Musical Meetups: a Knowledge Graph approach for Historical Social Network Analysis

Alba Morales T[1], Jason Carvalho[1], Paul Mulholland[1] and Enrico Daga[1]

[1]*Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom*

Abstract

The large-scale analysis of historical events data makes it possible to trace key points of cultural and social exchange in history. There has been research focused on facilitating the integration and interpretation of events from heterogeneous sources (such as memoirs, books, and biographies) mainly considering events as a sequence of spatiotemporal objects. However, exploring and discovering new connections (e.g., collaborations, interactions) between people does require characterising those events with dimensions that are relevant to the scholarly enquiry such as the actual participants and nature of the event. This paper describes the concept of *historical meetup* to represent the encounters (for instance, collaborations, exchanges, links) between personalities of European history and formalise its constituent parts as an ontology. Furthermore, we report on preliminary work undertaken to generate a Knowledge Graph of historical meetups extracted from encyclopedic sources, i.e. biographies collected from Wikipedia. We discuss our results and illustrate the challenges of extracting such type of knowledge from biographical sources. The current experimental setting explores historical meetups in the European musical culture between 1800 and 1945. Our work sketches the basis for applying event knowledge graphs to cultural and social history research, providing support for the analysis, and exchange of ideas and practices.

Keywords

Historical meetups, historical social network, Knowledge Graph, Polifonia, MEETUPS Ontology

1. Introduction

The scholarly study of cultural and social heritage is a dynamic flow of experiences, leaving heterogeneous traces difficult to capture, connect, access, interpret, and valorise. Social network analysis (SNA) has gathered the attention of researchers across sociology and computer science and can be a useful tool for cultural heritage research. By mapping out the connections, facts, and interactions between individuals, SNA can support the exploration and understanding of experiences, collaboration patterns and relationships, casting new light on aspects of cultural and social exchange [1]. For instance, social network analysis can be used to study the relationships between musicians, producers, and composers in the musical world and how these relationships influence the production and creation of musical pieces.

However, the interpretation and analysis of historical encounters using documentary sources can be a complex task. For example, sources are usually heterogeneous and complex datasets

SEMMES 2023: Semantic Methods for Events and Stories workshop ESWC 2023 Heraklion, Greece, May 28, 2023
alba.morales-tirado@open.ac.uk (A. M. T); jason.carvalho@open.ac.uk (J. Carvalho); paul.mulholland@open.ac.uk (P. Mulholland); enrico.daga@open.ac.uk (E. Daga)
0000-0001-6984-5122 (A. M. T); 0000-0001-6557-3190 (J. Carvalho); 0000-0001-6598-0757 (P. Mulholland); 0000-0002-3184-5407 (E. Daga)

[1, 2], such as books, memoirs, and biographies; the extraction of knowledge involves the implementation of various techniques and methods that typically requires technical expertise and skills that might not be common among scholars in the cultural heritage domain. Crucially, the main challenge researchers face is precisely the discovery of meaningful collaborations and influences among people. Beyond human participants, social interactions involve other entities such as organisations (e.g., record companies), places (location of the encounter), temporal relations (when and for how long) and types of events (e.g., festivals or concerts), including multiple and overlapping historical encounters [3]. Clearly, the analysis cannot be dedicated only to one entity (e.g., people) but extended to cover and complete the overall picture of such interactions.

In this paper, we introduce the concept of *historical meetups* to describe historical encounters, for instance, collaborations and exchanges between personalities in European history. The MEETUPS Ontology models the constituent elements of a historical meetup, including entities such as people, places, temporal relations and types of events. The ontology aims to support researchers in performing a thorough analysis of social networks and discovering potential new links, therefore, addressing challenges in terms of interpretation and analysis complexity. The design rationale of the ontology was motivated by considering the knowledge requirements of scholars exploring social-historical networks and formalised as Competency Questions (CQs) in Section 4. Furthermore, MEETUPS Ontology is used as the framework to create a Knowledge Graph (KG) of historical meetups extracted from encyclopedic sources, for instance, Wikipedia biographies. The experimental setting explores social interactions in the European music world between 1800 and 1945. We develop our work in the context of the MEETUPS Pilot[1] part of the Polifonia project. Our main contributions are:

- The MEETUPS Ontology[2] for representing historical encounters between personalities in European history.
- The preliminary work that implements a knowledge extraction pipeline[3] for building a KG of historical meetups from biographic sources, specifically Wikipedia.
- The first version of the MEETUPS Knowledge Graph[4].
- A set of SPARQL queries for validation of the MEETUPS Ontology.

The rest of the paper is organised as follows. Section 2, presents a review of relevant research on the construction of KG for cultural heritage and SNA of historical meetups crafted from biographies. Section 3 details the methodology used to build the MEETUPS Ontology. Section 4 gives an account of the knowledge requirements and the development process for the ontology. Section 5 presents the evaluation of the ontology. Section 6 describes preliminary work to build a KG of historical meetups, and Section 7 summarises the challenges found in building the KG. Finally, Section 8 presents the concluding remarks and future work.

[1]MEETUPS Pilot: https://polifonia-project.eu/pilots/meetups/
[2]MEETUPS Ontology repository: https://github.com/polifonia-project/meetups-ontology
[3]Knowledge extraction pipeline repository: https://github.com/polifonia-project/meetups_pilot
[4]MEETUPS Knowledge Graph repository: https://github.com/polifonia-project/meetups-knowledge-graph

2. State of the art

The collection of data about historical encounters makes it possible to trace key points of cultural exchange, dissemination and interesting connections between people, organisations and other entities. However, documentary sources (e.g., books, biographies, memoirs) are often unstructured and difficult to analyse, making it challenging for researchers to gain insights into the lives of individuals and the historical contexts in which they collaborated. The construction of knowledge graphs to represent structured data has the potential to shed light on the vast documentary resources (physical and digital) and provide an overview of events that is both comprehensive and easy to understand for scholars in the cultural heritage domain [1, 4], importantly supporting social network analysis methods.

Initial work in cultural heritage and social network analysis focused on using Linked Data (LD) and Knowledge Graphs (KG) to study general aspects of artist interaction. For instance, in the musical domain, research explored how certain musical pieces influenced musical genre development and aspects of music composition [5, 6]. However, the discovery of meaningful collaborations and cultural exchange among artists requires an analysis that should be expanded beyond general aspects and incorporate events as an essential component of cultural heritage research which brings together other entities such as participants, locations, and temporal relations [7, 3, 1, 2].

By drawing on the concept of historical events, current research expands on the development of ontological models and KGs for implementing event-centric approaches. For instance, work by [8] extracts events from a temporal perspective. In contrast, work by [9] prioritise the analysis of events from the perspective of social interactions by linking biographies of historical personalities, both studying specific elements of an event. The work presented by [2] aims to provide users with tools that facilitate the exploration of personalities' biographies. In their work, [2] follow a knowledge-based approach to connect other people's biographies and places, supporting knowledge discovery. However, their approach is limited to single document analysis, meaning events only gather one biographee's events from Finish personalities.

In contrast to these studies, our approach focuses on modelling events from the perspective of musical exchanges, including various entities, to bring a comprehensive view of meetings. Our approach aims to design a whole process to extract data and build a knowledge base of historical meetups. Unlike [8, 9] that expand their work from existing databases, we use biographies collected from Wikipedia. Furthermore, we expand the analysis of personalities across Europe from the 1800s to 1945 instead of being specific to a location, such as the [2] interested in Finish personalities. Importantly, we list encounters' categories built in collaboration with musical scholars, facilitating the study of knowledge exchanges and providing means for discovering interesting links between personalities.

3. The Historical MEETUPS Ontology

In what follows, we describe the methodology used to develop the MEETUPS Ontology and present the model for representing historical meetups. We follow ontology engineering good practices and methodologies [10, 11, 12], and devise the following steps:

- **Abstracting the scenario and identifying knowledge requirements.** In the first step, we identify the scope and purpose of the ontology. As described in Section 1, in this paper, we introduce the concept of *historical meetups* to represent encounters and collaborations that influenced personalities of European history. In this work, our experimental setting explores historical meetups in the musical culture of Europe between 1800 and 1945 and uses this scenario as a guide for identifying key ontology concepts and relationships that define our ontology. The knowledge requirements of scholars looking to uncover such encounters are identified and stated in the form of Competency Questions (CQs). The CQs will guide the ontology development process, providing the main framework to evaluate the expressiveness of the ontology.
- **Ontology design and construction - implementation.** The second step consists of a) identifying the concepts and relations constituting the ontology, b) using definitions to express the ontology in a formal language, and c) determining the reuse of established ontologies and how to integrate them into the MEETUPS Ontology. We built the ontology using the Protege ontology editor and the Web Ontology Language (OWL).
- **Knowledge Graph population.** The third step is dedicated to building the MEETUPS Knowledge Graph (KG). We developed a knowledge extraction pipeline that determines the entities part of a historical meetup and performs Natural Processing Language (NLP) tasks to extract and link such entities. We use SPARQL Anything [13] to generate RDF data and store the KG for evaluation in the next step.
- **Ontology validation.** The last step consists of the evaluation of the ontology. The main aspect to consider is the fulfilment of the CQs. We code the CQs into SPARQL queries and use them to interrogate the KG built in the previous step. By comparing the query results and the expected values from the CQs, we expect to obtain validation if the ontology fulfils the knowledge requirements.

A detailed account of the construction process for the MEETUPS Ontology is given in the following section.

4. Implementation of the MEETUPS Ontology

This section is divided into two parts. The first part introduces the process for identifying and capturing knowledge requirements. The second part is dedicated to defining the classes, relationships and ontology construction.

4.1. Knowledge requirements and CQs

First, we capture the knowledge requirements regarding support for historians and teachers in the exploration and visualisation of encounters between personalities in European musical history. The knowledge requirements were collected as part of collaborative work between scholars and researchers from the music department at The Open University and the Knowledge Media Institute (KMi); this work is part of the Polifonia project and the MEETUPS pilot.

Three Persona stories[5] were developed to exemplify the knowledge requirements from the

[5]Persona stories: https://github.com/polifonia-project/stories

perspective of target communities and scholarly users. The stories describe a series of scenarios regarding users' scholarly activities, their data needs and interest in terms of music and cultural heritage, for instance, their requirements to analyse interactions between music personalities and the discovery of patterns for social exchange. The predominant Persona story for our experimental works is Ortenz, selected by music scholars. In what follows, we give a brief description of the Personas.

- *Ortenz*: She is a research fellow at the Music Department. Her background is in literature and art history. Ortenz would like to have a system for visualising events (meetings of composers and musicians) in time and space in order to track musicians' careers, their overlap and intersections, gather trends in time and space and make emerging patterns of knowledge transmission.
- *David*: He is a professor and music historian working in the music department of a university. David is interested in understanding the social history of music, e.g. who were the musicians, who was the audience, and how a particular musical environment relate to the wider musical environment.
- *Sophia*: She is a musicologist and a practising musician. Sophia is interested in understanding the social-historical reasons behind how music was created and how it sounds.

We provide typical sentences from the historical domain that exemplify the information about personalities and events of interest for scholars. In this case, we take textual excerpts from Edward Elgar's[6] biography:

1. *To mark the coronation of **Edward VII**, Elgar was **commissioned** to set A. C. Benson's Coronation Ode for a gala concert at the **Royal Opera House** on **30 June 1902**.*
2. *In his first trips abroad, Elgar **visited** Paris in **1880** and Leipzig in **1882**. He **heard Saint-Saëns** play the organ at the Madeleine and **attended** concerts by first-rate orchestras.*
3. *Throughout his life, Elgar was often **inspired** by close women friends; Helen Weaver was succeeded by **Mary Lygon**, Dora Penny, Julia Worthington, Alice Stuart Wortley and finally Vera Hockman, who enlivened his old age.*

These examples are interesting for several reasons: (a) they contain details of events that influenced Elgard's work: he attended other personalities' performances (e.g., *Saint-Saëns* , he was inspired by contemporaneous personalities (e.g., *Mary Lygon*); (2) it contains explicit time references such as 1902 and 1880 which serve as guidance to search for other musicians that were working at the same time; (3) direct connections to other music personalities with whom he interacted; (4) and places he visited. For example, it is well accepted that European musicians influenced Elgard's work; however, questions such as *was he influenced by their surroundings (e.g., places he lived in, people he met, and shows he attended, **what** type of events)? Was his work part of emerging trends or patterns, **when** this happened? **who** were the contemporaneous emerging personalities?* are of interest to music scholars.

Based on the description of the Personas' stories, a list of CQs was generated. The list of relevant CQs is enumerated below:

[6]Edward Elgar biography, source Wikipedia: https://en.wikipedia.org/wiki/Edward_Elgar

Table 1: List of Competency Questions (CQs).

#	Story	Competency Question	Entity focus
1	Ortenz	What places did musician Z visit in her career?	Place where?
2	Ortenz	Where did musician X and performer Y meet?	
3	Ortenz	Did attendees travel to reach the place?	
4	David	Where were the musicians coming from?	
5	David	Where were the places in which musicians (X, Y, Z, etc.) played?	
6	Sophia	Who other musicians were working at the same time?	Participants who?
7	Sophia	What was the composer's network (patrons, institutions ...)?	
8	Ortenz	When did musician X and performer Y meet?	Time when?
9	Ortenz	Did musician X and performer Y ever meet?	
10	Ortenz	How attendees happened to be there?	Purpose / Event why?
11	Ortenz	Where did she perform?	
12	Ortenz	Where did she live?	
13	Ortenz	Why did musician X and performer Y meet?	
14	Ortenz	What is the nature of the event?	
15	Ortenz	Was it a celebration, a festival, a private event?	
16	Ortenz	Was it a religious or a secular event?	
17	Ortenz	Were attendees invited or was it an accidental meeting?	
18	Ortenz	How can we characterise the relationships among the participants (e.g., Patreon / Musician)?	

From this collaborative requirements acquisition exercise, it is clear that researchers expect to explore encounters that could potentially reveal unexpected connections and relationships. There is also a need for tools that support the analysis of social interactions from the perspective of "events", including layers such as time, places and participants. Therefore, the concept of historical meetups was introduced to describe the study of intersections and links among personalities from the perspective of the event and taking into account other dimensions, such as participants, place and time expressions.

The meetup type describes the *reason/purpose* for an encounter. According to scholars and researchers of the Music Department at The Open University, meetups can be grouped as business and career (e.g., contract, retirement), personal life (e.g., born, divorce, family), coincidence (e.g., meet, find, discover), education (e.g., learn, teach, conservatoire), public celebration or music-making (e.g., play, produce). Other dimensions of analysis include the *participants* involved, for instance, the person that is the subject of interest and the people interacting in the encounter, the *place* where it took place (e.g., city, country, venue), and the *date* when it took place. There could also be events that do not contain all the entities described previously.

4.2. Ontology construction

From this analysis, we identify the vocabulary, relationships, and properties of the entities of interest. Therefore, this section is dedicated to describing the core components of the MEETUPS

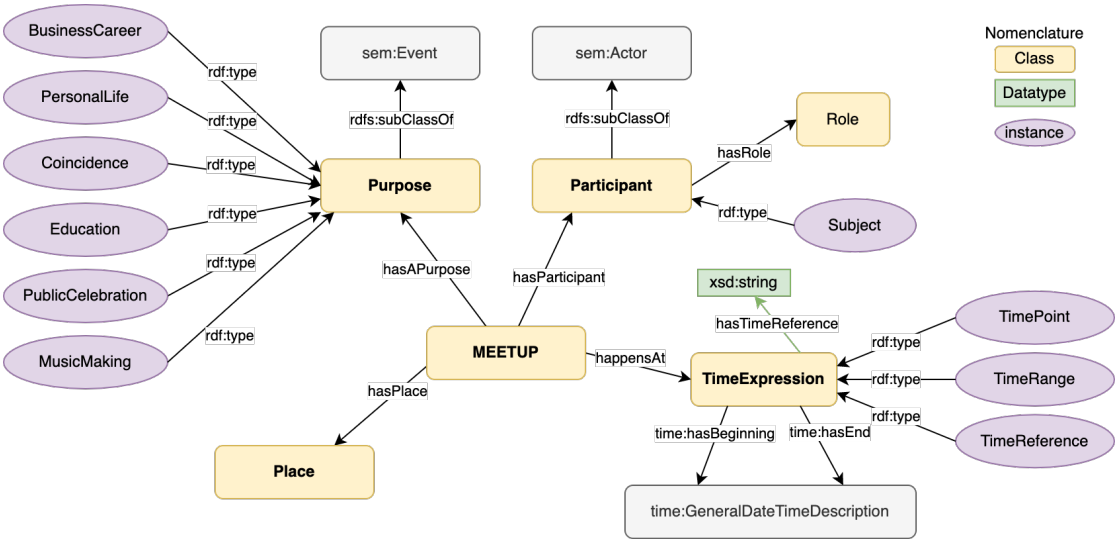

Figure 1: MEETUPS ontology, core classes.

Ontology. Furthermore, we explain how we incorporated established ontologies such as Time Ontology [14], PROV Ontology [15] and SEM (Simple Event Model) [16] in our design.

The preferred prefix for MEETUPS Ontology is `meetups:` and the namespace for all MEETUPS terms is `https://w3id.org/polifonia/ontology/meetups-ontology#`. The core class of MEETUPS ontology is `meetups:Meetup` (see Figure 1), which is the representation of a *historical meetup* between people in the musical world. The main classes of the ontology are `Purpose`, `Participant`, `Place` and `TimeExpression`.

A historical meetup is defined by the purpose of the encounter, a `meetups:Purpose` class, which could be classified into one of the types of encounters: `:BusinessCareer`, `:PersonalLife`, `:Coincidence`, `:Education`, `:PublicCelebration` or `:MusicMaking`. A meetup involves at least one person (a `meetups:Subject`) and one or more participants represented by the class `meetups: Participant` that could take a specific role (`meetups:hasRole`) in the musical encounter (`meetups:Role`). Historical meetups typically include a reference to the place (`meetups:Place`) where people or the activity developed and the date (`meetups:TimeExpression`) that indicates when this encounter took place.

Moreover, the ontology design includes classes for recording the provenance of the sources (see Figure 2). As described previously, typically data can be collected from encyclopedic resources such as biographies, and open public web pages; these sources are represented with the class `meetups:Sources`. At the same time, a source can `hasProvider`, represented by `meetups:Provider` class; for example, in our experiments, we use Wikipedia.

5. Evaluation of the ontology

As described in Section 1, this work aims to represent historical encounters between personalities in European history. We built the MEETUPS Ontology following the list of knowledge requirements collected in Section 4. We use a sample of randomly selected biographies to build

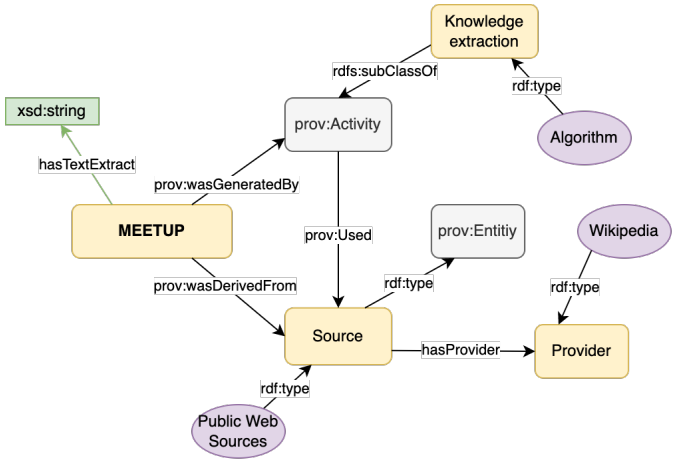

Figure 2: MEETUPS ontology, provenance description

the KG and provide data for evaluation.

We used Protégé and the Hermit reasoner to check the formal consistency of the ontology. In order to evaluate the expressivity of the ontology, we grouped the Competency Questions (CQs) listed in Section 4 according to the most relevant type of entity it questions about and then encoded them into SPARQL queries[7].

CQs (1 to 5) focus on place entities. These CQs inquire particularly about the locations where people met or visited. We take the example of the English composer Edward Elgar and build two queries (a) one that retrieves all the places he visited during his life, (b) one that retrieves all the places where he had an encounter with Adolf Pollitzer. As shown in Figures 3a and 3b, the ontology support such queries.

```
SELECT distinct ?o
WHERE
{
    ?s ont:place ?o .
    ?s ont:subject
       <http://dbpedia.org/resource/Edward_Elgar> .
}
```

(a) Places visited by a given artist

```
SELECT *
WHERE
{
    ?s ont:subject
       <http://dbpedia.org/resource/Edward_Elgar> .
    ?s ont:person
       <http://dbpedia.org/resource/Adolf_Pollitzer> .
    ?s ?p ?o .
}
```

(b) Places linked to two artist encounters

Figure 3: SPARQL queries for CQs focused on place entities

CQs (6 and 7) focus on participant entities. These competency questions refer to finding other contemporary artists and what were their social networks. We follow the example of Edward Elgar and built two queries (a) one that retrieves all contemporaneous artists given a specific date (see Figure 4a) and (b) a query that retrieves all people the artist met during their life and their role (see Figure 4b).

CQs (8 to 9) focus on time expression entities. These competency questions inquire about the time when artists met. We exemplify how the ontology answers this question by

[7]All available in the repository.

```
SELECT distinct ?subject ?date
WHERE
{
    ?s ont:subject ?subject .
    ?s ont:timeExpression ?date .
    FILTER contains(?date,"1902")
}
```

(a) People working at the same time

```
SELECT distinct ?o
WHERE
{
    ?s ont:person ?o .
    ?s ont:subject
       <http://dbpedia.org/resource/Edward_Elgar> .
    ?s ont:role ?role .
}
```

(b) People met during their life

Figure 4: SPARQL queries for CQs focus on participant

building a query (see Figure 5a) that retrieves the date specific personalities met.

CQs (10 to 18) focus on purpose/events entities. Finally, these CQs inquired about what was the main purpose of the meeting. For instance, CQ 11 asks for the places where an artist performed (e.g., the meetup type is `:MusicMaking`), or CQ 12 questions where they live (`PersonalLife`), questions that can be answered thanks to the list of categories of events gathered as part of the knowledge requirements task (see Section 4). In the example (see Figure 5b), we query the list of meetups and their type.

```
SELECT * #distinct ?o
WHERE
{
    ?s ont:subject
       <http://dbpedia.org/resource/Edward_Elgar> .
    ?s ont:person
       <http://dbpedia.org/resource/Edward_VII> .
    ?s ont:timeExpression ?date .
}
```

(a) Date at which personalities met

```
SELECT *
WHERE
{
    ?s ont:subject
       <http://dbpedia.org/resource/Edward_Elgar> .
    ?s ?p ?o .
    ?s ont:hasPurpose ?purpose .
}
```

(b) List of events by type

Figure 5: SPARQL queries for CQs 10 to 18

6. Preliminary extraction pipeline and Knowledge Graph generation

This section describes the knowledge extraction pipeline developed to build the MEETUPS Knowledge Graph (see Figure 6). As described in Section 1 our work focuses on facilitating the discovery of connections between personalities of European history. Therefore, our approach is focused on identifying the entities that are part of a historical meetup: people, places, time expressions and meetup type. The entity recognition and disambiguation process is performed at the sentence level. We start by collecting data that provides an overview of music personalities' lives, such as biographies in text format. Next, we process the corpus to link entities such as places and people, extracting time expressions and classifying them according to meetup types. The output of this step is a bag of entities at the sentence level, grouped by paragraphs and sections. Once the entities are extracted, we run the process to find encounters within the text, at sentence level or on a larger chunk of text using coreference resolution. The last step is dedicated to building the knowledge graph of historical meetups using as a framework the MEETUPS Ontology and the data output from the previous step. In what follows, we describe in detail the steps taken towards constructing the KG.

Data collection. The first step of the pipeline is dedicated to the collection of data that

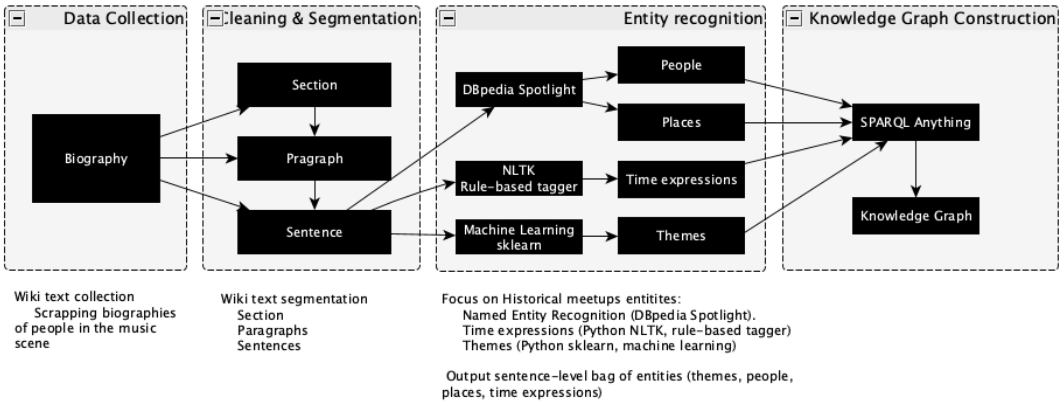

Figure 6: MEETUPS Knowledge Graph construction pipeline.

supports music historians' exploration of encounters between people in the musical world in Europe between the 1800s and 1945. Therefore, we rely on documentary sources such as biographies and select an open-access database such as Wikipedia as the primary data input. To obtain the list of targeted personalities, we built and queried the DBpedia SPARQL endpoint. We filtered biographies by date of birth in order to reflect our targeted sample and gathered a total of 33,309 biographies. Each biography was collected in text format from Wikipedia, using the identifier from DBpedia. In order to make experiments manageable, we randomly sampled a total of 1000 biographies that will be inspected in the following steps.

Cleaning & Segmentation. In the previous step, the biographies were collected in text format; therefore, the second step is dedicated to preparing the dataset to execute the Named-Entity Recognition and Disambiguation (NERD) tasks. First, we clean the text by removing line breaks, special characters and empty spaces. In order to facilitate the location and classification of the entities, the text is organised in sentences, grouped by paragraphs and sections. The final output is a corpus of biographies, each one containing a set of sentences identifiable by its position within paragraphs and sections.

Entity recognition. As explained previously, the main objective is to explore encounters by characterising them with dimensions that are relevant to scholars, specifically: people, places, temporal expressions and types of encounters. Therefore, we focus on locating and identifying the constituent parts of a historical meetup within sentences. In what follows, we detail the tasks developed to recognise each type of entity:

- *Named-entity recognition and disambiguation (NERD) for people and place entities.* People and places are two main elements that characterise a historical meetup. As listed in Table 1, knowing who was involved and where encounters happened is required. Therefore, this task is focused on determining people or places mentioned in the corpus and specifying them uniquely using a targeted knowledge base. We make use of DBpedia Spotlight[8] tool to automatically annotate mentions of people and places linking them to DBpedia resources.

[8]DBpedia-Spotlight: https://www.dbpedia-spotlight.org/

- *Identification of temporal expressions.* Another important requirement has to do with the date or moment in time when a particular encounter happened. To extract this information, this task uses a rule-based Time Expression recognition tagger based on research by [17] and SynTime software[9]. The authors implement a three-layer system that recognises time expressions using syntactic token types and general heuristic rules. Unlike SynTime, our software uses NLTK Toolkit and was developed using Python; also, the heuristic rules were revised and expanded for our experiments. Furthermore, this component classifies time expressions into three types of expressions: time range (e.g., from XX to XX, from XX, to XX, until XX), time point (e.g., exact date, 23/03/1294), and time reference (e.g., usually incomplete dates (19 April), two weeks, later this year).

Identification of meetup themes. The main element of a historical meetup is the theme, in other words, the topic/subject of the encounter. This task uses Machine Learning techniques and performs a semi-supervised classification process to annotate the meetup type (see Section 4.1) of each sentence in the corpus. In what follows, we describe the steps of the theme identification process in detail:

- *Building the training dataset.* In order to create the training dataset, we use a distant supervision approach. First, we built a curated list of keywords for each one of the meetup types; the list was developed with the advice of domain experts. Next, we randomly select a sample of sentences and annotate them automatically with one of the classes (meetup types) if it includes one of the keywords in the list. The final output is a dataset of sentences (a total of 16140) annotated according to one of the meetup types (see Table 2 for statistics).
- *Classification task - training and testing of Machine Learning (ML) algorithms.* In this step, we use the dataset built previously and implement a supervised classification approach. As shown in Table 2, the distribution of sentences by class is unequal; therefore, the first task is building a balanced dataset. We take as reference the lowest number of sentences per group, in this case, the class *Education with 629 sentences.* For each class, we select randomly 629 annotated sentences to ensure each class has the same number of annotations. The new training dataset has a total of 3774 annotated sentences. Then, we prepare the dataset for the training process; we use SBERT to capture the context of the sentences [18] and as input for the ML training. Next, we randomly divide the dataset into training and testing sets (80-20), both with the same proportion of classes. We train a Multilayer Perceptron (MLP) classifier (sklearn Python library), commonly used for natural language processing tasks, as it handles a large number of input features. We manually check a sample of 50 sentences and confirm results are sufficiently accurate (we leave a formal evaluation to future work). Finally, we apply the classifier to the entire corpus; the final output has a total of 73800 sentences annotated according to the meetup type.

The final output is a sentence-level bag of semantic entities: meetup types, people, places and temporal expressions. We aim to use this new corpus to build a knowledge graph of historical meetups.

[9]SynTime software: https://github.com/zhongxiaoshi/syntime

Table 2
Training dataset statistics

Class	No. Items
Education	629
Coincidence	952
Public celebration	1200
Business meeting	2113
Personal life	4074
Music making	7172
Total	**16140**

Table 3
Knowledge Graph statistics

Type	Total
Sentences	74445
Persons mentioned	51425
Places mentions	5595
Time expressions	79838
Events	73800

Knowledge Graph construction. In this last step, we focus our efforts on building a knowledge graph of historical MEETUPS, using the entities found at a sentence level. Crucially, we aim to identify missing elements at the sentence level and annotate them for resolution in the next version of the KG, for instance, implementing a coreference resolution task to determine if missing elements are located in previous sentences.

We built the KG following the MEETUPS Ontology model. The entities found at the sentence level represent a meetup candidate. The KG also stores provenance data (e.g., source, type of source, location identifier in the text). The KG was constructed from CSV files resulting from the process described so far, using SPARQL Anything [13] and is currently published in Turtle RDF format.

The MEETUPS KG contains data from 1000 artists' biographies. As summarised in Table 3, around 74445 historical meetups, there are 51425 mentions of people involved in different encounters. So far, the historical meetups gather around 5595 places and 79838 time expressions.

7. Challenges of KG generation and discussion

The evaluation of the MEETUPS knowledge graph has been made incrementally to ensure the data extraction process complies with the knowledge requirements (see Section 4). Here we present the input data validation using the first version of the KG. The data covers people, place, time expressions and meetup types. All these entities were extracted at the sentence level. We designed and executed a short survey as follows:

- Select randomly the sentences from musicians' biographies for evaluation.
  - The evaluation dataset was built from a random sample of 12 biographies from the total analysed in the first iteration (1000 biographies).
  - Due to time and resource constraints regarding the annotation and evaluation of the sentences, we selected one paragraph and its corresponding sentences from each biography. Some paragraphs contained only one sentence, others up to seven sentences, reaching a total of 40 sentences. Each sentence listed the total entities resulting from the knowledge extraction pipeline.
- For the data collection process:
  - For each sentence, we asked participants to analyse the correctness of the automatically identified entities in terms of the number of sentences identified as meetups, and the number of people, places, time expressions and meetup types correctly extracted.

Table 4
Knowledge Graph preliminary evaluation results - # entities at sentence level

	People	Place	Time	Theme
Entity in the text	21	23	24	34
No entity in the text	13	11	10	0
TOTAL	**34**	**34**	**34**	**34**

Table 5
Knowledge Graph preliminary evaluation results - evaluation entities quality

	People	Place	Time	Theme
Correct	0.62	0.87	0.42	0.74
Partially Correct	0.05	0.00	0.04	0.11
Incorrect	0.14	0.04	0.00	0.15
Missing	0.19	0.09	0.54	0.00

  - For each sentence, we asked participants to indicate the number of missing entities not identified by the automatic annotation process.

From the forty sentences, 34 (85%) were identified as potential meetups/encounters. Table 4 displays in detail the number of sentences that mention a type of entity. Of the 34 sentences identified as a meetup, nine sentences (26%) provide information about all four types of entities. Twenty-five sentences (74%) lack one or two of any of the types of entities.

We move now to analysing the entities extracted at the sentence level. First, regarding the completeness of sentences to represent a historical meetup:

- Results show that 26% of sentences (1 out of 4 sentences) are comprehensive (including all four types of entities), and 74% mention at least one out of the four types of entities. Therefore, it is possible that a historical meetup is not fully identifiable from one sentence. For example, the sentence: *"Born Al Albertini in Chester, Pennsylvania, United States, he went to South Philadelphia High School."* does not include a time entity but still provides data to answer the CQs. Future work regarding the completeness of historical meetups should explore cases of missing entities, such as searching for information in preceding sentences at paragraph and section levels.

Second, regarding the quality of the extracted entities:

- As mentioned in the previous point, sentences representing historical meetups include one or all four types of entities. Of the total of sentences analysed, 62% of them mention people and 38% of sentences do not (see Table 5). From the sentences that refer to people, the knowledge extraction process correctly identifies them in 62% of the sentences and missed people entities in 19% of sentences. People were incorrectly and partially identified in 19% of the sentences.
  - The cases in which people were incorrectly identified mainly pertain to homonyms names. For example, people's names are used to name prizes, and festivals, among other cases. Further validation includes querying DBpedia/Wikidata to determine if the entity is a person (for example, has a date of birth) and the entity type (for example, if it refers to a person or musician).

  - In cases where people entities were partially identified, we found most cases referred to entities named in a previous sentence or at the beginning of the paragraph. The next steps include exploring neighbouring sentences to identify such entities and implementing a coreference resolution task.

- From sentences that mention places, the knowledge extraction process correctly identifies place entities in 87% of the sentences, and only 9% of sentences were missed. Place entities were partially and incorrectly identified in 4% of the sentences. For the few cases in which places were missed, incorrectly and partially identified, the analysis shows the definition of places should be expanded. For instance, 'Conservatoire de Paris' can be identified as being in 'Paris'; this entity is of type `dbo:Campus` not `dbo:Place`. Future work should consider the level of detail required to recognise place entities.
- In both cases, people and place entities, a few instances were missing in DBpedia; for these cases it is essential to consider other databases such as Wikidata to link resources.
- From sentences that include references to time, the knowledge extraction process correctly identifies them in 42% of the sentences and 58% of sentences were missing. The analysis shows that in most cases, time expressions such as years (e.g., in 1923, 1899-1900), relative expressions (e.g., mid-eighties, late 1960) or indirect references (e.g., two months later, last year) are missing. The time expression identification process should expand the heuristic rules in the next iteration.
- Regarding the meetup type identification, all sentences are annotated with one of the classes identified in Section 4.1. The knowledge extraction process correctly identifies the correct type in 74% of the sentences. The analysis shows that in most cases, the meetup type is partially correct or incorrect when one sentence refers to more than one event in the same sentence. The time expression identification process should consider these cases in the next iteration.

8. Conclusions and Future work

In this paper, we presented the MEETUPS Ontology, a model that formally characterises the encounters between personalities of European history. We introduced the concept of the historical meetup to describe such encounters, for instance, collaborations and knowledge exchanges between people. The ontology describes the elements of a meetup, including entities such as participants, locations, time expressions and meetup types. We presented our approach as an essential tool to address the knowledge requirements of musical scholars and cultural heritage researchers who face challenges when analysing artists' social links and influences. Importantly, in collaboration with experts, we compiled a list of knowledge requirements expressed as Competency Questions (CQs). These CQs later served to validate the expressivity of the ontology.

Furthermore, we presented the preliminary work on developing the Knowledge Extraction pipeline for identifying the elements of historical meetups. The experiments were developed in the music cultural heritage domain context and used as data source biographies collected from Wikipedia. The execution of the pipeline produced the first version of the MEETUPS Knowledge Graph, crucially providing an overview of the main challenges to tackle to extract and represent

the constituent elements of a historical meetup. Therefore, future work includes addressing the challenges summarised in Section 7 and the data quality evaluation of the knowledge graph.

References

[1] J. Tuominen, E. Hyvönen, P. Leskinen, Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research (2018).

[2] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research, Lecture Notes in Computer Science, 2019, p. 574–589.

[3] S. Gottschalk, E. Demidova, EventKG: A Multilingual Event-Centric Temporal Knowledge Graph (2018). URL: http://arxiv.org/abs/1804.04526, arXiv:1804.04526 [cs].

[4] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web (2020) 187–193.

[5] N. J. Bryan, G. Wang, Musical Influence Network Analysis and Rank of Sample-Based Music, in: ISMIR, 2011, pp. 329–334.

[6] M. Kopel, Analyzing Music Metadata on Artist Influence, in: Intelligent Information and Database Systems, Springer International Publishing, 2015.

[7] A. Fokkens, S. Legêne, G. Schreiber, V. de Boer, Biographynet: Extracting relations between people and events (2018).

[8] S. Menini, R. Sprugnoli, G. Moretti, E. Bignotti, S. Tonelli, B. Lepri, RAMBLE ON: Tracing Movements of Popular Historical Figures, in: Proceedings of the 15th Conference of the European Association for Computational Linguistics, ACL, 2017.

[9] P. Beytía, J. Schobin, Networked Pantheon: a Relational Database of Globally Famous People (2018).

[10] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, The Knowledge Engineering Review 11 (1996) 93.

[11] V. Presutti, E. Daga, A. Gangemi, E. Blomqvist, eXtreme Design with Content Ontology Design Patterns, Proceedings Workshop on Ontology Patterns (2009).

[12] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The NeOn Methodology for Ontology Engineering, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, p. 9–34.

[13] E. Daga, L. Asprino, P. Mulholland, A. Gangemi, Facade-X: An Opinionated Approach to SPARQL Anything, IOS Press, 2021.

[14] J. R. Hobbs, P. Feng, Time ontology in OWL, 2006. URL: https://www.w3.org/TR/owl-time/.

[15] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O:The PROV ontology (2013).

[16] W. R. van Hage, V. Malaisé, R. Segers, L. Hollink, G. Schreiber, Design and use of the Simple Event Model (SEM), Journal of Web Semantics 9 (2011) 128–136.

[17] X. Zhong, A. Sun, E. Cambria, 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.

[18] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019.