Constitute: The World's Constitutions to Read, Search and Compare

Zachary Elkins[1], Tom Ginsburg[2], James Melton[3], Robert Shaffer[1], Juan F. Sequeda[4], and Daniel P. Miranker[4]

[1]Department of Government, The University of Texas at Austin
zelkins@austin.utexas.edu, rbshaffer@utexas.edu
[2]Law School, University of Chicago
tginsburg@uchicago.edu
[3]Department of Political Science, University College London
j.melton@ucl.ac.uk
[4]Department of Computer Science, The University of Texas at Austin
{jsequeda, miranker}@cs.utexas.edu

Abstract. A constitution forms the foundation of the virtually all governments around the world. A surprisingly large number of constitutions change each year. On average, 30 constitutions are amended and 5 are completely replaced each year. Despite this level of change, no country changes its constitution often enough for the country's officials to gain much experience as constitutional drafters. In order to address constitutional drafters need for systematic information a web portal, *Constitute*, has been created using semantic technologies. The portal provides searchable access to the work of the Comparative Constitutions Project. The Comparative Constitution Project has amassed on-line copies of over 700 constitutions dating from the present day back to 1789. Further, the project created an OWL onotology containing over 600 constitutional concepts, e.g. women's rights, and is well into a process to tag the text of their entire collection of constitutions. To build the Constitute portal, the Comparative Constituion database was converted to an RDF representation using R2RML mappings and Capsenta's Ultrawrap. The portal implements semantic search features such that constitution drafters may search, read and compare constitutions. Constitute is available at http://www.constituteproject.org

1 The Problem: Drafting New Constitutions

Constitutions empower, and limit the institutions that govern society. In doing so, they are intimately linked to the provision of public goods. Outcomes, like democracy, economic performance and human rights protection, are all associated with the contents of countries constitutions. It is little wonder that constitutions are often blamed for poor economic and political outcomes or that such outcomes, or that those outcomes are followed by constitutional change. Both domestic and international actors view constitutional change as a means to spur economic, political and social development.

Even as international events make news, few people are aware of the scope of constitutional change. On average, 30 constitutions are amended and 5 are completely replaced each year[3]. Despite this level of constitutional change, no country changes its constitution often enough that within its ranks of public officials are there people experienced in drafting constitutions. In fact, the most common scenario is, the people responsible for drafting constitutions or constitutional amendments have no prior experience in the matter, nor will they repeat their duties. Even basic background support, such as systematic information on the contents of other countries constitutions, and even previous constitutions in their own country, is lacking. Just the existence of such systematic information would form a basis for methodically approaching the most basic question of which topics should be addressed in a constitution. Access to corresponding text may serve to shape debate and enable more productive effort. In current practice, external advisors are frequently consulted. Even the most experienced advisors tend to rely on a small set of well-known models, and are only able to draw on anecdotal evidence when responding to specific questions.

2 The Solution: Constitute

In 2005, Elkins and Ginsburg launched a large-scale data-collection project: the Comparative Constitutions Project (CCP)[1]. This project identified and acquired almost every constitutional text within each country's series of constitutional laws. Additionally, the CCP categorized the content of national constitutions original texts and subsequent amendments for all independent states from 1789 to present day. Each constitution is organized by sections. Each section, if applicable, has been tagged with topics which represent the interpretations by domain experts. For example, Article 32 of the Constitution of Angola refers to the topic of *right to privacy*. These texts, in electronic and searchable form, represent a unique and comprehensive repository of constitutional legislation.

Given the scope and immediacy of the need, the first of what is anticipated to be scholarly and practical applications of the corpus, is Constitute (http://www.constituteproject.org). Constitute is a semantically enabled search portal, built using Semantic Web technologies. The primary purpose of Constitute is to address constitutional drafters need for systematic information on the contents of constitutions. Constitute allows users to access full texts of constitutions and excerpts from those texts on particular topics and geographic regions. For example, suppose a constitutional drafting committee wishes to review the constitutional language used on the topic of *right to privacy*, in a subset of constitutions written post WWII in Europe. Right to privacy is but one example; the set of constitutional topics forms an OWL ontology with over 400 concepts. The data for each constitution (i.e section name, content and topics) is represented in RDF. On average, each constitution consists of 9000 triples. *The number of constitutions hosted in Constitute is steadily increasing. At the time of this writing it hosts over 160 constitutions.*

[1] http://comparativeconstitutionsproject.org/

3 Constitute Architecture

Domain experts created the Constitution Ontology in OWL which represents the taxonomical relationship between constitutional topics and geographic regions. Subsequently, domain experts cleaned the constitution data from CCP and represented in a tabular format. Mappings between the tabular data and the Constitution Ontology are represented in R2RML, which is then used to generate the RDF using Ultrawrap. Finally, the RDF data and and Constitution Ontology is used to create a search portal built on top of the Google App Engine. Figure 1 represents the architecture of Constitute.

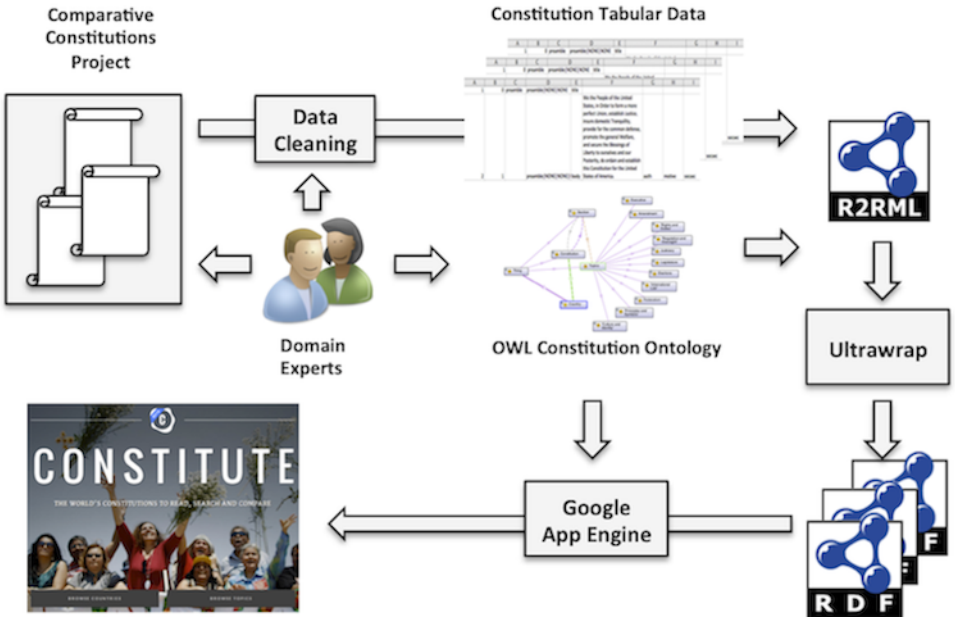

Fig. 1. Architecture of Constitute

The Comparative Constitutions Project The Comparative Constitutions Project (CCP) has focused on the systematic collection and interpretation of constitutional text since 2005. Through an exhaustive and costly process, the CCP has collected, scanned, and, when needed, translated into English the text for 95% of documents in the universe of constitutional systems. To date, the CCP repository consists of includes 789 of the worlds 839 constitutional systems, and 2877 of the 3234 amendments to these systems.

In the process of interpreting constitutions, the CCP tagged each section of a constitution (i.e Article 1), if applicable, with constitutional topics, such as *right to privacy* or *separation of church and state*. These topics come from a 669 question survey that CCP developed in order to interpret constitutions. The

interpretations of each constitutions are represented as a set of tuples which include the constitution name, a short description of the topic, a topic code, and a numerical reference to an organization header. A typical entry might read [Albania_2008, Official religion, offrel, 10.1]. This means that Section 10.1 of the Albanian constitution of 2008 has been tagged with the topic *official religion*.

Data Cleaning A team of political science and law graduate students (the domain experts) cleaned the CCP data. First, the domain expert selects a constitution, and downloads an uncorrected OCR scan of that constitution from the CCP repository. Subsequently the domain experts clean the document by fixing typos, errant line breaks, bad characters and formatting organizational headers (e.g. Chapter, Article, etc). Next, a Python script consisting of regular expressions merges the clean text with the tag data and generates a tabular representation of the data. The tabular data is represented in XLS so the domain experts could make use of the track change features. Around 10 domain experts in 9 months were able to clean and generate 180 constitution XLS files.

Constitution Ontology A domain expert created the Constitution Ontology which consists of two main parts: Topics and Geography. The topics part consists of a taxonomical relationship of the constitutional topics from the CCP. For example, the topic "Freedom of Religion" is a sub class of "Religion" and "Civil and Political Rights". Additionally, "Religion" is a subclass of "Culture and Identity" while "Civil and Political Rights" is a sub class of "Rights and Duties". The geography part is an import and extension of the FAO Geopolitical Ontology. New geographic sub-regions were added, such as the Balkans, Middle East, etc. Additional missing synonyms names for a countries were added(i.e. Netherlands, Holland). Figure 2 depicts a portion of the Constitution Ontology. We are currently in the process of extending the ontology to include time eras (Years, Decades, etc).

Generation RDF Each constitution XLS was loaded into a Microsoft SQL Server database, a table per constitution. By loading the XLS into a relational database, the mappings between the the constitution data and the Constitution Ontology could be represented in R2RML. R2RML, the Relational Database to RDF mapping language[2], in conjunction with the Direct Mapping[1], are two recently ratified standards by the W3C to expose relational databases to the Semantic Web. Capsenta's Ultrawrap[2], a productized version of a research prototype [4], was used to convert the constitutional data into RDF. Ultrawrap supports both W3C mapping standards, and both ETL and SPARQL execution on relational data. First, the Direct Mapping created an initial default mapping represented in R2RML. Subsequently, the R2RML was edited to use terms from the Constitution Ontology. In this current phase of Constitute, Ultrawrap generates periodic dumps of the constitution data as RDF.

Search Portal The Constitute search application was built using Google App Engine, Python and the RDFlib library[3]. Free-text search is powered by indexing raw constitution text using the Google App Engine search API. Semantic search

[2] http://www.capsenta.com
[3] https://github.com/RDFLib

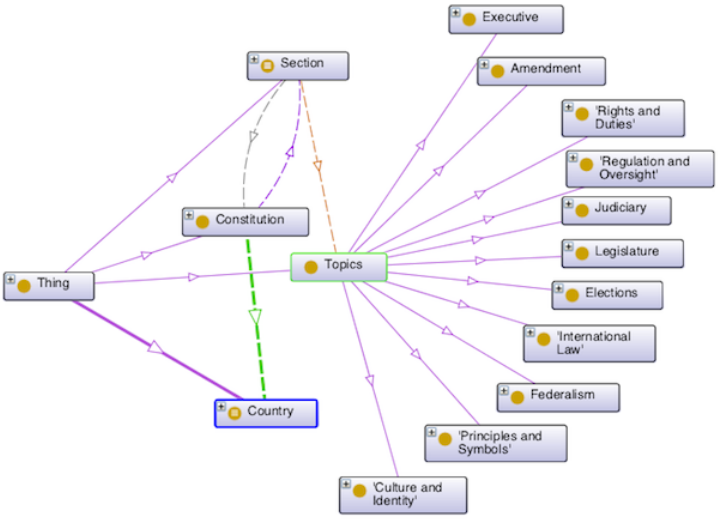

Fig. 2. The Constitution Ontology

is powered by the RDF triples and OWL ontology, which is stored in Google's DataStore. Currently, Constitute implements light weight inference. For example, Figure 3 shows a user typing "Religion" and the results are suggested topics that are semantically related to the Religion, such as "Separation of church and state" which is a sub class of Religion.

Search results are delivered to the front-end as JSON and browser-based computation is provided by AngularJS. Where possible HTML5 browser-based caching is used. CSS3 media queries are used to provide responsive design across desktop and mobile browsers.

4 Design Choices and Lessons Learned

Initially, Constitute was planned to be developed using non semantic web technologies such as XML to encode the constitution data. The switch to Semantic Web technologies was motivated by two features: linking data and reasoning. Representing the constitution data in RDF, enables the creation of links between constitutions through shared constitutional topics that are represented by URIs. Additionally, it opens the possibility of linking the constitutional data with other datasets in the LOD cloud, such as DBpedia, New York Times, etc. The taxonomical relationship between the constitutional topics allows to preform sub class reasoning, as shown in Figure 3.

Before Semantic Web technologies were used, the relationships between the constitutional topics were not formally represented in any format. After receiving a brief introduction (couple hours) on ontologies and Protege, a domain expert initiated the task of creating the Constitution Ontology. For the domain

Fig. 3. Subclass reasoning in topic search: User searches for "Religion" and gets related topics which are sub class of Religion.

expert, it was very intuitive to create the taxonomical relationships between the constitutional topics and import the FAO Geopolitical Ontology. The ontology consists of classes, sub classes, object properties and datatype properties. Starting from the constitutional topics, the Constitution Ontology was created in approximately 150 hours. Future work involves refactor the ontology in order to adhere to best practices. For example, we need to study if SKOS should be used instead of sub classes in OWL. Additionally, the domain expert did not reuse common vocabulary when possible

5 Next Steps

Currently, Constitute is wrapping up Phase 1: converting the constitution data into RDF and the creation of a search portal that supports light weight reasoning. In Phase 2, we plan to extended Constitute in two directions. First, we plan to make all the constitution data available as Linked Data and link to other datasets such as DBpedia and New York Times. Second, we plan to extend the semantic features of the search portal, including capabilities for exploratory search.

We anticipate that Constitute will improve constitution-making by allowing drafters, and their advisors, to consider the full array of possible choices when de-

termining the contents of their countrys constitution. We also anticipate the tool will empower domestic actors not directly involved in drafting the constitution but who are, nonetheless, integral to the success of that process. Increasingly, constitution-making processes ask the public to participate, for example by submitting suggestions to the constitutional drafting committee or approving the completed draft in a public referendum.

This would be of great interest to a wide variety of domestic actors in countries all over the world. Many constitutions are not available in digital form and tools to organize their provisions for a non-specialist are rare. However, there is substantial demand for such tools from public officials, lawyers, non-governmental organizations, students, etc. For instance, the Nigerian Constitution App[4] has been downloaded over 530,000 times[5], making it one of the most downloaded apps from the Google Play store in Nigeria. The plain texts generated by third projects will facilitate the creation of similar projects in other countries, which will empower citizens in those countries to challenge public officials who violate their constitutions tenets.

Acknowledgements Constitute is a partnership of the Comparative Constitutions Project and Google Ideas, with significant financial support from the Indigo Trust. Engineering and web-design support are provided by Psycle. The following organizations have made important investments in the Comparative Constitutions Project since 2005: the National Science Foundation (SES 0648288), the Cline Center for Democracy, the University of Texas, the University of Chicago, and the Constitution Unit at University College London. Juan F. Sequeda was supported by NSF Graduate Research Fellowship. Daniel P. Miranker was support by NSF Grant IIS 1018554.

References

1. M. Arenas, A. Bertails, E. Prud'hommeaux, and J. Sequeda. Direct mapping of relational data to RDF. W3C Recomendation 27 September 2012, `http://www.w3.org/TR/rdb-direct-mapping/`.
2. S. Das, S. Sundara, and R. Cyganiak. R2RML: RDB to RDF mapping language. W3C Recomendation 27 September 2012, `http://www.w3.org/TR/r2rml/`.
3. Z. Elkins, T. Ginsburg, and J. Melton. *The Endurance of National Constitutions*. Cambridge, 2009.
4. J. F. Sequeda and D. P. Miranker. Ultrawrap: Sparql execution on relational data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2013.

A Minimal requirements

– *The application has to be an end-user application:* The primary purpose of Constitute is to address constitutional drafters need for searching and accessing contents of constitutions. Additionally, anybody interested in learning about constitutions of the world can use Constitute.

[4] http://www.constitutionforall.com.ng/
[5] http://indigotrust.org.uk/2013/02/11/update-nigerian-constitution-app/

- *The information sources should be i) under diverse ownership, ii) heterogeneous and iii) contains real world data:* Each constitution is from a different country. The data is heterogenous because we start with text, convert into tabular form and then into to RDF.
- *The meaning of data has to play a central role: i) represented in Semantic Web technologies, ii) manipulated/processed in interesting ways to derive useful information and iii) achieve things that alternative technologies cannot do as well:* The constitution data is represented in RDF and the taxonomical relationship between the constitutional topics is represented in an OWL ontology. The ontology is exploited to do sub class reasoning in order to allow users discover topics that they were not aware of. The discovery capabilities would not be possible by having the data in XLS or in a relational database by itself.

B Additional Desirable Features

- *The application provides an attractive and functional Web interface (for human users):* Constitute provides a user interface which has been developed by expert web designers and evaluated at Google.
- *The application should be scalable:* By using Google App Engine and Datastore, we are able to scale.
- *Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained:* Constitute is currently in its initial launch phase. Next steps is to perform evaluation and improve the search features.
- *Novelty, in applying semantic technology to a domain or task that have not been considered before:* To the best of our knowledge, semantic technology has not been applied for search of constitutions.
- *Functionality is different from or goes beyond pure information retrieval:* Each section of a constitution has been tagged with concepts of an ontology. Thus search results are highly curated.
- *The application has clear commercial potential and/or large existing user base:* Constitution drafters around the world are the potential users. Additionally, any person interested in learning more about constitutions is a potential user.
- *Contextual information is used for ratings or rankings:* In this domain, we have not considered rankings.
- *Multimedia documents are used in some way:* The constitutions can be viewed in the browser as HTML or downloaded as PDF.
- *There is a use of dynamic data (e.g. workflows), perhaps in combination with static information:* Constitutions are static data. When new constitutions are created or amended, they can be added to the workflow.
- *The results should be as accurate as possible (e.g. use a ranking of results according to context):* The data comes from the CCP and has been highly curated by domain experts since 2005.
- *There is support for multiple languages and accessibility on a range of devices:* Constitute works on mobile phones. The next version of Constitute will support multiple languages (constitutions in their native language(s) and search).