Towards A Smarter Online Debates Analysis: A French Public Portal Use-Case

Ghislain Atemezing, Hacene Cherfi, Anh Huynh, Martin Coste

Towards A Smarter Online Debates Analysis:
A French Public Portal Use-Case

Ghislain Atemezing, Hacene Cherfi, Anh Huynh and Martin Coste

MONDECA, 35 boulevard de Strasbourg 75010 Paris, France.
<firstName.lastName@mondeca.com>

Abstract. This paper describes a semantic annotation workflow and visualization of politician speeches from the French Government Portal. The pipeline consists of an ontology modeling, a pre-processing and customized annotations of 7200 random speeches of public actors during the last 30 years using Content Augmentation Manager (CA-Manager) and a visualization component based on the central notion of a Topic. The implementation of the proof-of-concept shows that adding semantic annotations to the current publication of official speeches in RDF leads to efficient search of debates and better interconnection with external datasets.

Keywords: speeches, debates, semantic annotations, interoperability, TopicPages, ITM, CAM-Manager

1 Introduction

Information is flowing on the Web in many formats, such as CSV, PDF or HTML. For human readable formats such as HTML, the basic challenge is to add href links between pages as it was at the early days of the Web. In the recent years, the idea of publishing data and linking them has gone mainstream since the first draft by Sir. Tim Berners-Lee [1]. Almost twenty years after, publishers and governments bodies are now witnessing the benefits of using semantic technologies to publish and better discover their content on the Web. Although there exist some guidelines to transform legacy government data into linked data [5], there are some use-cases where a customized pipeline based on specific types of data provide efficient answer. It is the case for the French portal *vie-publique* service which contains more than 100,000 speeches in ATOM format, and is seeking to offer a novel approach to discover and explore its content based on semantic technologies.

We propose in this paper a semantic annotation workflow and a visualization based on the concept of a Topic. A TopicPage is the central element of the visualization obtained after annotating and enriching a pivot page with the concepts extracted by using the ontology for debates. The paper is described as follow: we first introduce the background and motivation (Section 2), and further detailed the components of the semantic annotation workflow (Section 3) and the ontology modeled for the debates online (Section 4). We then exemplify the

implementation in Section 5, followed by the steps for the automatic generation of the TopicPage (Section 6). A brief conclusion remark is provided at the end of the paper (Section 7).

2 Motivation and Background

The French portal `vie-publique.fr` service provides citizens with resources and useful data in order to understand the main topics which are subject to debate of the general public in France. The website `http://www.vie-publique.fr` also contains a collection of more than 100,000 speeches from public actors, such as the President, government, members of congress or mayors during the last thirty years. The main formats used by the portal is the atom syndication format (ATOM), which is an XML format, and which is in turn converted into HTML. Our aim is to extract relevant entities described in an ontology for speeches, and be able to extract persons, organizations, location and related information that can be useful to retrieve in a speech.

3 Semantic Annotation Workflow

The worflow used is based on a central component: the Content Augmentation Manager (CA-Manager). The content augmentation manager (CA-Manager) is in charge of processing XML and HTML documents. This module extracts the concepts and entities detected using text mining techniques with the resources coming from the ITM module. In the scenario presented in this paper, we use the GATE tool [4] for the entity extraction. In [3], Hacene et al., CA-Manager uses an ontology-based annotation schema to transform heterogeneous content (text, image, video, etc.) into semantically-driven and organized contents. Figure 1 represents the five components that support the building and customized management pipelines for semantic contents annotations. These components are the following:

1. Extraction: identify and tag domain-oriented knowledge (terms, named entities, relations) from content, performed by exiting Information Extraction (IE) tools (such as GATE, Luxid, GeolSemantics, EIDON, etc.);
2. Consolidation: reconcile extracted knowledge with the domain ontology and the content of the knowledge repository (instances and property values);
3. Storage: export and store the reconciled knowledge;
4. Validation: let the human user validate the suggested annotations and knowledge;
5. Enrichment: export new validated terms and entities into the IE's linguistic resources (gazetteers, grammars, named entity lists).

The rest of the components combined with CA-Manager to obtain a customized TopicPage visualization (Section 6) of any content are the following:

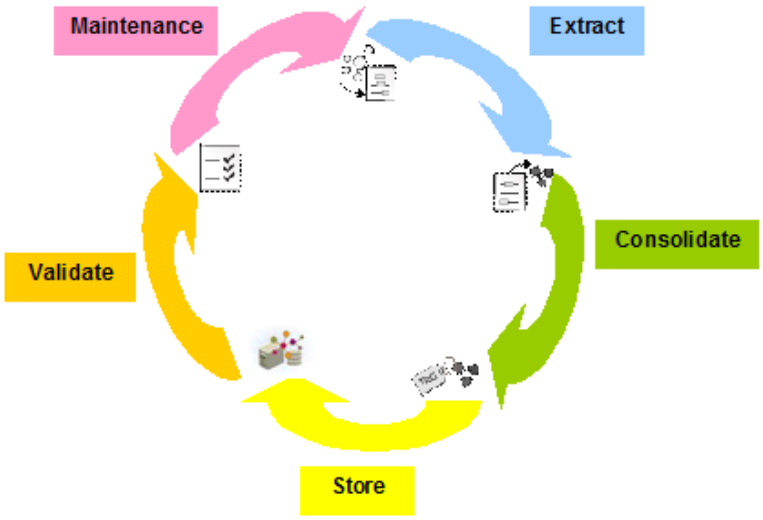

Fig. 1. CA-Manager's main components

- The ITM module to manage terminologies and the vocabularies. The model is stored and manipulated in this module containing entities (persons, organizations), authoritative thesaurus (geolocation, themes) and classification rules.
- The `Skos2gate` module, in charge of exporting in SKOS[1] the resources from the ITM module and generate the Gate Gazetteers.
- A classifier is the component used for applying rules in SPARQL-alike on generated RDF datasets for enriching the annotations. The rules are edited and exported from ITM.
- The RDF2XML[2] tool is used to generate on-the-fly HTML visualization of TopicPages based on a template defined by the developer or the data publisher.

4 Debatescore: A vocabulary for Online Debate

The debatescore ontology[3] is used during the annotation process. It is an extension of the Dublin Core [6] vocabulary with a unique class `Debate`[4], aiming at

[1] `http://www.w3.org/TR/skos-reference/`
[2] `https://github.com/MondecaLabs/rdf2xml`
[3] `https://raw.githubusercontent.com/MondecaLabs/vocabs/master/debatescore.ttl`
[4] `goo.gl/d712R4`

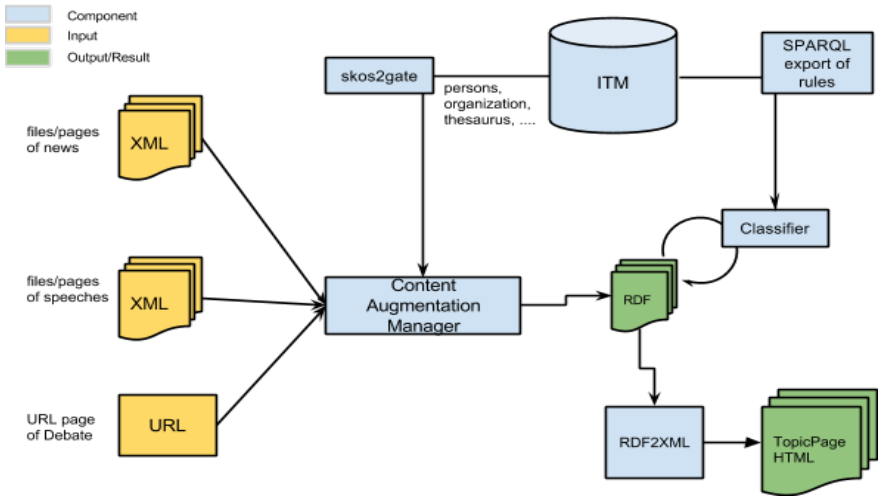

Fig. 2. Main components of the workflow for storing vocabularies, detecting and reconciling concepts in text, export and topicPage visualization.

giving to citizens and others to contribute or participate in taking public decision or political one. A debate should have the following requirements:

- An issued date captured with the property `dcterms:issued` with an xsd:date as range.
- A title captured with the property `dcterms:title`.
- A debate belongs to a category, which are five (05): mediation, concertation, public debate, consultation and public pool. Those values are modeled as SKOS [7] concepts.
- A debate is target to a specific audience with a status. Some controlled status are used to represent the status of a debate.

The ontology is loaded into ITM for annotating the speeches accordingly. Each speech is modeled in RDF as an individual of the Debate class.

5 Implementation

We extracted 7,100 XML documents that are first encoded in UTF-8. During the annotation, 11% of the documents were rejected (i.e., 789 documents) because they were malformed or were non-UTF-correctly encoded. For these documents, we need further investigation. However, we still have enough material for our experiment. The Figure 3 shows how one can manually add documents in our knowledge base in order to semantically annotate them against the terminology set shown on the left hand side of the page (Public information, organizations, persons, etc.).

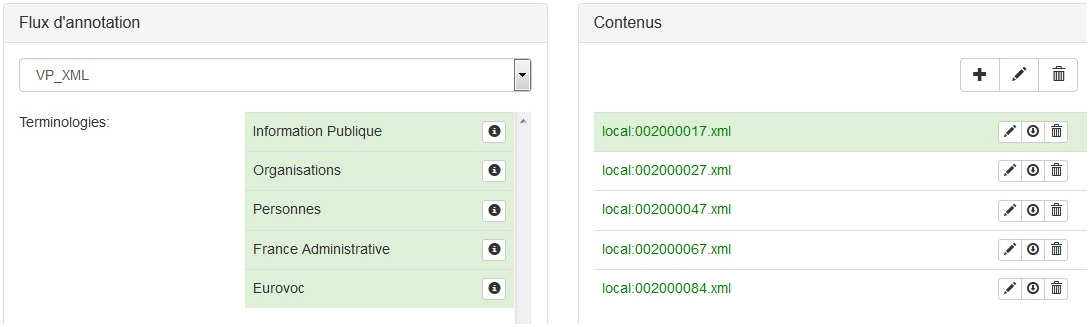

Fig. 3. An example of the annotation process for 5 speeches extracted from our corpus.

Each document annotation in CA-Manager leads to the creation of its corresponding object in the knowledge base management system (ITM). The most important part of each annotation shown in Figure 5 is the set of list of *subjects* identified from the terminologies (e.g., *Protocole*, *MINISTRE*) or suggested by the CAM system (e.g., *France* which exists in our base as an organization but not as a location) as shown in Figure 5.

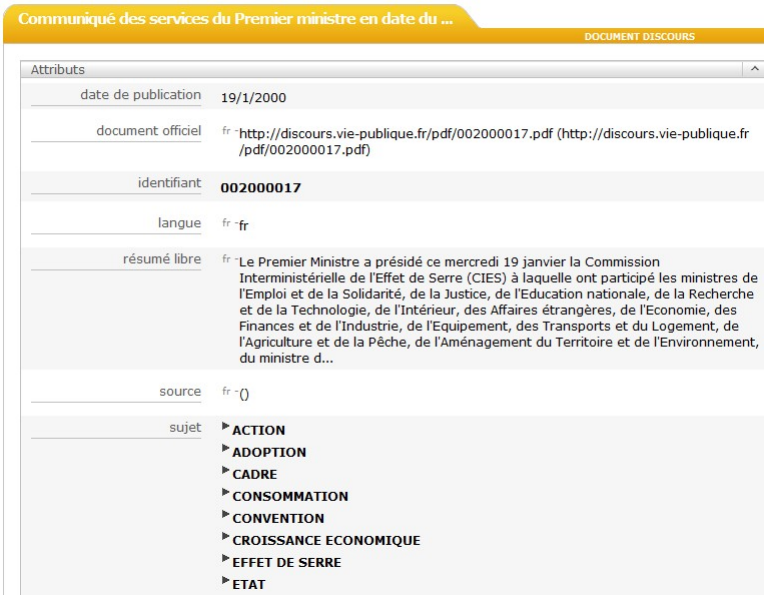

Fig. 4. An example of the annotations stored in ITM which have been extracted from a document Speech annotated in CA-Manager.

Figure 6 depicts for our use-case, the concepts retrieved associated respectively to EuroVoc multilingual thesaurus [5] reference data, specific concepts annotated in the portal website and the locations.

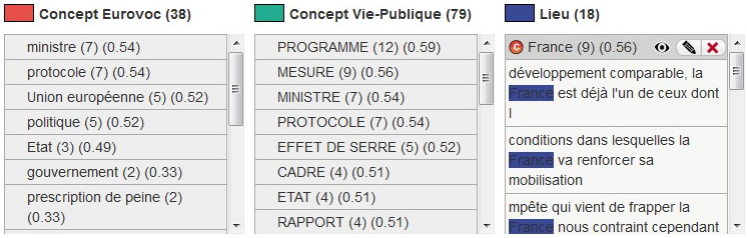

Fig. 5. An example of entities and concepts retrieved of a document in CA-Manager.

6 Automatic TopicPage Generation

The notion of topic here has its historical origin in the the Topic Maps standard [2], meaning a unique subject, concept or business object. A TopicPage is the visualization component of a "Topic" with relevant links to resources that are internal or external to the Enterprise knowledge base.

Four steps are needed to generate a TopicPage:

1. Selection of the pivot page. In this step, the user identifies which is the HTML or XML page containing the central topic, or the pivot page.
2. Convert into RDF all the corpus annotated and Pivot Page. The database obtained can be loaded in a SPARQL endpoint.
3. Creation of a new dataset based on the concepts from the Pivot page by using SPARQL filtering and CONSTRUCT queries.
4. Generate visualization, by using a templating framework (such as RDF2XML) to automatically generate relevant views of the TopicPage, and adding links to additional information.

The result of the visualization is available at `http://labs.mondeca.com/poc/vp/topicPageNosRegions.html` corresponding to the central topic "Our region" located at `http://www.vie-publique.fr/forums/nos-regions-demain-parlons-en.html`.

Note that for the TopicPage, the rendering of the speeches view is better interpreted using Chrome browser. We will further investigate the compatibility with other browsers.

[5] `http://eurovoc.europa.eu/`

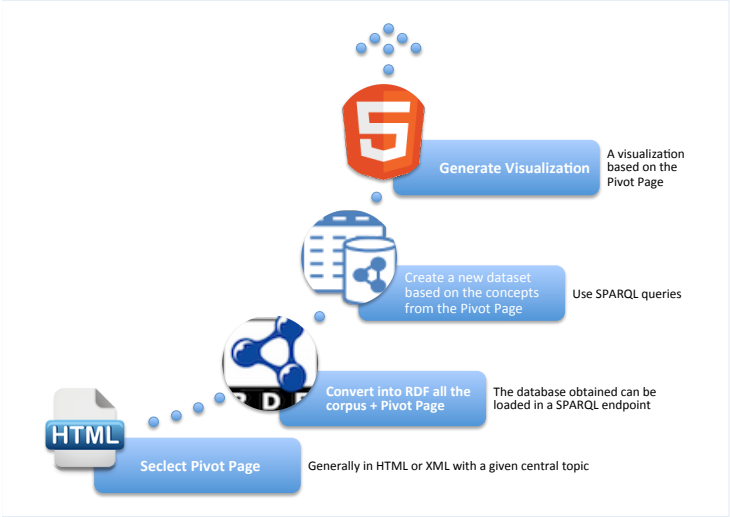

Fig. 6. Steps for generating customized TopicPage visualization.

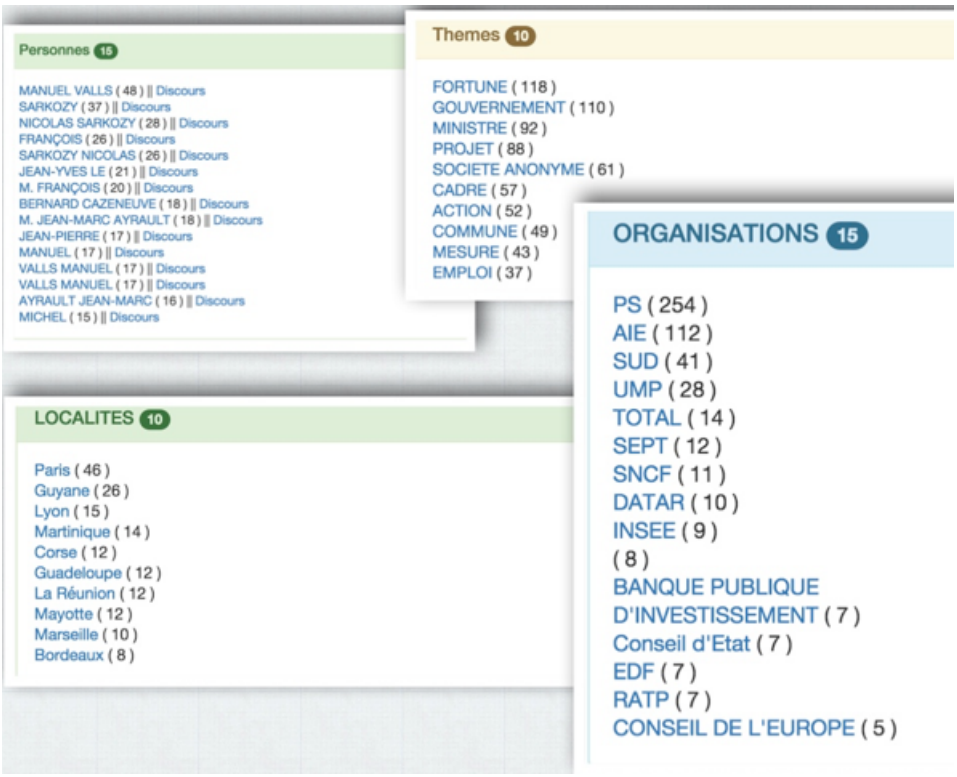

Fig. 7. Four categories detected in a politician speech. From left to right, persons, subjects, organizations and locations.

7 Conclusion and Future Work

We have presented in this paper a customized pipeline for annotating and enriching a data portal on public speeches and debates. We have first presented the motivation of our approach, followed by the description of the main components of the workflow, to extracting and generating knowledge based on the debatescore vocabulary. We have further implemented our approach on 7200 random speeches of public actors during the last 30 years in France. By doing so, we claim that such approach leads to a new way of publishing, discovering and navigating contents based on a pivot page or TopicPage, which aggregates personalized information surrounding a topic. Our next challenge is to index the RDF dataset to build a search index on-top of the RDF dataset. Another direction is to identify other similar datasets on debates to link to, in order to have fully 5-star Linked Data.

Acknowledgments. The authors want to thank Olivier Garry and team at DILA for the interactions to provide us with input and comments.

References

1. T. Berners-Lee. Linked data: Design issues, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`.
2. M. Biezunski. Topic maps at a glance. In *XML Europe. Conference*, pages 387–391, 1999.
3. H. Cherfi, M. Coste, and F. Amardeilh. Ca-manager: a middleware for mutual enrichment between information extraction systems and knowledge repositories. In *4th workshop SOS-DLWD "Des Sources Ouvertes au Web de Données*, pages 15–28, 2013.
4. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: an architecture for development of Robust HLT applications. In *40th Anniversary Meeting of the Assoc. for Computational Linguistics (ACL)*, 2002.
5. B. Hyland, G. Atemezing, and B. Villazon-Terrazas. Best Practices for Publishing Linked Data. W3C Working Group Note, 2014. `http://www.w3.org/TR/ld-bp/`.
6. D. C. Initiative. Dublin core metadata element set, version 1.1: Reference description, 2004.
7. A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. *W3C recommendation*, 18:W3C, 2009.