as event coreference resolution and temporal knowledge extraction from texts.

In historiography [7] and [9] present an event KG representing 690 thousand events enriched by a timeline generation system representing biographies. In [14] the available archive information of Finland's involvement in WWII is rendered as an event KG.

While relevant resources have been published in the domains as mentioned earlier, there is however evidence (Table 3) of a resource gap when presenting evidence of historical collaborations between personalities relevant to the music history domain. MMKG addresses this gap.

	P	**T**	**L**	**R**	**Description**	**Size**
BiographySampo [13]	✓	e	✓	✗	Finnish Biographies	13.100 biographies
AcademySampo [17]	✓	e	✓	✗	Finnish university student records	27,500 student records
Event KG [8]	✓	e	✓	✗	News, focus on events	18,510 news articles
Early Modern Scholarly Career KG [5]	✓	r	✓	✗	German biographies Professional development	Two sources, # of biographies unknown
Meetups Musical KG	✓	✓	✓	✓	Music personalities	33.300 biographies

Table 3: Knowledge Graph comparison. P - People, L - Location, T - Time, and R - Reason; e - related to an event only, r - related to a role only.

9 Conclusions and Future work

In this resource paper, we introduced MMKG, a knowledge graph of documentary evidence of social interaction for supporting research in music history. Our work shows the potential of hybrid methods for knowledge extraction, combining knowledge engineering with techniques from traditional NLP and current LLMs tools. Future work includes studying how to improve the coverage of the geospatial and temporal annotations of the biographies, for example, tackling

implicit, contextual time references. Further research regarding social iterations and network influence (e.g., musical and creative aspects) are natural directions to expand the use of the KG. A user interface for leveraging the potential of MMKG is under development in close collaboration with domain experts. MMKG and the related tools will be applied to teaching and learning content as well as scholarship of the Music Department of the OU.

Acknowledgements

This work was supported by the EU's Horizon Europe research and innovation programme within the Polifonia project (grant agreement N. 101004746).

References

1. Adamou, A., Brown, S., Barlow, H., Allocca, C., d'Aquin, M.: Crowdsourcing linked data on listening experiences through reuse and enhancement of library data. International Journal on Digital Libraries **20**(1), 61–79 (2019)
2. Asprino, L., Daga, E., Gangemi, A., Mulholland, P.: Knowledge graph construction with a façade: a unified method to access heterogeneous data sources on the web. ACM Transactions on Internet Technology **23**(1), 1–31 (2023)
3. de Berardinis, J., Carriero, V.A., Jain, N., Lazzari, N., Meroño-Peñuela, A., Poltronieri, A., Presutti, V.: The polifonia ontology network: Building a semantic backbone for musical heritage. In: International Semantic Web Conference. pp. 302–322. Springer (2023)
4. Beytía, P., Schobin, J.: Networked pantheon: a relational database of globally famous people. Available at SSRN 3255401 (2018)
5. Blanke, J., Riechert, T.: Towards an rdf knowledge graph of scholars from early modern history. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC). pp. 471–472. IEEE (2020)
6. Bryan, N.J., Wang, G.: Musical influence network analysis and rank of sample-based music. In: ISMIR. pp. 329–334 (2011)
7. Gottschalk, S., Demidova, E.: Eventkg: A multilingual event-centric temporal knowledge graph. In: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, Proceedings 15. Springer (2018)
8. Gottschalk, S., Demidova, E.: Eventkg - the hub of event knowledge on the web - and biographical timeline generation (May 2019), http://arxiv.org/abs/1905.08794
9. Gottschalk, S., Demidova, E.: Eventkg+ bt: Generation of interactive biography timelines from a knowledge graph. In: The Semantic Web: ESWC 2020 Satellite Events: ESWC 2020 Satellite Events, Heraklion, Crete, Greece. Springer (2020)
10. Guan, S., Cheng, X., Bai, L., Zhang, F., Li, Z., Zeng, Y., Jin, X., Guo, J.: What is event knowledge graph: A survey. IEEE Transactions on Knowledge and Data Engineering (2022)
11. Hyvönen, E.: Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web (2020)
12. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K., et al.: Linked data–a paradigm change for publishing and using biography collections on the semantic web. Biographical Data in a Digital World (2022)

13. Hyvönen, Eero and Leskinen, Petri and Tamper, Minna and Rantala, Heikki and Ikkala, Esko and Tuominen, Jouni and Keravuori, Kirsi: BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. Springer International Publishing (2019)
14. Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J., Hyvönen, E.: Warsampo knowledge graph: Finland in the second world war as linked open data. Semantic Web **12**(2), 265–278 (2021)
15. Kopel, M.: Analyzing music metadata on artist influence. In: Proceedings of ACIIDS 2015, Bali, Indonesia. Springer (2015)
16. Lata, K., Singh, P., Dutta, K.: Mention detection in coreference resolution: survey. Applied Intelligence (Jul 2022). https://doi.org/10.1007/s10489-021-02878-2
17. Leskinen, P., Hyvönen, E.: Using the academysampo portal and data service for biographical and prosopographical research in digital humanities. In: ISWC-Posters-Demos-Industry 2021 International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks. CEUR-WS. org (2021)
18. Liu, Y., Hua, W., Zhou, X.: Temporal knowledge extraction from large-scale text corpus. World Wide Web **24**, 135–156 (2021)
19. Lu, J., Ng, V.: Event coreference resolution: A survey of two decades of research. In: IJCAI. pp. 5479–5486 (2018)
20. Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., Lepri, B.: Ramble on: Tracing movements of popular historical figures. In: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics. pp. 77–80 (2017)
21. Morales Tirado, A., Carvalho, J., Mulholland, P., Daga, E.: Musical meetups: a knowledge graph approach for historical social network analysis. In: SEMMES 2023: Semantic Methods for Events and Stories Workshop co-located with 20th European Semantic Web Conference (ESWC 2023) (May 2023), https://ceur-ws.org/Vol-3443/ESWC_2023_SEMMES_Meetups-CR.pdf
22. Olieman, A., Azarbonyad, H., Dehghani, M., Kamps, J., Marx, M.: Entity linking by focusing dbpedia candidate entities. In: Proceedings of the first international workshop on Entity recognition & disambiguation - ERD 14. p. 13–24 (2014). https://doi.org/10.1145/2633211.2634353
23. Polkinghorne, D.E.: Narrative knowing and the human sciences. Suny Press (1988)
24. Tamper, M., Leskinen, P., Apajalahti, K., Hyvönen, E.: Using biographical texts as linked data for prosopographical research and applications. In: Euro-Mediterranean Conference. pp. 125–137. Springer (2018)
25. Tukey, J.W., et al.: Exploratory data analysis, vol. 2. Reading, MA (1977)
26. Tuominen, J.A., Hyvönen, E.A., Leskinen, P.: Bio crm: A data model for representing biographical data for prosopographical research. In: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017). CEUR Workshop Proceedings (2018)
27. Zhong, X., Sun, A., Cambria, E.: Time Expression Analysis and Recognition Using Syntactic Token Types and General Heuristic Rules. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017)