Exploring Linked Data For The Automatic Enrichment of Historical Archives

Gary Munnelly, Harshvardhan J. Pandit, and Séamus Lawless

Adapt Centre, Trinity College Dublin, Dublin, Ireland
{firstname.lastname}@adaptcentre.ie

Abstract. With the increasing scale of online cultural heritage collections, the efforts of manually adding annotations to their contents become a challenging and costly endeavour. Entity Linking is a process used to automatically apply such annotations to a text based collection, where the quality and coverage of the linking process is highly dependent on the knowledge base that informs it. In this paper, we present our ongoing efforts to annotate a corpus of 17^{th} century Irish witness statements using Entity Linking methods that utilise Semantic Web techniques. We discuss problems faced in this process and attempts to remedy them.

Keywords: entity linking, ontology creation, automatic enrichment

1 Introduction

The promise of Semantic Web [1] is an attractive one for any individual who is involved in cultural heritage research. It is a promise of powerful search, seamless integration and informed, reasoned decision making between the many siloed instances of data which prevail across domains. Unfortunately, before a collection can take advantage of the benefits gained by being a part of the Semantic Web, it must be annotated with a suitable vocabulary. This annotation process can be an expensive and challenging task for a number of reasons, not least of which is the time and labour cost of employing people to read and manually annotate the contents of the collection. Given the exponential effort of a manual annotation process, it is not surprising that there has been some interest in the effectiveness of applying automatic annotation tools to cultural heritage collections [2–4].

Of the variety of tools and methods that exist for extracting information from a collection, this paper focuses on those concerned with the problem of Entity Linking (EL) [5]. While EL has seen much research in recent years, the lack of suitable semantic web resources to inform the EL process is often a notable weakness which undermines efforts to apply EL methods to cultural heritage resources. To some extent we may accept this as an inevitable limitation due to the immense variety of cultural heritage collections that exist. It is improbable that we will ever have a single centralised source of information which covers all aspects of cultural heritage. However, as the scale of digitised cultural heritage collections grow (Europeana[1] alone currently curates more than 50 million

[1] https://www.europeana.eu/portal/en

different items and DPLA[2] hosts almost 21 million resources), it becomes increasingly worthwhile to consider how we might deal with these limitations for collections both great and small.

In this paper we present a discussion on the role of semantic web resources in the task of automatically enriching digitised cultural heritage collections using EL methods. Our discussion is motivated by ongoing efforts to annotate a collection of 17th century Irish witness statements so that they may be integrated as semantic web resources and avail of benefits such enrichment provides. We present some of the challenges faced and lessons learned in the course of this endeavour. The contribution of this paper is to demonstrate and emphasise the importance of structure in Semantic Web resources. With due consideration it is possible to create new ontologies which may help to facilitate the EL process.

2 Related Work

2.1 Entity Linking

Entity Linking (EL) refers to a specific challenge in computer science whereby a series of unknown textual mentions of entities (commonly termed "surface forms") are provided as input to a disambiguation service. The service is tasked with mapping each of the surface forms to an unambiguous referent entity. To provide a concrete example, given the input sentence, "*I Henry Jones Doctor in Divinity in obedience to his majesties Commission...*" and a request to identify the entity "Henry Jones", an EL service might return a reference to the URI `http://dbpedia.org/page/Henry_Jones_(bishop)`, identifying the subject of the reference as the 17^{th} century Anglican Bishop, as opposed to the fictional character played by Séan Connery in the 1989 film "Indiana Jones and the Last Crusade".

In order to perform this mapping process, an EL system fundamentally requires two components:

1. a knowledge base that stores information about all the entities of which the system is aware, and
2. a referent selection method, which uses evidence extracted from the knowledge base and present in any prevailing information surrounding the surface form to arrive at a set of likely referents for each ambiguous mention.

Given an ambiguous set of mentions, the EL system retrieves from the knowledge base, a set of candidate referents to which an entity mention many be referring. This is usually based on some fuzzy retrieval method. A variety of heuristics are applied and the system eliminates candidate referents which are unlikely to be the subjects of the mentions. Eventually it arrives at a set of mappings from textual mentions to knowledge base URIs which unambiguously identifies the referents.

[2] https://dp.la/

Numerous different methods and approaches to EL may be freely found in the literature [6–8]. Almost universally, these methods use some form of graph based measure as one of the heuristics in the referent selection process. After the candidates have been retrieved from the knowledge base, a graph derived from the relationships between the entities may be constructed. The nature of these relationships varies, but usually it is based on links between corresponding Wikipedia pages. If a strong network exists between a number of candidate referents, then there is a good chance that they are the correct disambiguation choice for the given set of mentions.

It is also common to augment the graph weights using contextual cues derived from the words surrounding an entity mention [8, 9]. We note this because systems which consider this feature are based on the assumption that some contextual description of each entity exists in the knowledge base.

The structure and content of the knowledge base is crucial, not only for informing the disambiguation service that an entity exists, but also for providing information which helps the the disambiguation algorithm to distinguish good referents from poor referents. Many modern systems make use of DBpedia[3] [10] and YAGO[4] [11] for this task. These are a good choice for most problems due to the prevalence of links between entities and the long form descriptions of entities obtained from their corresponding Wikipedia articles. However, for cultural heritage collections it is often the case that the range of information contained in these Semantic Web resources are not complete enough to capture the variety of entities we see in cultural heritage collections.

EL systems have the potential to be extremely helpful when enriching cultural heritage collections with semantic data. These fully automated systems are capable of deducing suitable annotations for raw, flat, textual documents based on information that is fed to them via a knowledge base. It is easy to see how a suitably informed EL system might dramatically ease the process of semantically linking new cultural heritage artifacts as they are digitised.

Further discussion could be had surrounding the precise point in the digitisation process at which EL is applied. Are we linking metadata which has already been normalised by an expert, or is the system capable of dealing with the noisy, original, primary source content from which the digital artifact is derived? In the case of the latter, how does the system manage archaic references, evolving entities and other such anomalies present in the source collection?

2.2 Automatic Enrichment in cultural heritage

There have been a number of efforts to investigate the effectiveness of EL methods in the automatic enrichment of cultural heritage collections.

A Europeana led task force produced a series of reports in 2015 which document their experience with evaluating different cultural heritage enrichment

[3] `http://wiki.dbpedia.org/`

[4] `https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/#c10444`

services and sourcing different descriptive vocabularies as targets for the annotation process. Their focus was on annotating metadata for digitised artifacts. The content of this metadata ranged from specific fields comprised of a single entity e.g. `dc:creator`, `dc:publisher` to more general, free-form data such as `dc:description`.

As part of the investigation, a comparative evaluation of seven cultural heritage EL services was conducted [12]. Each service used a different vocabulary for enrichment, however the investigators were able to normalise the annotations by exploiting the fact that many of them made reference to corresponding DBpedia and Geonames entities. Using an evaluation dataset that was developed based on a combination of automatic enrichment tools and manual human investigation the report showed that the accuracy of the targeted EL tools was extremely high for the chosen collections.

However, as a variety of previous studies have shown, while the accuracy of EL methods may be high, quite often only a very small percentage of entities contained in cultural heritage datasets may actually be linked with a referent. A recent study we performed on the 1641 depositions (see Section 3.1) showed that a human annotator could only identify referents for 33% of the people and locations in the depositions [13]. We would compare this to efforts by other scholars such as Agirre [14], who attempted to link Europeana artifacts to Wikipedia articles and discovered that only 22% of entities that he identified could be annotated in this manner. This is an important limitation of which we must be aware.

One aspect of the problem is simply that cultural heritage collections are so incredibly diverse, complicated and unique that finding a suitable Semantic Web resource with adequate coverage for all purposes is nigh impossible. This presents the question, how should we annotate a cultural heritage collection when an appropriate Semantic Web resource cannot be found? Moreover (and of particular importance to our own research) how should these new Semantic Web resources be structured in order to aid the automatic enrichment process?

Of particular note for this discussion is the work of Brando et al. [15] on the REDEN project which investigated methods of using multiple knowledge bases for disambiguation. This is an interesting approach which may help to fill the gaps in popular knowledge bases using the information contained in more tailored ones. In their experiments DBpedia was used in conjunction with the Bibliothèque Nationale de France (BnF) ontology on a collection of French literary works.

REDEN's candidate selection phase is based on a literal string comparison between the surface form and entities in the knowledge base. All candidates from all source ontologies are retrieved and a resolution step based on `owl:sameAs` and `skos:exactMatch` properties resolves duplicate mentions into a single reference. Once the candidates have been appropriately pruned, a degree centrality measure is used to select the referents.

REDEN demonstrated that developing EL methods which can avail of multiple knowledge bases may help with poor coverage, but this requires that it

be possible to establish reliable, accurate mappings between ontologies. Indeed, this property also facilitated the evaluation conducted by the Europeana task force. This is an important consideration when developing new vocabularies for cultural heritage collections.

3 Enriching the 1641 Depositions

Our own research has focused on attempts to automatically annotate a collection of 17^{th} century manuscripts using EL methods. This has been extremely challenging for a variety of reasons. However, we believe our experiences are a reasonably typical example of problems faced in this field. Below, we document our observations and experience of working with the collection to date.

3.1 The 1641 Depositions

The 1641 depositions are a collection of letters and witness statements taken from the people of Ireland during the 1641 Irish rebellion. The physical manuscripts are comprised of approximately 19,000 pages bound in 31 volumes. Ireland in 1641 was a tumultuous place, and while the accuracy of some of the witness statements may be questionable, the depositions provide an unparalleled window into this dark chapter in Irish history.

The depositions have been digitised, transcribed and annotated by a team of historical scholars who extracted references to people and locations, tagged depositions based on the nature of their contents, and preserved as much information about the physical manuscripts as possible including margin notes, original spelling etc. The resulting documents are stored in a combination of TEI annotated files and an SQL database. This data rich digital resource presents many interesting and exciting opportunities for computer scientists to begin experimenting with methods of analysing and extracting new information from this historical collection.

Working with the digital versions of the depositions comes with a number of challenges, not least of which is the inconsistent nature of the spelling and grammar used throughout. English was still a developing language in 1641, which means that a vast array of variant spellings for names and common words exist across the documents. The below extract from *The Deposition of Phillip Sergeant*[5] provides an example of these anomalies:

> "*And by those faire promisses the said ffitzpatrick getting possession both of their persons & goodes, they there behoulding daily cruelties & murthers vpon other English and belike suspecting the like to be exercised against themselues,* ~~*desired*~~ *fled away secretly* ~~*o n to*~~ *to Mountrath*"

From the historians' work, we find that the depositions contain references to more than 60,000 people and 7,000 locations. The people in question range

[5] `http://1641.tcd.ie/deposition.php?depID=815351r406`

from individuals of great historical importance such as Sir Oliver Cromwell, Sir Phelim O'Neill, and King Charles I, to individual servants and common folk who were affected by the rebellion. Locations similarly range from cities such as Dublin which still flourish today, to small plots of land which have been lost either as their names changed or borders shifted.

We know that several of the entities extracted from the depositions are duplicates. However, the huge range in spelling variations and naming conventions makes it extremely difficult to determine which mentions of entities in the depositions might be references to the same person or place. Compounding this problem is the fact that the severity of textual noise means that standard NLP tools can struggle with simpler tasks such as sentence chunking or Named Entity Recognition. Performing reliable analysis based on the language of the depositions is, to say the least, difficult.

We are not the first to attempt to decipher the contents of the depositions using computational methods. The CULTURA project [16] developed and applied a range of tools to provide a personalised experience for individuals who are interested in exploring the collection. This project was extremely successful and produced a number of valuable utilities for working with collections of this nature. However, the depositions' content remains in its original SQL database, disconnected from the Semantic Web.

An earlier study conducted on a manually annotated subset of the depositions attempted to assess the feasibility of automatically enriching the collection using standard Entity Linking tools [13]. From this study it was shown that only 33% of entities in the annotated subset had a corresponding referent in DBpedia. It was also observed that, of the ten Entity Linkers evaluated, no single tool could satisfactorily annotate the test corpus. Individually some did show promise on specific aspects of the linking problem, but these were undermined by weaknesses elsewhere. For example, AGDISTIS [7] often correctly abstained from annotating where no referent existed in DBpedia, but was generally incorrect in its choice of referent where one did exist. KEA [17] and Dexter [18] performed respectably in choosing the correct referent where one existed, but were often overzealous and applied labels when they should have abstained.

3.2 Challenges in Adopting Semantic Web

While there are many aspects of the depositions that we may choose to model into a potential ontology to generate linked data, we focus on people and places due to their perceived importance in the documents. For historians there are still a number of unanswered research questions about the motivations and influences behind the rebellion. A linked data solution may assist them when investigating these questions. However, modelling the depositions is extremely challenging for a number of reasons.

First, there is no true, definitive list of people and places on which to base the ontology. Ideally any ontology that we create would be populated with a set of distinct entities that can be found in the depositions. While there are resources which can help us to determine this set of entities (as discussed below), there is

still noise present in these sources. Sometimes people are referred to by lineage rather than their actual name, e.g. "the heirs of Mr. Gale", or even by title e.g. "Bishop of Meath". Given these ambiguities, there is much risk of accidentally omitting or conflating entities when the ontology is being constructed.

Second, the inconsistent language of the depositions means that multiple variant spellings for people and places can be found throughout the collection. If a suitable ontology can be constructed to represent each entity, discovering all the possible variant names by which it may be referenced would be a monumental task. Sometimes these variations are minor spelling differences e.g. "Florence FitzPatrick" being referred to as "Fflorenc Ffitz Patrick", but some are more severe, such as the "Barony of Fassadinin" being referred to as the "Barrony of ffassa and Dyninge". Detecting such differences is difficult through an automated process and requires an expert to assess its correctness.

Third, if we are to construct this ontology with an eye to automatic enrichment, then the inclusion of links between entities and how to establish them is an important consideration. We could use familial connections, but we are not aware of any reliable sources which document these in a readily adoptable manner. On what basis then are we to establish relationships between our entities? Currently, there is no reliable way to specify that a relation is likely without stating it as a fact in an ontology.

We must also exercise some degree of caution in our attempts to annotate the depositions. If the intention is to assist scholars with their research, then the information conveyed by the proposed solution must be accurate. This can be a subtle problem. For example, if we consider the entity "the Pope", should this be used to describe the role of the head of the Catholic Church, or should it describe an individual who held that role? If we assume the latter, then we must be sure to refer to the correct pope for the source document, which involves additional knowledge that may not be readily available in the knowledge base. Pope Urban VIII held the position until 1644 when he passed away and was replaced by Pope Innocent X. Modelling evolving entities such as these is a common problem in the cultural heritage domain.

In spite of these challenges, resources do exist which can help us to generate lists of distinct entities. Three resources at our disposal are:

- The Down Survey: A complete national survey of land in Ireland after the rebellion. The survey was conducted in order to establish which lands should be forfeited as penalty for crimes during the rebellion.
- The Statute Staple: A record of transactions between individuals. The staple documents goods bought and sold, and provides information about debts owed between various parties before the rebellion
- The Books of Survey and Distribution: A list of properties held by various land owners. These documents were used to determine taxes based on land ownership.

These documents have been the subject of historical research for a number of years and were some of the major contributing sources for the Petty Maps project [6].

Given the three resources above, we have begun the process of constructing an ontology to model the entities present in the depositions. Using standard record resolution methods based on the DICE coefficient [19] and Jaro-Winkler distance [20], we are able to resolve entities across the three collections. This work must be checked for integrity by a historian, which is a slow process. Yet once this resolution has been completed it will be possible to uplift the resulting records to an RDF representation. This makes it possible to construct an ontology which describes unique instances of land owners, geographic regions and some of the relationships between them and can be used to enrich the process of annotating the depositions.

Our goal is to eventually use the information in this new ontology as a knowledge base for an EL system. Using the information derived from the three sources, it may be possible to automatically enrich the depositions with semantics using standard EL methods. Furthermore, if the entities described by the ontology are found to be recorded in other historical sources as they are digitised, it may be possible to use this knowledge to automatically link and integrate them with the depositions and the greater web of knowledge.

In order to facilitate EL methods, we are attempting to capture features that are commonly used by EL algorithms. The variety of surface forms which may link to an entity will be an important consideration. As yet we are contemplating which relationships are likely to be the most helpful. Our initial focus is on attempting to model ownership of land and debts owed between individuals as monetary records are often well documented based on the resources available to us.

Where possible we will attempt to reuse existing knowledge by linking entities in our ontology to corresponding entries in DBpedia, Geonames or other similar ontologies using properties such as `owl:sameAs`. Existing commonly used vocabularies will be used to describe the entities, though it may be necessary to extend or create additional properties for accurate representations of entities. For example, vCard[7] has a number of useful properties for describing people, and while it does capture the concept of honorifics applied to people e.g. “Mr.”, “Mrs.”, “Dr.” etc., it does not quite distinguish between an honorific and a title e.g. “Earl” is a title held that is passed on to other people, rather than an honorific applied to an individual person’s name. As yet we are assessing the suitability of vocabularies such as FOAF, vCard and SIOC[8] for describing such properties.

[6] http://downsurvey.tcd.ie/index.html
[7] `https://www.w3.org/TR/vcard-rdf/`
[8] `https://www.w3.org/Submission/sioc-spec/`

4 Discussion

Performing automatic enrichment of cultural heritage collections is challenging for a variety of reasons. As evidenced by our own experience, and the documented experience of other researchers, finding knowledge bases with adequate coverage for a given cultural heritage resource is extremely difficult. While developing an entirely new ontology that does not reuse existing knowledge is a solution, if not done properly it can lead to inaccurate or incompatible knowledge representations that negate one of the greatest benefits of linked data i.e. connectivity among disparate collections. It is of far greater benefit to the community if these new vocabularies can be integrated with existing semantic web resources in a seamless fashion.

Due to the issue with well-known knowledge bases not covering a large percentage of the entities in specialised cultural heritage collections, it is likely that curators of such resources will need to develop their own ontologies in order to accurately represent the semantics of their data. While it is good to expand the web of knowledge with this new information, we suggest that due care be given to the structure of these resources and to how this structure may lend itself to informing automatic enrichment processes going forward. Methods such as REDEN may exploit `owl:sameAs` or similar relationships between a new ontology and more established ones in order to knit together various knowledge bases for the EL process. If automatic enrichment services can make use of the information in new linked data resources, then future annotation processes may be expedited as new collections are digitised and made available.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5) (2001) 34–43
2. Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities **30**(2) (2015) 262–279
3. Wilde, M.D.: Improving Retrieval of Historical Content with Entity Linking. In Morzy, T., Valduriez, P., Bellatreche, L., eds.: New Trends in Databases and Information Systems. Communications in Computer and Information Science, Springer International Publishing (September 2015) 498–504
4. Stiller, J., Petras, V., Gäde, M., Isaac, A.: Automatic enrichments with controlled vocabularies in europeana: Challenges and consequences. In: Euro-Mediterranean Conference, Springer (2014) 238–247

5. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2) (2015) 443–460
6. Ganea, O.E., Ganea, M., Lucchi, A., Eickhoff, C., Hofmann, T.: Probabilistic bag-of-hyperlinks model for entity linking. In: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee (2016) 927–938
7. Usbeck, R., Ngomo, A.C.N., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A.: Agdistis-graph-based disambiguation of named entities using linked data. In: International Semantic Web Conference, Springer (2014) 457–471
8. Yosef, M.A., Hoffart, J., Bordino, I., Spaniol, M., Weikum, G.: Aida: An online tool for accurate disambiguation of named entities in text and tables. Proceedings of the VLDB Endowment **4**(12) (2011) 1450–1453
9. Zwicklbauer, S., Seifert, C., Granitzer, M.: Robust and collective entity disambiguation through semantic embeddings. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16, New York, NY, USA, ACM (2016) 425–434
10. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2) (2015) 167–195
11. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07, New York, NY, USA, ACM (2007) 697–706
12. Manguinhas, H., Freire, N., Isaac, A., Stiller, J., Charles, V., Soroa, A., Simon, R., Alexiev, V.: Exploring comparative evaluation of semantic enrichment tools for cultural heritage metadata. In: International Conference on Theory and Practice of Digital Libraries, Springer (2016) 266–278
13. Munnelly, G., Lawless, S.: Investigating entity linking in early english legal documents. In: Digital Libraries (JCDL), ACM/IEEE Joint Conference on. (in press)
14. Agirre, E., Barrena, A., Lacalle, O.L.D., Soroa, A., Fern, S., Stevenson, M.: Matching Cultural Heritage items to Wikipedia. In: LREC. (2012) 1729–1735
15. Brando, C., Frontini, F., Ganascia, J.G.: REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. Complex Systems Informatics and Modeling Quarterly (7) (July 2016) 60 – 80
16. Steiner, C.M., Agosti, M., Sweetnam, M.S., Hillemann, E.C., Orio, N., Ponchia, C., Hampson, C., Munnelly, G., Nussbaumer, A., Albert, D., et al.: Evaluating a digital humanities research environment: the CULTURA approach. International Journal on Digital Libraries **15**(1) (2014) 53–70
17. Waitelonis, J., Sack, H.: Named entity linking in# tweets with kea. In: # Microposts. (2016) 61–63
18. Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., Trani, S.: Dexter: an open source framework for entity linking. In: Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval, ACM (2013) 17–20
19. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3) 297–302
20. Winkler, W.: String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods. (1990)