20

Digital Semantics

Formal Methods of Concept Mapping

Zachary Elkins

INTRODUCTION: THE PROMISE OF DIGITAL SEMANTICS

The text-as-data revolution has led to eye-opening analyses of staggering amounts of text of all kinds – from print news to historical documents to social media posts.[1] The range of new technologies has given new life to English and history departments in the cloak of "digital humanities." In the social sciences, text-as-data tools now account for a large and growing share of what was already a sizeable set of research methods. One might expect that this technology would yield dividends for methods of *concept formation and analysis* – the highly conscious approach to the vocabulary that structures and represents scientific ideas. Presumably, the standard practices of that approach – analyses of concept meaning, dimensionality, properties, related concepts, and the cases categorized by concepts – could be performed on a larger scale, more systematically, and more precisely if machine-processed.

We will call this shift to a more computerized analysis of concepts *digital semantics*. Though promising, the path forward for digital semantics is not obvious, or at least not well mapped. As such, what exactly would such concept-analytic technology look like, and what is the analyst's role in it? And what can be accomplished once we have such technology in place? I outline answers to these questions in the context of comparative law, which has undergone extensive experimentation with such methods. In particular, I evaluate a ten-year effort to systematize concepts about constitutional ideas using some methods proposed for organizing information in online environments.

[1] Thanks to David Collier and Julian Svedosh for comments and suggestions on this text, to the National Science Foundation for three grants to fund the research, and to my collaborators at the Comparative Constitutions Project for almost twenty years of discovering and documenting concepts in constitutional law.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

BASIC INPUTS IN DIGITAL SEMANTICS

The building blocks of digital semantics involve, principally, the formalization or "data setting" of concepts. The particular formalization that I have in mind goes under the names *controlled vocabulary*, *concept map*, *schema*, *ontology*, and *taxonomy*.[2]

These terms may have slightly different connotations across researchers and fields, but the differences are small, or at least not relevant here. In this chapter, I will mostly use *controlled vocabulary* (or just *vocabulary*), a neutral term that is less alien to most of us. Regardless of the term, the core idea is (1) to be *explicit* about terms that represent a domain of knowledge and (2) to record information about these terms; but also to record the information in a way that other researchers (and machines) can analyze and expand the set of terms. What kind of information to record? Traditional concept analysis suggests at least four elements with which to start:

(1) *Related concepts*. What historical and contemporary terms are related to the concept, either as synonyms, antonyms, subtypes, supertypes, or other relationships (e.g., diminished subtypes)? Think of this as mapping the "semantic field," following Giovanni Sartori's (1984) advice.
(2) *Properties*. What are the concept's characteristics, whether they are defining or associated properties?
(3) *Dimensionality and classes*. What, if any, are the subcomponents of the concept? What is the relationship between the concept and its components (e.g., hierarchical or not)? Note that this exercise is highly related to (1).
(4) *Cases*. What cases or instances serve as prototypical examples of the concept? These might operate as ostensive definitions, as in "look at that, that's what I mean."

How *exactly* are researchers to record such information in a way that will map onto insights from other researchers? And, more to the point, why would they do so? A preliminary approach is to formalize one's conceptual data tabularly, as one would a standard data set. Imagine rows of concepts tabulated against columns of information about the four areas just identified. In the research on constitutions that I describe later, we distill the ideas enshrined in 840 historical national constitutions to roughly 330 concepts (topics).

This is obviously an extremely large *N*. Importantly, the concepts (ideas/topics) and their attributes (definition, related terms, example text, etc.), are published in a standalone data set alongside the core data set – the yearly characteristics of each constitutional system. Relatedly, another example of

[2] Concepts about concepts can themselves be quite opaque, though absolutely essential to mapping methods. Many such terms are defined in the glossary at the end of the chapter.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

a large-*N* approach to concepts is Diana Kapiszewski et al.'s (2024) study of 1,621 instances of "constitutions with adjectives," a data set the authors analyze in order to understand the aspects of the constitutional order represented by the adjectives. One can also base the analysis on other information about concepts, such as their degree of contestedness, an approach adopted by Gerring and Cojocaru (2025) in their study of 383 concepts.[3] This approach of analyzing a large *N* substantially extends the horizon of what may in fact be "qualitative" research on concepts.

In these three examples, the research product is a data file of information about concepts that any data scientist could analyze with their preferred data analytic software. If such files take on a somewhat standard form and are deposited in a data archive, one can begin to integrate ideas.[4] Analytic and visualization software can illuminate such information to encourage further conceptual refinement and translation. For example, concept diagrams play a central role in David Collier's extensive work on concept analysis. Indeed, the many ways that creative users and their applications will employ concept data are – as with data of any kind – limitless and ever evolving.

But what other dividends result from organizing concepts in this way? Recall the constitutions example, which runs through this chapter. In 2005, Tom Ginsburg and I conceived the Comparative Constitutions Project (CCP), our ongoing effort to collect and analyze historical constitutions from around the world. Without a standard vocabulary to consult (no constitutional Linnaeus had yet appeared), we devised a set of some 650 attributes, drawn from our reading of a sample of texts from the genre, which we used to code (interpret) constitutional texts. In 2013, our team partnered with Google Ideas (now Jigsaw) to leverage these data in building a public repository of constitutional texts, which we called Constitute.[5] Importantly, we indexed the texts with some 330 constitutional topics, drawn from the 650 attributes included in our data. The repository's goal is to allow/encourage constitutional drafters to call up a set of representative excerpts on any provision (topic) of interest. The thought is that drafters would now understand their options.

Since 2013, the site has grown steadily; as of June 2025, it hosts 7,000 visitors a day and serves constitutional drafting teams worldwide. That the site has become a research tool for drafters of the texts themselves imposes an additional set of responsibilities on the data set. If drafters use the site's topics to identify the ideas that *should* be included in their text, then it is imperative that the topics represent the latest in constitutional ideas. Indeed, some of the ideas in modern constitutional drafts, such as those from Chile's recent

[3] See Chapter 14, this volume.

[4] A repository of vocabularies with a method to search and compare terms would be a basic and helpful tool in concept integration.

[5] Online at constituteproject.org.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

	Provisions
Constitute	42,461
Amendment	322
Constitution amendment procedure	229
Unamendable provisions	93
Culture and Identity	2,510
Citizenship	852
Citizenship of indigenous groups	7
Conditions for revoking citizenship	155
Equality regardless of nationality	35
Equality regardless of origin	120
Power to deport citizens	54
Requirements for birthright citizenship	190
Requirements for naturalization	148
Restrictions on entry or exit	64
Right to renounce citizenship	79
Indigenous Groups	46
Citizenship of indigenous groups	7
Indigenous right not to pay taxes	1
Indigenous right to illegal activities	0
Indigenous right to political parties	0
Indigenous right to representation	14
Indigenous right to self governance	17
Indigenous right to vote	7
Language	469
Equality regardless of language	95
Official or national languages	173
Protection of language use	112
Trial in native language of accused	89
Race and Ethnicity	278
Equality regardless of race	159
Equality regardless of tribe or clan	20
Integration of ethnic communities	57
Provisions for wealth redistribution	42
Religion	865
Equality regardless of creed or belief	113
Equality regardless of religion	158
Establishment of religious courts	15
Freedom of religion	217
God or other deities	137
Official religion	102
Separation of church and state	85
Status of religious law	30
Tax status of religious organizations	8

FIGURE 20.1A Select topics from Constitute: amendment and culture/identity.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

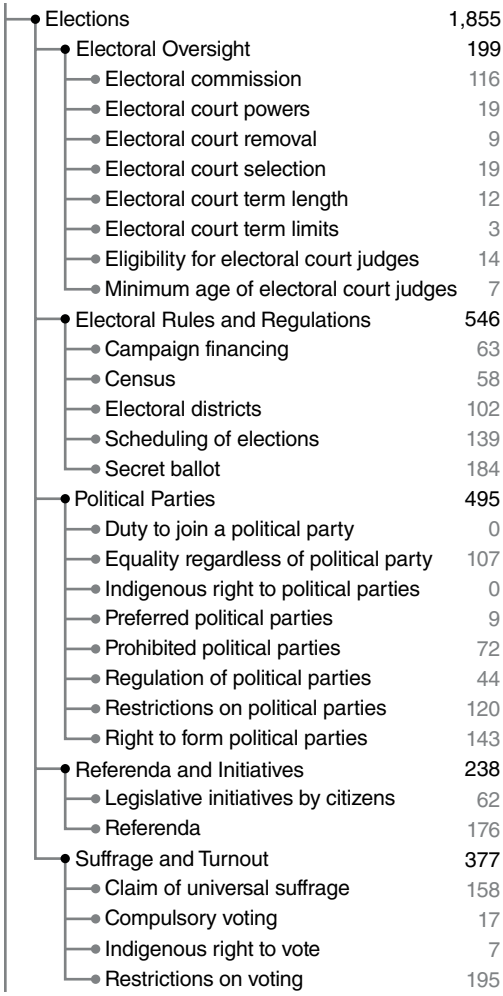


FIGURE 20.1B Select topics from Constitute: elections.

efforts, are not as well indexed by the Constitute vocabulary as one would like; vocabulary needs to keep up with innovations in ideas. For example, Figure 20.1a and 20.1b depict some of the topics in the initial set (v.1.0) of topics indexed on Constitute. Figure 20.1a includes topics related to *amendment* and *culture/identity* and Figure 20.1b includes those related to *elections*, a fraction of the 330 topics included in v.1.0 of the vocabulary. Small updates to this overall set, sourced from constitutional reform and academic projects, are released periodically. One analog to this updating process is the regular "editioning" of the *Diagnostic and Statistical Manual of Psychiatric Disorders* (*DSM*), which attempts to standardize ideas in the mental health

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

field. The controversial *DSM* operates in a particularly delicate area of research. It may be imperfectly representative of the field's ideas, which are evolving and sometimes highly contested, but the manual is still an invaluable benchmark for coordinating research and treatment.

The set of topics on Constitute is a controlled vocabulary – a data set of concept labels, definitions, related terms, and examples – that users can download from the website and from various data repositories (e.g., the Qualitative Data Repository at Syracuse University). The idea, at the outset, was to facilitate the use of the data by machines, as well as to coordinate the use of terms among researchers in constitutional law. In most research domains, an initial data set forms the core vocabulary, and is expanded and enriched when merged with other vocabularies. We decidedly do not presume ours to be the central node in the network, nor ourselves as the second coming of Linnaeus. Indeed, an important part of the collection and publication of relevant terms and data sets is to expand the vocabulary beyond Constitute itself.

CONCEPTUAL ALIGNMENT, DISCOVERY, AND ENRICHMENT

One of the challenges of scholarly collaboration is aligning two competing, or even complementary, vocabularies. Imagine if Linnaeus were to compare (and integrate) his biological vocabulary with that of his competitors. Comparing multiple vocabularies can result in knowledge transfer, accumulation, and translation. However, only once vocabulary is formalized can one begin to integrate them meaningfully.

For example, consider the vocabulary used to categorize articles, books, and data sets. The major cataloguers of these three sets of works (JSTOR for articles, the Library of Congress for books, and ICPSR for data sets) use different vocabularies to classify the works, seemingly without any connections among them. Certainly, the origins and purposes of the schemes are different. For example, the Library of Congress' categorization scheme grew from Thomas Jefferson's initial set of categories (Jefferson's donation of his library would become the Library of Congress) and pretends to represent every domain of knowledge and to catalog every published book.

One could manually build a crosswalk file that integrates these three (and other) vocabularies. However, text-as-data methods offer tractable ways to integrate large vocabularies, albeit with human oversight. In other work, my collaborators and I use some promising natural-language-processing methods of semantic similarity (vector semantic models, specifically) to align (initially) these three vocabularies, followed by further stitching by hand (Elkins, Gardner, and Moran 2022). The result of these aligned vocabularies is that one can understand which topics and categories exist in one collection or another. For example, the Library of Congress "K Class" (Law) is deep and extensive mostly because the one and only K-class librarian worked for over

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

four decades to develop and populate a nuanced set of categories.[6] Not only can one evaluate the reach (conceptual "extension") of the vocabularies themselves, but also, since these categories are tied to actual objects (books, articles, and data sets), one can calculate which topics are most populous in those collections.[7] I am not sure that we scholars have a clear understanding of the relative population of books and articles across topics. Is there more work on, say, "populism" than on "clientelism"? As it happens, the answer is yes, especially these days, with five times as many books and articles tagged as "populism." I have come to call such analyses *conceptual ecology*, by which I mean the assessment of the prevalence of a particular idea in a given population (era, region, discipline, etc.). One might also think of it as Zeitgeist studies.

In the analysis of constitutions, this sort of alignment of concepts across scholars is invaluable. Scholars of constitutional law have produced multiple data sets on various aspects of constitutions, including data on the constitutional text itself (Elkins and Ginsburg 2005/2022); data on particular topics of the text (e.g., Lambert and Scribner 2009 on gender; Koenig, Tsutsui, and Crabtree 2023 on minority incorporation); and data on judicial decisions (e.g., Gabel et al. 2024). Each data set has its own conceptual vocabulary, some more formal than others. And for some of these projects, the vocabulary from the CCP has functioned as something of a baseline vocabulary from which the others build their concepts.

An example of this kind of alignment (and enrichment) of constitutional topics is a project by Jill Lepore (2023), who has collected the text of the almost 14,000 proposed amendments to the US Constitution since 1789. Lepore used the topics in Constitute to categorize the various proposals. Not surprisingly, the topics in the Constitute vocabulary did not always fit each of the US proposals. Sometimes, the US corpus introduced an idea that was not in the Constitute vocabulary (e.g., the right to water, the rights and duties of parents, and honorary titles for retired office holders) or that was more nuanced (e.g., a special district for the national capital). And, of course, there are many more topics in the Constitute vocabulary that do not appear in the US proposal set. Lepore and the CCP team worked together to enrich the vocabulary. As a result, twenty-nine topics were added to the Constitute vocabulary, which enriched both projects. Note that in this case the classification of the amendment proposals was performed by humans, not machines; the "digital" component here is the sharing of a digital collection of concepts. The advantages of a mutually tagged data set are significant. One basic analysis facilitated by the data would be one of conceptual ecology (Elkins

[6] Part of our exploration in this area has involved a series of interviews with the K Class librarian, Dr. Jolande Goldberg, who has worked on the K Class for forty-five years and counting.

[7] The actual classification of these materials may be processed by humans, machines, or a collaboration of the two; the important step that I emphasize here is to stipulate, and coordinate along, the classes themselves.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

and Lepore, forthcoming). That is, which topics populate one set but not the other, and to what degree? But one could imagine many other studies. This scholarly collaboration, in this case between historians and political scientists, points to the value of a unified vocabulary.

Sometimes, one has data (text or otherwise) but no vocabulary with which to classify the data. Consider, for example, the mountain of comments often collected in public consultations held in advance of constitutional assemblies. In the preparation of the Chilean constitutional draft of 2022, the government collected some 250,000 comments in a series of town hall meetings, as well as in open public comment periods. These comments represent something of a citizens' wish list for the new constitution (which failed in a referendum in September 2022). But wishes for what, exactly? Without a working vocabulary, one cannot categorize citizen remarks by topic. And even with a vocabulary, one would not likely have the bandwidth to categorize the comments by hand. These challenges of vocabulary and labor explain, in part, why public consultation records go largely unanalyzed. However, a centralized vocabulary and machine-mediated methods render the task more tractable. For example, in an analysis using vector semantics, we matched most of the 250,000 comments to a topic in the CCP vocabulary (Cruz et al. 2023). We then compared the prevalence of these ideas against their prevalence in national constitutions. Perhaps not surprisingly, we find that most of the Chileans' comments had something to do with rights, a result consistent with the attention to rights over institutional structure in modern constitutional design (Gargarella 2013). But then, this kind of conceptual ecology was possible only with a representative set of topics and ideas.

AUTOMATED INTEGRATION: THE CASE OF LINKED OPEN DATA

In the examples given, the format of the vocabulary files is straightforward, sometimes involving simple text files, which scholars can read and analyze with their software of choice. Some file formats, however, provide opportunities for machines to assess and match concepts more easily, and then connect these concepts to data on concrete instances of the concept. One such solution is Linked Open Data, sometimes called the Semantic Web, and more aspirationally Web 3.0. It is a data structure that is increasingly common on the web and that supplies the information panels and carousels that search engines now regularly present alongside internet search results. Social science data is only sporadically available in this format. As it happens, the CCP was one of the first data projects in the social sciences to employ this structure, which makes for a helpful case study. The project's use of Linked Open Data has yielded some modest but potentially significant benefits.

The data for Constitute, both concepts and text, take a very different form, one built for the online environment and one that both human beings and machines can easily access. Consider some early dividends of this data

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

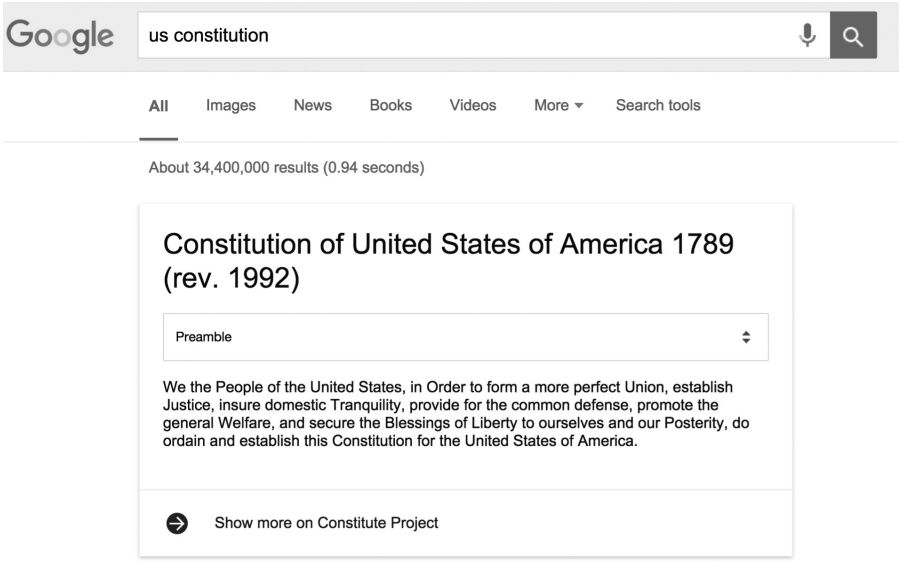


FIGURE 20.2 Results from a Google search on "US Constitution."

strategy, since sometimes it can be helpful to see the research target. In 2015, we made our data available on a SPARQL endpoint (a data hub that machines can consume). Shortly thereafter, Google's search engine began to ingest this structured data and surface it in different ways in search results. For example, a Google search on "constitution" returned the 4,500-word US Constitution on a "card" (a "one-box," in Google-speak) at the top of the search results (Figure 20.2). A drop-down menu on the card allowed the reader to navigate sections. The text and data on these cards came directly from our online Constitute data. A number of other constitutions would show up too. A Google search on the "Bhutan Constitution" allowed readers to read the country's commitment to Gross National Happiness in Article 9(2). All of these cards pulled directly from the Constitute repository and updated as Constitute updated. The US Constitution on an index card is, quite literally, a small thing, but something that represents a huge advance for information science, and for the social scientists who depend on it.

Where do these infographics come from? The data are stored in Google's Knowledge Graph, a curated set of highly connected and machine-readable data (also known as *linked data*). Google began delivering results from the Knowledge Graph in 2012, but the concept of linked data had been simmering for some time. Tim Berners-Lee, known to many as the inventor of the World Wide Web and to many of his followers as simply "TimBL," had begun championing the concept as early as 2009 as the heart of his Web 3.0. Berners-Lee runs an organization devoted to building and standardizing the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

relevant technology. Linked data is also sometimes identified as "graph data" (highlighting its interconnectedness) and is a core part of what technologists describe as the Semantic Web.

Linked data are simple to understand, and their utility is immediately obvious. One key feature is that each data element, whether a concept (right to privacy) or a concrete "thing" (e.g., the US Constitution), has its own unique location on the web. These locations (Uniform Resource Identifiers) can be web addresses (URLs) that human beings read, but more often they are places where data resides for machines. Each of these entities is linked to other entities through some relationship, which itself is labeled with a unique URI.

So, a typical linked data file comprises seemingly endless lines of subject–predicate–object "triples," each of whose elements is a distinct URI. For example,

<*http://constitute/constitution/sudan2005/article2*>
<*http://constitute/hastopic*>
<*http://constitute/torture*>

is one of many triples in the Constitute data set. Linked data files have the suffix .nt (as in, *N* triples).[8] 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 data set alone would have *other* links to each of these entities (i.e., other links to "Article 2," "Sudanese Constitution," and "torture"). And other data sets would have things to say about these entities as well. As you can imagine, data files with endless triples marked up in this way are utterly forbidding to browse directly. Editing and visualization tools allow analysts to work with and understand various relationships in these files. However, the lengthy lines of code are instantly intelligible to machines. The beauty of linked data is that these "entities" (concepts, properties, data elements, etc.) can be linked to an infinite number of other things and concepts (hence the graph analogy, which refers to a network graph). And every entity lives at a unique address on the World Wide Web. The consequence is that machines and their human analysts can draw connections easily and exponentially as network ties grow. There are real challenges in relating such data to those of others, which is why many semantic data sets (such as Google's Knowledge Graph) exist in a semi-open state – consuming public data but storing and managing it more privately.

One virtue of data in this form is that human beings and machines can consume each other's data more easily. Just as Google can put up Constitute's texts, structured as their designers see fit, so too can we import dynamic and fascinating data that appears and updates by the minute. One source is DBPedia, a collection of the structured part (the information boxes) of Wikipedia. All of the data included in info-boxes in Wikipedia articles are included in DBpedia. Another source is the *New York Times*, which has made

[8] A variety of other data formats, such as JSON-LD, have evolved in addition to the .nt file.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

available its data, in linked-data format, for machines and humans to extract. In the same way that constitutions on Constitute are tagged with topic tags, so are *New York Times* articles. Data sets in this format are multiplying and evolving, growth that renders each such data set that much more valuable. This result follows from the basic law of network externalities.

Apart from data integration, perhaps one of the most interesting aspects of the Semantic Web is its utility for concept formation and enrichment. In particular, Semantic Web methods have the distinct advantage of being able to share, and collaborate on, vocabularies. With tools developed for the Semantic Web, one can edit a data set's vocabulary to update or expand any part of this conceptual structure. My sense is that Linnaeus and colleagues and competitors would ideally have developed their rival classification systems using Semantic Web technology.

So, for example, consider the 300 or so topics in the Constitute vocabulary (Figure 20.1). One can browse these topics on the Constitute website and return a complete set of excerpts related to each topic. See Figure 20.3, a snapshot that shows the arrangement of just a few of these concepts on Constitute. In this example, the first-level topic ("Culture and identity") is expanded to depict several subtopics under which "Indigenous Groups" is expanded to see a set of constitutional provisions, including "Indigenous right not to pay taxes."

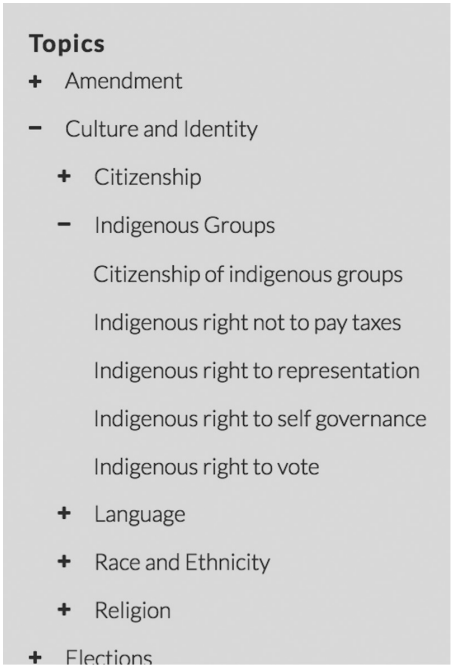


FIGURE 20.3 A snapshot of Constitute's topic tree.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

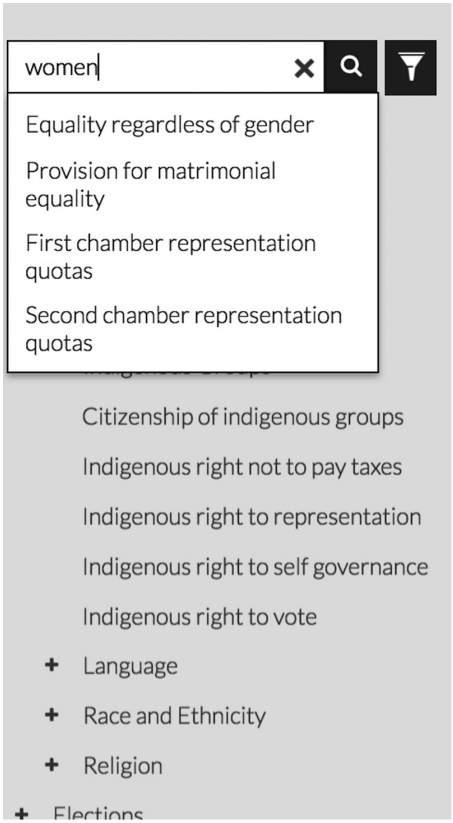


FIGURE 20.4 Entering "women" in Constitute's search box triggers topics related to gender.

As we have emphasized, any given categorization is only one view of constitutional ideas and is not suitable or useful for all users. For example, none of the topics has "women" in its label. There are certainly topics that are related to women in our vocabulary, such as inheritance laws and marriage equality. These are important concepts closely related to the status of women. Another researcher might have included them under the category "women" or "gender." So how would one integrate this new category?

In the case of Constitute, we introduced the keyword "women" and linked relevant topics to that concept. So now if one types "women" in the search bar, Constitute auto-suggests each of the topics related to that concept (see Figure 20.4). Adding keywords essentially allows Constitute to integrate different conceptualizations and, as such, enrich the underlying vocabulary. What this means practically is that one can access constitutional texts and data associated with concepts other than those stipulated by the CCP.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

CONCLUSION: A PATH FORWARD

This chapter takes on a perennial question for social scientists: How do we formalize our conceptual map of the world, and then how do we compare our map with that of others?

Scholars in earlier eras had much greater challenges in coordinating than we do today. The tools of concept analysis are ideally suited to recent advances in text-as-data methods and to web-based data structures. The domain of constitutional law is just one of many domains suited to this kind of methodological intervention. The case I describe in this chapter – that of the concepts developed to represent constitutional ideas – suggests that a little bit of formalization goes a long way. Cataloging concepts in simple, standardized data files allows researchers to align concepts and thus to translate ideas from one data project to another. Researchers can use such a standardized set of concepts to index their own data objects and even expand the vocabulary, thereby enriching the original set of concepts. These simple concept sets are easily transformed into machine-readable data files that allow for more seamless connections on the web. Importantly, the creation, organization, and enrichment of such files still require significant concept curation. This exercise preserves and incentivizes the kind of conceptual work that many scholars find intellectually satisfying and that originally motivates their engagement in concept analysis.

GLOSSARY

Comparative Constitutions Project (CCP) A data project founded by Zachary Elkins and Tom Ginsburg in which the authors have identified each formal change to each of the world's constitutions since 1789 and recorded a large number of attributes of these constitutions in order to test hypotheses regarding the origins and effects of constitutional ideas. The project also hosts a repository (Constitute) that includes a sample of these constitutions indexed with some 330 topics. Related: **Constitute**.

Concept integration Combining two vocabulary sets from the same or related domain. Related: **concept merging**; **concept translation**.

Conceptual ecology A study of the prevalence of certain ideas in a certain population (time, space, or literature, for example).

Conceptual extension The degree to which a particular controlled vocabulary indexes the full set of ideas in a domain.

Concept mapping Specifying and organizing the concepts in a domain of knowledge. See **knowledge representation**.

Concept translation Matching a term from one researcher to those of others. Related: **concept integration**; **concept merging**.

Constitute An online indexed repository of the world's constitutions, indexed with topics derived from data from the CCP. Online at constituteproject.org. Related: **Comparative Constitutions Project**.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

Data setting (of concepts) The idea of treating concepts as units of analysis, by organizing information about concepts systematically.

Digital semantics The use of computerized methods to organize and analyze ideas (concepts) and their meanings.

Embeddings Representations of words or phrases as vectors of numbers. Related: **vector semantics**.

Graph database A data structure in which attributes of entities are stored as nodes and edges in the style of network (or graph) data, such that the edges represent relationships between the nodes. The data format is intended to facilitate semantic queries. One example is Google's Knowledge Graph, which contains the information used to present the info-boxes that appear alongside web search results.

Knowledge representation In general, the process of organizing and representing information in a particular domain of knowledge, often so that computerized methods (e.g., artificial intelligence) can perform some sort of task but also so that nonexperts can understand the ideas in a field. Related: **concept mapping**.

Natural language processing Machine-learning technology in which computers interpret human language.

N-triples A data format understood by machines in which concepts and entities are encoded as URLs and organized as triples of subject–predicate–object form.

Ontology The terms and categories, and their properties, that are used widely in a particular domain of knowledge. Facilitates classification and knowledge representation. Related: **taxonomy; schema; controlled vocabulary**.

Ostensive definition The expression of meaning by illustration or example; that is, by pointing (from the Latin *ostens*, stretched out to view).

Schema A set of standardized vocabulary, often used to mark up web pages for machines to analyze. Closely related to the project schema.org. Related: **ontology; taxonomy; controlled vocabulary**.

Semantic Web An extension of the World Wide Web, in which concepts and entities on web pages are marked up with standardized format to allow machines to process and interpret. Related: **Web 3.0; Linked Open Data**.

SPARQL (endpoint) Computer language used to query data marked up in a standardized form, usually as Linked Open Data, and deposited on the web (endpoint). Related: **Semantic Web; Web 3.0**.

Taxonomy A system of classification for a particular domain of knowledge, especially one that is hierarchical. Related: **ontology; controlled vocabulary; schema**.

Text-as-data methods A general term for the various techniques and methods of analyzing digital text, often at a large scale. Related: **natural language processing**.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031

Uniform reference identifier (URI) A unique sequence of characters that identifies entities (such as a book, a person, or a concept) for referencing on the web. URLs, which identify a location on the web, are a kind of URI.

Vector semantics The method of representing words or phrases with a vector of numbers, often in order to assess their proximity in meaning. These representations are often called "embeddings."

Vocabulary enrichment Adding or refining terms and meanings to a set of concepts, often from a comparison of other vocabularies or corpora.

REFERENCES

Gabel, Mathew, Clifford J. Carrubba, Gretchen Helmke, Andrew Martin, Jeffrey K. Staton, Dalston Ward, and Jeffrey Ziegler. "CompLaw: A Coding Protocol and Database for the Comparative Study of Judicial Review." *Journal of Law and Courts*. Published online 2024:1–27. https://doi.org/10.1017/jlc.2024.4.

Cruz, Andres, Zachary Elkins, Roy Gardner, Matt Martin, and Ashley Moran. 2023. "Measuring Constitutional Preferences: A New Method for Analyzing Public Consultation data." *PLoS ONE* 18(12): e0295396. https://doi.org/10.1371/journal.pone.0295396.

Elkins, Zachary, Roy Gardner, and Ashley Moran. 2022. "What Do You Work On? The Conceptualization of Intellectual Products." Working Paper, University of Texas.

Elkins, Zachary, and Tom Ginsburg. (2005) 2024. Comparative Constitutions Project. https://comparativeconstitutionsproject.org.

Elkins, Zachary, and Jill Lepore. Forthcoming. "The Public Appetite for Constitutional Amendment in the United States." *Political Science Quarterly*.

Gargarella, Roberto. 2013. *Latin American Constitutionalism, 1810–2010: The Engine Room of the Constitution*. Oxford: Oxford University Press.

Gerring, John, and Lee Cojocaru. 2025. "Conceptual Contestation: A Large-*N* Analysis." (See Chapter 14, this volume.)

Kapiszewski, Diana, Deborah Groen, and Katja Newman. 2024. "Constitutionalism with Adjectives: Conceptual Innovation in the Comparative Study of Law." *Law & Social Inquiry* 49(1): 178–208.

Koenig, Matthias, Kiyoteru Tsutsui, and Charles Crabtree. 2023. "From Nationalism to Multiculturalism? Constitutional Provisions on Ethnic, Linguistic and Religious Diversity, 1789 to 2010." Working paper.

Lambert, Priscilla A., and Druscilla L. Scribner. 2009. "A Politics of Difference Versus a Politics of Equality: Do Constitutions Matter?" *Comparative Politics* 41(3): 337–57.

Lepore, Jill. 2023. Amend: Rewriting the Constitution. https://amendmentsproject.org/.

Sartori, Giovanni. 1984. *Social Science Concepts: A Systematic Analysis*. Beverly Hills, CA: Sage Publications, Inc.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 14 Nov 2026 at 12:09:27, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009551250.031