Challenging knowledge extraction to support the curation of documentary evidence in the humanities

Enrico Daga
enrico.daga@open.ac.uk
The Open University
Milton Keynes, United Kingdom

Enrico Motta
enrico.motta@open.ac.uk
The Open University
Milton Keynes, United Kingdom

ABSTRACT
The identification and cataloguing of documentary evidence from textual corpora is an important part of empirical research in the humanities. In this position paper, we ponder the applicability of knowledge extraction techniques to support the data acquisition process. Initially, we characterise the task by analysing the end-to-end process occurring in the data curation activity. After that, we examine general knowledge extraction tasks and discuss their relation to the problem at hand. Considering the case of the Listening Experience Database (LED), we perform an empirical analysis focusing on two roles: the listener and the place. The results show, among other things, how the entities are often mentioned many paragraphs away from the evidence text or are not in the source at all. We discuss the challenges emerged from the point of view of scientific knowledge acquisition.

CCS CONCEPTS
- Information systems → Information extraction;
- Computing methodologies → Information extraction;
- Applied computing → Arts and humanities.

KEYWORDS
documentary evidence, knowledge extraction, named entity recognition, DBpedia

1 INTRODUCTION
The identification and cataloguing of documentary evidence from textual corpora is an important part of empirical research in the humanities. An increasing number of recent initiatives in the digital humanities have as primary objective the curation of a database collecting text excerpts augmented with fine-grained metadata, mentioned entities, and their relations, often in the form of knowledge graphs developed adopting the linked data paradigm. These databases are developed following controlled processes, in the spirit of digital library management, where the identification and onboarding of relevant information is substantially entrusted to research students, librarians, and similar domain experts. The Listening Experience Database Project (LED)\(^1\), for example, is an initiative aimed at collecting accounts of people’s private experiences of listening to music [4]. Since 2012, the LED community explored a wide variety of sources, collecting over 10,000 unique experiences.

\(^1\)https://led.kmi.open.ac.uk/

These are catalogued through a sophisticated workflow but more importantly by means of a rich ontology covering a variety of aspects related to the experience, for example, the time and place it occurred, the source where the evidence has been retrieved, and the entities involved, such as, a performer, a composer, or a creative work [1]. Another example is the UK Reading Experience Database (RED). UK RED includes over 30,000 records of reading experiences sourced from the English literature. The curatorial effort required to populate these databases was significant and the size and quality of these databases is a major achievement of these projects.

In this position paper we ponder the applicability of knowledge extraction techniques to support the data curation activity. Initially, we introduce the case study and analyse the data curation activity. After that, we examine general knowledge extraction tasks and discuss their relation to the problem at hand. Considering the case of the Listening Experience Database (LED), we perform an empirical analysis of a portion of the database, focusing on the role "listener" and "place". Specifically, we elaborate on the hypothesis that the related entities can be automatically retrieved from the source. Finally, we discuss a set of challenges for knowledge extraction related to supporting the curation of this type of evidence databases.

2 DATA CURATION ACTIVITY
In general, the discovery and selection of documentary evidence is an activity that may not be conducted systematically. However, in the context of enterprises such as the LED project, there is an attempt to objectively select, extract, and curate documentary evidence from texts. From the curator’s perspective, it is not about searching archives or repositories but exploring specific sources of value, for example, specific books. In [8] we developed an approach for retrieving textual excerpts relevant for a certain theme of interest in a book by combining language analysis, entity recognition, and a general purpose knowledge graph (DBpedia) and showed that many of those pieces of evidence are characterised by implicit information. In addition, once the text is found, populating all the metadata is a long and difficult task.

To illustrate the problem, let’s consider two examples from the LED project:

\(E_1\) "Music is certainly a pleasure that may be reckoned intellectual, and we shall never again have it in the perfection it is this year, because Mr. Handel will not compose any more! Oratorios begin next week, to my great joy, for they are the highest entertainment to me."\(^2\)

excerpt refers to Mrs Delany’s report of a (series of) live performances of Operas and Oratorios by George Frideric Handel, happened in March, 1737.

E2 ‘I then went to Amsterdam to conduct Oedipus at the Concertgebouw, which was celebrating its fortieth anniversary by a series of sumptuous musical productions. The fine Concertgebouw orchestra, always at the same high level, the magnificent male choruses from the Royal Apollo Society, soloists of the first rank - among them Mme Hélène Sadoven as Jocasta, Louis van Tulder as Oedipus, and Paul Huf, an excellent reader - and the way in which my work was received by the public, have left a particularly precious memory that I recall with much enjoyment.’ Stravinsky, in the beginning of 1928, celebrates the high level of the Concertgebouw orchestra and singers performing his Oedipus Rex. All of them are listed as entities in the LED database.

In both examples, several of the entities involved are not mentioned in the excerpt and are derived from the curator’s knowledge of the source (for example, Mrs Delany is the author of the letter in E1) and the domain (e.g. the full name of the work is Oedipus Rex in E2).

Here we focus on the challenge of automatically populating the record and support an expert in identifying, collecting and inputting the relevant information. In other words, we aim at automatically populating (as many as possible) roles of the ontology. For instance, a listening experience specification can be derived from the available graph on data.open.ac.uk [7]. The type ListeningExperience includes the following properties, among others (we omit namespaces for readability):

• agent (who is the listener)
• time (when the listening event occurred)
• place (where it occurred)
• subject (what was listened)
• is_reported_in (a link to the source)
• has_environment (e.g. was it a public or a private place, indoor or outdoor)

A ListeningExperience is related to other relevant items, notably Performance, WrittenWork, MusicArtist, and Country. The knowledge extraction system should be able to derive the requirements from the ontology specification, primarily the data values and roles involved. For example, it should derive the requirement to find the agent of the ListeningExperience, its place and time, and that there may be a specific musical work to be identified and, eventually, the author of the musical work, filling the roles associated to the path subject -> ? a Performance -> performance of -> ? a MusicExpression.

3 KNOWLEDGE EXTRACTION

Knowledge extraction is a branch of artificial intelligence covering a variety of tasks related to the automatic or semi-automatic derivation of formal symbolic knowledge from unstructured or semi-structured sources.

The area comprehends research in a variety of problems related to lifting an unstructured or semi-structured source into an output described using a knowledge representation formalism. Entity extraction and classification are two related tasks referring to the location of mentions of entities in an input text and their categorization, as in the following example: “We went to the rehearsal of JoshuaPerson last Tuesday time”. Entity Linking, instead, refers to finding mentions of entities from a database into a natural language resource or, similarly, to appropriately disambiguate words by associating a knowledge base identifier. Often, the three tasks are performed together and labelled Named Entity Recognition and Classification (NER) [12]. Linked Data and NER together have been extensively employed in a number of knowledge extraction and data mining tasks (e.g., the work of H. Paulheim [21]).

Relation extraction refers to the identification of n – ary relations (for n ≥ 2) within the source, usually addressed with a combination of NLP and machine learning techniques [22]. The relations Composer (Oedipus Rex, Starvinsky) and Performed (Oedipus Rex, Concertgebouw, 1928) are two examples. Event extraction is a special case of relation extraction where the focus is on identifying an event, usually an action being performed by an agent in a certain setting. This task is extensively studied in domains such as Biomedicine [5], Finance and Politics [15], and Science [26]. Approaches dedicated to the detection and extraction of historical and biographical events are designed in [25, 29]. The notion of event is generally considered as something happening at a specific time and place, which constitutes an incident of substantial relevance [14]. Therefore, the objective is to identify the action triggering the event (e.g. the verb perform) and then the associated roles. Data-driven approaches usually involve statistical reasoning or probabilistic methods like Machine Learning techniques. In contrast, knowledge-based methods are generally top-down and based on pre-defined templates, for example, lexicosemantic patterns [15]. The two approaches can be combined and machine learning methods used to learn such patterns [23]. However, the notion of event is still ill-defined in NLP research and this makes it hard to develop methods which are portable, effectively, to multiple domains [14]. Research in open domain event extraction focuses essentially on social media data [24] where the task is the extraction of statements for summarization purposes, similar to the one of key-phrases extraction [28]. Ontology-based information extraction (OBIE) uses formal ontologies to guide the extraction process [17, 27]. Relevant work in the area is surveyed in [9, 19]. In 2013, Gangemi provided an introduction and comparison of fourteen tools for knowledge extraction over unstructured corpora, where the task is defined as general purpose machine reading [10]. A machine reader transforms a natural language text into formal knowledge, according to a shared semantics. State of art methods include FRED [11] and PIKES [6]. These approaches are based on a frame-based semantics that is at the same time domain- and task-independent. Instead, a domain-oriented solution would identify knowledge components of interest in the text, similarly to what explored, for example, in the work of Alani [3]. This task is also considered as an automatic ontology instantiation [2] or semi-automatic creation of metadata [13]. A suitable approach should be able to detect the requirements from a domain-specific ontology and, having as input the text excerpt, the source metadata, and potentially other knowledge bases, generate suitable hypotheses of values and entities on any relevant role.
were memories, diaries, and collection of letters. In addition, this is not surprising as one of the most researched types of resources (for instance, in 2019 include a place (88.3%) and in 72.2% of them the place points to a DBpedia entity (79.9%). The agent is specified in 825.8 of them (91.2%) but only 29.9% refer to a DBpedia entity (33.1%). In all other cases the agent is created as a novel entity.

Finally, only 11.3% of the sources could be retrieved as open texts, half of the agents exist in DBpedia (2130 times, 23.5% of the total). Deriving information from the source metadata. However, less than 64.8% of the agents exist in DBpedia (79.9%). The agent is specified in 825.8 of them (91.2%) but only 29.9% refer to a DBpedia entity (33.1%). In all other cases the agent is created as a novel entity.

64.8% of the listeners are also the authors of the text - 5874 cases. This is not surprising as one of the most researched types of resources were memories, diaries, and collection of letters. In addition, this answers our Q4 and shows how important it could be to intelligently retrieve information from the source metadata. However, less than half of the agents exist in DBpedia (2130 times, 23.5% of the total). Finally, only 11.3% of the sources could be retrieved as open texts, referring to 1026 of the documentary evidence in the database. Of these, 7.3% includes DBpedia entities as place or agent, 690 excerpts from 26 books. These are the objects in our analysis.

Results are summarised in Figure 2. Charts display the distance of the entity mentions, measured in number of paragraphs. This analysis is partial as it only covers DBpedia entities being used as places or agents (listeners) with relation to books which sources we could retrieve from the Web. However, the answers to the remaining questions are quite interesting. (Q1) The DBpedia place was mentioned in the textual excerpt only in 25.9% of the observed cases (179). The listener was mentioned in the excerpt only in 13 cases, 13.4% of the observed population (97). (Q2) 10% of the times the place mention is less than 5 paragraphs from the evidence text. The agent is mentioned within 5 paragraphs from the evidence text.

Text segmentation is itself a difficult task. In our analysis, we measured distances in number of characters, considered one word to be 5 characters (the approximated average length in English) and one paragraph to amount to 200 words.

**Listing 1: Detect the location of an excerpt in a source.**

```plaintext
excerpt, Source; best[t, b, e, s] := // text, begin, end, score
words[] = tokenize(excerpt)
words[] = sortByLengthDesc(words[]); // Longest on top
Foreach word in words[]:
  occurrences[[b, e]] = find(word, Source)
  position[b, e] = find(word, excerpt)
  begin = occurrence.b - position.b
  end = occurrence.e + len(excerpt) - position.e
  possible = substring(Source, begin, end)
  score = levenshtein(excerpt, possible)
  if(score < best[s])
    best[t, b, e, s] = [possible, begin, end, score]
End
return best
```

**Figure 2: Distance of entity mention, in paragraphs.**

(a) Places.  (b) Agents.
An important characteristic is the amount of implicit information necessary to characterise the documentary evidence that is not derivable from the reference text. As a result, a typical knowledge extraction approach may fail at performing an inference that is normally the result of user’s expertise. A domain-independent machine reader could produce a formal representation of the text with entities and roles linked together. Theoretically, processing a text through a machine reading system would reduce the problem to one of ontology alignment. However, as we have seen, the needed entities may not be mentioned in the text excerpt at a reasonable proximity. In addition, having to deal with an ontology alignment problem does not necessarily reduces the distance to the goal.

Crucially, metadata about the sources should be used to derive information such as the time span of the documentary material or information about the author(s). Determining who is the person reporting the event could contribute to populate the agent (for first-person reports) but also on deriving more contextual information, for example, related to the historical period or the interests of the author. Linking an author to a knowledge graph (such as DBpedia) could provide insight on the validity of the hypotheses for assigning certain roles, for example, by deriving that Stravinsky is the author of Oedipus Rex (E2). Therefore, a general solution should be able to reason upon contextual knowledge. Intuitively, the system should be capable of fitting within the constraints of the domain specific ontology and exploit it to tailor the approach. The ontology specification would provide information about the main types and relations of interest, and those can be used to derive contextual information from existing commonsense knowledge bases (e.g. ConceptNet\textsuperscript{6}).

Although it may seem that these databases have a limited domain of interest, there are few chances that the variety of types and entities useful could be found in a single, encyclopedia-like knowledge base. In the case of the LED project, part of the Linked Open Data, the documentary evidence links to a variety of external resources (e.g. MusicBrainz\textsuperscript{7} and Geonames\textsuperscript{8}). The system should be able to work across distributed and heterogeneous datasets in search for relevant resources. These may include common-sense knowledge and linguistic resources, textual corpora, gazetteers, thesauri, and specialised digital libraries. Ultimately, the system should be able to recognise entities and their roles despite the fact that they can be linked to any reference database.

Ultimately, cultural studies like the ones performed in the LED and RED projects often coin novel concepts, such as Listening Experience, whose structure and features cannot be found in pre-existing databases. In fact, the definition of a concept of interest is itself a scientific output for which the database constitutes the empirical proof of relevance to scholarship in the related field. It is an open question to what extent learning from one of such databases could help in supporting a new, coming one.

REFERENCES