Up-cycling Data for Natural Language Generation

Up-cycling Data for Natural Language Generation

Amy Isard, Jon Oberlander, Claire Grover
Language Technology Group,
Institute for Language, Cognition and Communication,
School of Informatics, University of Edinburgh
10 Crichton St, Edinburgh, UK
{amy.isard, claire.grover}@ed.ac.uk

Abstract

Museums and other cultural heritage institutions have large databases of information about the objects in their collections, and existing Natural Language Generation (NLG) systems can generate fluent and adaptive texts for visitors, given appropriate input data, but there is typically a large amount of expert human effort required to bridge the gap between the available and the required data. We describe automatic processes which aim to significantly reduce the need for expert input during the conversion and up-cycling process. We detail domain-independent techniques for processing and enhancing data into a format which allows an existing NLG system to create adaptive texts. First we normalize the dates and names which occur in the data, and we link to the Semantic Web to add extra object descriptions. Then we use Semantic Web queries combined with a wide coverage grammar of English to extract relations which can be used to express the content of database fields in language accessible to a general user. As our test domain we use a database from the Edinburgh Musical Instrument Museum.

Keywords: Natural Language Generation, Cultural Heritage, Semantic Web

1. Introduction

There are large collections of cultural heritage data which are currently not able to be widely shared and exploited because they are stored in databases whose structure and format are inaccessible to the general public. We use a number of Natural Language Processing techniques to bridge the gap between existing databases and Natural Language Generation systems to create varied, adaptive texts which can be tailored to particular visitors on a journey through an exhibition.

A number of research projects have focussed on using NLG systems to create multilingual adaptive texts from structured cultural heritage data. ILEX (O'Donnell et al., 2001) and M-PIRO (Isard et al., 2003; Oberlander et al., 2008) worked from hand-authored resources. They used language-independent databases in a specific format containing up to a fifty objects and a few hundred triples describing attributes of the objects, created in collaboration with curators. The linguistic resources for each language available (English for ILEX, and English, Italian and Greek for M-PIRO) were also hand-created by computational linguists. The generated texts could be tailored to a museum visitor's progress through an exhibition, allowing for comparisons between exhibits, and preventing the repetition of background information. More recent systems have generated texts from Semantic Web ontologies; NaturalOWL (Galanis and Androutsopoulos, 2007; Androutsopoulos et al., 2013) generated texts in English and Greek from OWL ontologies, and Dannélls et al. (2013) generated texts from Semantic Web data in 15 languages, but in both cases expert input was still required to create the necessary linguistic resources.

Sun and Mellish (2007), Mellish and Pan (2008), Mellish (2010) and Androutsopoulos et al. (2013) have experimented with performing NLG using OWL/RDF ontologies which do not have domain-dependent linguistic resources, using the relations provided by the ontologies as a starting point for the English presentation of the facts represented, and Gardent et al. (2017) have used DBpedia (Lehmann et al., 2015) crowdsourcing methods to extract large numbers of linguistic resources which can be used by NLG systems.

However, many museum databases contain information which is structured, but less regular than that found on the Semantic Web. Data may have been annotated over a number of years by multiple authors before being collected together, and the relation names used cannot always be relied upon to contain the information necessary to derive resources suited to NLG. From an NLG point of view, the museum data is often *inconsistent*, for example where the same date or company appears in multiple versions, *insufficient*, for example where it is not clear how to express a given relation, and *incomplete*, in that there is further information which could be added from other sources to enrich the texts presented to a visitor. We aim to bridge this gap by using automatic methods which can be applied to any museum database in any domain to provide all of the resources needed to generate texts using the Methodius NLG system (Isard, 2016), which generates texts from structured data (described in Section 2.2.).

We use the Edinburgh Musical Instrument Museum (MIMEd) as an example domain throughout the paper, but the techniques are designed to be used with any Cultural Heritage dataset. In Figure 1 we show parts of the current MIMEdh web pages for two cornets and a bassoon - the information displayed comes directly from the database and is not very engaging for a museum visitor. In contrast, Figure 2 shows a mock-up of a potential visitor experience of a virtual web musem visit using texts generated by Methodius using the techniques we will describe below. In this example, the visitor has first selected a cornet, followed by another cornet and then a bassoon. The current web page gives the information about dates and makers, but the Methodius texts put the facts in context and also contain some background information about the instruments

and companies involved. This paper describes the methods used to automatically acquire the extra data necessary for the generation of the texts with the minimum of expert input.

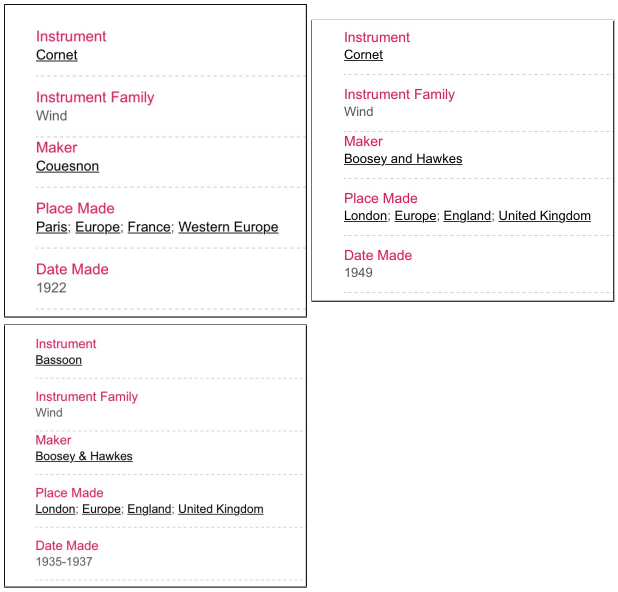

Figure 1: Screenshots of text from current Musical Instrument Museum web page, including canned text snippets

In the rest of the paper, we first describe the Edinburgh Musical Instrument Museum data which we have used as our test domain (see Section 2.1.) and the Methodius NLG system (see Section 2.2.). Figure 3 shows a summary of the automatic processing which we carry out in order to create the resources used as input by the Methodius system. We then describe the three stages of automatic domain processing which we have carried out. Firstly we perform data and name normalization, described in Section 3.1. We then extract relations which are used to link objects to their descriptions (see Section 3.2.). Finally we add descriptions of objects where information is available from DBpedia (see Section 3.3.). We conclude with some ideas for evaluations and future work (see Section 4.).

2. Background

2.1. Musical Instruments Museums Edinburgh

We are using as our test domain the Musical Instruments Museums Edinburgh Collection (MIMEd)[1], which was originally part of the European Musical Instrument Museums Online project (MIMO)[2]. The MIMEd collection contains photos and metadata information for about 5000 instruments. The MIMEd data is stored in an XML format based on the Dublin Core Metadata Initiative[3]. We have filtered the data in order to select a subset of the fields which can be used for NLG, and to select only the 4232 exhibits which have a specified instrument type.
An example of part of the data for an instrument is shown in Figure 4. In order for this data to be used with the Methodius system, we need to normalize the representation of dates

[1]http://collections.ed.ac.uk/mimed
[2]http://www.mimo-international.com/MIMO/
[3]http://dublincore.org

This instrument is a cornet, made by Couesnon in Paris in 1922. The cornet is a brass instrument very similar to the trumpet, distinguished by its conical bore, compact shape, and mellower tone quality. The most common cornet is a transposing instrument in Bb. It is not related to the renaissance and early baroque cornett.

This instrument is another cornet, made by Boosey & Hawkes in London in 1949. Boosey & Hawkes is a British music publisher purported to be the largest specialist music publisher in the world. Until 2003, it was also a major manufacturer of brass, string and woodwind musical instruments. Formed in 1930 through the merger of two well-established British music businesses, the company owns the copyrights or agencies to much major 20th century music, including works by Bartók, Leonard Bernstein, Britten, Copland, Kodály, Prokofiev, Richard Strass and Stravinsky.

This instrument is a bassoon, manufactured in 1946. Like the last cornet you saw, this bassoon was made in London by Boosey and Hawkes. The bassoon is a woodwind instrument in the double reed family that typically plays music written in the bass and tenor clefs, and occasionally the treble. Appearing in its modern form in the 19th century, the bassoon figures prominently in orchestral, concert band, and chamber music literature. The bassoon is a non-transposing instrument known for its distinctive tone color, wide range, variety of character and agility. Listeners often compare its warm, dark, reedy timbre to that of a male baritone voice. Someone who plays the bassoon is called a bassoonist.

Figure 2: Web museum mock-up displaying Methodius generated texts

(Section 3.1.), and acquire linguistic information so as to be able to generate sentences like “this bassoon was made by Buffet Crampon” (Section 3.2.). In addition, we can add information which is not present in the original database (Section 3.3.).

2.2. Methodius NLG system

The Methodius NLG system (Isard, 2007; Marge et al., 2008; Isard, 2016) is a descendant of the Exprimo system, which was developed during the M-PIRO project (Isard et al., 2003). The M-PIRO web interface allowed users to navigate through a small collection of ancient Greek artefacts by clicking on thumbnail images of the objects. Methodius was designed to be a more robust and modular NLG system, which can deal with collections of at least a million objects, and can be used for any domain in which an ontology of objects and attributes exist.
The system uses a typical NLG architecture based on the pipeline model described in Reiter and Dale (2000), which appears on the right in Figure 3. Once a user has chosen an object in which they are interested, the first phase of the generation is Content Selection where an algorithm is used

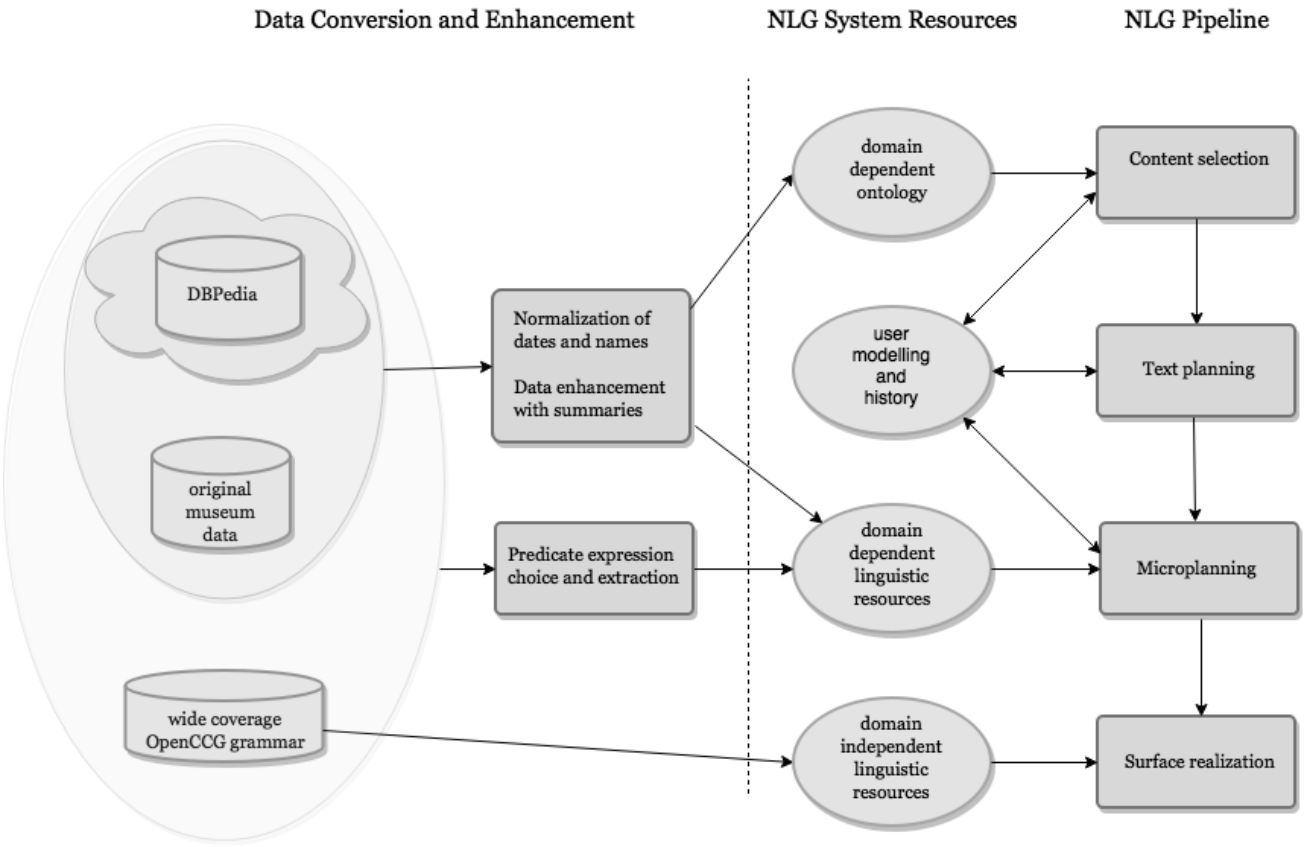

Figure 3: Normalization and enhancement techniques used to create the resources used by the NLG system

```
<dublin_core>
 <dcvalue qualifier="" element="identifier">
  164</dcvalue>
 <dcvalue element="type" qualifier="">
  Bassoon</dcvalue>
 <dcvalue element="contributor"
          qualifier="author">
  Buffet Crampon</dcvalue>
 <dcvalue element="coverage"
          qualifier="temporal">
  1921</dcvalue>
 <dcvalue element="coverage"
          qualifier="spatial">
  Paris
 </dcvalue>
</dublin_core>
```

Figure 4: MIMEd Object Specification

to select a subset of the available facts about the object, based on user modelling information, including a history of previously viewed objects. The next stage is Text Planning, where Rhetorical Structure (Mann and Thompson, 1998; Isard, 2016) is used to group and reorder the selected facts, adding comparisons with previous objects where available (Isard, 2007) and using aggregation rules to combine multiple facts as fluently and coherently as possible. This is followed by Microplanning, during which a logical form representing the sentences is built, and then sent to the Surface Realization component, which outputs the finished text using an OpenCCG grammar (see Section 2.2.2.).

The generated texts can be tailored to a museum visitor's progress through an exhibition, allowing for comparisons between exhibits (Isard, 2007), and preventing the repetition of background information. The resources used by the system consist of a domain ontology containing a hierarchical structure of the types of entities included in the domain, a set of domain-dependent linguistic resources, a set of domain-independent linguistic resources for each language for which texts are to be produced, and a user model and history, which stores a representation of each user's progress through a collection of objects in a virtual or real museum.

2.2.1. Methodius Domain Files

The information for a particular Methodius domain ontology is stored as a set of XML files which represent information about entities and their attributes, and the relationships between entities. The files consist of:

a hierarchy of entity types as in the example in Figure 5, which states that in this domain, the parent of type "bassoon" is "wind" and the parent of type "wind" is "instrument". It also states the noun to be used for

```
<type name="bassoon">
 <parents>
  <parent name="wind"/>
 </parents>
</type>
<type name="wind">
 <parents>
  <parent name="instrument"/>
 </parents>
</type>
```

Figure 5: Extract of Methodius Type Hierarchy

```
<defobject type="bassoon"
           is="object164">
 <role slot="maker" filler="Buffet Crampon"/>
 <role slot="creation-time" filler="1921"/>
 <role slot="original-location
       filler="Paris"/>
 </defobject>
```

Figure 6: Methodius Object Specification

each type, which links to the OpenCCG grammar (see Section 2.2.2.).

a set of entity instances with associated type, each of which has a number of fields containing attribute-value pairs, such as the ones shown in Figure 6 - this example contains the information extracted from the MIMEd data in Figure 4, which states that this exhibit is of type bassoon, was made by Buffet Crampon, was created in 1921 and was originally from Paris.

a set of linguistic text plans , such as the one shown in Figure 7, which states that the "maker" field applies to any object of type exhibit, and is expressed using the verb "make-verb" in the passive voice and the preposition "by". This information is used to build the appropriate logical form for the OpenCCG grammar, as described in Section 2.2.2.

2.2.2. OpenCCG Grammar

During the Surface Realization stage of the NLG process, Methodius uses the OpenCCG library, which provides parsing and realisation tools based on the CCG grammar formalism (White, 2006; White et al., 2007). It depends on a lexicon, a set of grammar rules, and a logical form which describes the structure of the content to be generated. In Methodius, the logical form is created during the Microplanning stage, described above, and then passed to the OpenCCG generation component, which produces the final text. Previous Methodius domains have relied on hand-written lexicons, but as part of this research we will use the wide coverage grammar of English provided with OpenCCG (White, 2014) after extracting the necessary domain relations, as described in Section 3.2.

```
<expression id="maker">
 <arg-one type="exhibit"
          refexp="default"/>
 <arg-two type="entity"
          refexp="default"/>
 <verb tense="past" voice="passive"
       pred="make-verb"/>
 <preposition id="by"/>
</expression>
```

Figure 7: Methodius Linguistic Expression

3. Data Wrangling and Up-Cycling

From an NLG point of view, the museum data is often *inconsistent*, for example where the same date or company appears in multiple versions, *insufficient*, for example where it is not clear how to express a given relation, and *incomplete*, in that there is further information which could be added from the Semantic Web to enrich the texts presented to a visitor.

We apply automatic processing in a number of stages in order to:

- normalize dates and names
- extract a Methodius type hierarchy
- extract modifier terms
- create a list of entities with fields and types
- extract common nouns (types) and proper nouns (entity names) for the OpenCCG lexicon
- automatically add descriptions of instrument types and entities such as companies or people

3.1. Date and Name Normalization

Many of the objects in the MIMEd data have fields which contain a date, but these have been annotated over many years by different authors, and are expressed in a wide variety of formats. There are date recognition software packages such as HeidelTime (Strötgen and Gertz, 2010) and SUTIME (Chang and Manning, 2012), but they were not able to parse many of the dates we found, as they are not geared towards historical dates with mis-spellings and expressions for uncertainty and vagueness. Blog posts from the Bentley Historical Library at the University of Michigan (Pillen, 2015a; Pillen, 2015b) describe techniques similar to ours.

In addition, during the processing we have created a hierarchy of time expressions so that where possible we can use the Methodius algorithms to generate further meaningful comparisons between objects. For example, if a visitor looks at a clarinet made in 1874 and then another made in 1888 we could say "like the last clarinet you saw, this one was made in the late 19th century".

The date normalization was implemented as a Python date processing module. The number of individual dates is greatly reduced by this process; for example the following seven strings from the Mimed data set, are all assigned to the date "second half of the 19th century", with the last three being assigned "possibly"' or "probably" modifiers to be used during generation.

```
Second half of the 19thcentury
Second half of 19thcentury
second half of 19th century
Second half of the 19th century
Probably second half of the 19thcentury
Probably second half of the 19th century
Possibly second half of the 19th century
```

As well as individual years, we process decades, centuries and date ranges, and in addition to the modifiers above, we search for the various ways in which approximate dates are expressed, such as "Circa 1840-1860", "c1860-1879", "1960s or a little later", "1855 or shortly before".

The MIMEd corpus contains a total of 4005 dates, from which we have extracted 941 unique date expressions, with 91 remaining unprocessed at present. Some of the remainder could be processed with the addition of more rules, but others contain so much free text that automatic processing is not possible, for example "1778-1830. The maker's mark is from the earlier part of this period but the hallmark dates from between 1809 and c1819."

The proper names in the database can also be rationalized, by using text processing techniques as well as by taking advantage of the DBpedia `redirect` facility. We perform DBpedia queries on all of the names we find in the corpus in order to add description texts as described in section 3.3., and this also allows us to gather together names which are considered to be synonymous by DBpedia. For example, we have many references to the manufacturer Boosey & Hawkes. As with the dates, we begin by extracting "probably" and "possibly" modifiers. We then have the following names, ordered by the number of occurrences in our data:

Boosey & Co 104

Boosey & Hawkes 77

Boosey and Hawkes 27

Boosey and Co. 21

Boosey & Co. Ltd 2

The first three all redirect to "Boosey & Hawkes" on DBpedia, allowing us to group them together. The third and fourth do not initially find a match on DBpedia, but if we apply some further text processing to remove the string "Ltd" and then any trailing full stops, we can capture all five versions. We would like to capture as many variations and typos as possible without writing individual rules for every possible eventuality, so we concentrate on the most frequent instances.

3.2. Extracting Relations

Previous work has looked into extracting domain-dependent linguistic resources from an ontology where the resources are not already available (Androutsopoulos et al., 2013; Han et al., 2015). These techniques relied on the ontology terms being linguistically related to the appropriate English terms, but In contrast, in the MIMEd data, the Dublin Core roles do not always directly relate to the English meaning.

For example, in the example from the MIMEd data in Figure 4, the role "author" is used for instrument manufacturers, and this role cannot be used directly to express the relationship in English. We therefore use a two-stage process to attempt to automatically find possible ways of expressing the relationship between instruments and manufacturers. First, we take a list of all of the instrument and manufacturer pairs in the domain, and also use the instrument type hierarchy to include all of the direct parents of each instrument. We then use SPARQL[4] queries to DBpedia (Lehmann et al., 2015) to retrieve all texts which describe one of the entities, and select all the sentences which contain the other half of the pair in question. Because we have many objects which have the same role, we have many opportunities to find suitable expressions, and if the same verb occurs repeatedly with different instruments and manufacturers, it can be considered to be an excellent candidate. For example, using the pair "bow" and "James Tubbs", we first retrieve the text below from DBPedia.

> James Tubbs (b 1835-d 1921)- one of the most celebrated English bow makers, and is considered "The English Tourte". Together with his son Alfred (d. 1912) they produced more than 5000 bows. It is universally accepted that James Tubbs ranks among the five or six most important bow makers in history. The Tubbs family made bows and instruments as early as the 1800s, and five generations have practiced the craft. In 1885 he won a Gold medal for his bows at the Inventions Exhibition held that year in London, after which he was made bowmaker by Special Appointment to HRH the Duke of Edinburgh.

We then parse each sentence in the text using the OpenCCG wide coverage grammar of English (White, 2014), and look for sentences in which there is a main verb which has the instrument as object and another noun phrase as the subject. We allow all subject noun phrases as there are many instances where the subject of the sentence is a different phrasing of the original name, a combination of names, or a pronoun. Although this means that some of the sentences will not in fact refer to the desired subjects, we make the assumption that there will be enough results in total to ensure that the incorrect verbs which may be captured will be infrequent and therefore not chosen as the top candidates.

As an example from the MiMEd data, the OpenCCG parse of a partial sentence "they produced more than 5000 bows" is shown in Figure 8. Here we have the main verb "produce.01" with subject "they" and object "bows" (with several modifiers, which we ignore). From the text above we find two potential main verbs - make.01 and produce.01, and we can then create Methodius expressions using just the subject, object and main verbs which will be used to generate sentences for this role, which in this case would eventually result in the generated sentences "James Tubbs produced bows" and "James Tubbs made bows". Because many instruments share the same role, we will also be able to generate for any other instrument which has an "author" field.

[4] https://www.w3.org/TR/rdf-sparql-query/

```
<node id="w10" pred="produce.01"
      tense="past">
 <rel name="Arg0">
  <node id="w9" pred="they" />
 </rel>
 <rel name="Arg1">
  <node id="w14" pred="bows" det="nil"
        num="sg">
   <rel name="Mod">
    <node id="w13" pred="5000">
     <rel name="Mod">
      <node id="w12" pred="than">
       <rel name="Arg1">
        <node id="w11" pred="more"/>
       </rel>
      </node>
     </rel>
    </node>
   </rel>
  </node>
 </rel>
</node>
```

Figure 8: OpenCCG parse

At the moment we are only investigating verbal expressions, but we plan to include noun phrases in future, so that we can also capture phrases such as "bow maker" or "instrument manufacturer".

3.3. Adding Description Snippets

The MIMEd metadata does not contain any descriptions of types of instruments, or of any entities mentioned in descriptions of individual instances, such as companies or people. We have already retrieved descriptions of some entities from DBpedia during the extraction of relations, and we also collect descriptions of instrument types where available, and add them to the Methodius data to provide richer descriptions. This not only allows us to add the text snippets to the MIMEd data, but also provides a method for entity disambiguation as described in Section 3.1.. We used SPARQL queries to download the rdfs:comment fields for all the instrument types in the ontology, where there is a DBpedia page available for the instrument, creating a generic instance for the instrument with a link to the text snippet. The snippet itself will be stored in the OpenCCG lexicon and retrieved through a reference in the logical form. For example, the comment field retrieved for bassoon is:

> The bassoon is a woodwind instrument in the double reed family that typically plays music written in the bass and tenor clefs, and occasionally the treble. Appearing in its modern form in the 19th century, the bassoon figures prominently in orchestral, concert band, and chamber music literature. The bassoon is a non-transposing instrument known for its distinctive tone color, wide range, variety of character and agility.

```
PREFIX rdfs:
  <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?name ?comment
WHERE {
  SERVICE <http://dbpedia.org/sparql>
  ?instrument rdfs:label ?name .
  FILTER(?name = "Bassoon"@en) .
  ?instrument rdfs:comment ?comment .
  FILTER(langMatches(lang(?comment), "EN"))
  }
}
```

Figure 9: SPARLQ for instrument names

4. Conclusions and Future Work

We have described a number of automatic processing techniques which aim to bridge the gap between existing cultural heritage databases and NLG systems, to allows the generation of fluent and adaptive texts for musem visitors. We described our methods for normalizing dates and names, and adding object description text snippets, and for using the Semantic Web to extract relations which can be used to express the content of database fields. We provided examples from our test domain from the Edinburgh Musical Instrument Museum.

We will carry out a number of evaluations on parts of the extraction process and the output of the Methodius system using the automatically acquired resources. First we will test the acceptability of the extracted relation expressions using a crowdsourcing platform. When the system is complete, we will generate texts as part of a virtual museum experience, and evaluate the acceptability of the texts on a number of levels including fluency and coherence, and also relating to a number of other new features currently in the process of being added to Methodius but not part of the work described here.

5. Acknowledgements

This work is funded by a UK EPSRC PhD studentship. We would like to thank Vasilis Karaiskos for helpful comments on drafts of the paper. None of this would have been possible without Jon Oberlander and many years of entertaining discussions, encouragement, enthusiasm and support.

6. Bibliographical References

Androutsopoulos, I., Lampouras, G., and Galanis, D. (2013). Generating natural language descriptions from OWL ontologies: the NaturalOWL system. *Journal of Artificial Intelligence Research*, 48:671–715.

Chang, A. X. and Manning, C. D. (2012). SUTIME: A Library for Recognizing and Normalizing Time Expressions. In *8th International Conference on Language Resources and Evaluation (LREC 2012)*, May.

Dannélls, D., Ranta, A., Enache, R., Damova, M., and Mateva, M. (2013). Multilingual access to cultural heritage content on the Semantic Web. *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 13)*, August.

Galanis, D. and Androutsopoulos, I. (2007). Generating Multilingual Descriptions from Linguistically Annotated OWL Ontologies: The NaturalOWL System. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 143–146, Stroudsburg, PA, USA.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017). Creating Training Corpora for NLG Micro-Planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada.

Han, X., Ioris, A. A., and Lin, C. (2015). Web as Corpus Supporting Natural Language Generation for Online River Information Communication. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 363–364. International World Wide Web Conferences Steering Committee.

Isard, A., Oberlander, J., Matheson, C., and Androutsopoulos, I. (2003). Speaking the users' languages. *Intelligent Systems, IEEE*, 18(1):40–45.

Isard, A. (2007). Choosing the best comparison under the circumstances. In *Proceedings of the International Workshop on Personalization Enhanced Access to Cultural Heritage (PATCH07)*, Corfu, Greece, June.

Isard, A. (2016). The Methodius Corpus of Rhetorical Discourse Structures and Generated Texts. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., and others. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Mann, C., W. and Thompson, A., S. (1998). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 3:243–281.

Marge, M., Isard, A., and Moore, J. (2008). Creation of a new domain and evaluation of comparison generation in a natural language generation system. In *Proceedings of the Fifth International Natural Language Generation Conference (INLG08)*, pages 169–172, Ohio, USA, June.

Mellish, C. and Pan, J. Z. (2008). Natural language directed inference from ontologies. *Artificial Intelligence*, 172(10):1285–1315, June.

Mellish, C. (2010). Using semantic web technology to support NLG case study: OWL finds RAGS. In *Proceedings of the 6th International Natural Language Generation Conference (INLG10)*, pages 85–93, Dublin, Ireland, July.

Oberlander, J., Karakatsiotis, G., Isard, A., Androutsopoulos, I., and others. (2008). Building an adaptive museum gallery in Second Life. *Proceedings of Museums and the Web, Montreal, Quebec, Canada*.

O'Donnell, M., Mellish, C., Oberlander, J., and Knott, A. (2001). ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(03):225–250.

Pillen, D. (2015a). ArchivesSpace Dating Advice. In *Bentley Historical Library Blog*.

Pillen, D. (2015b). Normalizing Dates with OpenRefine. In *Bentley Historical Library Blog*.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Strötgen, J. and Gertz, M. (2010). HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Los Angeles, California. Association for Computational Linguistics.

Sun, X. and Mellish, C. (2007). An Experiment on "Free Generation" from Single RDF Triples. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, Stroudsburg, PA, USA.

White, M., Rajkumar, R., and Martin, S. (2007). Towards broad coverage surface realization with CCG. In *Proc. of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+ MT)*.

White, M. (2006). CCG chart realization from disjunctive inputs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 12–19. Association for Computational Linguistics.

White, M. (2014). Towards Surface Realization with CCGs Induced from Dependencies. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, Philadelphia, U.S.A. Association for Computational Linguistics.