Multilingual Online Generation from Semantic Web Ontologies

Dana Dannélls
University of Gothenburg and GSLT
Department of Swedish Language
SE-405 30 Gothenburg, Sweden
dana.dannells@svenska.gu.se

Mariana Damova
Ontotext AD
Tsarigradsko Shosse 47A
Sofia 1784, Bulgaria
mariana.damova@ontotext.com

Ramona Enache
University of Gothenburg and Chalmers
University of Technology
Department of Computer Science and Engineering
SE-412 96 Gothenburg, Sweden
ramona.enache@chalmers.se

Milen Chechev
Ontotext AD
Tsarigradsko Shosse 47A
Sofia 1784, Bulgaria
milen.chechev@ontotext.com

ABSTRACT

In this paper we report on our ongoing work in the EU project Multilingual Online Translation (MOLTO), supported by the European Union Seventh Framework Programme under grant agreement FP7-ICT-247914. More specifically, we present work workpackage 8 (WP8): Case Study: Cultural Heritage. The objective of the work is to build an ontology-based multilingual application for museum information on the Web. Our approach relies on the innovative idea of Reason-able View of the Web of linked data applied to the domain of cultural heritage. We have been developing a Web application that uses Semantic Web ontologies for generating coherent multilingual natural language descriptions about museum objects. We have been experimenting with museum data to test our approach and find that it performs well for the examined languages.

Categories and Subject Descriptors

[**Information Systems**]: Information Storage and Retrieval; [**Software**]: Software Engineering; [**Computing Methodologies**]: Artificial Intelligence | Natural Language Generation

Keywords

Semantic Web, Multilingual Web, Linked Open Data, Reasonable View, Ontology, Natural Language Generation, Cultural Heritage

1. INTRODUCTION

The work described in this paper is developed within the Multilingual Online Translation (MOLTO) project.[1] More specifically, we present workpackage 8 (WP8): Case Study: Cultural Heritage. The objective of this workpackage is to build an ontology-based multilingual grammar for museum information using natural language generation technologies.

We have developed a Web application that applies natural language generation techniques to generate multilingual descriptions about museum objects from ontologies. Our approach is to utilize discourse structures that capture how concepts and relationships are realized linguistically. We have been experimenting with museum data to test our approach and find that it performs well for the examined languages.

The remainder of this document presents the motivation and the goals of WP8 (section 2). We describe the knowledge representation framework (section 3). In section 4, we describe the grammar implementation and present some generation results. We end with conclusions and directions for future work (section 5).

2. THE MOTIVATION AND GOALS OF WP8

The general motivation of WP8 is the increase of Cultural Heritage (CH) information on the Semantic Web. Today there exist millions of collections and thousands of applications providing a wide range of users direct access to cultural heritage material. This has brought up a need to develop tools that are capable of searching and presenting different kinds of information to end-users in their language of preference.

The goals of WP8 are to:

- build an ontology-based multilingual grammar for museum information for artefacts at Gothenburg City Museum (GCM) starting from the Conceptual Reference Model (CIDOC-CRM);
- build a prototype of a cross-language retrieval and representation system to be tested with objects in the museum, and automatically generate Wikipedia articles for museum artefacts in 5 languages;
- cover 15 languages for baseline functionality and 5 languages with a more complete coverage.

[1] http://www.molto-project.eu/

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

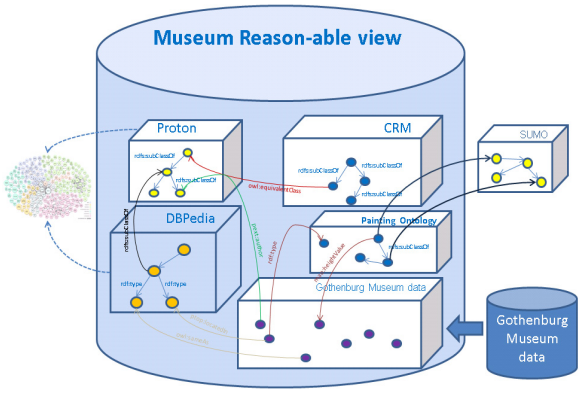

Figure 1: The Museum Reasonable View.

This paper describes the implementation of the prototype for retrieving and representing information about museum objects on the Web. It also describes the grammar that has been developed to automatically generate coherent object descriptions in two languages: English and Swedish.

3. THE MUSEUM REASON-ABLE VIEW

The Museum Reason-able View is an assembly of independent datasets, which are used as a single body of knowledge with respect to reasoning and query evaluation. Each data set in the Museum Reason-able View is aiming at lowering the cost and the risks of using specific linked datasets for specific purposes. This approach to linked data techniques has been discussed and implemented as a Reason-able View of the web of data [14].

The Museum Reason-able View environment, described in [6] is built as an instance of BigOWLIM triple store [1]. It contains: DBPedia 3.6,[2] Geonames,[3] PROTON [18], CIDOC-CRM [4],[4] the painting ontology [9], their mappings, and the triplified Gothenburg City Museum data [10].

Figure 3 shows the architecture of the Museum Reason-able View, which includes interconnected schemata and links to external datasets of the Gothenburg City museum data, such as the entire DBpedia. The Museum Reason-able View contains 245,365,883 explicit statements and 70,704,053 entities of which close to 10 thousand are museum artifacts from the Gothenburg city museum database.

3.1 Integrating Museum Data

Integrating datasets into linked data in RDF usually takes place by indicating that two instances from two datasets are the same by using the built in Web Ontology Language (OWL) predicate: `owl:sameAs`.[5] However, recent research [5, 7, 13] has shown that interlinking the models according to which the datasets are described is a more powerful mechanism of dealing with large amounts of data in RDF, as it exploits inference and class assignment.

We have adopted this approach when creating the infrastructure for the museum linked data, including several layers of upper-level ontologies. They provide a connection to different sets of linked data, for example PROTON for the Linked Open Data (LOD) cloud [2]. They also provide an extended pool of concepts that can be referred to in museum linked data that do not directly pertain to the expert descriptions of the museum objects, and the strictly expert museum knowledge is left to CIDOC-CRM. This model of interlinked ontologies offers a flexible access to the data with different conceptual access points.

3.2 Accessing Museum Linked Data

The data in the Museum Reason-able View is accessible via SPARQL [11] end-point and keywords.[6] The queries can be formulated by combining predicates from different datasets and ontologies in a single SPARQL query, retrieving results from all different datasets that are part of the Reason-able View.

A query example about museum objects from Swedish museums is given below.

```
select ?museumObject ?museum where {
 ?museumObject
 core:P109_has_current_or_former_curator ?museum .
 ?museum ptop:locatedIn ?location .
 ?location ptop:subRegionOf dbpedia:Sweden }
```

The above query returns the results that are depicted in figure 2. Note that the returned location is the DBpedia resource about the city of Gothenburg.

Other queries can be asked about the types of artwork preserved in the museum, their material, or about artwork from a certain period of time, etc.

4. NATURAL LANGUAGE GENERATION

The grammar formalism utilized for generating natural language descriptions from semantic web ontologies is the Grammatical Framework GF [16]. It is a grammar formalism, based on Martin-Löf's type theory [15]. The key feature of the grammar is the division of an abstract syntax, which acts as a semantic interlingua and concrete syntaxes, representing linearizations in various target languages (natural or formal).

GF comes with a resource library [17], covering the syntax of more than 20 languages.[7] The resource library aids the development of new grammars for specific domains by providing the operations for basic grammatical constructions, and thus making it possible for users without linguistic background to generate syntactically correct natural language.

4.1 Translation of the Museum Reason-able View to GF

The output result from BigOWLIM is a set of triples in the form of Resource Description Framework (RDF) statements consisting of constructs of the shape `<subject, predicate, object>` describing a resource. Each resource in the museum reason-able view is linked to its corresponding lexical unit in the GF lexicon. For example, GIM8165Obj is defined as a painting object in the abstract syntax and it is linearized as a person name with its title in the English dictionary:

```
GIM8165Obj : PPainting ;
GIM8165Obj = mkPN "Kliché";
```

[2] `http://dbpedia.org/`
[3] Geonames website: `http://www.geonames.org/`
[4] `http://www.cidoc-crm.org/`
[5] `http://www.w3.org/TR/owl-ref/`
[6] The data is available at: `http://museum.ontotext.com`
[7] `www.grammaticalframework.com`

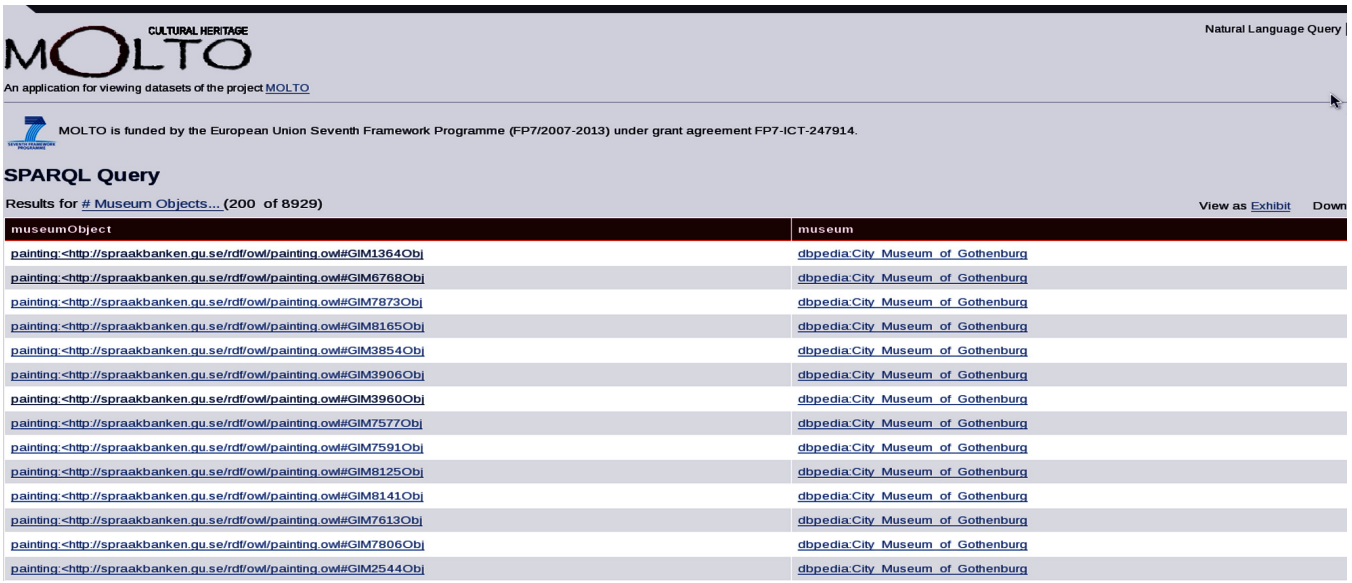

Figure 2: Query results about museum objects from Swedish museums.

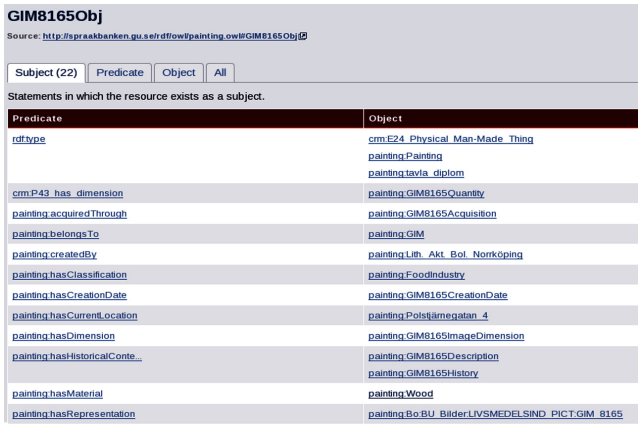

Figure 3: A set of triples describing the painting object GIM8165Obj.

Some lexical units such as paper, wood, etc. are already available in existing lexicons that have been imported to GF. Two of the lexicons that we are currently utilizing are the Oxford dictionary for English and the Swedish Association Lexicon (SALDO) [3], which is also available in LMF [12], for Swedish.

Painting resources are encoded in the abstract grammar as a sequence of semantic categories instead of a set of statements. For example the description of the GIM8165Obj depicted in figure 3 has the following semantic representation in GF.

```
fun GIM8165ObjDescription : PaintingDescription
  GIM8165Obj LithAktBol Y1916 NoMuseum GIM
  NoColour NoSize GIM8165ObjRepresented Wood ;
```

In the above example, the function *PaintingDescription* contains the following semantic concepts: painting, painter, year, museum, collection, colour, size, represented, and material. We should note that the retrieved information from the SPARQL query (figure 3) contains additional semantic concepts that are not covered by the discourse patterns yet.

4.2 Discourse Structures

Through linguistic analysis we have observed how the domain representation is encoded in a large set of well-formed object descriptions. We then followed the discourse structure to learn how the ontology statements are composed in English and Swedish [8]. Below we summarize some of the discourse patterns and the semantic concepts presented as functions in the GF abstract grammar.

- DP0 : painting painter year -> Text
- DP1 : painting museum painter size -> Text
- DP2 : painting painter represented museum -> Text
- DP3 : painting material year painter -> Text
- DP4 : painting painter year museum colour size -> Text

The discourse patterns are manually encoded in the application's abstract grammar. By optimizing the grammar we are able to generate several examples for each description.

```
def GenDP4 NoPainting _ _ _ _ _ _ _ _ _ = noText ;
def GenDP4 _  NoPainter _ _ _ _ _ _ _ _ = noText ;
def GenDP4 _ _  NoYear _ _ _ _ _ _ _     = noText ;
def GenDP4 painting painter year _  _  _  _  _  _ _ =
DP0 painting painter year ;
```

The basic idea behind the above implementation rules is that although there are several semantic concepts available for a certain object we can match its description with simpler patterns containing fewer semantic concepts.

4.3 Generation Results

Using the above discourse pattern constructions we are able to generate the following descriptions:

- (DP1-eng) Sommer Joy was painted in 1886. It measures 349 by 776 cm.
- (DP1-swe) Sommarnöje blev målad år 1886. Den är av storlek 349 och 776 cm.
- (DP2-eng) Sommer Joy is a painting made by Anders Zorn. The work depicts a view from Lilla Bommen at Hisingen.
- (DP2-swe) Sommarnöje är en målning av Anders Zorn. Den föreställer en utsikt från Lilla Bommen mot Hisingen.
- (DP3-eng) Sommer Joy is painted on paper in 1886 by Anders Zorn.
- (DP3-swe) Sommarnöje blev målad på papper år 1886 av Anders Zorn.
- (DP4-eng) Sommer Joy was painted by Anders Zorn in the year 1886. It is of size 349 by 776 cm and is painted on paper. The painting is displayed at the Museum of World Culture.
- (DP4-swe) Sommarnöje blev målad av Anders Zorn år 1865. Den är av storlek 349 och 776 cm och är målad på paper. Målningen återfinns på Världskulturmuseet.
- (DP4-eng) Sommer Joy was painted by Anders Zorn.
- (DP4-swe) Sommarnöje blev målad av Anders Zorn.

5. SUMMARY AND FUTURE WORK

In this paper we present a prototype developed in the context of the MOTLO project. We outline the entire infrastructure of the Museum Reason-able View and show its connection to existing infrastructures such as Dbpedia. We present the multilingual grammar application that is being developed to generate multilingual museum object descriptions from the described resources and demonstrate how the generation results are obtained.

This work is about an automatic work-flow of sharing data infrastructures that is explicitly targeted towards the Semantic Web. The primary goal of this effort is to support question answering and automatically generate short Wikipedia-like articles for museum artifacts in 5 languages with extensive coverage. We are currently extending the grammar to support more patterns and more languages including Finnish, French and German. The generation results will be evaluated using native speakers of the language.

6. ACKNOWLEDGMENTS

This work is supported by MOLTO European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement FP7-ICT-247914.

7. REFERENCES

[1] B. Bishop, A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, and R. Velkov. OWLIM: A family of scalable semantic repositories. *Semantic Web Journal, Special Issue: Real-World Applications of OWL*, 2011.

[2] C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas. 4th linked data on the web workshop (ldow2011). In *WWW (Companion Volume)*, pages 303–304, 2011.

[3] L. Borin, M. Forsberg, and L. Lönngren. SALDO 1.0 (Svenskt associationslexikon version 2). Technical report, Språkbanken, Göteborg universitet, 2008.

[4] N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff. *Definition of the CIDOC Conceptual Reference Model*, 2009.

[5] M. Damova. *Data Models and Alignment*, May 2011. Deliverable 4.2. MOLTO FP7-ICT-247914.

[6] M. Damova and D. Dannélls. Reason-able view of linked data for cultural heritage. In *Proceedings of the third International Conference on Software, Services and Semantic Technologies (S3T)*, 2011.

[7] M. Damova, A. Kiryakov, M. Grinberg, M. K. Bergman, F. Giasson, and K. Simov. Creation and integration of reference ontologies for efficient lod management. In *Semi-Automatic Ontology Development: Processes and Resources, IGI Global*, Hershey PA, USA, 2011.

[8] D. Dannélls. *D.8.1 Ontology and corpus study of the cultural heritage domain*, 2011. Deliverable of EU Project MOLTO Multilingual Online Translation.

[9] D. Dannélls. An ontology model of paintings. *submitted to the journal of applied ontology*, 2011.

[10] D. Dannélls, M. Damova, R. Enache, and M. Chechev. A framework for improved access to museum databases in the semantic web. In *Proceedings of RANLP, Workshop of Language Technologies for Digital Humanities and Cultural Heritage*, page 8, 2011.

[11] P. Eric and S. Andy. SPARQL. the query language for RDF, January 2008. W3C Recommendation.

[12] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. LMF for Multilingual, Specialized Lexicons. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, 2006.

[13] P. Jain, P. Z. Yeh, K. Verma, R. G., M. Damova, V. P. Hitzler, and A. P. Sheth. Contextual ontology alignment of lod with an upper ontology: A case study with proton. In *Proceedings of 8th ESWC, Extended Semantic Web Conference*, Heraklion, Greece, May 2011.

[14] A. Kiryakov, D. Ognyanoff, R. Velkov, Z. Tashev, and I. Peikov. LDSR: Materialized reason-able view to the web of linked data. In *Proceedings of OWL: Experiences and Directions (OWLED) 2009*, Chantilly, USA, 2009.

[15] P. Martin-Löf. Constructive mathematics and computer programming. In Cohen, Los, Pfeiffer, and Podewski, editors, *Logic, Methodology and Philosophy of Science VI*, pages 153–175. North-Holland, Amsterdam, 1982.

[16] A. Ranta. Grammatical Framework, a type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189, 2004.

[17] A. Ranta. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2(2), 2009.

[18] I. Terziev, A. Kiryakov, , and D. Manov. *D.1.8.1 Base upper-level ontology (BULO) Guidance.*, 2005. Deliverable of EU-IST Project IST.