Faust.rdf - Taking RDF literally

Timm Heuss
University of Plymouth,
Plymouth, United Kingdom
`Timm.Heuss@{plymouth.ac.uk, web.de}`

Abstract

This paper undertakes the modelling experiment of translating excerpts of the natural language play - "Faust" by Johann Wolfgang von Goethe - into a RDF structure, so that it is accessible by machines on a word or concept level. Thereby, it is crucial that statements made in the logic of the play can be distinguished from the usual, general purpose Linked Open Data. The goal is to find a standard compliant solution, stressing RDF's central role in the Web of Data as a format for arbitrary data.

1 Introduction

The Resource Description Framework (RDF) is meant to be the ideal data format for arbitrary information in the Web of Data, since it is open, machine-readable, non-proprietary and a World Wide Web Consortium (W3C) standard. In Berners-Lees popular five star ranking system, rankings above three stars can only be archived if the data is published in RDF (Berners-Lee, 2009).

Thus, this format plays an important role in many data portals[1], which publish data according to the Linked Open Data (LOD) paradigm.

RDF is the approved and commonly known method of making information of any kind available to machines, so they can access it in a structured way.

2 Motivation

However, when talking about publishing natural language content as LOD, the simple but very strict design goal of structuring information for machines is often violated. Whenever the natural language itself is not the modelling subject, datasets usually treat RDF as a meta data format, where accompanying natural language content remains either semi-structured (like in the Bible Ontology[2]) or entirely unstructured (like in Gutenberg in RDF[3]), in fields of the type `xs:string`. In both cases, natural language content is still not readable (in terms of accessible in a structured way) by machines - despite the fact that RDF is used. It's just displayable, just like it is displayable in the "eyeball Web" (Breslin et al., 2009, p. 82).

3 Idea

This paper is about the experiment Faust.rdf. Thereby, the author tries to strictly apply the central design principle of the RDF format, namely structuring information in a machine readable way, to natural language content, that would otherwise just be stored as `xs:string`. This means that natural language needs to be converted in a very structured variant, that formalizes the content on word or concept level. To choose a realistic scenario, the source material that is being formalized is the play "Faust: The First Part of the Tragedy" (Faust I) from Johann Wolfgang von Goethe[4].

The idea and the selected kind of source text is heavily inspired by the work of Richard Light, who expressed the works of Shakespeare as LOD (Light, 2013). In contrast to this paper, there is one important difference: The resulting LOD of Light semi-structures Shakespeare's texts, having selected an actual text line as smallest modelling atomicity. When taking RDF's principles literally, this goes, in the opinion of the author, not far enough.

[1] Popular Open Data portals are `http://data.gov.uk/` (accessed 2013-07-05), `http://data.gov.uk/` (accessed 2013-07-05) or `https://www.govdata.de/` (accessed 2013-07-05).

[2] `http://datahub.io/dataset/bible-ontology` (accessed 2013-07-05).

[3] `http://wifo5-04.informatik.uni-mannheim.de/gutendata/directory/texts` (accessed 2013-07-05).

[4] `http://en.wikipedia.org/wiki/Faust:_The_First_Part_of_the_Tragedy` (accessed 2013-07-06).

Faust.rdf is about trying to model selected statements of Faust with RDF's capacities, with the atomicity of words or concepts. The goal of identifying the status quo of RDF for this kind of modelling subjects and documenting experienced challenges. Resulting RDF statements would, of course, not replace or constitute the entire source text. Instead, it would form a novel kind of secondary source in the spirit of the Open Data movement. This is comparable to Wikipedia, where there are articles in English and, alternatively, in Simple English[5].

4 Requirements

The figures 1a and 1b show the definition the two central entities, that are considered in the following formalization of a literature work:

$$verse(time, source, speaker, type, act) \quad (1a)$$

$$content[verse](statements) \quad (1b)$$

The basic unit of a text is a *verse*, that refers to a certain *time* measure (which is, in poetry, usually the verse number). It's also constituted by the *source*, the human readable original text, an optional definition of the *speaker*, the *type* of the verse (whether it is a question, a proposition or a scene description) and the *act* in which the verse is subordinated to.

This structure is very generic and is comparable to other, semi-structured approaches, like Light's Shakespeare experiment (2013). In contrast to those approaches, the entity *content* is defined, consisting of actual *statements* about the natural language content of a play and the referenced *verse*, as defined above.

The goal is to produce five-star LOD (Berners-Lee, 2009) that represents the entities in the script of the play as close as possible. Thereby, links to external references are important and the underlying poetry needs to be understood properly. This is in contrast to automatic RDF extraction approaches like FRED[6] or the controlled vocabularies Attempto Controlled English[7] and Processible English[8], that try to extract and to represent logical relations out of a given natural language text. As this is a novel endeavor, the translation of Faust into RDF statements is a manual process.

5 Design

The solution consists of two design decisions: The exact way of how the two defined entities verse (1a) and content (1b) are converted into a RDF structure, and which LOD datasets are employed in that process.

5.1 RDF structure

In this modelling approach, context is very important. A verse in poetry does not contain general-purpose world knowledge, but very play- and actor-specific, subjective views on a fictional world - that could even turn out to be wrong at a later point.

To respect this fact in the realm of RDF, N-Quads (Cyganiak et al., 2008) can be used. N-Quads extend the *subject*, *predicate* and *object* of RDF triples with the fourth component *context*, which allows an optional definition of context for those triples.

RDF triples about the verse are within the context of the entire play (1a), and it is sufficient to rely on the basic building blocks of the Semantic Web (Allemang and Hendler, 2011, p. 9) by choosing an adequately unique Uniform Resource Identifier (URI). Instead, statements about the actual content (1b) of a certain verse are where protagonists claim, ask or lie. Thus it is a good idea of having a special handling for this kind of statements, i.e. putting the RDF triples about the content in a quadruple, in the context of a certain verse.

5.2 LOD datasets

LOD portals, in this case Datahub[9], make it very easy nowadays to find the linguistically grounded and interlinked data sets for the given use case.

In this project, entities are linked with lemonUby[10], one of the most comprehensive resources, especially in linking verbs. In addition, DBpedia[11] has an entry for entire play[12], which is used as namespace for all Faust.rdf statements.

[5] http://simple.wikipedia.org/wiki/Main_Page (accessed 2013-07-06).

[6] http://wit.istc.cnr.it/stlab-tools/fred/ (accessed 2013-07-07).

[7] http://attempto.ifi.uzh.ch/site/description/ (accessed 2013-07-07).

[8] http://web.science.mq.edu.au/~rolfs/peng/ (accessed 2013-07-07).

[9] http://datahub.io/ (accessed 2013-07-07).

[10] http://lemon-model.net/lexica/uby/ (accessed 2013-08-30).

[11] http://dbpedia.org/About (accessed 2013-07-07).

[12] http://dbpedia.org/page/Faust:_The_First_Part_of_the_Tragedy (accessed 2013-07-07).

Besides these LOD datasets, the source texts in natural language are taken from eBooks@Adelaide[13] (English) and Wikisource[14] (German).

6 Implementation

In this section, the findings of the preceding sections are tested with an exemplary RDF translation in the N-Quads notation. Thereby, the verses 1323 to 1325 of Faust are excerpted. In the English source text, they read (Goethe, 2005):

```
      FAUST
1323   This was the poodle’s real core,
1324   A travelling scholar, then?
       The casus is diverting.
      MEPHISTOPHELES
1325   The learned gentleman I bow
       before
```

In the following, the RDF translation of worthwhile parts is documented. Please note that triples are abbreviated using prefixes[15], even though the N-Quads notation format does not allow them. The full translation of verses 1323 to 1325 is available at GitHub[16].

6.1 Common statements

First, the protagonists Faust and Mephistopheles are introduced as instances of person respectively devil:

```
<:Faust> <rdf:type>
  <ubywn:WN_LexicalEntry_15513> .

<:Mephistopheles> <rdf:type>
  <ubywn:WN_LexicalEntry_134036> .
```

6.2 Verse Metadata

Translation of the verses follow a given, straightforward pattern, as introduced 1a on the preceding page. First, the source text is defined as human readable label:

```
<:verse1323> <rdfs:label>
  "This was the poodle’s real core"@en .
```

Line number and act are defined in a similar way and are omitted in this paper. However, translating the certain kind for verse 1324 is especially notable, as this verse contains a question as well as an assertion. After defining verse 1324 in the fashion of verse 1323 above, there are two variations:

```
<:verse1324a> <rdfs:subClassOf>
  <:verse1324> .
<:verse1324b> <rdfs:subClassOf>
  <:verse1324> .
# "verse1324a is a question"
<:verse1324a> <rdf:type>
  <ubywn:WN_LexicalEntry_153777> .
# "verse1324b is a statement"
<:verse1324b> <rdf:type>
  <ubywn:WN_LexicalEntry_81754> .
```

The per-verse meta data is completed by the assignment of the according speakers, e.g.:

```
# "Faust asks verse1324a"
<:Faust> <ubyvn:VN_LexicalEntry_1993>
  <:verse1324a> .
```

6.3 Verse content

As mentioned, the verse content is defined with N-Quads, having the individual verses as respective context. This allows to distinguish between general purpose world knowledge, like the fact that the play Faust has a certain verse 1324, from elements of the play, like the fact that the devil is a poodle. This way, it is also possible to encode a lie: Just like in the previous section, a certain verse would not be defined as a statement, but as a lie. Thanks to the context notation, further RDF statements can be made within the context of this verse, respectively in the context of this lie.

The following statements reflect the content of the verses 1323 to 1325:

“This was the poodle’s real core”

```
# "Poodle is a disguise"
<:Poodle> <rdf:type>
 <ubywn:WN_LexicalEntry_48830>
  <:verse1323> .
# "Poodle transforms into Mephistopheles
<:Poodle> <ubywn:WN_LexicalEntry_90692>
    <:Mephistopheles>  <:verse1323> .
```

**“A travelling scholar, then?
The casus is diverting.”**

```
# "Mephistopheles is a travelingScholar"
<:Mephistopheles> <rdf:type>
  _:travelingScholar <:verse1324a> .
# "travelingScholar is a scholar"
_:travelingScholar <rdf:type>
  <ubyvn:WN_LexicalEntry_99198> .
# "travelingScholar is a traveler"
_:travelingScholar <rdf:type>
  <ubywn:WN_LexicalEntry_115017> .

# "Verse 1324a amuses Faust"
<:verse1324a>
```

[13] http://ebooks.adelaide.edu.au/ (accessed 2013-08-30).

[14] http://de.wikisource.org/wiki/Faust_I (accessed 2013-07-07).

[15] Besides the common rdf and rdfs prefixes, the following are used: ubywn = http://lemon-model.net/lexica/uby/wn/, ubyvn= http://lemon-model.net/lexica/uby/vn/. Default namespace is http://dbpedia.org/page/Faust_I

[16] https://github.com/heussd/faust.rdf/blob/master/src/main/resources/faust.nq (accessed 2013-08-19).

```
  <ubyvn:VN_LexicalEntry_2516>
    <:Faust>  <:verse1324b> .
```

“The learned gentleman I bow before”

```
# "Verse 1324a is true"
<:verse1324a> <rdf:type>
  <ubywn:WN_LexicalEntry_53631>
    <:verse1325> .

# "Mephistopheles appreciates Faust"
<:Mephistopheles>
  <ubyvn:VN_LexicalEntry_731>
    <:Faust> <:verse1325> .
```

7 A critical view

This is a very first step and the translation might be neither perfect nor complete. While modelling the verse metadata is a not very exciting task, the crucial thing is to have a working context pattern for the actual content RDF statements.

It can be stated that RDF is, together with the N-Quads notation, indeed able to represent a fictional play, including the conditional statements that it involves. It is notable that N-Quads can considered to be a shortcut for a number of reification statements, as stated by an early mailing list posting (Palmer, 2001). Therefore, even pure RDF could able to cope with contexts, even though the resulting code would be much more complicated.

In the previous attempt[17], YAGO2[18] in combination of DBpedias subproject Wiktionary[19] was used to interlink entities. However, because both datasets still certain words, some translation results didn’t reflect the play properly.

Not being able to use prefix-namespaces, however, bloated up results unnecessarily and affected human readability. Having available assisting editor tools during the manual translation would have made things easier or at least faster. There is a clear lack of assistive, cross-disciplinary Natural Language Processing (NLP) and LOD tools, that are both user friendly and still can cope with giant, distributed datasets like YAGO2 and DBpedia.

8 Conclusion

When storing natural language, RDF still today plays a metadata role: Texts remain un- or semi-structured, stored in `xs:string`-fields, inaccessible to machines just like before the “eyeball Web”.

This paper undertakes the experiment of translating a natural language script of play, excerpts of Faust by Johann Wolfgang von Goethe, on a word or concept level into a RDF structure, so that it is accessible by machines, in the spirit of five-star LOD. Thereby, it is crucial that fictional statements made by protagonists of the play can be distinguished from the other, general-purpose statements containing world knowledge.

With N-Quads, in association with a number of datasets like lemonUby and DBpedia, a convenient solution is successfully designed and exemplary tested. This demonstrates the maturity of the used datasets as well as the RDF format, confirming it as a credible backbone of the LOD movement. Nevertheless, some issues are identified, regarding interdisciplinary tool support for the NLP and LOD domain.

9 Outlook

Having available large amounts of natural language texts, structured like proposed in this paper, would have a number of benefits.

As mentioned, in the role of a secondary source, RDF statements could give users hints in understanding the idea of the original text. This is especially important for very old texts, that use old-fashioned variants of natural languages that, eventually, only historians can understand.

Another benefit is that it could enable even non-experts to answer in-depth questions on the text, e.g. in case of Faust I: “In which scene does the devil appear the first time?”[20]

Also, knowledge-based NLP applications could become more common, like a Machine Translation approach, for example, which relies on the ability to extract human readable stories from RDF (Harriehausen-Mühlbauer and Heuss, 2012).

10 Acknowledgements

The author wants to thank Richard Light, Bernhard G. Humm and Kerstin Reinking for the inspiring discussions. He also thanks the LDL2013 reviewers for the very useful suggestions, which significantly improved this work.

[17] See GitHub diff page at `https://github.com/heussd/faust.rdf/commit/93f06b43c4212f0835171ab17ca89f22719aa2e4` (accessed 2013-08-30).

[18] `http://www.mpi-inf.mpg.de/yago-naga/yago/` (accessed 2013-07-07).

[19] `http://dbpedia.org/Wiktionary` (accessed 2013-07-07).

[20] In Faust I, answering the question of the first appearance of the devil requires a deeper text understanding, as he first appears in form of a dog, that later transforms into the devil.

References

[Allemang and Hendler2011] Dean Allemang and James A. Hendler. 2011. *Semantic Web for the Working Ontologist - Effective Modeling in RDFS and OWL, Second Edition*. Morgan Kaufmann.

[Berners-Lee2009] Tim Berners-Lee. 2009. Linked Data. Webpage, June.

[Breslin et al.2009] John G. Breslin, Alexandre Passant, and Stefan Decker. 2009. *The Social Semantic Web*. Springer, Berlin.

[Cyganiak et al.2008] Richard Cyganiak, Andreas Harth, and Aidan Hogan. 2008. N-Quads: Extending N-Triples with Context, July.

[Goethe2005] Johann Wolfgang Von Goethe. 2005. *Faust*. The Project Gutenberg.

[Harriehausen-Mühlbauer and Heuss2012] Bettina Harriehausen-Mühlbauer and Timm Heuss. 2012. Semantic Web based Machine Translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 1–9, Avignon, France, April. Association for Computational Linguistics.

[Light2013] Richard Light. 2013. Open Data on the Web position paper. In *W3C Workshop on the Open Data on the Web, 23 - 24 April 2013, Google Campus, Shoreditch, London, United Kingdom*.

[Palmer2001] Sean B. Palmer. 2001. Nquads. mailinglist, 08.