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 knoweldge 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. 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] The

[1] https://led.kmi.open.ac.uk/

In: *Proceedings of the 3rd International Workshop on Capturing Scientific Knowledge (Sciknow), November 19th, 2019. Collocated with the tenth International Conference on Knowledge Capture (K-CAP), Los Angeles, CA, USA.*.

[2] Source: Mary Granville, and Augusta Hall (ed.), Autobiography and Correspondence of Mary Granville, Mrs Delany: with interesting Reminiscences of King George the Third and Queen Charlotte, volume 1 (London, 1861), p. 594. https://led.kmi.open.ac.uk/entity/lexp/1444424772006 accessed: 30 September, 2019.

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**.

E_2 *"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."*[3] 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 E_1) and the domain (e.g. the full name of the work is *Oedipus Rex* in E_2).

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[4].

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* $\textbf{\textit{Joshua}}_{Person}$ *last* $\textbf{\textit{Tuesday}}_{Time}$*"*. **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 (NERC)** [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 $n \geq 2$) within the source, usually addressed with a combination of NLP and machine learning techniques [22]. The relations `Composer(Opedipus Rex,Starvinsky)` and `Performed(Opedipus 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, lexico-semantic 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 knowldge bases, generate suitable hypotheses of values and entities on any relevant role.

[3] Igor Stravinksy, Igor Stravinsky: An Autobiography (1936), p. 139. https://led.kmi.open.ac.uk/entity/lexp/1435674909834 accessed: 30 September, 2019.
[4] For a general introduction, see [16].

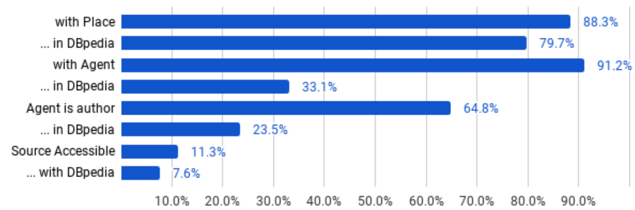


Figure 1: Statistics on the LED database to illustrate the coverage of DBpedia entities and the scope of our analysis.

4 EMPIRICAL ANALYSIS

To discuss the feasibility and difficulty of the task, we relax the problem and verify to what extent the entities that are part of the curated metadata could potentially be automatically derived from the sources. Specifically, we want to answer the questions: ($Q1$) Could a system find the target entities in the excerpt? ($Q2$) Could a system find the target entities in the text surrounding the excerpt? ($Q3$) How far from the excerpt the entity is? ($Q4$) Could it be found in the metadata of the source?

We consider the case of the LED database and focus on two relation and roles: the listener and the place of the listening event. The LED curation workflow reuses entities from DBpedia, MusicBrainz, and also defines new entities in the Linked Data. Our analysis is limited to books from archive.org annotated with a listener or place from DBpedia. We use DBpedia Spotlight [20] as entity recognition and linking system.

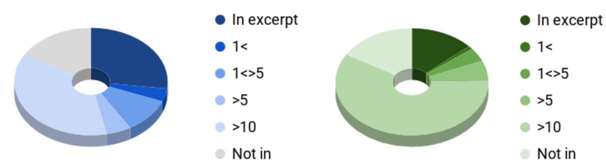
First, we need to find the position of the evidence text back in the original source. Identifying the position of LED items in the original book is not an easy task. In fact, the process of reporting an excerpt from the book involves a number of modifications in the format that makes it very rare the chance that a precise text match would work. In addition, the reported text includes often *omissis* or rephrasing in order to include co-references derived from previous paragraphs. To solve the problem, we developed the algorithm presented in Listing 1. The method is based on using the longest words as *locators*. The algorithm selects the occurrences of the longest words and isolate the surrounding portion of text using the length of the excerpt as heuristic. The resulting candidates are then ranked according to their similarity against the excerpt using the well-known Levenshtein distance [18]. The candidate with the lowest score is elected as the original text.

Figure 1 illustrates the features of the corpus. Of the 9059 listening experiences in the database with a textual excerpt reported, 7999 include a place (88.3%) and in 7222 of them the place points to DBpedia (79.9%). The agent is specified in 8258 of them (91.2%) but only 2996 refer to a DBpedia entity (33.1%). In all other cases the listener 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 type of resources were memories, diaries, and collection of letters. In addition, this answers our $Q4$ and shows how important it could be to intelligently derive 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.

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

```
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)
 Foreach occurrence[b,e] in occurrences[][b,e]:
   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]
   fi
 End
End
return best
```



(a) Places. **(b) Agents.**

Figure 2: Distance of entity mention, in paragraphs.

Results are summarised in Figure 2. Charts display the distance of the entity mentions, measured in number of paragraphs[5]. 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 then 5 paragraphs from the evidence text. The agent is mentioned within 5 paragraphs from the evidence in 4% of the observed cases. ($Q3$) 83.2% of the times the DBpedia place was explicitely mentioned at least once in the source (574). In 79 cases (11.4%) the place hasn't been found either in the excerpt or anywhere else in the source. A similar result is observable for agents. Finally, there is good chance the entity is somewhere away from the evidence text.

5 CHALLENGES

There are several aspects that make the task of automatically supporting the acquisition of knowledge about documentary evidence particularly interesting from the point of view of scientific knowledge acquisition.

[5]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.

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 typycal 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 (E_2). 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 commons sense knowledge bases (e.g. ConceptNet[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, encyclopedic, 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[7] and Geonames[8]). The system should be able to work across **distributed and heterogeneus datasets** in search for relevant resources. These may include common-sense knowledge and linguistic resources, textual corpora, gazetteeres, 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

[1] Alessandro Adamou, Mathieu d'Aquin, Helen Barlow, and Simon Brown. 2014. LED: curated and crowdsourced linked data on music listening experiences. *Proceedings of the ISWC 2014 Posters & Demonstrations Track* (2014).

[2] Harith Alani, Sanghee Kim, David E Millard, Mark J Weal, Wendy Hall, Paul H Lewis, and Nigel Shadbolt. 2003. Web based knowledge extraction and consolidation for automatic ontology instantiation. (2003).

[3] Harith Alani, Sanghee Kim, David E Millard, Mark J Weal, Wendy Hall, Paul H Lewis, and Nigel R Shadbolt. 2003. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems* 18, 1 (2003), 14–21.

[4] Helen Barlow and David Rowland. 2017. Listening to music: people, practices and experiences.

[5] Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at PubMed scale. *Bioinformatics* 26, 12 (2010).

[6] Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. 2016. Frame-based ontology population with PIKES. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (2016), 3261–3275.

[7] Enrico Daga, Mathieu d'Aquin, Alessandro Adamou, and Stuart Brown. 2016. The Open University Linked Data - data.open.ac.uk. *Semantic Web* 7, 2 (2016).

[8] Enrico Daga and Enrico Motta. 2019. Capturing themed evidence, a hybrid approach. In *18th Int. Conference on Knowledge Capture (K-CAP)*. ACM, To appear.

[9] Dejing Dou, Hao Wang, and Haishan Liu. 2015. Semantic data mining: A survey of ontology-based approaches. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015)*. IEEE, 244–251.

[10] Aldo Gangemi. 2013. A comparison of knowledge extraction tools for the semantic web. In *Extended semantic web conference*. Springer, 351–366.

[11] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic web machine reading with FRED. *Semantic Web* 8, 6 (2017), 873–893.

[12] Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review* 29 (2018), 21–43.

[13] Siegfried Handschuh, Steffen Staab, and Fabio Ciravegna. 2002. S-CREAM—semi-automatic creation of metadata. In *International Conference on Knowledge Engineering and Knowledge Management*. Springer, 358–372.

[14] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska De Jong, and Emiel Caron. 2016. A survey of event extraction methods from text for decision support systems. *Decision Support Systems* 85 (2016), 12–22.

[15] Wouter IJntema, Jordy Sangers, Frederik Hogenboom, and Flavius Frasincar. 2012. A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics: Science, Services and Agents on the World Wide Web* 15 (2012).

[16] Jing Jiang. 2012. Information extraction from text. In *Mining text data*. Springer.

[17] Vangelis Karkaletsis, Pavlina Fragkou, Georgios Petasis, and Elias Iosif. 2011. Ontology based information extraction from text. In *Knowledge-driven multimedia information extraction and ontology evolution*. Springer, 89–109.

[18] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.

[19] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2018. Information extraction meets the semantic web: a survey. *Semantic Web* (2018).

[20] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. ACM, 1–8.

[21] Heiko Paulheim. 2013. Exploiting Linked Open Data as Background Knowledge in Data Mining. *DMoLD* 1082 (2013).

[22] Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. 2017. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191* (2017).

[23] Jakub Piskorski, Hristo Tanev, and Pinar Oezden Wennerberg. 2007. Extracting violent events from on-line news for ontology population. In *International Conference on Business Information Systems*. Springer, 287–300.

[24] Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1104–1112.

[25] Roxane Segers, Marieke Van Erp, Lourens Van Der Meij, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga, Johan Oomen, and Geertje Jacobs. 2011. Hacking history via event extraction. In *Proceedings of the sixth international conference on Knowledge capture*. ACM.

[26] Maria Vargas-Vera and David Celjuska. 2004. Event recognition on news stories and semi-automatic population of an ontology. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*. IEEE, 615–618.

[27] Daya C Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches.

[28] Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. 2005. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. IGI Global, 129–152.

[29] Kalliopi Zervanou, Ioannis Korkontzelos, Antal Van Den Bosch, and Sophia Ananiadou. 2011. Enrichment and structuring of archival description metadata. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. ACL.

[6] http://conceptnet.io/

[7] https://musicbrainz.org/

[8] https://www.geonames.org/