Formal Methods of Concept Formation*

Zachary Elkins
zelkins@austin.utexas.edu

Jessie Baugher
baugher@illinois.edu

Robert Shaffer
rbshaffer@utexas.edu

September 12, 2017
v.3.0

Abstract

The building blocks of any science are its concepts. But concepts are, by definition, personal (and hence subjective) abstractions of an individual's observations. The accumulation of knowledge requires inter-subjective understanding, which implies either conceptual *standardization* or, at least, *translation* across scholars. Unpacking and relating concepts can be a delightful process that fosters a sense of scientific engagement and accelerates discovery, but requires substantial effort from researchers across the disciplinary spectrum. In this paper, we propose a data structure designed to formalize and share conceptual frameworks (i.e., terms, classes, properties, etc.). We elucidate this structure with an application to constitutional law, one of many research domains that are ripe for such exploration. The simple premise of this project is that such researchers should coordinate *somehow*, and the methods proposed below are as promising as any. We demonstrate some positive dividends of such for both scholars and their various audiences.

*Prepared for presentation at the *Poliinformatics Workshop.* Thanks to the National Science Foundation, Google Ideas (now Jigsaw), the Cline Center for Democracy, the University of Texas, and the University of Chicago for their early and/or continued investment in the Comparative Constitutions Project (CCP). Thanks to Tom Ginsburg and James Melton for ideas and inspiration. Thanks to David Collier, Diana Kapiszewski, Matt Ingram, David Waldner, Matthias Koenig, Juan Sequeda, and Dan Miranker for guidance in various parts.

Why are cows and whales are in the same taxonomical class but sharks and whales are not? Mostly because Linneaus organized them that way in his twelve-page work, *Systema Naturae*. But certainly *other* biologists, using different methods or focused on different distinguishing traits, have come up with other taxonomic schemes. Think, for example, of the two grand traditions in biological classification that have followed since Linneaus, phenetics and cladistics. Phenetic methods classify organisms by their appearance (morphology), whereas cladistics does so based on their evolutionary lineage. Each of these approaches has advanced significantly over the years. Still, Linneaus' scheme seems to have captured many hearts and minds, and it has established a vocabulary for the rest of us, for better or worse.

But what if we could incorporate the categorization schemes of Linneaus' competitors and successors? What if these schemes could co-exist, and be accessed, in some rationalized system of knowledge? So, a researcher might query some database of concepts about "shark" and learn about its characteristics and the various scientific (or even folk) categories in which scientists have seen fit to group the animal. She might use the category that makes the most sense for the purpose at hand or, if she was content with the standard categorization, she could at least understand (and, importantly, visualize) what other observers are thinking. That is, she could see the entity's properties, categories, synonyms, and related entities. At least to us, this approach seems like a more equitable and systematic way to organize knowledge.

Linneaus and his 18th century colleagues have some excuse – other than outright competitiveness – for not stitching these views and labels together. However, modern data systems offer a more systematic alternative. In particular, some of the data-standardization suggestions of *Web 3.0* architects point to a workable (and already working) structure. What is required is implementation (and evangelization) of the system in the social sciences. Here, we explore what such a formalization of concept formation might look like in the social sciences and, in particular, in the field of constitutional law. A growing number of scholars are engaged in the systematic analysis of national constitutions, but with different purposes and approaches. We argue that a coordinated system of conceptualization and knowledge

organization would be more fruitful, and suggest a specific and actionable set of methods and activities designed to implement such a scheme.

1 The Domain of Constitutional Text

Consider the study of national constitutions. Just as biologists gaze at the living world and see different species, students of constitutional law scan the very unnatural world of text. Legal documents are famously ambiguous, and lend themselves to multiple interpretations; for example, a Linneaus in this field might see a leader that is selected *by the people* to serve for a *fixed term* and label the system that enshrines those attributes as "Presidentialist." Another observer might look to still other properties, such as veto power or nominating power, and categorize the system quite differently.

These sorts of decisions hit very close to home. For over ten years, we have been intensively involved in the Comparative Constitutions Project (CCP), an effort to identify, excavate, and interpret each national "constitution" that has come into force since 1789 (Elkins et al. 2009). To assist with the project's goals, CCP's founders (including one of our co-authors) created a conceptual inventory that we have used to read and interpret constitutions in order to record their contents. Our conceptual scheme – starting with the identification of what, exactly, a country's constitution *is* – represents merely *our* view of the world. Of course, our view is informed by the many scholars who have come before us. After all, the classification of constitutional elements goes back at least as far as Aristotle, who assembled a dataset of the constitutions of Greek city states (*Politics*, especially Books IV and VI). But our working assumption is that other researchers should, and will, read texts differently from us and will record different properties about them. Still, in releasing our data, we have assertively propounded our own perspective, and perhaps, discouraged the elaboration of others. Ideally, however, our interpretations should speak to and *with* the voices of others. So, for example, other scholars should be able to see that we recorded Article 38, section 1 of the Cape Verdian constitution of 1982 as expressing, among other things, the right to privacy. That particular section reads:

> "The right to personal identity, to civil rights, to a name, honor, and reputation, and to personal and family privacy shall be guaranteed."

"Right to privacy" is a concept that we find useful for capturing an element of that clause, but other researchers might find additional or other, perhaps more refined, concepts just as relevant. Indeed, the field of human rights has been particularly fertile, with respect to concepts. And so it may be that the Cape Verdian clause speaks to evolving concepts such as the "right to be forgotten" (Rosen 2011), or even the "right to the city" (Harvey 2008). And, of course, the clause is a manifestation of multiple other concepts even in our own CCP taxonomy.

2 Relationships Among Conceptual Schemes

In order to understand how different conceptual schemes can "converse," we need to understand how the schemes might vary. It is only then that we can understand how to translate and share them. Of course, these schemes can vary in different ways, but consider three kinds of relationships among datasets in the constitutional domain, each of them related to the CCP data in roughly three orders of proximity (see Table 1).

2.1 First Order Extensions and Refinements

In the first category, we should consider those projects in which scholars are essentially observing and coding the same thing, albeit with different purposes and theoretical frameworks and, importantly, different levels of generality. Some schemes are simply more refined than others, depending upon observers' tastes and interests. This is the phenomenon of the proverbial Eskimos and their sixty kinds of snow. Our categorization scheme, in some areas, stops at a relatively high level of generality (perhaps comparable to the genus level in biology), but other researchers might prefer to make finer distinctions (say, at the species level). For example, in the CCP, we identify six characteristics related to the treatment of women in constitutions. Scribner and Lambert (2010), however, have since coded a smaller set of constitutions across a much wider (and deeper) set of characteristics, a set that

was motivated by their own theoretical agenda. Specifically, these authors are interested in comparing constitutions that "emphasize women's different needs and provide gender-based protections [*difference*] as compared to countries with constitutional structures that emphasize equality or gender neutrality [*equality*]" (Lambert and Scribner 2009: 337, bracketed insertions are ours, but concepts are theirs). This particular distinction between their concepts of *difference* and *equality* is not fully specified in our data, though one might certainly measure something about equality provisions in our framework.

Or consider another case of refinement. In a budding project, Koenig and Tsutsui are reading and coding constitutions across a set of characteristics related to ethnic identity and minority incorporation. To be sure, the idea of ethnic group accommodation is something that we sought to capture in our data collection, but not exactly in the way that Koenig and Tsutsui conceptualize the variation. Koenig and Tsutsui propose to record a large set of characteristics, some of which are redundant with our schema (duplication that may well be helpful), and some of which deepen the domain of study in important ways.

Or take Mila Versteeg's study of constitutional Rights. Versteeg codes the content of constitutions for rights, and only rights. Our inventory of rights overlaps extensively with hers, but also differs in important ways. Each dataset includes some rights that the other does not, and specifies some rights at different levels of generality. On some rights that are common to both datasets, the Versteeg dataset captures more systematically the *qualifications* (conditions for derogation) that constitutions sometimes attach to rights' restrictions. There are also differences in caseload. Our sample, which begins in 1789, is more extensive in coverage than Versteeg's, which starts in the post-WWII era. In general, however, the two datasets are very closely matched, and one can imagine all sort of benefits (both substantive and methodological) to systematizing and comparing the concepts, measures, and observations of the two datasets.

The discussion above implies a sequential, generational, relationship in which data collectors build on one another. It could be, of course, that analysts till the same soil independently (sometimes willfully so) of one another. We might even think of some projects as competing with one another – carried on in parallel but with glances at each other. Gould

Type	Unit	Topic(s)	Source
First Order	Constitution	Gender	Scribner and Lambert (2010)
		Minority Protection	Koenig and Tsutsui
		Rights	Versteeg
		Judicial Authority	Brinks and Blass
		Judicial Independence	Ríos-Figueroa and Staton (2012)
		Executive power	Shugart and Carey (1992), Elgie (2005), Alvarez et al. (1996)
		Environment	Boyd (2011)
		Judicial Councils	Garoupa and Ginsburg (2009)
		Constitutions c. 1978	van Maarseveen and van der Tang (1978)
		Constitutions c. 1999	Harutyunyan and Mavčič (1999)
Second Order	Jurisprudence	Various	CompLaw
		Free Speech	Keck (2007)
		Gender	Women's Link Worldwide
		Various	New York Times
		Rights	Cichowski (2007)
		German Cases	Honnige et al.
Third Order	Country	Development	World Development Indicators
		Political Authority	Varieties of Democracy
		Minority Status	Minorities at Risk
		Rights Enforcement	Cingranelli and Richards (2010)
		Ethnic Relations	Wimmer et al. (2009)

Table 1: Selected Datasets in the Constitutional Domain

(2010)'s account of Brontosaurus comes to mind. In his telling, two zealous paleontologists presented competing names for a group of large, quadrupedal dinosaurs: *Brontosaurus*, and *Apatosaurus*. Early work by Riggs (1903) advocated lumping the two terms into a single genus, rendering *Brotonsaurus* the "junior synonym" due to its later date of proposal. However, Tschopp et al. (2015) split the group into at least two distinct genera based on an updated survey of the fossil record, restoring both *Apatosaurus* and *Brontosaurus* to proper taxonomic status.

Political science, for better or worse, does not have any sort of taxonomical police force. This lack of centralization is largely a practical issue; taxonomic reorganization - biological or otherwise - is often contentious, requiring researchers to reconcile and eliminate synonymous and partially-overlapping names advanced by competing research groups. Because of these practical constraints, we suggest that *translation* (maintaining multiple useful names) is often both more practical and more useful for scientific inquiry than *standardization* (retiring names). Both methods are forms of de-cluttering, but one is more neutral than the other. It seems to us that the most productive approach to integrating multiple perspectives on the world is to record and relate them, ideally in a formalized (machine-readable) fashion. In the CCP, for example, the project's principal investigators incorporated ideas from van Maarseveen and van der Tang (1978) into their conceptualization scheme, and subsequently became familiar with the work of Harutyunyan and Mavčič (1999). But, had the earlier schemes been articulated in a formalized fashion, the labor involved in discovering, developing, and relating these schemes would have been substantially reduced.

2.2 Second Order Extensions

Another task of conceptual translation is to connect schemes in one domain (or set of units) to those in another. Take national constitutions; constitutional texts can be inherently interesting and informative, but require outside interpretation to be implemented in practice (for example, by judges, academics, or ordinary citizens). The indeterminacy and enforceability of text is a matter of dispute *ad nauseum* at nearly every forum on constitutional law, and not worth belaboring here. But regardless of one's own interpretation of a par-

ticular document, a reader of constitutional text *will* wonder how the relevant courts or officials have interpreted and implemented the law. Conversely, a reader of a court's decision regarding a particular law will want to read the accompanying law in question. By way of analogy, the same is true of religious texts and the doctrine surrounding them. In order to read scripture and its attending doctrine efficiently, one needs several basic pieces of information. One is, of course, a systematic recording of what the texts say. Another is the decoding of the *decisions and opinions* regarding that text. But importantly – and the point of this essay – one also needs a *relatable* set of concepts in order to connect these pieces of data. Relatability implies either a common (perhaps standard) lexicon, or translation tools that allow one to connect different conceptual vocabularies.

Several empirical projects are in progress that offer some promise in this regard (see the second-order group in Table 1. One is the Complaw project (Carrubba et al. 2012), in which a group of political scientists has collected data on a year's worth of judicial opinions from some 75 countries. Another promising project is one led by Keck (2007), who is coding decisions on free speech by a sample of international, national, and subnational courts. Another example is a project organized by Women's Link Worldwide, which hosts an online repository of court decisions related to gender. A basic goal might be to connect first-order projects with second-order projects. That is, text on the right to free speech in the South African constitution would be connected to that country's high court interpretation of that provision in cases.

2.3 Third Order Extensions

A third extension, even further from the constitutional text, is to data on the relevant unit of analysis (e.g., the country, which is the jurisdictional unit of observation in the CCP). For example, one might want to connect constitutional data in a given year and country to country-year data on human rights enforcement, economic conditions, or demographic characteristics. This kind of merging is, of course, common in any data-analytic paper on the origins or consequences of constitutions. Often, however, systematic conceptual mappings can ease this task. Consider Gurr (2000)'s Minority-at-Risk (MAR) data, which include

characteristics of groups that have experienced marginalization in their state of residency. How do MAR concepts such as "political discrimination" relate to some of the elements of ethnic accommodation in constitutions? It might useful to think of a constitutional idea such as "constraints on party formation" as an example of "political discrimination". If so, the relationship between those two concepts is worth recording.

3 Concept Formation and Enrichment

The discussion thus far assumes a world in which multiple scholars independently analyze similar phenomena with similar concepts. This state of affairs is nothing new, though increasing data availability may make it more frustrating. We might instead imagine a different world, in which researchers could see, understand, and connect to one another's conceptual schemas (or ontologies). Understanding how one's concepts fit with another's concepts is tremendously valuable and enlightening in and of itself, as we elucidate. However, building bridges among concepts has other, perhaps even more powerful, downstream effects. If one researcher can connect her ontology to another, related scheme, she can then connect the manifestations of each ontology (that is, their data) to one another. Mapping concepts allows us to combine data creatively, which is critical for the analysis of measurement and hypothesis testing.

An example might clarify the benefits of living in this utopian world. Recently, one of us set out to write a paper on political polarization across contexts. His specific objective was to understand whether the variance in constitutional structure across countries has led to more or less polarized politics. Political polarization takes many interesting forms, but one form seemed particularly relevant to this paper: the degree of *social distance* between those who identify with different parties. Social distance (Bogardus 1925) is a familiar concept to many sociologists and political scientists, which measures the degree to which the average member of a given group is comfortable living in proximity to members of another group. Like many individual-level latent variables, social distance is assessed using a battery of survey questions, which ask respondents to indicate whether they would be

comfortable having a member of an out-group as a close family member, a neighbor, and so on, all the way down to the most distant – co-existing in the same country.

We had hoped to find data on this measure of *social distance*, or a related concept or measure, from various institutional contexts. But first we wondered where *social distance* sat with respect to other concepts in its "semantic field". What are its synonyms? Which attributes do these related terms share, or not share? Are there any recognized sub-dimensions of the concept? And what were the major works related to this set of concepts? If a map relating the idea of "social distance" to other ideas were available, we could query it to retrieve information regarding definitions, presumed attributes, and related concepts – all rooted in the relevant academic literature. However, at least in political science, data repositories such as Dataverse or the ICPSR do not make their content available in this fashion.

Other disciplines have made greater efforts in this area. For example, psychiatrists developed the *Diagnostic Statistical Manual of Mental Disorders (DSM)* in order to harmonize their concepts and observable attributes for just the purposes described previously. Admittedly, it is not clear that this authoritative classification of mental disorders *deserves* to be the standard any more than does Linneaus' version of the natural world. Certain concepts (disorders) are raised to a perhaps undeserved legitimacy, and the degree of reification is sometimes alarming (at least to an outside observer). In an ideal world, we might imagine multiple DSM-like volumes, developed for differing purposes but linked using a shared conceptual mapping and data structure.

It might be most illustrative to think more concretely about the information demands of such a pluralist conceptual system. If we were to formalize our understanding of political concepts, and their relationship to others, what information – exactly – would we wish for? Fortunately, concept formation is a mature domain of study in political science (Collier and Adcock 1999; Sartori 1970). Drawing on their work, we propose three fundamental tasks, which are at the heart of what we will variously refer to as concept formation and concept enrichment.

1. *Mapping the semantic field.* What historical and contemporary terms are related to the concept, either as synonyms, antonymns, subtypes, supertypes, or other relationships (e.g., diminished subtypes)?

2. *Properties.* What are the concept's characteristics, either defining or elective?

3. *Dimensionality and classes.* What, if any, are the sub-components of the concept? What is the relationship between the concept and its components (e.g., hierarchical or not)?

Enumerating these three tasks is one thing, but how might accomplishing these tasks look in practice? And, what data structures are most appropriate to these goals? Of course, one could formalize one's conceptual data in tabular form, the most familiar structure of such information. Imagine rows of concepts described by columns of information about classes, properties, synonymns, antonyms, and the like. The result is a table of information about concepts that any computer can interpret. Moreover, analytic and visualization software applications can render such information in a manner useful for human interpreters. And, of course, one would like to combine that table with other tables, linked by any of the items in the data.

4 One Promising Solution: Linked Open Data

One data paradigm that meets these criteria is termed "Linked Open Data," sometimes called the "Semantic Web," and more generally, "Web 3.0." The easiest way to introduce this paradigm (and to demonstrate its utility) is to show its results in practice. We present such an example using Constitute, a linked open data repository built to house the Comparative Constitutions Project (CCP) data.

4.1 Linked Open Data in Action

First, some background information: the core intellectual product of the CCP is a set of data on some 600 characteristics of the world's constitutions (and revisions to those constitutions)

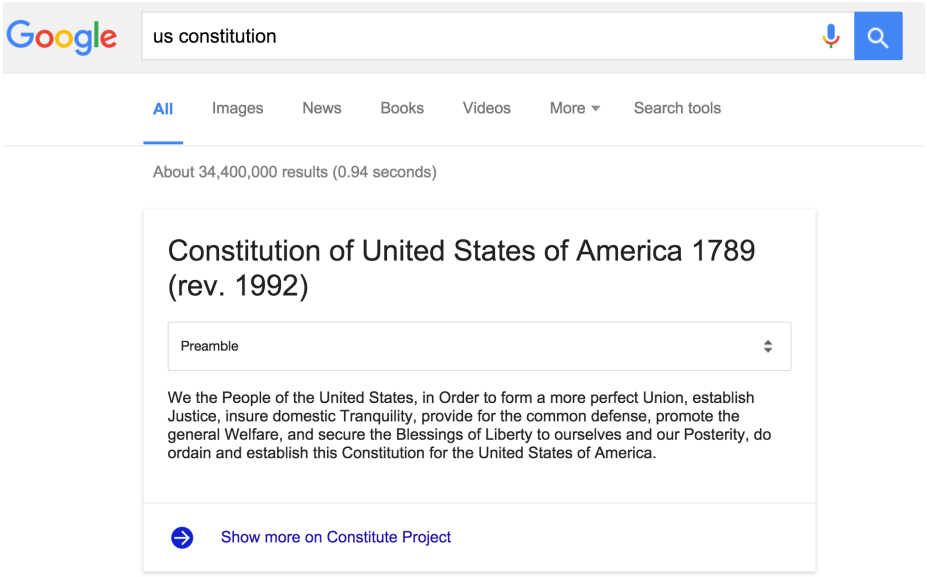

Figure 1: Results from a Google search on "US Constitution"

since 1789. In 2013, we partnered with Google Ideas (now Jigsaw) to leverage these data in order to build Constitute, a public repository of constitutional texts currently in force. Constitute consists of a public-facing front-end that allows users to search constitutional texts across a simplified subset of the original conceptualization scheme, containing some 300 unique characteristics.

The datafiles for CCP consist of a recognizable tabular format: a matrix of rows of constitutional events and text against columns of constitutional attributes. The data for Constitute, however, take a very different form built for the online environment – one that can be easily accessed by both human beings and machines. Much more on that format shortly. But, first we fast-forward to what we see as some early dividends from this strategy.

As of September 17, 2015, a Google search on "constitution" returns the 4,500-word U.S. Constitution on a card at the top of the search results (and a 3x5-inch card at that, at least on most screens). A onebox on the card allows the reader to jump through sections. The text and data on these cards comes directly from online Constitute data. A number of other constitutions show up too. Try a Google search on "Bhutan Constitution" and read about that country's commitment to Gross National Happiness in Article 9 (2). All of these cards pull directly from the Constitute repository.

The U.S. Constitution on an index card is, quite literally, a small thing, but something that represents a huge advance for information science, and the social scientists like us who depend on such. Most of us by now are at least semi-conscious of the info-boxes that surface in and around internet search results. Googling "things to do in Austin" returns a carousel of images depicting top-rated Austin outings, and "MLB standings" returns a onebox that shows the state of play in major league baseball. Click any of these items (from within the carousel or onebox) and a "knowledge panel" on the right margin materializes with more information about the item (the Yankees' knowledge panel, for example, lists the current 25-player roster).

4.2 From Linked Open Data to One-boxes and Knowledge Panels

The underlying source for these kinds of search results is Google's "knowledge graph," a curated set of highly-connected and machine-readable data ("linked data"). Google began delivering results from the knowledge graph in 2012, but the concept of linked data emerged much earlier (Berners-Lee et al. 2001). Linked data are simple to understand and their utility is immediately obvious. One key feature is that each data element, whether a concept or a concrete "thing" (e.g., the U.S. Constitution) has its own unique location on the web. These locations (URIs) are not websites (URLs), but places where a single data point resides. Each of these entities are linked to other entities through some relationship, which is itself labeled with a unique URI. So, a typical Linked data file comprises a seemingly endless lines of subject-predicate-object "triples," each of whose elements is a distinct URI. For example,

<http://constitute/constitution/sudan2005/article16>
<http://constitute/hastopic>
<http://constitute/torture>

is one of many triples in the Constitute dataset. This particular triple tells us that Article 2 of the Sudanese Constitution deals with the topic of torture. It should be clear that the Constitute dataset alone has many *other* links to each of these entities.

Data files containing this kind of information are forbidding to browse directly. However, this data structure is highly useful for machine interpretation. Editing and visualization tools allow analysts to work and understand various relationships in these files (more on that below), and database tools can execute arbitrary queries quickly and efficiently on data that are structured in this fashion. Moreover, the entities contained in a linked data structure (e.g. concepts, properties, data elements) can be connected to an array of other other entities and concepts.

Consider, again, the constitution-on-index-card idea, which represents a very simple use of linked data. To produce these cards, Google's "knowledge graph" queries data on the world's constitutions that our website Constitute makes available as linked data on a SPARQL endpoint (a data hub that machines can hit). Google's engineers can then program its search engine to reproduce the textual data as an info-box with the text indexed by the section headers, which are also identified in the Constitute data. More complex applications are also possible; for example, imagine a box that lists provisions on "cruelty" in constitutions from countries about which Human Rights organizations have made allegations of torture. Linked data on all of those elements exists; one needs only to construct the appropriate query.

Still, it seems to us that open-linked data maintains a frontier-like quality. At time of writing, a relatively small group of high-profile web users have adopted the Semantic Web paradigm. Prominent adherents include DBpedia (a collection of the structured components of Wikipedia), the New York Times, and knowledge graphs maintained by Google and Bing. Academic applications of the Semantic Web have been sparser still; to our knowledge, Constitute is the only major dataset in political science that uses the Semantic Web data paradigm.

4.3 Ontology Enrichment

One of the most intriguing aspects of the Semantic Web is its utility for concept formation and enrichment. Datasets structured in an open linked format must include a formalized

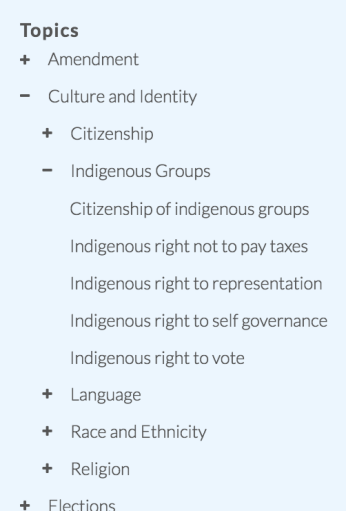

Figure 2: A Snapshot of Constitute's Topic Tree

Constitute ontology, with the first-level topic "Culture and Identity" expanded to see several sub-topics. The subtopic "Indigenous Groups" is also expanded to see a set of constitutional provisions including "Indigenous right not to pay taxes."

"ontology", which relates the data, concepts, and properties present in the given domain. These relationships - which are structured in the subject-verb-predicate triple structure described in the previous section - can be intimidating to manipulate, but provide substantial downstream analytical and theoretical payoffs.

In many applications, the information necessary to generate an appropriate ontology is already available. For example, the 600-topic Comparative Constitutions Project survey instrument was designed in a hierarchical fashion, organizing topics into high-level categories (e.g. those referring to the Head of State, or to civil and political rights), with a series of lower-level categories and variables embedded within. To create the 300-topic Constitute ontology, we drew on this structure to form a three-level hierarchy, with equivalence relationships defined between some concepts in order to collapse them into more general ideas (see Figure 2 for an example).

To reiterate, the Constitute categorization is only our view of constitutional elements and by no means suitable or useful for all users. For example, none of the topics has "women" in its label, which at one point frustrated some users who had understandably searched for constitutional provisions related to gender. There *are* certainly topics that are related to

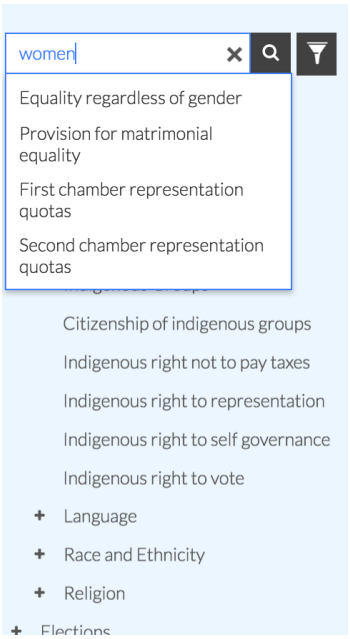

Figure 3: Entering "women" in Constitute's Search Box Triggers Topics Related to Gender

women in our taxonomy; for example, inheritance laws, marriage equality, etc. These are important concepts closely related to the status of women. Another researcher might have included them under the category, "women," but we had not done so. So, how would we integrate that researcher's category? In our case we introduced the keyword "women" into our ontology, and linked each relevant topic to that concept using an equivalence relationship. In pseudocode, the relationship we define appears as follows:

[Women] – [is a keyword associated with] – [gender quotas]

Using this information, Constitute can auto-suggest topics related to women's rights (see Figure 3). Adding equivalence relationships of this kind allows us to integrate different conceptualizations and, as such, enrich our underlying ontology. What this means practically is that one can access constitutional texts and data associated with concepts other than our own.

4.3.1 Editing Ontologies

Ontologies can be constructed in many languages (for example, XML or HTML), but most Semantic Web ontologies are defined using the Ontology Web Language (OWL). As noted

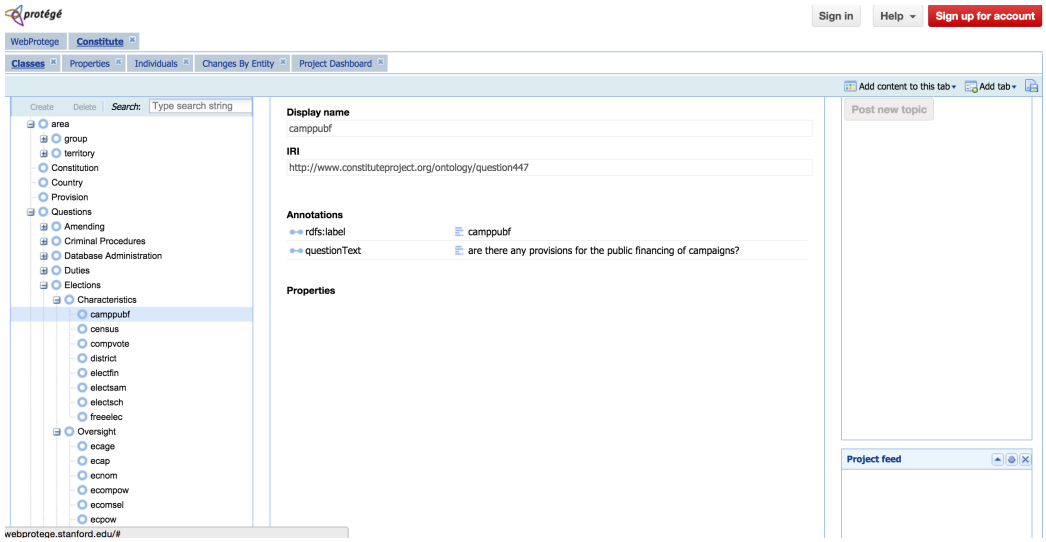

Figure 4: View of the dashboard of WebProtege, an ontology editor

previously, OWL is not a human-friendly format, but tools have evolved to ease viewing and editing of raw OWL data. One widely-used example is Protege (Knublauch et al. 2004), an open-source ontology editor designed specifically for the Semantic Web community. Figure 4 gives a screenshot of the Constitute ontology, as viewed in Protege's app-as-a-service alternative WebProtege (Tudorache et al. 2008). Usefully, WebProtege allows for collaborative ontology construction; so, a researcher interested in making "first-order" modifications to the Constitute ontology could easily fork our ontology, make context-specific edits, and submit the revised ontology to our platform for approval. Once uploaded to the Constitute server, this updated ontology would then result in an enriched Constitute that includes these researchers' terms, which can be associated with other concepts and attached to relevant textual excerpts.

From our perspective, the payoffs from these kinds of ontology enrichment efforts are both immediate and substantial. Researchers who wish to incorporate their view of the world could update the ontology and search through constitutional texts using their conceptual schema. Actually, they can do so in two ways. First, by enriching the ontology by adding information on conceptual relationships (for example, that inheritance laws have something to do with "women"). But second, researchers could add new connections *from concepts to data*. That is, researchers could relate their new concepts to constitutional

excerpts. Suppose that some researchers' understanding of Article 38, section 1 of the Cape Verdian constitution of 1982 (discussed above) was different from ours. They might see the clause as expressing not only a "right to privacy" but also a "right to identity", defined based on their own conceptualization of constitutional content. Traditionally, such "ontology merging" tasks have been conducted by hand, but researchers have also proposed machine learning-oriented approaches to the problem (see, e.g., Ngo and Bellahsene 2012)

4.3.2 Visualizing Ontologies

Displaying ontologies in a human-readable fashion is an important but non-trivial task. Imagine that we wanted to combine taxonomies from Linneaus and his competitors. How would be actually view and compare these worldviews? Since ontologies can be viewed as networks relating concepts and data, a logical approach would be to view them as graphs, with entities (again, concept, label, property, classification relationship, etc.) as nodes and ties labeled with relationships (e.g., "property of" or "sub-class of").

Network visualization is a mature field in its own right, with a plethora of tools and algorithms available for a broad class of use cases (e.g. Csardi and Nepusz 2006; Hagberg et al. 2008). Unfortunately, few of these tools are adapted for use with Semantic Web data formats. Partly, this limitation is practical; most real-world ontologies are composed of many nodes and connections, with a large number of unique values (e.g. "isAncestor" or "hasReligion") associated with the edge set. As a result, network-style visualizations of Semantic Web ontologies often must be simplified or restricted to a subset of the network (see Katifori et al. 2007, for an overview and additional discussion). Future work should attempt to bring these two streams of work more closely together; however, for now, the visualization options for Semantic Web data are limited.

As an illustration, Figure 5 provides a visualization of the Constitute ontology using LodLive (Camarda et al. 2012), a lightweight web-based ontology viewer. Here, we show an egocentric network giving connections to the Campaign Finance concept. This case provides a good example of the complexities involved in real-world ontologies. Because Constitute's

Figure 5: A view of part of the Constitute data on the LODLIVE visualizer

ontology is multilingual, each concept contains synonyms from other languages, as well as overlapping hierarchy and equivalence relationships between the ontology's conceptual components. These relationships are difficult to display with any level of detail when more than a few nodes are present.

5 Conclusion and Discussion

This essay tackles a pair of perennial, if background, questions for scientists. How do we formalize our conceptual map of the world, and how do we compare this map with that of others? Most of our applied examples come from the field of constitutional law, but (we suggest) many of the problems we face are common to all academic disciplines. The Semantic Web-based approach we outline in this paper provides a standardized infrastructure for mapping concepts, which offers both theoretical and methodological benefits. Indeed, this solution has already paid dividends. Google and other search engines are already interacting with the data that my collaborators and I have disseminated in this form.

But we see other dividends as well. One is in the systematic description of the con-

ceptual map of any given concept. In any given inquiry, researchers are at various points interested in understanding the semantic relationships among concepts, particularly in the presentation and development of theory. If there is any hallmark of good science, it is that it is highly conceptual. Things mean something only to the extent that we are able to put them into meaningful bins of knowledge – i.e., concepts. These bins are all the more compelling if they are shared and understood with respect to other bins.

Another, perhaps more immediately appreciated, dividend has to do with the retrieval of data. Connecting concepts allows one to find and surface data from associated concepts. This step is critical in any sort of empirical testing and can be important in crafting original empirical designs that merge data from different sources and domains.

It may seem to the reader that this essay has an evangelical tone. Such an interpretation would be well-founded, for we recognize that it makes no sense to expound upon the virtues of formalizing concept formation without actually doing something about it. So, what to do? One reasonable path forward, it seems to us, would be to explore this data framework in the context of constitutionalism. Constitute has already benefited substantially from the Semantic Web data framework, but we have yet to incorporate ideas from related research efforts into our ontology. In future work, we plan to expand substantially in this area, adding ideas, data, and connections drawn from some of the projects we cite throughout this paper. We hope our efforts will offer an example to other researchers engaged in large-scale research efforts, reducing startup costs and increasing incentives to adopt the approach we describe.

References

Alvarez, M., Cheibub, J. A., Limongi, F., and Przeworski, A. (1996). Classifying political regimes. *Studies in Comparative International Development (SCID)*, 31(2):3–36.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

Bogardus, E. S. (1925). Measuring social distance. *Journal of applied sociology*, 9(2):299–308.

Boyd, D. R. (2011). *The environmental rights revolution: a global study of constitutions, human rights, and the environment*. UBC Press.

Camarda, D. V., Mazzini, S., and Antonuccio, A. (2012). Lodlive, exploring the web of data. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 197–200. ACM.

Carrubba, C., Gabel, M., Helmke, G., Martin, A., and Staton, J. (2012). An introduction to the complaw database. *Unpublished Manuscript, Emory University. Available at http://perma. cc/MMA6-UBVR*.

Cichowski, R. A. (2007). *The European court and civil society: litigation, mobilization and governance*. Cambridge University Press.

Cingranelli, D. L. and Richards, D. L. (2010). The cingranelli and richards (ciri) human rights data project. *Human Rights Quarterly*, 32(2):401–424.

Collier, D. and Adcock, R. (1999). Democracy and dichotomies: A pragmatic approach to choices about concepts. *Annual Review of Political Science*, 2:537–565.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

Elgie, R. (2005). From linz to tsebelis: three waves of presidential/parliamentary studies? *Democratization*, 12(1):106–122.

Elkins, Z., Ginsburg, T., and Melton, J. (2009). *The endurance of national constitutions*. Cambridge University Press.

Garoupa, N. and Ginsburg, T. (2009). Guarding the guardians: Judicial councils and judicial independence. *The American Journal of Comparative Law*, 57(1):103–134.

Gould, S. J. (2010). *Bully for brontosaurus: reflections in natural history*. WW Norton & Company.

Gurr, T. R. (2000). *Peoples versus states: Minorities at risk in the new century*. US Institute of Peace Press.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL).

Harutyunyan, G. and Mavčič, A. (1999). *Constitutional Review and Its Development in the Modern World:(a Comparative Constitutional Analysis)*. Hayagitak.

Harvey, D. (2008). The right to the city. *The City Reader*, 6:23–40.

Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., and Giannopoulou, E. (2007). Ontology visualization methods—a survey. *ACM Computing Surveys (CSUR)*, 39(4):10.

Keck, T. M. (2007). Party, policy, or duty: Why does the supreme court invalidate federal statutes? *American Political Science Review*, 101(2):321–338.

Knublauch, H., Fergerson, R. W., Noy, N. F., and Musen, M. A. (2004). The protégé owl plugin: An open development environment for semantic web applications. In *International Semantic Web Conference*, volume 3298, pages 229–243. Springer.

Ngo, D. and Bellahsene, Z. (2012). Yam++: A multi-strategy based approach for ontology matching task. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 421–425. Springer.

Riggs, E. (1903). The vertebral column of brontosaurus. *Science*, 17:393–394.

Ríos-Figueroa, J. and Staton, J. K. (2012). An evaluation of cross-national measures of judicial independence. *The Journal of Law, Economics, & Organization*, 30(1):104–137.

Rosen, J. (2011). The right to be forgotten. *Stan. L. Rev. Online*, 64:88.

Sartori, G. (1970). Concept misformation in comparative politics. *American political science review*, 64(4):1033–1053.

Scribner, D. and Lambert, P. A. (2010). Constitutionalizing difference: A case study analysis of gender provisions in botswana and south africa. *Politics & Gender*, 6(1):37–61.

Shugart, M. S. and Carey, J. M. (1992). *Presidents and assemblies: Constitutional design and electoral dynamics*. Cambridge University Press.

Tschopp, E., Mateus, O., and Benson, R. B. (2015). A specimen-level phylogenetic analysis and taxonomic revision of diplodocidae (dinosauria, sauropoda). *PeerJ*, 3:e857.

Tudorache, T., Vendetti, J., and Noy, N. F. (2008). Web-protege: A lightweight owl ontology editor for the web. In *OWLED*, volume 432.

van Maarseveen, H. T. J. and van der Tang, G. F. (1978). *Written constitutions: a computerized comparative study*. Brill.

Wimmer, A., Cederman, L.-E., and Min, B. (2009). Ethnic politics and armed conflict: A configurational analysis of a new global data set. *American Sociological Review*, 74(2):316–337.