PreMapper: Improving Entity Extraction Accuracy in the Digital Humanities

Cormac Hampson
Knowledge & Data Engineering Group
Trinity College Dublin
Ireland
+353 (1) 896 8431
cormac.hampson@scss.tcd.ie

Ella Rabinovich
IBM Research
Haifa
Israel
+972 4 829 6184
ellak@il.ibm.com

Sara Porat
IBM Research
Haifa
Israel
+972 4 829 6309
porat@il.ibm.com

Maya Koleva
Commetric
Electronica, Floor 3, 63 Shipchenski Prohod Blvd,1574 Sofia, Bulgaria
+359 (2) 4913 110
maya.koleva@commetric.com

Ivan Uzunov
Commetric
Electronica, Floor 3, 63 Shipchenski Prohod Blvd,1574 Sofia, Bulgaria
+359 (2) 4913 110
ivan.uzunov@commetric.com

Owen Conlan
Knowledge & Data Engineering Group
Trinity College Dublin
Ireland
+353 1 896 8431
owen.conlan@scss.tcd.ie

ABSTRACT

Cultural heritage collections that have been digitised present a great opportunity for engagement by users with different levels of expertise. Unfortunately, the metadata around these collections can often be sparse, limiting user exploration. To help counteract this, the CULTURA project is using Entity-Relationship extraction to great effect in enriching cultural archives with a new layer of semantics. Furthermore, CULTURA is enabling curators to manually improve the accuracy of the outputted entities through the use of its graph based PreMapper tool. This paper gives an overview of PreMapper and details how changes made to the entity graph impact the data flow within the CULTURA architecture.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries - *collection, dissemination, systems issues, user issues.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services - *data sharing, Web-based services.*

General Terms

Algorithms, Management, Design, Experimentation, Human Factors.

Keywords

Entity-Relationship Extraction, Network Analysis, Digital Humanities, Cultural Heritage, CULTURA, PreMapper.

1. INTRODUCTION

CULTURA [1] is a digital humanities portal that is supporting the exploration of cultural heritage collections by a range of different users; from professional researchers and historians, to those with little or no experience of a particular archive. To date, three digitised collections have been successfully enhanced and integrated into the CULTURA environment; the 1641 Depositions[1], the IPSA Collection[2] and the Bureau of Military History[3].

[1] http://cultura-project.eu/1641
[2] http://cultura-project.eu/ipsa
[3] http://cultura-project.eu/1916

Entity-Relationship extraction is a powerful technique, employing natural language processing, which can be used to inject semantics into unstructured text. This can involve training a data set and/or using prior knowledge e.g. pre-built dictionaries so that specific entities (people, locations, organisations, dates etc.) can be identified within the text. In CULTURA, both IBM LanguageWare [2] and the open source GATE[4] architecture (General Architecture for Text Engineering) are being used to perform such entity extraction.

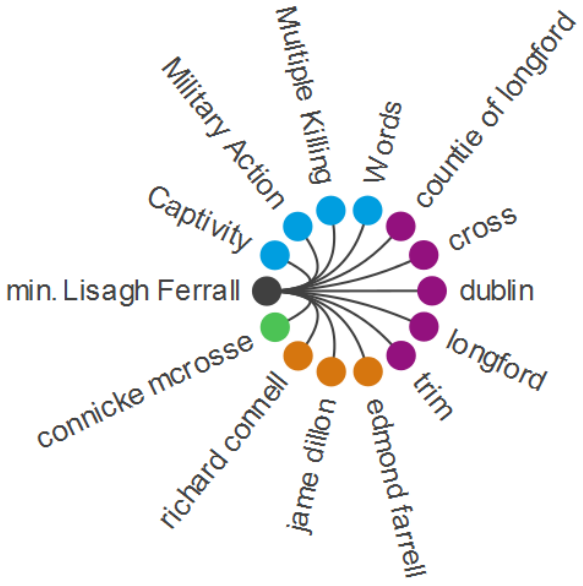

Figure 1. PreMapper Visualisation of a Deposition's Entities

Depending on the variety and structure of the data collection, the specific algorithms used, and the extent of training that occurs, the accuracy of the entity extraction process will vary considerably. Moreover, the 1641 Depositions, which contain noisy text,

[4] http://gate.ac.uk/

inconsistent grammar and spelling, introduce an additional difficulty to the already challenging task of entity extraction. Despite this, the benefits of entity extraction are many, and the outputted entities can be used to improve services such as personalisation, search, visualisation and exploration. However, no system will provide full accuracy and errors will be introduced into outputted data set of entities. Importantly, these errors can damage a curator's trust in the automatic processing as well as an end user's overall confidence in the system. Hence, approaches to improve the accuracy of the entity extraction process are of major benefit to the CULTURA environment.

One such approach implemented within the CULTURA project is to provide a graph-based user interface for domain experts to manually correct entities that have been automatically extracted. This approach is especially useful for finite cultural collections, where the accuracy of the entities can be iteratively and collaboratively improved over time. The PreMapper Tool which supports this process is described in Section 2, Section 3 outlines the feedback loop that occurs within the CULTURA environment when edits are made to the entity graph, and Section 4 summarises the paper.

2. PreMapper Tool

PreMapper comes in both desktop and web-based versions, but this paper will limit its discussion to the web-based implementation that is integrated directly with CULTURA's infrastructure as a Drupal[5] module (see Figure 1). Social Network Analysis (SNA) studies, maps and evaluates the relationships and flows between people, groups, organisations, computers, websites, and other connected information/knowledge entities. The nodes in the network are the people or organisations while the links show relationships or flows between the nodes. The network perspective provides insights in both numeric and visual formats and helps to bring to the surface information that cannot be made with other quantitative methods. Importantly, these same network visualisations can be used to help manage and correct entities extracted from cultural heritage collections.

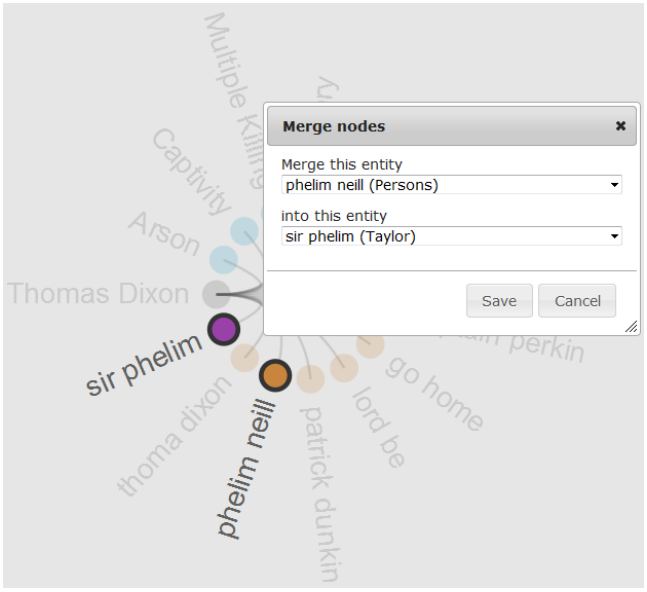

Figure 2. Merging Two Entities within PreMapper

Within the CULTURA project, the 1641 Depositions have had their text normalised and entity extraction implemented over them. The entity data outputted includes data on people (with information on occupation and religion where applicable), locations and events. This data is outputted as XML and stored in Drupal's default MySQL database as JSON. Originally the XML outputted was stored in a Neo4J[6] graph database, but this was found to be less efficient in terms of performance.

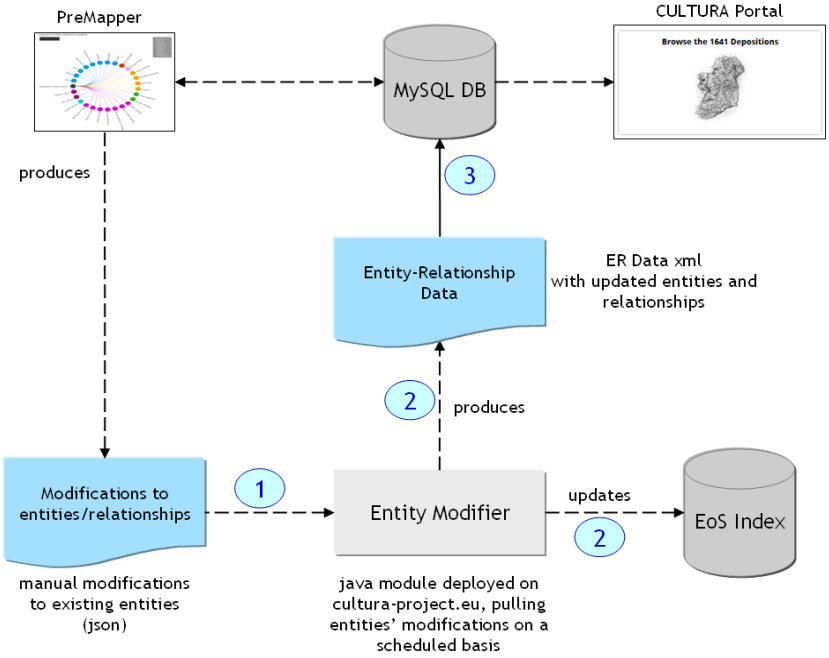

Figure 3. Feedback loop architecture and workflow

[5] http://www.drupal.org

[6] http://www.neo4j.org

A key aspect of PreMapper is that it allows users to visualise the entity graph for individually artefacts that have been exported after the entity extraction process occurs. For instance Figure 1 shows the various entities that have been extracted for a single witness statement, within the 1641 Depositions collection. Among its features, PreMapper enables curators of the collection (or those with the requisite permissions) to add, delete, merge, disambiguate and edit entities using a GUI. Figure 2 shows how the entity "phelim neill" can be merged into the entity "sir phelim" if an expert deems that these entities refer to the same person.

The transparency of this tool helps experts to identify errors in the entity graph (false positives, omitted entities, duplicate entities etc.), and allows them to manually correct the output of the automatic entity extraction process. This has a positive impact on the various services (personalisation, exploration etc.) that use these entities to enrich their understanding of the collection.

3. PreMapper Feedback Loop

Once a curator makes changes to the entity graph using PreMapper, this can have impact on other services within the CULTURA Environment. Hence, this section describes the design of the PreMapper Feedback Loop that accounts for these updates.

3.1 Architecture

The high-level architecture of the feedback loop within CULTURA is depicted in Figure 3. As described in section 2, the PreMapper module allows manual modification of entities, which then reports all these updates to a predefined repository that is accessible via a URL. Entity Modifier is a component that initiates the Entity-Relationship (ER) Data updates on a scheduled basis (every hour, once a day etc.). This is a pull-based module that retrieves the updates and generates an updated ER Data.

The basic feedback loop workflow in Figure 3 is illustrated by consecutive numbers: (1) a request for entity updates is issued by the Entity Modifier, then (2) the new ER Data is generated and the EoS index is updated accordingly, and finally, (3) the new ER Data may need to be uploaded to the MySQL DB serving both PreMapper and the CULTURA portal. For details and clarifications on step 3, please see Section 3.3.

3.2 Updating ER Data

This section describes various types of modifications one can do with PreMapper, how these modifications are mapped to changes in ER Data, and semantic inferences entailed by them. The CULTURA Entity-Relationship Diagram (ERD) consists of description of entities, their attributes and relationships, as detailed in Figure 4. One point we should consider with regards to modifications made by PreMapper users is entity identifiers (id).

3.2.1 Entity identifiers

Each extracted entity is assigned with an "id" attribute which uniquely identifies this entity. This attribute exists for each entity type, namely Person, Location, Event and Deposition, as depicted in Figure 4. The id attribute is immutable in a sense that once it is set, its value is constant for a specific entity's lifetime. The procedure of setting an id for an extracted entity is typically done by concatenation of a subset of its attributes. A manually added entity's id can be set with any value, as long as it is unique in the entire system. Entity's id attributes are underlined in Figure 4.

3.2.2 Entity Modifications with PreMapper

A user can perform the following entity modifications with the PreMapper tool:

- Add a new entity
- Update an entity's attributes and relationships
- Merge two entities into a single entity
- Delete an entity
- Add a link (relationship) between two entities
- Delete a link between two entities

3.2.3 From ER Manual Modifications to Enhanced ER Data

The set of manual entity modifications must be translated into corresponding changes in the underlying ER Data. This is achieved by contacting the API for the ER Data modification

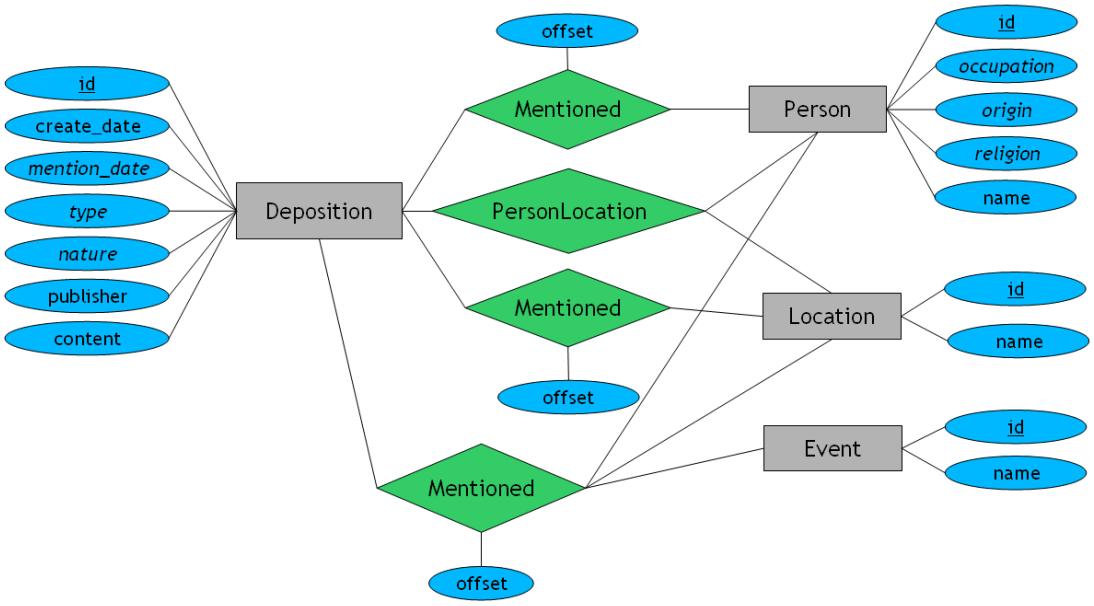

Figure 3. Entity-Relationship Diagram (ERD) for the 1641 Depositions' Collection

module, with each manual modification involving invocation of a function or a sequence of functions. Consider the following concrete examples.

Deleting an entity from PreMapper's deposition view will cause removal of this entity from the specific deposition, i.e. what is actually removed is the relationship connecting this specific entity to the deposition. In this case the *deleteRelationship* API call will be activated.

Merging two (main and secondary) entities triggers several underlying modifications: (1) enriching the multi-valued attributes of the main entity with values of the secondary one. PreMapper does not include this step as part of its merge. A user will manually update an entity's attributes upon a merge action, (2) adding all secondary entity's relationships to the main entity, and (3) deleting the secondary entity. This merge will, therefore, result in the following sequence of function calls (along with intermediate processing): *updateEntityAttribute, addRelationship*, and *deleteEntity*.

3.2.4 Entity Disambiguation via Expert-oriented tools

In some cases changes made by a user to a certain entity may result in an entity that seems similar to a pre-existing one. Consider a change of "Rob. Meredith" to "Robert Meredith". In case a person named "Robert Meredith" already exists in the system, a reasonable assumption would be that these two people *might be* the same person. However, since the system allows for the existence of two different people with the same name, automatic merging is not appropriate. Instead an expert in the subject matter must make an explicit request regarding the merging of these two entities. This is vital in helping to maintain the integrity of the data.

As an example, a researcher may want to search for an existing entity with the new name ("Robert Meredith" in our example) via the PreMapper tool, and if found, merge these two entities to a single one (even if found in different depositions). Consider entity A and B that are identified for merging. The PreMapper "merge" action will move all the relationships from entity A to entity B and "delete" entity A by removing it from the view. A researcher can further refine the attributes of entity B according to A's attributes, if needed.

3.2.5 From changes in ER Data to coherent exploration experience

Changes in ER Data are simultaneously reflected in the Entity-Oriented Search (EoS) index, as depicted in Figure 3, thus maintaining consistency in the entire CULTURA environment. EoS offers enhanced means for faceted retrieval and exploration of entities, their attributes and relationships, utilizing the Lucene search engine. Any change of entity in the ER Data would immediately be materialized in the EoS index, supporting the fully integrated solution.

3.3 Challenges and Future Work

Some scenarios will call for explicit generation of new ER Data, e.g., changes in the procedure of entity extraction in the NLP module. Such a case may potentially produce data that substantially differs from its original version: entities dropped/changed/merged etc. This case poses a challenge for applying modifications made by a user, since there could be an inherent inconsistency between the ER Data modified by the user in the first place, and the newly created version.

There is a module for importing new ER Data to the MySQL DB that is used by PreMapper. However, manual modifications made by users will not be applied to the uploaded data. We suggest further investigation of this issue to devise more comprehensive ways of accounting for this within the architecture.

The high-level design presented in this paper leaves much space for further investigation. As an example, a common data model (i.e., the Entity Relationship Diagram) should be shared by the Entity Extraction and PreMapper tools, so as to produce a seamless way of applying experts' updates to the ER Data. Also, an explicit invocation of Entity Modifier (Figure 3) should be explored, providing means for almost real-time updates in the ER Data upon PreMapper modifications, thus constantly maintaining end-to-end system's consistency.

4. SUMMARY

This paper has described the role of PreMapper within the CULTURA project and how it supports domain experts to improve the accuracy of the ER data CULTURA consumes. This is of particular importance, as the entities generated are key components of many useful services that CULTURA offers its users. Evaluations of PreMapper with curators of cultural archives are ongoing and will provide vital feedback in refining the user interface and the exchange of data between PreMapper and the rest of the CULTURA infrastructure.

5. ACKNOWLEDGMENTS

The work reported has been funded by the Seventh Framework Programme of the European Commission, Area “Digital Libraries and Digital Preservation” (ICT-2009.4.1).

6. REFERENCES

[1] Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., and Wade, V. (2012). The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. In *Progress in Cultural Heritage Preservation - 4th International Conference (EuroMed*, Limassol, Cyprus, 2012), LNCS 7616, Springer, Heidelberg, Germany, 668-675.

[2] Carmel, D., Zwerdling, N., & Yogev, S. (2012). ‘Entity oriented search and exploration for cultural heritage collections: the EU cultura project’. In *Proceedings of the 21st international conference companion on World Wide Web*, New York, USA, pp. 227-230